SPECTRAL COMPONENT ANALYSIS ON DISTORTED DATA

A thesis submitted to the University of Manchester
for the degree of Doctor of Philosophy
in the Faculty of Engineering and Physical Sciences

2009

By
Caleb Antonio Rascon Estebane
School of Electrical and Electronic Engineering

Contents

Words in text: 31,816
Words in headers: 393
Words in float captions: 671

Final word count: **32,880**

List of Tables

List of Figures

Abstract for **Spectral Component Analysis on Distorted Data**
For the award of the degree of **Doctor of Philosophy**
In the **University of Manchester**
To the candidate **Caleb Antonio Rascon Estebane**. *October 27, 2009*

A spectrum sampled from a material or product contains important and relevant information that can be used in many areas of Science and Engineering and is being used with greater frequency in the Industry, specifically in the areas of Quality Monitoring. Currently, many quality measurements are taken in an off-line manner which are costly and time-consuming. On-line instrumentation, coupled together with automate monitoring and control systems offers enormous benefits.

Extracting information from spectral data, and using this information in an automated quality control framework offers potential for faster response times to disturbances, as well as continually delivering high quality products. However, using spectral data has proven to be a challenge because of their location-to-magnitude inconsistency that can be attributed to sensor de-calibration or external influences, such as temperature and foreign materials. Although these effects can be accommodated by frequent calibration, it requires a high investment in time and money. Modelling methods have addressed the issue, providing insight into how external influences affect the shape of a spectral measurement, but only for particular processing factors.

This thesis presents a framework that extracts information from spectral data that suffers various forms of distortion. The developed techniques incorporate knowledge how a spectral sensor reacts to specific external factors, and compensate during the information extraction process. The proposed techniques achieve this by approaching the process as an optimisation task. The contribution of each component is estimated by finding the amount that each pre-defined external factor affects each component contained within the sampled spectrum. A major benefit of the proposed techniques is that the user is given, not only estimated information relevant to the process, such as the contribution of each component, but the techniques estimate the extent to which the external factors are interfering with the measurements. To test the feasibility of this framework, two types of generic distortion were considered: frequency displacement and spectral warp.

In many applications, such as the framework proposed in this thesis, a precise set of reference component signatures are required. This information will not always be available and to address this, novel techniques are proposed in this thesis which are able to extract the components in affected data sets. The capabilities of all the techniques proposed in this thesis are demonstrated through their application to several simulated and benchmark industrial data sets.

Declaration

No portion of the work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright

Acknowledgements

First and foremost, I would like to thank my greatest teacher of all: God. I know that I am here and that I am able to write all of this for a reason. I will do my best in never forgetting what a great fortune I have had in just being here, and that it comes with a lesson and a responsibility. I hope I am doing the work you have planned me to do.

I would like to thank the Mexican National Council of Science and Technology (CONACYT). Without their grant, this PhD would not have been possible.

I would like to thank my supervisor, Prof. Barry Lennox, for putting in me the idea of shooting directly for a PhD and for creating an environment of humour around this whole ordeal. Whilst the sensation of doing something that would impress everybody was short-lived, those first few days of uncertainty that you pulled with me are ones that I will not ever forget. Whatever the reason of why you offered me this opportunity, may that have been my good looks or my impossible-to-ignore charm: thank you for believing in me, even if it only was for just a few moments. Oh, and I will hunt you down, wherever you are, for making me suffer; I will leave a tip though.

I would like to thank my friends and colleagues that I have met in this my home far away from home called Manchester. Specially Oskar and Marisa, who, even though have reduced me to a third wheel in our relationship, have blossomed into a partnership that will not be forgotten. Whatever happens with you two, do know that, throughout these last couple of years, your relationship has provided me with an impressively beautiful site to see, as it is when two friends fall in love with each other.

I would like to acknowledge Nihil and NihilBack, my two Apple laptops. NihilBack is a PowerBook G4 with a 17” screen, 1.67 Ghz, 1 Gb RAM. Nihil is a MacBook Pro with a 17” screen, 2.5 Ghz, 4 Gb RAM. I want to thank you for your essential support in this project, as, without it, all of my work, from the thought process, to the simulations, to the actual writing up, would not have been possible. Both of you have granted me the possibility to do whatever I want, to experiment on my thoughts, to write my ideas down, and to share them with the world. You are the bridge to my loved ones, and my wings to my endeavours. I know you are just pieces of hardware to anybody else, but to me, you are an extension of my being, and without you, it is hard to be me. Thank you.

I would like to thank my beloved, Maria del Carmen Valle Lira. Our relationship was born in a very odd way, but I would not have expected otherwise, as both of us are odd in our own beautifully weird world. You portray the symptoms of my shortcomings, and the celebration of my virtues. I have become a better man because of the mirror you hold up for me. Thank you and I love you.

Finally, my parents: Luis Antonio Rascon Mendoza and Virginia Estebane Ortega. They gave me my name, they gave me my life, and everything else in between. I pride myself in having words for everything, but they truly shut me up when it comes down to describing how much I love them and appreciate the efforts they have put into giving me the life I have now. They are the reason I did this; they are the reason I thrive to be better. Their pride for me is my main goal in life. As I have said many times before, the only thing I aspire for is that when they lay in their deathbed they would think, ”I am proud of my son.” *Thank you, thank you, thank you.*

Qualifications & Work Presented/Published

- **Journal:**
  - Rascon, Lennox, Marjanovic. Recovering Independent Components from Shifted Data using FastICA and Swarm Intelligence. *Applied Spectroscopy.* To be Published. 63(10). October, 2009.
  - Rascon, Lennox, Marjanovic. Extracting Fundamental Components from Distorted Data by Sample Substraction. *Chemometrics and Intelligent Laboratory Systems.* In preparation.
  - Rascon, Lennox, Marjanovic. Review of Curve Resolution Methods with Distorted Data. *Chemometrics and Intelligent Laboratory Systems.* In preparation.
  - Rascon, Lennox. Effects of Particle Fidelity in Particle Swarm Optimisation. *Swarm Intelligence.* In preparation.
- **Conference Papers/Posters:**
  - Rascon, Lennox, Marjanovic. Extraction of Fundamental Components from Distorted Spectral Measurements. *Advances in Process Analytics and Control Technology 2009 Conference.* May 5-7, 2009.
  - Rascon, Lennox, Marjanovic. Using Lagged Spectral Data in Feedback Control Using Particle Swarm Optimisation *Proceedings of UKACC International Conference in Control 2008.* ISBN: 978-0-9556152-1-4.
  - Rascon, Lennox, Marjanovic. Effects of Frequency Displacement in Independent Component Analysis. *Proceedings of the ICA Research Network International Workshop 2008.*
  - Rascon, Lennox. Effects of De-Tuning on Current Note Detection Algorithms. *DMRN+2: Digital Music Research Network One-day Workshop 2007.* December 18, 2007.

We believe nut things because it is part of our little monkey brains to try desperately to make patterns. That is the genius of humans, the quality that lets us learn. ***Pattern recognition has moved us off the hostile savanna and into the much safer condominiums.*** *When you see your cavemate die shortly after a snake bite, it is probably a good idea to avoid all snakes. Of course, this over-simplification also leads to racism, religion, and all kinds of magical thinking.*

Penn Jillette: magician, actor, political commentarist, all-around nice guy.

You are me, and I am you, and you are listening to our song right this instant, but you don't know it. Whatever you think you're hearing isn't there right now, it was prefabricated and melted into your ear a long time ago... try to go beyond the frequencies, beyond the sound of your surroundings and listen to whatever is coming out. You'll found out that I'm waiting for you on the other side, being you, being me, and our song, our real song, was playing all along.

Anonymous

Chapter 1

Introduction

Spectral Component Analysis (SCA) is an important element of data analysis in various fields throughout the scientific community. This is due in large because a sampled spectrum provides a concise visualisation of the fundamental properties of a process or material. However, in an automated environment, where the calibration of spectral sensors is significantly affected by external influences, such as temperature, pressure, or other foreign components, measurement distortion needs to be considered.

The vast majority of the SCA literature assume a perfectly calibrated sensor, completely ignoring the possibility of distortion taking place. This assumption is understandable, since the areas where SCA is applied are typically in a controlled laboratory setting. However, interest in using spectral data as a feedback measure in automated quality control is increasing, as there is a vast amount of information that resides inside the spectrum that could enhance the process as well as the product.

Enabling SCA to be robust against spectral distortions is a difficult challenge. The characteristics of these distortions differ from application to application, and from batch to batch, which complicates the task of characterising every possible distortion source in a single model. However, there has been an increase in interest in this issue, and it has been addressed in certain specific instances, by modelling particular distortions in spectral data [15], such as frequency displacement and warping[7, 116, 74]. It would be of great interest to apply the knowledge obtained from these models to estimate the distortions that spectral data is suffering.

A framework is proposed that aims to utilise the knowledge provided in literature about spectral distortions. It introduces the effects of a pre-defined set of known distortions into a set of reference spectra such that they will best fit a sampled spectrum. In this approach, the contribution of each component is retrieved, as well as the amount of distortion suffered by the sample. Frequency shift and spectral warping are the types of distortions that are most frequently reported in literature and are able to be detected in a spectrum by the naked eye. Hence, these were chosen to be investigated in this project. However, the task is carried out as an optimisation algorithm, with the aim of it being flexible and able to incorporate new knowledge about the spectral sensor and the way that it reacts to external factors if necessary.

The main objective of this project is to provide a method able to cope with the distortions taking place in a specific process. Even though the algorithm aims to be applied to generic data, it requires specific knowledge of the sensor from which the spectral data is sampled. This is because the effects that are to be considered need to be congruent to the composition and physical effects of the external factors to both the material and spectral sensor. Therefore, a plant expert is required to provide indication of the suspected types of distortions that may be occurring, as well as verify the way the distortions are being introduced.

The proposed technique aims to provide a framework from which different types of distortions can be considered without the need for developing a specific global model to characterise them all. This approach provides the user with the flexibility of considering several 'suspected' distortions in the data, from which the framework indicates which are in fact taking place.

The proposed framework requires a set of reference spectra to be available. Although it can be assumed that for many processes this information will be available, this may not always be the case. In addition, even if a set of reference spectra is available, identifying foreign undesired components will also be of interest. Component extraction can be applied to a set of spectral samples to retrieve this information. However, the measuring devise may be subject to distortion, which current component extraction methods are sensitive to. To circumvent this situation, several tools are proposed in this thesis that can extract the components, or information about the components, from a spectrum in the presence of different types of measurement distortion.

All the proposed methods approach the task as if it were an optimisation problem, 'finding' the component inside a given frequency area, as well as estimating the distortion being suffered. Thus, an important aspect of this project is the use of optimisation algorithms, specifically Particle Swarm Optimisation, for which observations and improvements have been made.

1.1 Problem Statement

Spectral distortion is a challenging problem to overcome in Spectral Component Analysis (SCA). In most scientific fields where it is experienced, it has been either addressed only in very specific manners or ignored completely. In fact, the nature of a distortion in a large quantity of applications is not known and, thus, almost always left for future work or disregarded by assuming an undistorted spectrum.

Some of the problems caused by spectral distortions are being addressed. However, this is only on a case by case basis, if they are being resolved at all. Although the fields in which SCA has been applied are quite disparate because of the standardised nature of spectral sensors throughout these fields, the problem of frequency distortion is similar in each of them. Furthermore, the effects of specific external factors have been studied, providing insight in how a deviation of these factors affects the resulting spectrum. Although a generalised solution may be impossible for all types of distortions, a framework that facilitates the resolution for a particular process would benefit a large range of fields.

Such a framework could be developed to extract different types of spectral components that may suffer from different types of spectral distortion. It could serve not only to obtain information required for quality monitoring, but also as a way to counter the influence of sensor distortion in a process. The algorithms, during their application, could either inform the user about the situation and let the user decide how to proceed, or extract the components automatically, regardless of the experienced distortion. When enough information about the distortion taking place has been gathered, the framework could serve as a starting point to create a solution specific for the process.

1.2 Objectives/Aim

The global aim of this research project was to devise an analytical framework, in which tools would be developed that could extract the diverse spectral components from a composite signal, and to do this in a way that is robust against spectral distortions. The framework should be flexible to incorporate new knowledge about the distortions taking place. A list of global objectives for this undertaking were as follows:

1. Create a generalised algorithm that would estimate the concentrations of the underlying components in a composite frequency spectrum, given a set of reference spectra of the components. This algorithm must be:
    (a) Robust against sensor distortion in the measured spectrum. Initially frequency shift and warp should be considered, but this robustness should be flexible to consider other types of distortions.

    (b) Able to identify, with a high success rate, the intensity or concentration of each of the reference spectra.

    (c) Capable of processing reference spectra of different shapes, each suffering from different amounts and types of spectral distortions.
2. For situations where the set of reference spectra may not be available, a tool will be developed that, given a set of sampled spectral data, will estimate the set of reference spectral signatures of the underlying components. This procedure will be:
    (a) Robust against sensor distortion.

    (b) Able to handle diverse types of spectral component shapes.

Chapter 2

Literature Review

The aim of this chapter is to define the type of data known as a spectrum, as it is essential in the methodologies of Spectral Component Analysis (SCA), as well as to define the types of distortions suffered by spectral data, and considered in this project. The reader is then introduced to a review of literature describing different areas in which SCA is applied and the different ways that spectral distortion is handled in each field. Finally, a set of techniques that are currently being used or are relevant to the project are reviewed.

2.1 Spectral Component Analysis Background

A spectrum is a very vague concept that goes beyond the scientific community. A spectrum can be anything that denotes *range*: a spectrum of pitches a tenor can sing, a spectrum of colors in a light beam, or a spectrum of frequencies a signal can be compounded of. In this document, for practical reasons, the concept of *spectrum* will be restricted to the field of frequencies, and, consequently, the wavelengths of such signals.

A frequency spectrum can be calculated in several ways from a signal (which can be a recorded sound, an image, a movie, etc.). The most recognized of all is the Fourier transform, a description of which can be found in literature [24, 13]. For the sake of completeness, a brief description follows.

The Fourier transform is the conversion of one function from the time domain to the frequency domain. It can be divided in two different categories: continuous and discrete. In practice, a great amount of signals are processed in discrete time, thus, for the sake of brevity, only the Discrete Fourier Transform (DFT) will be discussed. However, it is important to note that the Continuous Fourier Transform (CFT) is an essential tool in the area of Pure Maths, and the CFT and DFT have an intrinsic relationship, as the latter is a specific instance of the former [98] (ch. 8).

In a discrete system, the DFT can be considered a type of mapping between domains, in which a point from one domain can be calculated using the values from the other domain. A frequency point from an integrable function sampled in time can be obtained as shown in (2.1).

$$X(k) = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N}kn} \tag{2.1}$$

where X is the DFT of the integrable time-function x; k is the number of one of the various coefficients of X; and N is the number of samples of x. The absolute value of the coefficients in X will provide the frequency spectrum for the duration of the signal. In many cases, expecting a finite signal is not practical, as in live recordings or radio signals from outer-space. A Short-Time Fourier Transform (STFT) can therefore be used instead, in which a signal is divided into *windows* of a specific length that are transformed into the frequency domain. The result is a *time-frequency plot*, which gives the frequency components of the time signal for each window through time. Unfortunately, a trade-off between frequency and time resolution is given by the window size. Using a long window will result in a poor time resolution and a good frequency resolution, and using a short window will produce the opposite.

Another way to obtain a spectral view of a signal is through *Wavelet Transforms* (WT), a detailed description of which can be found in the literature [17, 12]. In summary, it uses a template function, called *mother wavelet*, to create functions serving as dynamic-width filters, arranged to cover the optimal resolution for each of the sections of the spectrum.

Both STFT and WT have been used extensively in almost all aspects of signal processing. Two specific examples being speech recognition and music analysis (see Appendix A). However, it has been found that in these fields *Gabor Analysis* is much more efficient and simpler to use than STFT [26]. It differs from STFT as it obtains a time-frequency plot in a ‘reverse’ manner: essential areas of the frequency domain are calculated using information obtained from relevant parts of the time signal, providing a good resolution throughout the domain.

In the algorithms discussed above, the spectrum is *calculated* from a time-domain signal. There are, however, other cases in which the spectrum is derived directly from the sensor. A good example of this is the case of Raman Spectroscopy [9, 95], where a laser is directed at an object and the light reflected from it is captured. Since some photons are scattered inelastically (an effect known as *Raman Scattering*), the energy (measured as frequency) from the reflected light will be different from the one originating from the laser. If this difference in energy is analyzed correctly, the information contained in the resulting spectrum may be used to characterise the object without the need of invasive methods. Another technique used for spectral sampling is Infrared Spectroscopy, which is used extensively in various fields [28, 5, 16, 40, 21, 41, 51, 36]. The chemical bonds of any material vibrate at specific frequencies and almost all of them vibrate at frequencies inside the infrared frequency range. A photon vibrating at the same frequency that the chemical bond vibrates will be absorbed by the material. This means that if a beam of infrared light is directed at a material, inspection of the energy absorbed at each wavelength will provide information about its chemical composition.

Independently of how the spectrum is obtained, it is often desirable to identify the independent features in it, as they may contain information essential for measuring the quality of a product. For example, the technique has been used for detecting the presence of a malign disease (see Section 2.2.2 for more examples). Analysis of the spectrum is preferable to the raw data itself because, even in the unlikely case in which the raw data is accessible, it is easier and more appropriate to apply pattern recognition and other techniques to it than the raw data from which it is calculated [42].

Many different methods have been used to obtain information from measured spectra. For example, in speech recognition [77] and, more recently, music analysis [70, 72], Mel-Frequency Cepstral Coefficients (MFCC) have been used to obtain features from the spectrum that serve as indicators for recognizing a speaker or a chord progression. Typically, these coefficients are obtained by calculating the DFT of the signal, then warping the frequency domain by using Mel-based spacing (to better accommodate the human ear). The warped spectrum is then segmented into sub-bands, and finally the MFCC vector is calculated by applying a Discrete Cosine Transform to the resulting spectrum. Other ways to obtain these coefficients have been developed [77], but the aforementioned scheme is the most popular.

Another way to analyze a spectrum is to use it as if it were a generic composite signal. There is a great amount of flexibility that comes with this notion (partially applied by Gabor-Analysis-based algorithms [26]), as it is possible to use the different types of component analysis algorithms that are already developed for these kinds of signals. Blind Source Separation (BSS) techniques, such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), have been extensively described [78, 37] and widely used in various scientific fields [9, 42, 114]. These algorithms, through their corresponding methodologies, identify important tendencies inside the examined data and assume them to be the different components which that data is composed of. The detailed description of these algorithms are provided later in this thesis, but it is important to note that these algorithms are very sensitive to frequency displacement.

Another set of algorithms relevant to this study, are the Self-Modelling Curve Resolution methods (SMCR) [55, 22]. These methods have a similar objective to those of BSS, and, although their application is bound to the domain of the spectrum, as their original focus was in the field of Chromatography, they are now beginning to gain popularity in the field of Spectroscopy [55]. To estimate the underlying components in a set of spectra, these methods employ techniques, such as Singular Value Decomposition (akin to PCA), or use the values in locations common in all the sampled spectra. All require the number of components to be known *a-priori*. Hence, a preliminary estimate of this is necessary and this can be carried out in a variety of ways [55, 22]. However, the presence of noise and other influences can make this procedure non-trivial, which suggests that pre-processing steps are required before applying any of these methods in experimental data [55]. As will be seen later, these steps are also sensitive to frequency distortion.

Because of the proven usefulness of both BSS and SMCR sets of techniques, there is an important need for them to be robust against spectral distortion. It would greatly contribute to the development of not only the fields that are currently using them, but to other areas in which these algorithms have not been applied, because of the sensitivity to measurement distortions shown later in this thesis. To do this, however, it is necessary to first understand the origins of spectral distortions, as well as the diverse ways in which it occurs in spectral data. The following section provides insight into these issues.

2.2 Spectral Distortion

2.2.1 Causes and Types of Distortion

A spectrum, as described earlier, can be sampled from various sources. Spectral instrumentation requires precise and complex calibration methods, if it is to to obtain spectra that can be used in spectral analysis [58, 110, 52, 115, 84]. Continuous calibration of spectral sensors is necessary as spectral data acquisition methods, such as Raman, Mass and Infrared, are sensitive to external influences, which include temperature changes [21, 41, 53, 15, 14, 68, 51, 36, 30], pressure changes [14, 51], foreign materials [58, 52] and hidden factors in the material [110].

It can be argued that calibrating a sensor inside a sterile laboratory is all that is necessary to overcome the problem of spectral distortion. However, it has been observed that specific types of sensors respond non-linearly to different materials as well as to interactions between different compounds inside the sample [11], meaning that a single calibration may not resolve the problem. Also, there is a growing need to use different laboratories for sampling similar material and for this a complex and costly calibration method, called "Calibration Transfer", is required [83, 68]. The problem is further complicated by the fact that using the same equipment, in the same laboratory, with the same sampled material, has been shown to produce inconsistent results [3]. In addition, the advent of spectral analysis in real-time quality monitoring [88] highlights the need for aligned spectra in sensors located inside a plant in which conditions are time-varying. All of the described factors illustrate the difficulty of satisfying the fundamental assumption that the spectra being analysed are perfectly aligned.

Examples of different types of distortions are described in the following sections, and are shown in Figure 2.1. This figure shows samples of a mixture of two components. One component is made up of the two peaks in the range of 200 to 350 Hz. and the other component is made up of the peaks in the range of 550 to 750 Hz. Each figure shows two samples, one with and another without distortion.

The distortions observed in literature can be categorised in two families: Global and Local distortions. Global distortions are those that affect the spectra uniformly. Local distortions are those that affect each component differently. This distinction is made as a way of simplifying the task of component extraction in the project. However, it is important to note that there are many types of distortions whose behaviour depend on the type of spectral sensor being used, the type of material being inspected, the external factors involved, etc. Hence, the following should not be regarded as an extensive list of all the types of distortion a spectrum may exhibit, as it only describes those that were found to occur the most regularly, have been the most studied, and are able to be detected by the naked eye.

2.2.1.1 Global Shift

A global shift is a linear displacement of all the spectrum to the right or the left. As shown in Figure 2.1a, the spectrum in this example is shifted uniformly to the right.

Global shift is probably the most common type of distortion and the causes vary widely depending on the type of sensor being used. In IR spectroscopy, for example, a global shift may occur as a result of adding a component that reacts specifically with the active component [36] and by changes in pressure and/or temperature [51, 68]. In mass spectrometry, however, temperature is the relevant factor [11, 54].

2.2.1.2 Local Shift

A local shift is a linear displacement of each component that is independent from each other. In Figure 2.1b, the peak in the lower frequencies is shifted towards the right, while the peak in the higher frequencies is shifted towards the left.

Local shift is prevalent but relatively easy to confuse as a global shift, because the components may be shifting a similar but not identical amount in the same direction. In [35] a series of spectra where obtained by sampling a mixture of ice and a type of carbon molecule, using UV rays at different temperatures. A shift, as expected, occurred in each spectrum, but the spectral signatures of both components shifted by different amounts. Different compounds react differently at varying temperature, so it is expected that their spectral signatures will behave differently.

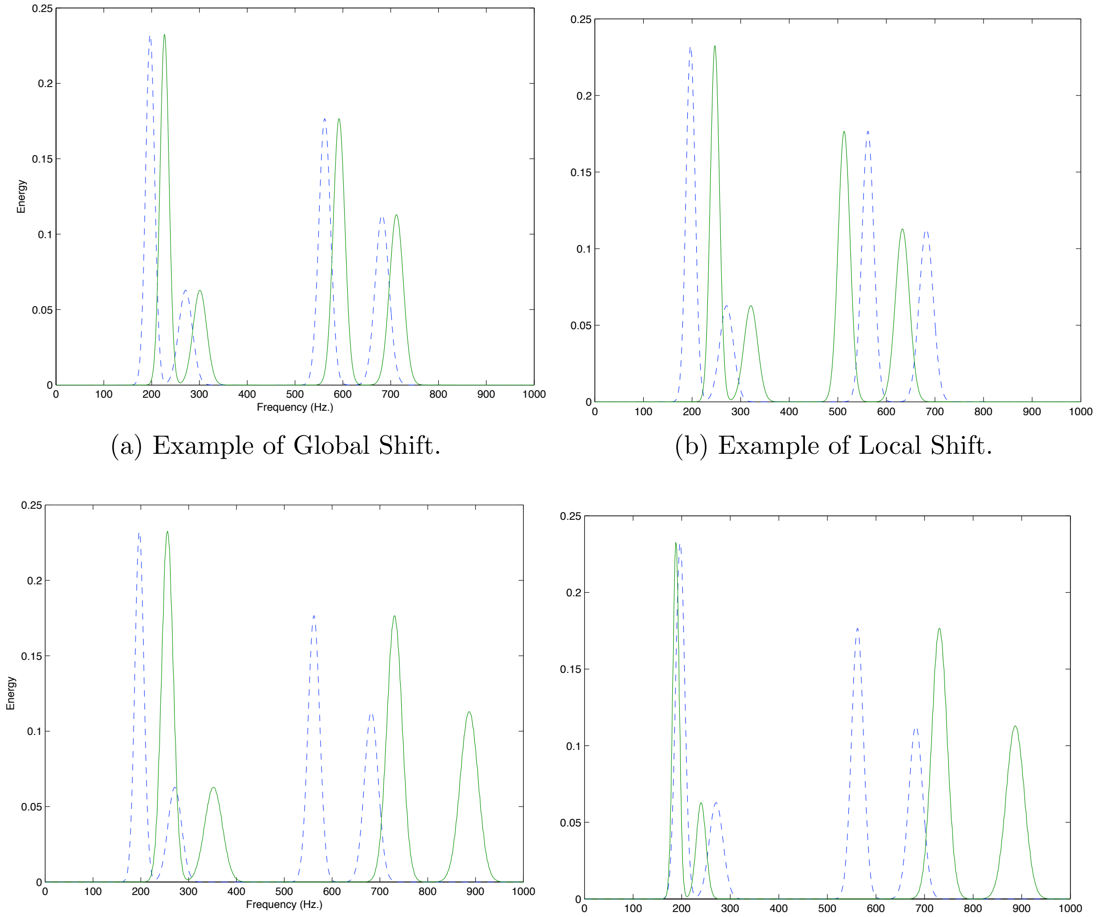
2.2.1.3 Global Warp

A global warp is a type of “stretch” or “compression” of the whole spectrum. In Figure 2.1c, all the spectrum is stretched by 30% which can be observed by the increased width of all the peaks. Global warp has been identified in both Raman [30] and mass spectroscopy [43], caused by temperature changes.

2.2.1.4 Local Warp

A local warp is the “stretching” or “compression” of the components independent of each other. In Figure 2.1d, the component in the lower frequencies was compressed by 30%, while the component at the higher frequencies was stretched by 30%.

Local warp is not particularly common in Industry. However, it is the primary type of distortion found within the field of Music Information Retrieval, because of the nature of a musical note [61, 8]. When it is not “calibrated” or not in tune, the spectral signature of a note is warped. If various de-tuned notes are played at the same time, as the calibration of each is independent of each other, each note will warp by different amounts.



(a) Example of Global Shift.

(b) Example of Local Shift.

(c) Example of Global Warp.

(d) Example of Local Warp.

Fig. 2.1: Examples of Different Types of Distortions

The types of distortions described here were able to be simulated by consideration of the physical attributes of spectral sensors and from generic chemical compositions that were gathered from various sources during this study and are described in the following section. A detailed description of how their simulation is carried out is provided in Appendix C.

2.2.2 Applications Observed

Spectral measurements are being used with increasing frequency in many different areas of science and engineering. Spectral measurements are typically used to extract information from industrial compounds [116]; or in the Signal Processing field where radio and/or television signals are emitted, recorded and analysed [80, 1]. However, it is also common to find them being used in other areas as well. The following sections describe examples of where Spectral Analysis has been used and its importance in each field. Additionally, some limitations and/or advances in each field, specifically in the area of robustness to distortion, are described.

2.2.2.1 Pharmaceutical Industry

Spectral measurements are beginning to be used to analyse the composition of drug tablets, as discussed in [116, 23, 103, 28, 32, 31]. A pharmaceutical tablet can be analysed by comparing the expected frequency spectra of the ingredients with the composite spectra measured from the tablet. From this analysis it is possible to determine the distribution and concentration of the ingredients. This information is crucial if pharmaceutical companies are to satisfy the increased quality assurance that is being demanded from the US Food and Drugs Agency. However, extracting this information is a challenging task, even when not considering spectral distortion.

A difficulty introduced by spectral distortions is the mis-alignment between the sampled and the reference spectra. There is a vast amount of literature dedicated to calibration [83, 103, 28, 31], because of the need for this alignment. There have even been approaches that have considered calibration variations based on temperature-dependent models [68, 14, 15, 41]. In the approach carried out in [68], where NaCl is the mixture tested at different temperatures, the authors admit that “the simplicity of the calibration transfer in this study can [...] be attributed to the fact that the NaCl solution is a simple system and that the spectral shift caused by an increase in temperature [...] is small”. Meaning that a model heavily relies on the information it was based upon, and such models can only be used reliably if this information is recorded (in this case, temperature) for each sample, and this information is not available in many applications.

Modelling techniques, such as that presented in [68, 14, 15, 41], provide insight into how to simulate a spectral distortion. This information is essential in the development of the algorithms proposed in this thesis, and studies like these are important when incorporating new types of distortions into the proposed framework.

For example, the modelling technique used in [15] was successfully applied to a spectral data set that suffered from temperature variations. The effects of the temperature change in the measured spectra, given the specific spectral sensor and material, was modelled well as a second-order system. If this spectral sensor and material would be used again, the information provided in [15] could be used to simulate the distortion created by temperature changes.

2.2.2.2 Biomedicine

There is a vast amount of literature describing the situation in which the components of a particular sample are contained in a mass or Raman spectrum. Work has also been conducted in finding if certain components may be of use for the detection of a particular pathology.

For example, in [89], the authors discuss the use of Particle Swarm Optimisation (PSO), using Support Vector Machines (SVM), to identify the biomarkers that were the most sensitive in detecting the presence of liver cancer from a mass spectrum. They applied Wavelets as a form of smoothing the spectrum, noise reduction schemes, and a Simple Peak Finding (SPF) algorithm to obtain viable biomarkers from the spectra. The latter is fully explained in [20], where they deal with the detection of breast cancer from the mass spectrum of ductal fluid samples. The algorithm lists where all the potential peaks are in a mass spectrum, which indicate the presence of a certain element or substance in each sample.

The results presented in [89] were reasonably successful and the PSO-SVM algorithm was able to detect the importance of the same biomarkers independently of the subset of biomarkers given to it, indicating great strength in using PSO as an optimisation algorithm. However, the authors admit that they did not perform any sort of procedure to overcome a misalignment in the spectra and left this topic as future work.

A major benefit of Raman spectroscopy is that sampling can be carried out in a non-invasive manner. For instance, in [95], the authors used Raman spectrometry and Principal Component Analysis (PCA) to find the presence of oesophagus cancer in a patient. This form of cancer is treatable at an early stage if detected, but can be mortal at advanced stages. They showed that detecting changes in specific frequencies in the Raman spectrum was a good indication of the presence of this form of cancer in patients.

The authors of [114] aimed to identify an unknown tissue (liver, kidney, lung, or gland) from the spectrum given by a non-invasive Raman sensor. The Raman spectra of several normal and tarnished samples were inspected and it was found that the spectrum was not considerably affected by the tarnish. However, to overcome the effects of the fluorescence background, the linear-quadratic baseline was subtracted from the data. In addition, to eliminate the influence of scatter effects of particle size, crystal state, etc. on the frequency response and measure signal intensity, the spectra was altered to have a mean of zero and a standard deviation of one. PCA was then applied and provided four distinct clusters, one for each type of tissue. A sample could then be classified depending on the cluster it belonged to.

Results presented in the references presented here suggest that the spectrum can be influenced by factors independent of the sample, and that these factors need to be considered if machine learning or other algorithms are to be applied to it. Unfortunately, little mention of robustness to distortion was mentioned in [20, 89, 95, 114] and no way of countering it has been proposed.

2.2.2.3 Textile Industry

Raman spectroscopy has recently gained interest in the Textile Industry. In [86], the authors compared diverse ways of applying Raman spectroscopy when judging the orientation, shrinkage and quality of silk fibers, without the need for destructive testing. The ratio between the magnitude of two wavelengths in the spectrum was used as a measure of the fibers orientation. Although no component analysis was employed, it was found that if the measurements were taken in wet-deformed circumstances a wavelength shift and overall shape-change occurred. The authors, with the measurements discrepancies as evidence, concluded that the fibers internal characteristics changed because of the wetness. While this may be true to some degree, it is also possible that the measurement itself may have been influenced by the presence of the wetting substance, thus deforming the spectrum shape. If this is true, there is a need for a analytical tool that is robust against these types of distortions to obtain an accurate analysis.

2.2.2.4 Pigment Identification in Paints

In [9], the presence of a library of pigments inside a mixture was identified using Raman spectroscopy and Fuzzy Logic. The correlation between a reference signal, representing a specific pigment, and the unknown spectrum, representing the mixture, was used by the system to propose the pigment as a good candidate to be part of the mixture. If more than one candidate was proposed, a new library was built upon them and graded again by a second Fuzzy Logic system. The combination that was closest to the unknown spectrum was then proposed to be the one describing which pigments were in the mixture.

The system appeared to work well, however it could only recognise mixtures comprised of up to two pigments in them, and the task to recognise more pigments was left for later work. There was no mention of distortion robustness, but, because the correlation coefficient was the only decision parameter used by the system, it could potentially be modularised to use another measure that is robust against global shift, which seems to be the present distortion.

2.2.2.5 Face Recognition

The use of spectral analysis in face recognition techniques is not common. However, the authors of [42] showed that using the Fourier Transform of photographs of faces, instead of the actual raw data, resulted in better performance. The recognition process used a combination of features obtained from Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA), which were then used by a Radial Basis Function Network (RBFN). It is noteworthy that some ilumination problems were encountered, and the resolution of these was left for future work. The authors did not expand on the problem, but it is reasonable to suspect that the lighting influenced the Fourier transform which must have changed significantly. A tool robust to this shift would grant even greater performance to this algorithm.

2.2.2.6 Music

Musical Information Retrieval (MIR) has attracted significant attention in the scientific community, as many of the identified research areas in this field are very challenging, and can arguably be applied in other scientific fields.

The acquisition of musical information via a spectrum is based upon the identification of musical notes [71] which, in turn, are identified by specific frequencies [2]. However, a musical instrument can be de-tuned from a standard configuration because of human error or temperature changes. Meaning that every note may be subject to different degrees of de-tuning, resulting in a spectrum that is highly arduous to correct.

It is necessary, then, to identify each note as an individual component in the recording, a process known as Polyphonic Music Retrieval [61, 25]. Unfortunately, this technique uses methods that rely on the specific nature of the note spectral signature, making them non-generalisable, and, thus, difficult to export to other scientific fields.

However, given the relative ease in which musical data can be obtained[1] and the complex type of spectral distortion that it suffers, the Music arena provides a practical testbed for the algorithms developed in this thesis.

An important amount of time was invested in this field during the development of this project. And, although, irrelevant to the other fields reviewed in this chapter, the methods proposed in this project did provide good results when applied to musical recordings. A further description of these results can be found in Appendix A.

2.2.3 Previous Work with Spectral Distortion

The identification of underlying components within a spectral measurement is a complex problem that is typically compounded by the fact that frequency displacement often occurs as a result of poor sensor calibration and temperature changes [117, 21, 41, 53, 15, 14, 68, 51, 36, 30, 43]. It can be argued that an obvious method to tackle this problem is to re-calibrate the spectral sensors frequently. In [117], the authors briefly describe several ways to calibrate gas sensors, specifically by either exposing them to the gas that is causing the shift and normalising from there or by comparing the result with that of a reference gas. However, they assert that continually calibrating gas sensors is too expensive and time-consuming.

The authors of [117] also proposed a new method to recover a shifted signal. The method is based on the following assumption: if a wavelet decomposition is performed on a signal that has a shifted spectrum, information regarding this shift is contained in the wavelet coefficients at the highest level (i.e. the lowest frequencies). In other words, if the coefficients at the highest levels are disregarded from the wavelet basis tree, the restored signal would be un-shifted with very low information loss. The method was shown to be accurate when applied to a signal with an artificially created shift.

Another example of a shifted-signal-recovery algorithm was described in [62], where a Frequency-Shift-Invariant Discrete Wavelet Packet Transform of a signal was calculated by building a basis tree for the wavelet decomposition and its

[1] The author of this work is a music hobbyist, as is his supervisor.

coefficients for each type of frequency shift possible. Then, by using a depth-first binary-tree search and a given cost function, the optimal frequency shift was found. With this optimal shift, a basis tree was created that represented the unshifted version of the signal. The algorithm was shown to be efficient, compared to other best-basis algorithms.

In [7] a wavelet-packet-based transceiver was simulated and tested against Carrier Frequency Offset (CFO), where the frequency of the carrier signal was not at the predisposed one. The main objective of the paper was to estimate a model in which the error and the performance of a transceiver, based on Wavelet Packet Modulation (WPM), could be evaluated. The model was found to successfully estimate the error, as well as being useful in deciding which tree basis to apply, such that the WPM became robust against a specific CFO.

Although the methods proposed in [117, 62, 7] are useful, the work undertaken was based on transforming the time-domain signal in various manners, which implicitly assumes access to a possibly none existent signal. They do provide insight of one of the types of distortions investigated: Global Shift, as it was assumed that the displacement suffered by the signal was the same across it. However, no other type of distortion was considered.

Another possible approach to cope with distortions is to align the spectral data by using using Dynamic Time Warping (DTW). This method is used primarily in Speech Recognition to stretch or shrink one data timeline to meet the length requirements of another in such a way that their sample points meet and, thus, can be analysed simultaneously. If the data of one timeline is of a different scale to that of another, which is normal in, for example, Batch Processing, it is possible to overcome it by applying a Wavelet Transform during the process [66]. In theory, this approach could be applied to the shifted spectrum to scale it to its unshifted state. However, the degree of shift would need to be known beforehand, and, as discussed earlier in this thesis, this information is rarely available accurately.

Furthermore, if two spectra are of the same size, DTW can be applied to align them with each other, but, as will be seen later in this thesis, pre-aligning with DTW can deteriorate the results obtained using spectral analysis tools. This is because DTW, as well as other aligning algorithms, modify the overall shape of the spectrum during the alignment procedure, severely complicating the component extraction process afterwards.

Aligning procedures could be simplified by automatically detecting the shift from each spectral sample. *Cross-correlation* is widely used in Digital Signal Processing [24] to detect shifts in a signal by comparing it to other 'reference' signals. Its main objective is the localization of high correlation points throughout the length of the signals being compared, providing a practical way to find the amount that a signal is shifted relative to another, as well as a first step towards a measure of 'similarity' between the two signals regardless of the shift. For discrete functions, a Cross-Corretion vector is created by artificially shifting *linearly* one signal, and calculating the correlation coefficient for each shift using (2.2). The correlation coefficient shown is also known as the *Pearson product-moment coefficient*.

$$CC(k) = \frac{\sum_i (x_i - m_x)(y_{i-k} - m_y)}{\sqrt{\sum_i (x_i - m_x)^2}\sqrt{\sum_i (y_{i-k} - m_y)^2} \tag{2.2}$$

where x and y are the two discrete signals being compared; k is the point at which y is being linearly shifted and the correlation is being calculated; m_x and m_y are the mean values of x and y, respectively; and CC is the resulting cross-correlation vector. From this vector it is trivially known if the examined signal is shifted or not by observing if the highest correlation point is not located on $k = 0$.

An application of cross-correlation is explained in [49] where the authors used cross-correlation to inspect the relationship between the spectrum of the image of handwritten arabic numbers and the spectrum of the expected patterns of such numbers. A set of features were obtained which were effective at identifying both well-written and incomplete numbers. Another field that uses the Pearson correlation coefficient is Genetics [79], where a correlation statistic could be used to discern different types of gene expressions, producing fewer false-positives than K-Means and other segmentation techniques.

One major disadvantage of cross-correlation is that it assumes that the shift between the signals is uniform. Although this may not be sufficient, in this project it was found that it is a good starting point towards a solution to the problem of analysing distorted spectra. Specifically, the shift operator could be modularised to be any type of distortion. An important aspect of the Pearson coefficient is that it does not consider the magnitudes of the two compared signals, just their respective shapes, which, in Spectral Component Analysis, is of great importance.

It is important to note that coping with any distortion is not the only issue that must be overcome. Understanding the source and influence of a distortion is an important step that needs to be carried out before simulating its effects in the spectral data. By using this knowledge in an appropriate way, existing algorithms can benefit by artificially modifying the data it is using. For example, the authors of [69] introduced an enhanced version of Wavelet Decomposition (WD) which was robust against local orthogonalities, which cause information loss and is the prime reason why WD is considered unreliable for spectral analysis. The methodology involved artificially shifting the phase of each wavelet such that the orthogonalities were avoided, while keeping track of the wavelet shifts. The methodology was applied to a magnetic drive system, where the WD variation was used to monitor undesirable energy consumption, which presented itself as high frequencies in the spectrum. The primary objective of the experiment was met when an increase in the reliability of the WD was observed, however it also presented an interesting use of shift-analysis: fault detection and diagnosis. It was carried out by modifying the signal in a way that was congruent with the physical characteristics of its application.

2.3 Conclusion

Spectral Component Analysis (SCA) is of interest in various fields both in Academia and Industry. Spectral data is essential to the diverse range of SCA methodologies. This chapter explained how the sampling procedures with which spectral measurements are obtained are plagued by external influences that modify and distort the resulting spectra, degrading the performance of SCA techniques.

The various sources of distortion are dependent on the sampling procedure and the material being used. The effects that two of the most frequently observed distortions that occur in spectral measurements were identified as frequency displacement (shift) and spectral warping. These two types of distortion were then described in detail.

This chapter explained that that there is an important requirement for a spectral analysis tool that is robust to distortion. Such a tool would be beneficial in various fields, including the Pharmaceutical Industry, Biomedicine, Textile Industry, Pigment Identification, Face Recognition, and Music Information

Retrieval. Unfortunately, in most of the fields discussed here, the spectral distortion topic has been either addressed only briefly or incompletely, left for future work, or just completely ignored. However, some modelling techniques for characterising distortions were found in the literature, and these models provided some insight into how the two chosen distortions, as well as others, affect the shape of spectral data.

A tool that meets the objectives of SCA methods whilst being robust against a pre-defined set of distortions would contribute greatly in the aforementioned fields. It would also open up possibilities of application of SCA in fields in which it was not applicable because of the distortions encountered.

Chapter 3

Background to Optimisation

In later chapters, a Spectral Analysis Framework, as well various component extraction methods are proposed. These methodologies are formulated as optimisation problems. Hence, the topic of Optimisation needs to be discussed.

An optimisation algorithm aims to find the global optimum of a given problem. Such a problem is defined by what is called an *objective function*, that involves all the variables considered in the problem. The objective function is such that its minimum or maximum value is calculated using the values of the variables that are the solution to the problem.

An objective function also describes a solution space, which are all the solutions given by the objective function using all the possible combinations of values of the variables considered. The shape and type of a solution space is important, as different optimisation algorithms are not able to cope with certain types of solution spaces. For example, a solution space with one big hill, also known as *convex*, is desirable, as simple optimisation algorithms such as Hill-Climb can find the optimum value. However, solution spaces with several hills are prevalent in practical situations, and using a Hill-Climb approach may lead to a solution that is non-optimal. Such a solution is known as a *local optimum*.

In the following sections, a short review of the different approaches for solving optimisation problems is presented.

3.1 Gradient Descent-Based Methods

Used solely with continuous solution spaces, Gradient Descent-Based methods calculate the gradient of the solution space at its current state to choose the direction of the next 'jump' [91]. Which implies:

$$x \leftarrow x - \alpha \nabla f(x) \tag{3.1}$$

where x is the current state of the search, f is the solution space defined by the problem, $\nabla f(x)$ is the gradient in the state x, and α is a small constant that defines the magnitude of the 'jump'. The value of α and its variation is the differentiator to many alternative versions of gradient-based optimisation. Originally, the value of α was tuned specifically to work with a given solution space; not large enough to overshoot the global maximum, and not small enough to get stuck in local hills. Other versions vary the value of α depending on the variation of the value of the current state, $f(x)$, by increasing it if $f(x)$ slows down and vice versa [91]. Another approach 'nudges' the current state to another random location if $f(x)$ has stopped changing [91], a tell-tale sign of it getting 'trapped'.

Other implementations employ different methods for calculating the direction of the next jump. A very popular technique is the Newton-Raphson method (N-R), introduced in the 17th century [91]. N-R assumes that the global optimum of f is located at a state in which the gradient $\nabla f(x)$ equals 0. To locate the roots of $\nabla f(x)$, the second partial derivatives of f are calculated. A Hessian matrix, H, of second derivatives is calculated from:

$$H_{ij} = \partial^2 f / \partial x_i \partial x_j \tag{3.2}$$

in the case of a two-dimensional solution space, where x_i and x_j are the two dimensions. A Newton-Raphson step is calculated as follows:

$$x \leftarrow x - H_f^{-1}(x) \nabla f(x) \tag{3.3}$$

However, if the dimensionality of the solution space is increased, the computational time spent in the calculation of H at each step increases exponentially.

The successful application of these methods is bound by having intrinsic knowledge of the solution space, as the gradient value at each step needs to be known. Such a value may be unavailable or difficult to obtain, as in non-continuous solution spaces or non-linear problems. Another concern is that the algorithm may become trapped in local minima. Enhancements to the algorithm which alleviate this problem have been proposed but no global solution has yet been found. If this method is applied to sets of continuously-changing data, such the data sets encountered in this study, then there would be a need to re-tune and re-validate the algorithms. Thus, approaches that require the least amount of information from the data sets, such as Black-Box Oriented algorithms, were concluded to be better suited for this project.

3.2 Black-Box Algorithms

Black-Box Oriented Optimisation algorithms aim to find the maximum or minimum value of a given problem given only a limited amount of knowledge of the system. To compensate for the lack of knowledge, stochastic measures and randomness, as well as simulations of phenomena observed in Nature, are used to extract information from the data. Unfortunately, the random nature of these algorithms also introduces a certain amount of unpredictability in the algorithms, that can range from not knowing if the value found is in fact the global optimum, to not knowing if the algorithm is going to converge at all. However, because of their generalised approach, they have few restrictions on what type of solution space can be optimised. These algorithms have therefore been shown to be useful in many scientific fields [89, 4, 83].

3.2.1 Simulated Annealing

Simulated Annealing was introduced in 1985 by Černý [10], and can be considered to be a random hill-climb method. A normal hill-climb is the discrete version of the Gradient Descent method, although less sophisticated. Also known as a greedy or naive search, the normal hill-climb, given its current state, 'jumps' to the neighbouring state with the highest value that is higher than its current state; if it does not find such state then the search is over. Simulated Annealing adds two concepts to this approach:

- When the hill-climb has reached a minimum, the search will be re-started by 'nudging' the current state to another random location.
- The 'energy' in which this 'nudge' takes place is reduced throughout the process, until no energy is left and, therefore, no 'nudging' can occur.

Annealing is the slow cooling-down process of a high-temperature material, used to harden metals. A simulated annealing process works in the same terms, as the initial high 'energy' or randomness of the search is forcefully reduced, resulting in a search process that, given an appropriate amount of initial energy, will eventually converge to the global optimum [91]. It is important to note that this process employs several sequential hill-climbs, which means that it will converge, but only after a long time, which is defined by the initial amount of energy introduced.

3.2.2 Population-Based Algorithms

Population-Based algorithms are a set of methods that locate a global optimum by sharing information between several sub-searches, that are taking place at the same time. The methodologies with which the information is shared and how it is used differ from one algorithm to another, but the principle of the global optimum being found by a multitude of small searches is common to all of them.

3.2.2.1 Genetic Algorithms

The Theory of Evolution, introduced by Charles Darwin, defined how Nature finds the best path for a species to adapt to the environment. Genetic Algorithms (GAs) simulate such processes by creating a population of *genomes*, each being a set of *genes*, that represent a possible solution to the optimisation problem [91]. A sub-set of the best genomes of a population are chosen to create a new population, from which another sub-set is selected, and the process is repeated until the solution converges to a global optimum.

A new genome is created by two elder genomes exchanging genes, a process known as *crossover*. A random point inside the genome is chosen, and all the genes below the point are crossed over between the genomes. The new genomes can also randomly suffer from *mutation*, which is the change of the value of a gene inside the genome.

There are several variations of GAs, differing in the process by which crossover occurs, which elder genomes are transferred to a new population and the criteria used to judge convergence. Parameters which need to be specified within a GA include the population size, the amount of elder genomes used, the rate of mutation, and the structure of the genome.

The theory underpinning the GA approach is that every new population will be fitter than the previous one, and that in a certain amount of iterations, the population will evolve into the best state possible. It can be shown that, given a properly structured genome, the evolution process will, over time, create an increasing number of close-to-optimal genomes [91]. However, the successful convergence of a GA relies heavily on how the genome represents the variables in the objective function and creating a properly structured genome is not a trivial task.

The GA approach has gained popularity in several fields and has been applied successfully in many applications, such as multi-instrument spectroscopic calibration [83] and solving non-linear and complex problems [4].

3.2.2.2 Differential Evolution

Differential Evolution is similar to Genetic Algorithms in that it uses the information of two members of a population to create another. However, their implementations differ, as the evolution of the members of the population, called *parameter vectors*, in Differential Evolution is spatial, not generational [102]. This means that a parameter vector defines a 'location' in the solution space, and this location is modified by the information of other randomly selected vectors.

The basic principle for every vector inside the population is: the weighted difference of two randomly selected vectors is added to the current vector and if the resulting vector provides a better objective function value then the vector is replaced by the new one. This process is repeated for every vector in the population, until the best vector of the population converges.

Having a small amount of a parameters to tune, a simple framework is provided from which various versions of Differential Evolution have been developed. One version applies a crossover function before the weighted difference is added onto the vector, with the intention of diversifying the population [102]. Another version weights by a tuneable factor the best vector of the population when calculating the weighted difference, resulting in the population moving towards one point [101].

Differential Evolution has been applied in real world applications, such as in the design of an IIR filter [100], with good results. However, even though the algorithm is considered simple, the modification of the vector can be considered as a type of 'mutation', and, considering the now prominent use of the crossover variation of the algorithm, it is difficult not to see the similarities between Differential Evolution and Genetic Algorithms. Therefore, as with GAs, attention needs to be directed towards the structure of the vector/genome, so that convergence to a global optimum is achievable.

3.2.2.3 Particle Swarm Optimisation

Particle Swarm Optimisation (PSO) is a search algorithm that was introduced by James Kennedy and Russell Eberhart in 1995 [60]. It is based on the inner social behaviour of a flock or a school to find food.

A group of particles (or *swarm*) is randomly placed inside the solution space defined by an objective function. Each particle can 'move' towards different locations in the solution space, and each location is graded by the objective function. Every particle is able to remember the best-graded location it has found, and makes it known to a pre-defined number of neighbours. During each iteration, the velocity of each particle is modified by considering the best-graded location found by the particle and the best one found by its neighbours, i.e.

$$V_{x_i}(k+1) = V_{x_i}(k) + 2r(pbest_{x_i} - present_{x_i}) + 2r(gbest_{x_i} - present_{x_i}) \tag{3.4}$$

where k is the iteration index, V_{x_i} is the velocity of the particle in the direction x_i, $pbest_{x_i}$ is the best-graded location in direction x_i found by the particle, $gbest_{x_i}$ is the best-graded location in direction x_i found by the neighbours of the particle, $present_{x_i}$ is the current location of the particle in direction x_i, and r is a stochastic factor that prevents several particles being in the same location. The inclusion of r makes the particles 'spread out' in an *area*, rather than focus on a single point. This improves significantly the chances of finding the true global optimum. All the V_{x_i}s of all the particles are modified according to (3.4) until the best-graded location found by the whole swarm converges or the maximum number of iterations is exceeded.

PSO can incorporate the concept of a *time-decreasing inertia* [96], which forcefully decreases velocities later in the search. This technique is in fact an implementation of the temperature decrease of a Simulated Annealing search [10]. Applying it to PSO results in an initial exploration of the whole solution space, pinpointing the area where the global optimum is suspected of being located, and then evolves into an exploitation of this area for the remainder of the search. It has been shown that using time-decreasing inertia in PSO provides faster and more accurate results than without [96].

In PSO, each direction is treated individually. Therefore, it can be argued that the structure of the particle against the objective function is not relevant. However, as discussed with Differential Evolution, population-based algorithms need to consider how the particle/vector/genome is structured so as not to deviate significantly from the definition of the objective function.

3.3 Choosing an Optimisation Algorithm for the Proposed Framework

The Spectral Component Analysis Framework proposed in this thesis needs to incorporate different types of distortions. Thus, it is necessary for the optimisation algorithm to be able to cope with complex and unpredictable solution spaces, with very little initial information from the process. In this way, if new information from the process is found, it can be used to enhance the analysis, without requiring to rebuild the analytical tool. Thus, a Black-Box Oriented Algorithm is best suited. The framework, in fact, was designed such that any of the discussed Black-Box Optimisation algorithms were able to solve it. Nevertheless, to avoid unnecessary redundancy and confusion, one algorithm needs to be chosen as the one to be used whenever an optimisation problem is at hand.

In [111], it was shown that, given a set of optimisation problems, and a set of optimisation algorithms, the average performance of each of the algorithms was identical. In other words, for a specific problem, with no *a-priori* knowledge, all Black-Box Oriented Algorithms will have the same probability of convergence [111]. It does not mean, however, that all the algorithms will perform the same or reach the global optimum in the same amount of time. What it does mean is that choosing an appropriate optimisation algorithm based on its ability to identify the global solution is not appropriate.

The process of choosing an algorithm therefore needs to be based on other factors, such as ease of use or tuning complexity. To this effect, Particle Swarm Optimisation was chosen based on the relative ease with which it can be implemented and modfied. The latter is important because, by the definition of the problem being tackled, different types of spectral distortions need to be easily incorporated.

3.4 Observations on Particle Fidelity

During the implementation of Particle Swarm Optimisation in this work, the importance of particle fidelity was observed. When a particle 'asks' its neighbours for the best location they have found, the different variations of the algorithm do not specify if the neighbours are those closest to the particle at that precise moment, or if they are the initial neighbours of the particle when the algorithm was initialised.

A particle, in the initial stages of the algorithm, can gather its closest neighbours, consider them as its 'family', and only ask them throughout the search, rather than relying on the findings of the closest neighbours it has at each iteration. In this thesis, this behaviour is referred to as *fidelity*.

To test whether particles being faithful has any impact on the performance of the algorithm, two variations of PSO were implemented: one in which the particles relied on the findings of the closest neighbours at each iteration (Neighbour PSO) and another in which the particles remained faithful to their respective families (Family PSO). Each method was applied 100 times to each of four different two-dimensional optimisation problems. The optimisation problems studied in this section have been popular in testing optimisation algorithms [82, 76].

Optimisation Problem 1: De Jong's Sphere. [56] This is a convex solution space that features a single wide peak. It is a relatively simple problem that is applied here as a frame of reference, as all optimisation algorithms are expected to be able to solve it. It is described as:

$$Z = -\sum_{d=1}^{D} x_d^2 \tag{3.5}$$

where D is the number of dimensions (2 in this case) and x_d is a dimension. The maximum is located at $x_d = 0$ for every d, with a value of 0. A graphical representation of De Jong's Sphere is shown in Figure 3.1a.

Optimisation Problem 2: Rastrigin Function. [105] The Rastrigin Function features many local minima, but only one global optimum at its center. As shown in Figure 3.1b, the shape of the Rastrigin Function is similar to that of De Jong's Sphere, but much 'bumpier'. It is described as:

$$Z = -10 \cdot D - \sum_{d=1}^{D} x_d^2 - 10\mathbf{cos}(2\pi x_d) \tag{3.6}$$

It was modified such that the optimum was the maximum value of the function, with a value of 0, located at $x_d = 0$ for every d.

Optimisation Problem 3: Schaffer F6 Function. [93] As shown in Figure 3.1c, this function simulates a set of 'waves' similar to those that appear after throwing a rock into a pond. All the points in the top of each wave have values very similar to that of the global optimum, and the closer the wave is to the center, the closer these values are to the global optimum. This translates to having an infinite number of local minima to avoid. The function is described as:

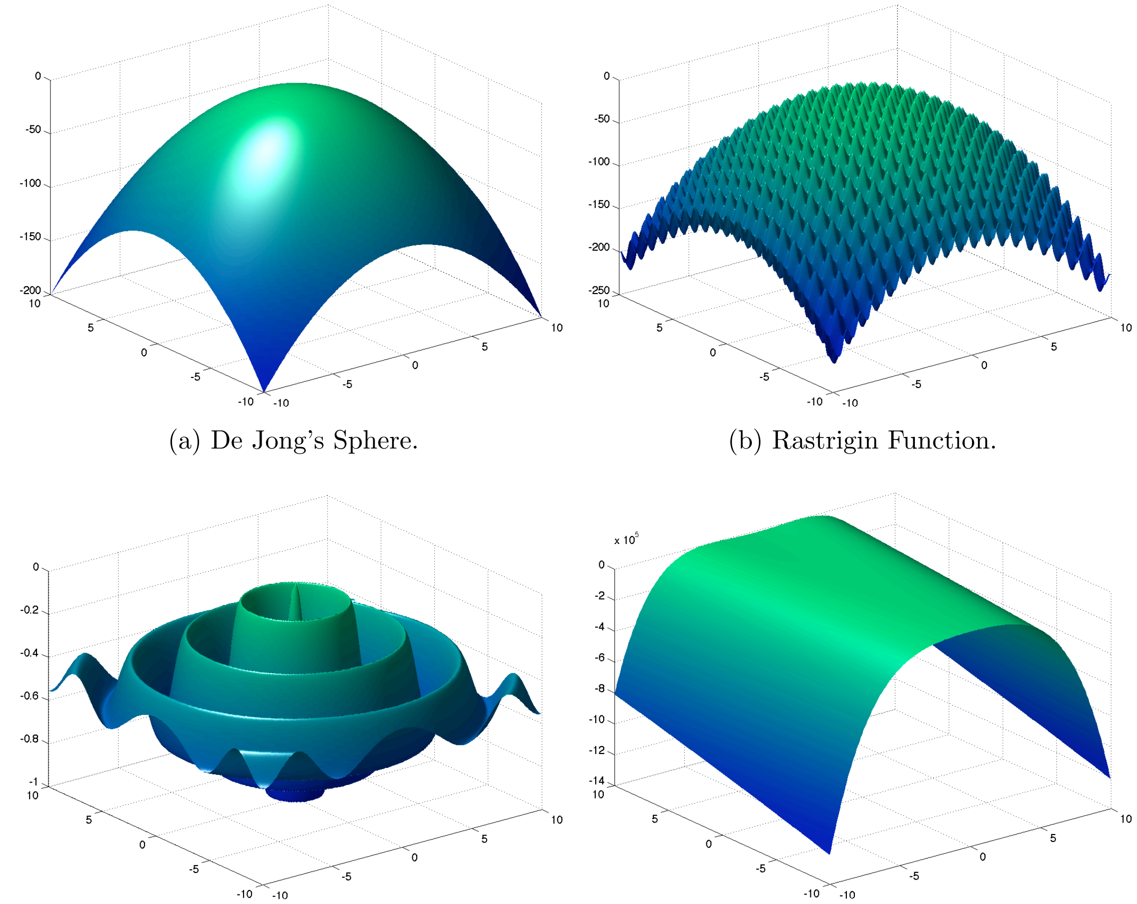
$$Z = -0.5 - \frac{\mathbf{sin}^2\sqrt{x_1^2 + x_2^2} - 0.5}{1 + 0.01(x_1^2 + x_2^2)} \tag{3.7}$$

It was modified such that its optimum was the maximum value of the function, with a value of 0, located at $x_d = 0$ for every d.

Optimisation Problem 4: Rosenbrock Function. [94] The Rosenbrock function aims to 'trick' the optimisation algorithm into finding an area of local minima from which it is difficult to 'jump out' of and find the global optimum. It is described as:

$$Z = -\sum_{d=1}^{D-1} 100(x_{d-1} - x_d^2)^2 + (1 - x_d)^2 \tag{3.8}$$

It was modified such that its optimum was the maximum value of the function, with a value of 0, located at $x_d = 1$ for every d. A graphical representation of this modified version of the Rosenbrok Function is shown in Figure 3.1d.



(a) De Jong's Sphere.

(b) Rastrigin Function.

(c) Schaffer F6 Function.

(d) Rosenbrock Function.

Fig. 3.1: Graphical representation of the solution spaces used in testing.

3.4.1 Results when Relying on Fidelity with PSO

In the tests, the number of particles in the swarm was set to 10, with each particle having 5 neighbours/family members. The optimal value in all the problems was 0, and, because the accuracy of the MatLAB workstation was of $2.2204 * 10^{-16}$, any point found with a fitness below this value was considered to be the global optimum and given the value of 0. Inertia was reduced from 1 to 0.2 in the first 300 iterations of the search. The number of iterations was limited at 30,000; if the maximum number of iterations was reached, the best solution found at this time was returned as the solution of the search.

The results when testing Neighbour PSO are shown in Table 3.1, and the ones when testing Family PSO are shown in Table 3.2.

	De Jong	Rastrigin	Schaffer F6	Rosenbrock
Mean Error from Optimum Value	0	0	0	0.4342E−7
Mean Error from Optimal Variable Values	0	0.0021E−6	0.0021E−6	0.7459E−6
Mean Number of Iterations	559.7	1189.4	604.8	6096.1
Standard Deviation of Number of Iterations	24.9	4138.8	106.9	3461.6

Table 3.1: Results using Neighbour PSO.

	De Jong	Rastrigin	Schaffer F6	Rosenbrock
Mean Error from Optimum Value	0	0.1776E−7	0	0
Mean Error from Optimal Variable Values	0	0.0865E−8	0.2230E−8	0
Mean Number of Iterations	617.8	966.7	958	5664.5
Standard Deviation of Number of Iterations	34.8	2936.5	261.8	761.6

Table 3.2: Results using Family PSO.

In all the tests, both PSO variations arrived at a value close to the optimal, and, as expected, the De Jong function was optimised consistently. However, because of the nature of the test functions, a close-to-optimal value may belong to a local optimum. An analysis of the mean error from the optimal values of each dimension shows that it decreased substantially when the particles were faithful to their family, from errors in the range of 10^{-6} to 10^{-8}.

Another factor that was considered was the number of iterations necessary to reach a global optimum. Although the mean number of iterations did not appear to deviate when using either of the two variations, the substantial decrease of the standard deviation when testing the Rastrigin and Rosenbrock function indicates that, when the particles were being faithful, the number of iterations needed to find the optimal solution tends to vary less from its mean. This suggests that, when applying Family PSO, the search is more likely to find an optimal solution in the expected number of iterations. When testing with the De Jong and Schaffer F6 test functions, the change in the standard deviations between using the PSO variations is small.

The tests described have shown that, when optimising functions with a high number of local optima, there is increased reliability and performance in the PSO search if the particles are faithful to their initial family. An explanation for this is that, because of their initial uniform distribution, all the particles will have a different family, resulting in all the particles communicating with each other. If one family finds an important location, the members of that family will communicate it to their respective families, and so forth. In addition, because of their fidelity, the flow of information remains constant, resulting in consistent findings. In the case of Neighbour PSO, there is a high probability that the neighbours of one particle are the same as another, as their close distance is the only factor that joins them, which can result in a high number of particles not communicating with the rest of the swarm.

In the rest of this work, when applying Particle Swarm Optimisation, the algorithm used is the Family PSO variation.

3.5 Convergence of Particle Swarm Optimisation

An important topic of discussion when applying any Black-Box Optimisation Algorithm is that of convergence. Because of their stochastic nature, it is difficult to prove that convergence will be reached. This issue is of interest because if an algorithm such as Particle Swarm Optimisation is to be applied, it is important to ensure that a solution will be obtained, particularly in a real-time application.

To address this, Clerc and Kennedy identified a surprisingly simple set of conditions in which a PSO search is guaranteed to converge [18]. Their approach was to reduce the movement model of one particle to a state-space description, from which it was concluded that convergence was reached if the eigenvalues of the matrix describing the system had real values, and the particle remained steady, i.e. its velocity reached zero. To force this to happen, a set of constriction coefficients were introduced into the algorithm, resulting in it not being necessary to limit the velocity of the particles by v_{max}, as described earlier. The one-particle system was then extrapolated to a full swarm search, and applied to several test functions (several of which are described in Section 3.4) with good results. Their approach was solution-space independent, as only the movement of the particle was considered, which meant that the convergence conditions were generalisable to any objective function.

Eberhart and Shi observed that the way that the constriction coefficients were calculated and applied in the PSO variation proposed by Clerc and Kennedy was reminiscent of the application of the inertia value in the velocity modification equations, when using the concept of time-decreasing inertia in PSO [29]. They showed that applying an inertia weight that is equal to the main constriction coefficient, while ensuring that the sum of the weights for each influential location is greater than 4, is equivalent to the converging PSO variation developed by Clerc and Kennedy [29]. Meaning that incorporating a variation of the concept of time-decreasing inertia into PSO does not only speed up the search process, it also ensures convergence.

In addition, although the limit of v_{max} is unnecessary, applying it as the maximum range of values in every dimension has shown to provide faster results [29].

3.6 Conclusion

The topic of Optimisation is of great relevance to this project. In this chapter, several optimisation algorithms were reviewed to choose one to be applied in the framework proposed in this thesis. It was concluded that Gradient Descent-based methods could not be used because of their tendency of converging in locally minima, as well as being applicable only on continuous solution spaces. Black-Box Oriented Algorithms were chosen, as the methodologies and algorithms proposed in this thesis require flexibility and must rely on only limited information of the process.

The proposed framework is built such that any Black-Box optimisation algorithm is applicable, as long as it can converge in complex and unpredictable solution spaces. However, to avoid confusion, only one Black-Box optimisation algorithm must be applied. Particle Swarm Optimisation (PSO) was chosen for its relatively easy implementation and visual aspect, as well as the fact that, in literature, it has been found to converge in very complex solution spaces.

Furthermore, research referenced in this chapter has shown that, given a simple set of parameters, PSO is guaranteed to converge, which is of great interest in the fields that the proposed framework is to be applied.

In this chapter, an improvement to the PSO algorithm was made. Particle fidelity was shown to reduce the mean error of the optimisation process. It was also shown that, when particles are faithful to their initial families, PSO is more likely to find the global optimum in the expected number of iterations.

Chapter 4

Review of Blind Source Separation and Spectral Distortion

An important requirement for the proposed framework is that of a reliable reference spectra set needs to be available. This set may be acquired from a commercial spectral library that has been created by laboratory analysis of pure components, following a stringent and standardised protocol [16]. Another possibility is to synthesise the spectral set by using software packages that can generate a reference spectrum, provided that the absorption values expected from the material are known *a-priori* [16].

Unfortunately, the options available to acquire a reference spectrum may be too costly, or the data necessary to generate one may not be available. In addition, an important part of quality monitoring is that of detecting the presence of foreign materials in the product. Hence, methodologies that are able to extract the underlying components from a spectral data set are of great interest, because a reference spectrum can be derived as a result, along with being used in identifying foreign materials in the product.

Nonetheless, as discussed before, spectral distortion should be expected in online measurements, and component extraction techniques are required to be robust against them. In this chapter, a review will be given of current component extraction algorithms, specifically those in the field of Blind Source Separation and Curve Resolution. Furthermore, the effects that typical spectral distortion have on these algorithms will be discussed.

4.1 Component Extraction

The algorithms discussed in this chapter are based on a data centric model, described as

$$X = CS \tag{4.1}$$

where X is as matrix containing the measurements from where the components are to be retrieved. X is comprised of m rows, each containing a spectrum of length n. C is the set of concentrations of the components inside the data and is a matrix of m rows and an amount of columns, k, which is equal to the number of components inside the data. S is the set of spectral signatures (a.k.a. profiles) of the components inside the data; it is a matrix of k number of rows, each representing a spectral signature of length n.

4.1.1 Principal Component Analysis

Principal Component Analysis was introduced in 1901 [85] with its objective being the reduction of the number of dimensions that describe a function or group of points, whilst still describing the function in great detail. The main procedure involves finding the direction of the largest variance inside the function or data set, and defining it as one dimension. It then proceeds to find the direction with the next largest variance that is orthogonal to the first direction, and defining it as another dimension. The iterative process continues until a number of dimensions equal to m or n, whichever is smaller, are found. To minimise the number of dimensions, the directions with the least variance are discarded, either manually, automatically by setting a cut-off variance value, pre-defining a number of desired dimensions, or using cross-validation [109]. The resulting number of dimensions can be considered as the components that are present in the data.

From its inception, several methods have been proposed for finding the directions with most variance in a data set, one of which is referred to as the Covariance Method. Considering the model defined in (4.1), X is centred by subtracting the mean, and the covariance matrix $cov(X) = XX^T$ is calculated. The eigenvectors (λ) and eigenvalues (V) of the covariance matrix are then calculated by an eigenvalue decomposition such that:

$$(XX^T)V = \lambda V \tag{4.2}$$

where the contents of V are not equal to zero. An eigenvalue will then represent the 'energy' that its respective eigenvector contributes to the data. The eigenvectors are ordered from largest eigenvalue to lowest. If the sum of all eigenvalues represent the total 'energy' of the data, the number of eigenvectors with the largest eigenvalues can then be retained such that their sum represents a significant portion of the total 'energy'. The percentage of total energy can be increased or decreased arbitrarily if needed. The resulting group of eigenvectors ($\hat{V}$) can be mapped back to the real domain to calculate a set of non-correlated vectors (S), also known as *loadings*, i.e.:

$$S = (\hat{V}X)^T \tag{4.3}$$

The *scores* or magnitudes of each loading can be calculated by solving for C in (4.1).

Another popular method is the Singular Value Decomposition (SVD), where a mean-centred X is decomposed into three matrices, as defined in (4.4).

$$X = U\lambda V^T \tag{4.4}$$

where U provides the coordinates of the data in X in the Principal Components space. Both λ and V can then be used to calculate the loadings and scores as described in (4.1) and (4.3).

PCA has gained popularity in a variety of fields [108, 75, 23, 53, 117, 42] because of its flexibility and intuitiveness. However, the assumptions of linearity and non-correlation in the sources are not sufficient for PCA to obtain a unique set of source estimates, and the flexibility it boasts also brings about difficulties for component extraction. PCA is not designed to find the sources that are confined within a data set, it aims only to identify the orthogonal variations inside a data set. Quantifying such variations provide important insight towards the mechanics inside the data, but they may not actually have any physical meaning, and, thus, may not always be assumed to be the actual components. In certain applications, such as the one being studied in this work, PCA is not able to find a unique set of components because, even if they are completely un-correlated[1], they have other features, such as non-negativity or unimodality.

In fact, a number of studies were conducted where PCA was applied to a

[1] Semi-overlapping sources are frequently analysed in the industry, such as Pharmaceutical and Medical, which obviously bare a small degree of correlation.

simple data set comprised of artificial spectra, with no noise or corruption applied to it. PCA was not able to obtain a consistent and sensible set of components that estimated the simple un-correlated sources. However, it was always able to estimate the *number of sources* that were inside the data set.

Estimating the number of components inside a data set of mixtures is an important initial phase in all algorithms that are used for component extraction. It is in this phase, rather than in component extraction, that PCA, and more specifically the SVD method, is considered a valuable tool. However, as will be discussed later, even in this part of the process, PCA has been shown to be very sensitive to spectral distortion.

4.1.2 Independent Component Analysis

Independent Component Analysis (ICA) was conceived by Jutten and Herault in 1986 [57] and formally defined by Comon in 1994 [19]. The method can be used to separate a multivariate measurement into a number of sub-components, or *sources*. The approach is ideally suited to the analysis of spectral measurements, which are often additive signals that contain the concentrations and spectral signatures of the independent compounds that comprise a mixture.

The independence between sources is used as the objective function within the ICA formulation. This is then optimised to identify the signals that are most independent of each other. The different ways that independence is calculated and how it is maximised have resulted in several implementations of the ICA concept [47, 90, 99]. A popular implementation is FastICA [50, 104, 48], introduced by Hyvärinen [45]. FastICA uses the amount of mutual information shared among the sources as a measure of independence, estimated using differential entropy or *negentropy* [48].

Considering the model described by (4.1) an estimate of S ($\bar{S}$) can be extracted from X by applying (4.5).

$$\bar{S} = BX \tag{4.5}$$

Where B is a de-mixing matrix that ICA seeks to obtain. B is acquired by using a fixed-point iteration scheme applied to an orthogonal matrix $\tilde{X}$ obtained from (4.6).

$$\tilde{X} = \mathbf{svd}(\bar{X}\bar{X}^T) \tag{4.6}$$

Where $\bar{X}$ is X mean-centred and whitened, and **svd** is the Singular Value Decomposition function. Being $\tilde{X}$ orthogonal, the number of parameters that must be identified is reduced [48]. In this step, the dimensionality of the problem can be reduced by discarding small eigenvalues of XX^T. It is important to note that this step is actually carried out by applying Principal Component Analysis, and is where the number of components in X is estimated.

After computing $\tilde{X}$, B is found by using a fixed-point iteration scheme as defined below:

$$\begin{aligned} b_i = {} & E\{\tilde{X}G(b_i^T\tilde{X})\} - E\{G(b_i^T\tilde{X})^T\}b_i && (4.7) \\ b_i = {} & b_i/\|b_i\| && (4.8) \\ & \text{repeat (4.7) and (4.8) until convergence criteria is satisfied} \end{aligned}$$

where b_i is the ith row in B, and G is the first order derivative of a non-linear function g that "does not grow too fast" [48] so it can converge at a minimal level of entropy. (4.7) was derived by applying a constraint on the expected value of $b_i^T\tilde{M}$ that satisfies Kuhn-Tucker conditions, making it possible to find its maximum value by a Newtonian method. The constraint applied is $\|b_i\| = 1$, which is met by applying (4.8). Both of these equations are applied until b_i converges.

The process repeats as many times as there are rows in $\tilde{X}$. All the created b_is are then concatenated[2] to form B. However, more than one b_is may reach the same maximum, resulting in several estimates representing the same source. To avoid this, (4.9) is applied after each iteration of the fixed-point schemes to ensure that the rows in B are *non-correlated.* This method of de-correlation is referred to as *symmetrical*, and is preferred for its equal weighting of all the b_is.

$$B \leftarrow B(\sqrt{B^TB})^{-1} \tag{4.9}$$

When B is calculated, the resulting $\bar{S}$ will hold the estimated sources that, because of the small amount of mutual information between them, can be

[2] It is not known beforehand which row will ultimately lead to which source, so the order of the rows is not important.

considered independent, hence the name Independent Components (ICs).

There are similarities between PCA and ICA, in that the identified components are non-correlated. In fact, as it is shown in Figure 4.1, it could be argued that PCA is a type of pre-processing step in the ICA process [19]. However, the result that each algorithm provides is different, as ICA calculates *independent* components, which not only implies non-correlation [48], but that the information shared between components is minimal. Thus, the components that ICA aims to extract come from different physical sources, providing components more congruent to reality [57, 48].

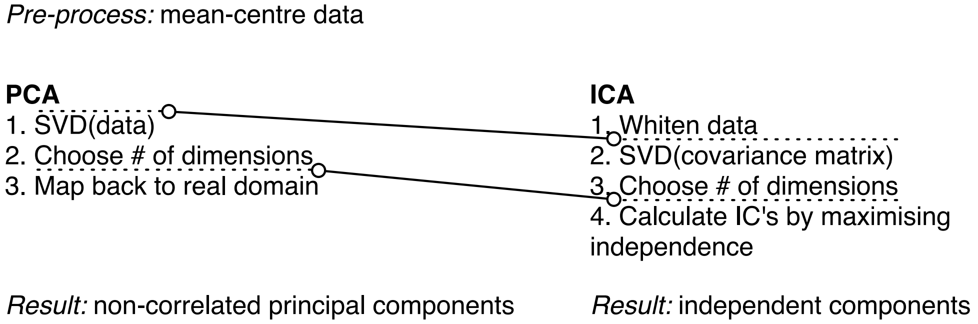

Pre-process: mean-centre data
PCA
1. SVD(data)
2. Choose # of dimensions
3. Map back to real domain
ICA
1. Whiten data
2. SVD(covariance matrix)
3. Choose # of dimensions
4. Calculate IC's by maximising independence
Result: non-correlated principal components
Result: independent components

Fig. 4.1: Comparison between PCA and ICA.

It is important to mention that throughout this work, unless noted otherwise, when using the term ICA, this refers to the FastICA implementation of this method. It could be argued that there are more ICA implementations to be reviewed, however, FastICA was chosen as it extracts components efficiently and the independence estimation it employs is considered better than others [44]. In addition, it will be seen in Chapter 7 that the FastICA implementation behaves in a way that is of interest to this project, as the results it provides when using locally shifted spectral data can be corrected by a post-processing method.

4.1.3 Self-Modelling Curve Resolution Methods

Self-Modelling Curve Resolution (SMCR) methods are considered to be a different field to Blind Source Separation (BSS). However, for practical reasons they are discussed here. It was originally intended that SMCR methods be used only in the field of Chromatography [55], as its application is bound to the domain of the spectrum (be they mass, reflectance, frequency, etc.). This is unlike BSS which was first applied to Sound Analysis, in the time domain. Having said that, it is easy to blur the line between the two fields as the use of SMCR is gaining popularity in fields outside Chromatography [97, 6, 55], fields where BSS is now also considered applicable. Another important reason for the confusion of the two fields is that their objectives are very similar. SMCR methods aim to find the 'pure' components from which several sampled spectra are comprised, an objective identical to that of some BSS techniques.

Unfortunately, SMCR methods require that the number of components be known *a-priori*, and it can be argued that BSS methods do not require this knowledge. Still, the methods proposed to estimate the number of components [55] before applying SMCR methods are similar to those used in PCA and ICA, further describing their similarity. It is important to note that the methods available for estimating the number of components have been reported to be sensitive to noise, and spectral distortions, such as shift, may introduce difficulties with the techniques [55].

SMCR methods can be divided into two groups or families [55]; Unique Resolution methods and Rational Resolution methods.

Unique Resolution methods aim to find a 'unique' solution to a problem defined by the data. The solution is usually given as a model of the mixing process of the material or the mathematical description of the different processes applied to it. This unique solution then provides the definition of the spectral components that comprise the material. To find such a solution, Unique Resolution methods apply Singular Value Decomposition (SVD), or similar rank analysis tools, to specific regions of the data. These methods therefore rely on procedures very similar to those used with Principal Component Analysis. Hence, if PCA is sensitive towards shift or other spectral distortions, then these methods will be also.

Rational Resolution methods employ a specific set of assumptions regarding the components and aim to find a set of components that are both consistent with the data as well as with these assumptions. Examples of assumptions that can be used are non-negativity, meaning that the components do not have negative parts; extreme dissimilarity, the components are completely different from each other; unimodality, one possible mixture; and maximised purity, the components have no noise [22]. The majority of the methods of this family are iterative, and are based on or can be enhanced by an iterative algorithm called Alternating Least Squares (ALS) [55]. In this regard, ALS seems to be an important contribution to the whole of SMCR, and, although the intent is not to reduce the entire field to just one algorithm, an inspection of ALS against spectral distortion will provide a good measure for the sensitivity of SMCR methods against it.

The ALS algorithm [59] was introduced in 1989. It assumes a linear behaviour, as described in (4.1) from which two equations can be derived. One of these equations is obtained by solving for the concentration matrix, C:

$$C = XS^T(SS^T)^{-1} \tag{4.10}$$

and the second by solving for the spectral signatures matrix, S:

$$S = (C^TC)^{-1}C^TX \tag{4.11}$$

The algorithm starts by randomly initialising the values within S. The iterative part of the process then proceeds:

1. C is calculated by applying (4.10).
2. Constraints are imposed onto C, such as non-negativity and unimodality.
3. S is re-calculated by applying (4.11).
4. Non-negativity and normalisation is applied to S.
5. $\hat{X}$ is calculated using the current C and S by applying (4.1).
6. A measure of distance from the actual data, X, is calculated.
7. The cycle repeats until $\hat{X}$ is within a specified tolerance of X.

The author of ALS proposed using the Sum of Squared Errors (SSE) between X and $\hat{X}$ as the distance metric, which is defined as:

$$SSE = \sum_{m}^{M}\sum_{n}^{N}(x_{mn} - \hat{x}_{mn})^2 \tag{4.12}$$

although other measure of distances can be used.

Ten years after the introduction of ALS, another algorithm was proposed called Non-Negative Matrix Factorisation (NNMF)[63]. This algorithm can be considered to be of the Rational Resolution family of SMCR methods, as it assumes non-negativity in both the concentrations and spectral signatures. It has been compared to Independent Component Analysis, yielding improved results when applied to spectra with overlapping features (a characteristic that breaks the independence assumption of ICA) [6].

Various methods have been proposed to achieve the objectives of NNMF [6], and it has been pointed out that one of these methods is in actuality the ALS algorithm, using the non-negativity constraint [6]. In fact, of all the variations of NNMF, the 'Alternating NNMF' version is the fastest to converge and the least likely to encounter difficulties with local minima [6].

The authors of NNMF have stated that the algorithm may not be sufficient for all types of decomposition or factorisation problems, as it may not be able to handle very complex types of data, such as images being viewed from different perspectives [63]. This means that even if the process that estimates the number of components in the data could overcome a spectral distortion, it is still possible that it may be an issue during the extraction process.

4.2 Artificial Data Creation

In the following sections, several experiments were carried out using artificial data sets. These data sets contained simulated spectral samples from mixtures of reference components. The spectral signatures of the reference components were created as a combination of 'peaks', which are defined using (4.13).

$$c_i = \sum_{n}^{N} \frac{1}{\frac{0.4}{h_n}\sqrt{2\pi}} e^{\frac{-(f_i - p_n)^2}{\frac{w_n}{4}^2}} \tag{4.13}$$

where c is the reference spectra for one particular source; c_i is the spectral intensity at frequency f_i; N is the number of peaks that component c contains; h_n, p_n, and w_n are the height, frequency location, and width of the nth peak respectively.

The equation (4.13) was developed during the project to create artificial spectral peaks, based on gaussian-shaped functions. It is relatively easy to create different types of spectra from a gaussian function, by modifying three variables (height, width, and location) as shown in Figure 4.2, where the peak is located at 1 Hz ($p_n = 1$), has a width of 1 Hz ($w_n = 1$), and has a width of 1 ($h_n = 1$).

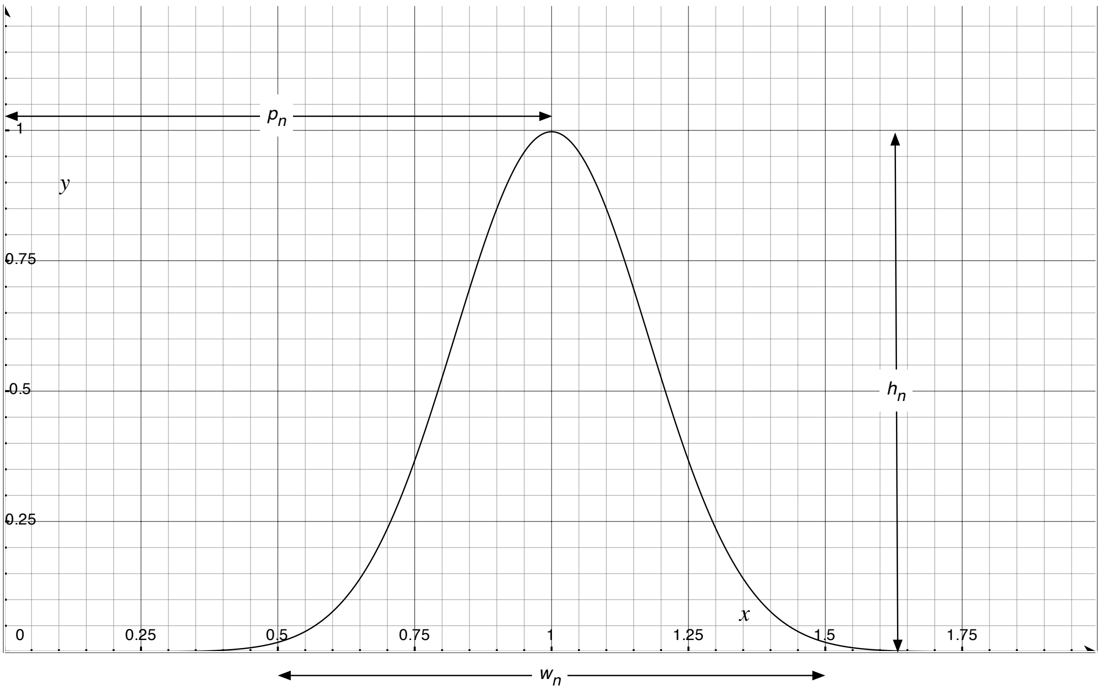


Fig. 4.2: A peak defined by in (4.13), located at 1 Hz, with a width of 1 Hz and a height of 1.

A reference spectrum can then be created by defining the number of peaks (N) and their respective values. A set of four reference spectra was created to be used in this study by choosing the aforementioned values in a random manner. The reference set is shown in Figure 4.3. The x-axis in this figure is frequency (Hertz) and the frequency resolution of these spectra is 0.1 Hz per frequency point (fp).

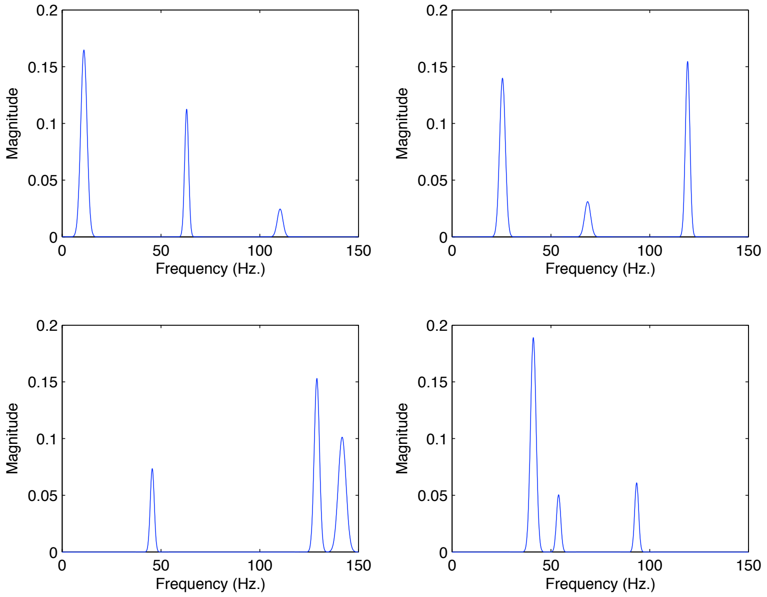


Fig. 4.3: Reference spectra randomly generated for use in experiments.

The structure of these spectra were defined such that they were consistent with data obtained in the Pharmaceutical and Biomedical Industry [32], as well as other fields [8, 9].

A series of sampled data sets were then created using the defined reference spectra. Each data set consisted of 100 samples, with each sample containing a spectra of a simulated mixture defined as follows:

$$d_k = \sum_m c_{k_m} \mathbf{shift}(S_m, l_{k_m}) \tag{4.14}$$

where d_k is the kth sample in data set d; c_{k_m} and l_{k_m} are a randomly-generated concentration and shift values, respectively, that are applied to the mth component in the reference spectra S. The shift function displaces the information of the spectrum from f_i to $f_{i+l_{k_m}}$ where l_{k_m} can be negative.

The data sets had a pre-defined constraint on the range of concentration and shift values they were subjected to. All concentrations were specified to be between [0.2, 1] for all data sets. Each data set was given a different maximum shift value (max_shift) in frequency points (*fp*), and shift values were randomly-generated to be between $[-max\_shift, max\_shift]$.

It is to be noted that the manner in which the shift function is being applied is a simulation of the Local Spectral Shift distortion, described in Section 2.2.1. If a Local Spectral Warp were to be applied, the warp function can be used, the implementation of which is described in detail in Appendix C.

4.3 Shift Effect on Component Extraction Algorithms

Understanding the effects that spectral distortion has on the algorithms described in the Section 4.1 is crucial to the project detailed in this work. However, exploring the effects of every type of spectral distortion is a monumental task that may lead to unnecessary redundancy. The focus of the work reported in this thesis is in large part coping with Spectral Shift. This particular type of distortion was investigated because it frequently appears in a broad section of Industrial applications [11, 14, 106, 36, 39, 38].

4.3.1 Principal Component Analysis

As described earlier, PCA was not designed specifically for component extraction. However, it has been found to be well suited for estimating the number of components within a data set. In the experiment described in this section, PCA was applied to an array of data sets, each suffering from different degrees of shift, with the objective of estimating the number of sources in each data set. All of the data sets had 100 samples and were created with the artificial spectra described earlier, meaning that only four components were expected to be identified.

In Figure 4.4, each tick in the x axis represents a data set that suffers from that specific degree of shift. The y axis represents the number of components that PCA identified from that specific data set. Cross-validation was used to decide how many Principal Components to keep. Meaning that, for each data set, the number of components identified was considered to be the number of PCs that minimised the Predicted Residual Sum of Squares (PRESS) statistic [109].

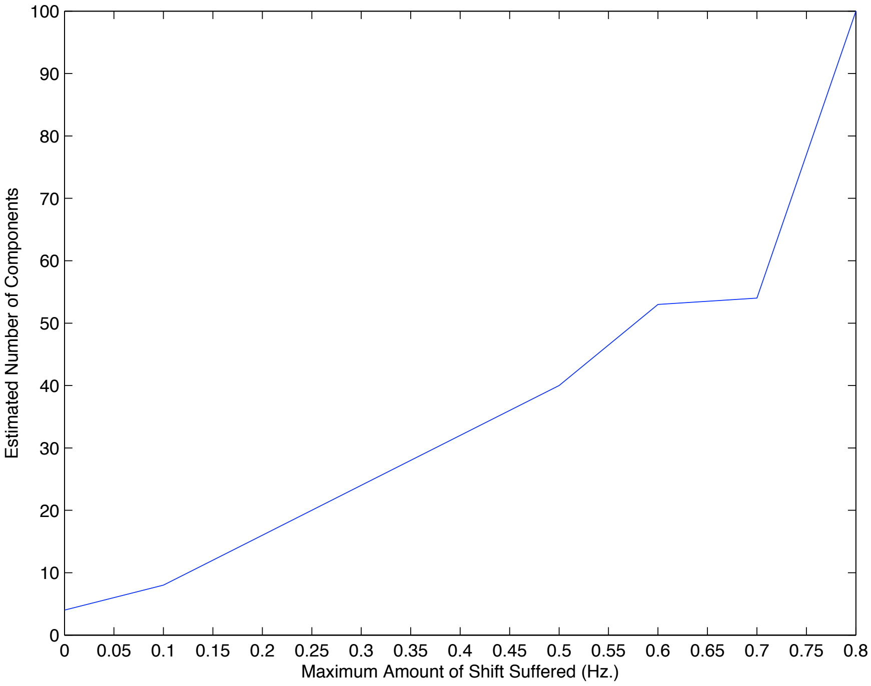


Fig. 4.4: Estimated number of components from data sets suffering from different degrees of Spectral Shift using PCA.

Figure 4.4 shows that Spectral Shift degrades the ability of PCA to estimate the correct number of components within a spectral data set. In fact, PCA demonstrated no useful reduction of dimensions when moderate degrees of Spectral Shift occurred, as the number of identified PCs was equal to the number of samples within a data set suffering from a shift of 0.8 Hz.

It is important to note that this does not imply that PCA is not able to correctly estimate the number of components when *any* amount of shift is applied to *any* spectral data set. As it will be seen later, the value of shift that PCA is tolerant to is heavily influenced by the shape and specifically the width, of the components.

4.3.2 Self-Modelling Curve Resolution Methods

Because the Unique family of SMCR methods rely on rank analysis tools, such as Singular Value Decomposition/PCA [55], their performance will all be similar to that reported in Section 4.3.1. From the other family of methods, the most popular is Alternating Least Squares, which is often the basis of other algorithms of the same family or is used for increasing their performance. Should ALS fail against using shifted data, it would indicate that the remaining methods from this family will fail also.

For the sake of illustrating the sensitivity of the ALS algorithm towards spectral distortion, the number of components will be assumed known. However, it is important to point out that rank analysis tools such as SVD/PCA are used to estimate the number of components in the data set before applying SMCR methods. As observed in the previous section, PCA is very sensitive towards shifted data, implying that, in a practical situation, ALS would not have the correct information to produce accurate results. Therefore, the following demonstration should be considered as a theoretical exercise only, and the reader should keep in mind that the results will be better than normal when applying ALS to data containing spectral shift.

The experiment involved the creation of artificial data sets, as described in Section 4.2, applying a different maximum shift value for each one, ranging from 0 *fp* (no shift applied) to 40 *fp.* (4 Hz). ALS was then applied to each data set with *a-priori* knowledge that 4 components should be extracted. The components estimated using ALS were each compared to the original sources, using the *Pearson Correlation Coefficient*. The performance measurement of ALS for a given data set is the mean of the highest correlation values that were recorded for each estimated component. Therefore, if the algorithm performs perfectly, a value of 1 will result. The lower the value of the mean correlation, the worse the results. The performance of ALS plotted against the shift applied to a data set is given in Figure 4.5.

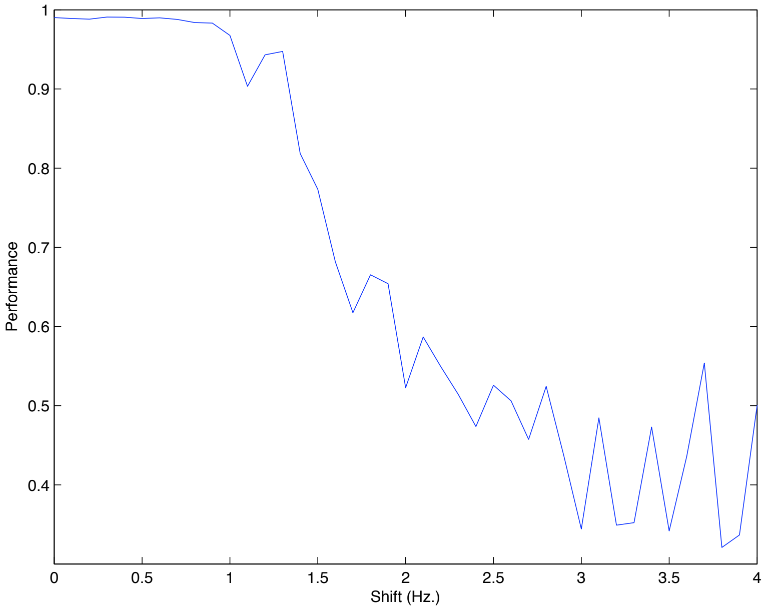


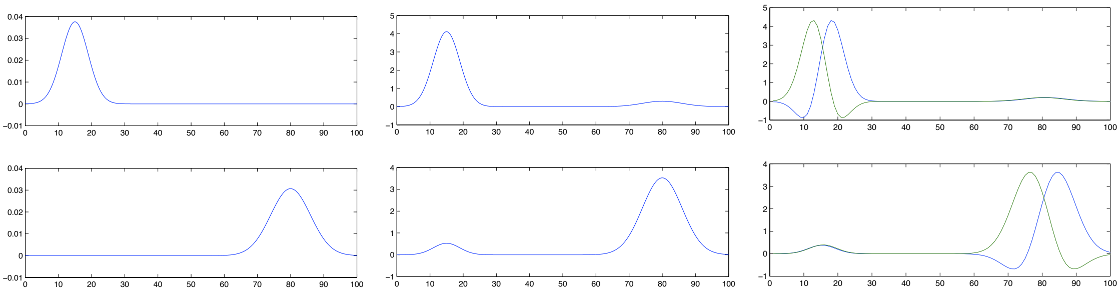
Fig. 4.5: Performance of ALS with different shifts applied.

Figure 4.5 shows a clear deterioration of the performance of ALS as the amount of shift applied to the data set is increased. It is noticeable that there is a considerable drop starting at a shift of 1 Hz. Given that the data used in this test did not contain any measurement noise, a performance below 90% should be considered unacceptable. Figure 4.5 therefore suggests that the ALS algorithm is unable to cope with shifts greater than approximately 1.4 Hz. Furthermore, since in this test it has been assumed incorrectly that PCA would accurately determine the number of underlying components prior to the application of ALS, it is reasonable to conclude that ALS is unlikely to yield good results in the presence of shift.

4.3.3 Independent Component Analysis

ICA relies heavily on Principal Component Analysis, as it not only estimates the number of components, but also uses PCA when whitening the data. This suggests that if PCA is sensitive towards spectral distortion, then ICA will be expected to be so too. In this section, the effect that spectral shift has on ICA will be demonstrated. In particular, the effect that it has on the components identified using ICA will be analysed.

Experiment 1: Effect of Shift on ICA. Two components were created, with a frequency resolution of 1 Hz per *fp*: one with its peak at 15 Hz and a width of 20 Hz, and another with its peak at 80 Hz and a width of 30 Hz, both shown in Figure 4.6a. A data set of 1000 signals was created with concentrations varying between 0.2 and 1. A second data set was created using the same concentrations, but in this data set the components in the samples were locally shifted by ±1 *fp*.



(a) Components used. (b) ICs obtained without shift. (c) ICs obtained with a 1 Hz shift.

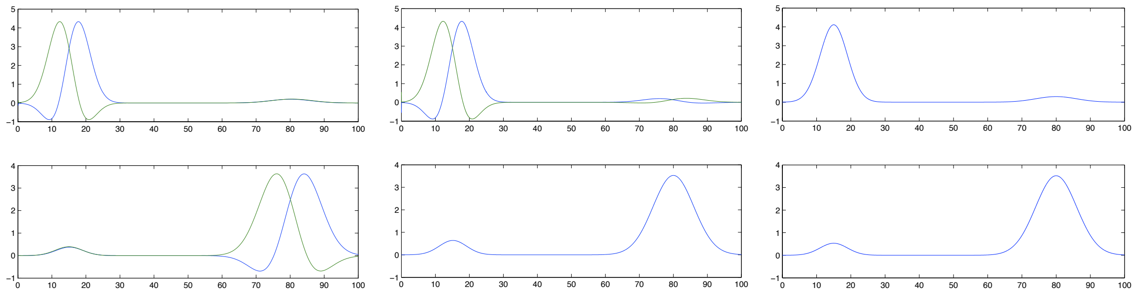
Fig. 4.6

In Figure 4.6b the components extracted from the non-shifted data are shown, and, as expected, they are close to the original sources in Figure 4.6a. In Figure 4.6c the components extracted when ICA was applied to the shifted data are shown. Both upper and lower Figure 4.6c each show 2 components that are similar to each other. Meaning that this figure shows a total of 4 identified components, rather than 2, and that these components appear to be grouped. It appears as if each component was *divided* into two; this effect, for reference, will be referred to *component division* in the remainder of this thesis, and, as will be observed later, is a common feature when applying ICA to shifted data. This simple example has demonstrated that ICA is unable to cope with spectral distortions. The subsequent experiments explore this further.

Experiment 2: Maximum Shift Handled by ICA. The same sources as in Experiment 1 were used, although now with a frequency resolution of 0.01 Hz per *fp*. The objective was to find the maximum shift with which ICA could extract the correct components. Different data sets were created, each having their components shifted by a random amount between in $[-max\_shift, max\_shift]$, and each data set had a different value of max_shift applied to it.

In Figure 4.7, three sets of ICs are shown. Figure 4.7a shows the extracted components from data shifted at 10 *fp*; values from 10 to 4 *fp* provided similar results.

Figure 4.7b and 4.7c show the extracted components from data shifted at 3 *fp* and 2 *fp* respectively. The ICs extracted from the un-shifted data set and that with a maximum shift value of 1 *fp* are similar to those shown in Figure 4.7c.



(a) ICs obtained at 0.1 Hz shift. (b) ICs obtained at 0.03 Hz shift. (c) ICs obtained at 0.02 Hz shift.

Fig. 4.7: ICs obtained at different shift.

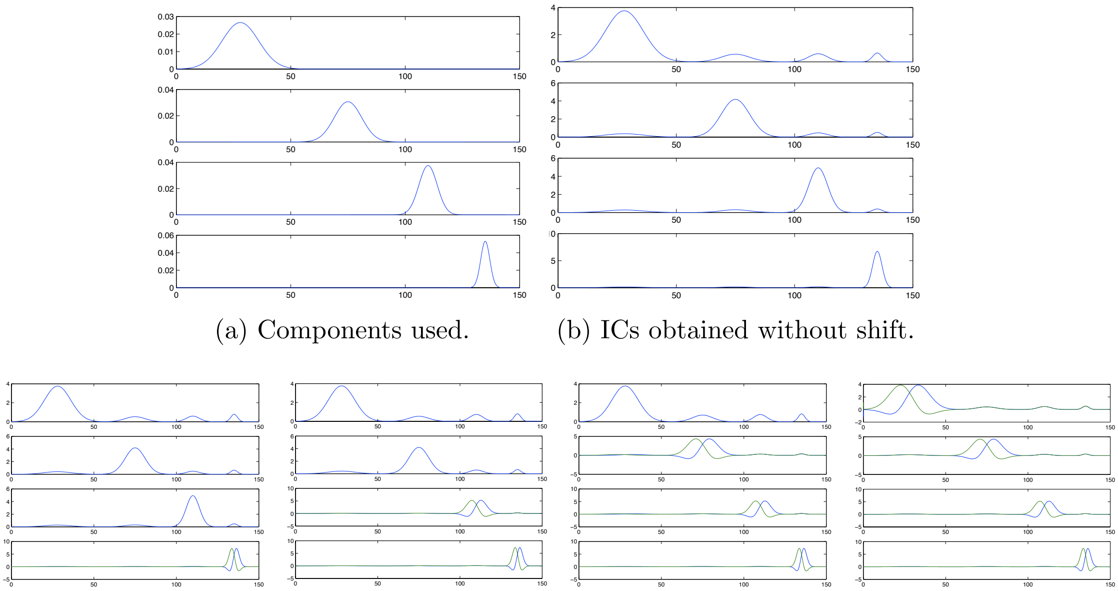
This result indicates that ICA can only extract the components when there is very little, or no shift.

Experiment 3: Influence of Component Width. It can be seen in Figure 4.7b that when shifting the data by 0.03 Hz, only the component at 80 Hz was able to be extracted correctly. The only difference between the components is their width, so another experiment was developed to test if the amount of shift that would affect the correct extraction of a component is dependent on its width.

Four components with widths of 40, 30, 20 and 10 Hz were created and are shown in Figure 4.8a, with a frequency resolution of 0.01 Hz per *fp*. The process of creating data sets explained in Experiment 2 was repeated and the results are shown in Figure 4.8.

When shifting the data at 0.02 Hz, only the thinnest component suffers from component division. When increasing the range of shift values to 0.03 Hz, the second thinnest component suffers from component division. This tendency continues until the widest component is divided when using values in the range of 0.07 Hz.

This result indicates that the width of the component does affect its extraction. In Figure 4.9 a grey area is shown, the border of which was obtained by plotting the maximum shift applied to a data set against the width of the widest component that was not properly identified. Any point residing in this area represents a component not identified corretly using ICA.



(a) Components used.

(b) ICs obtained without shift.

(c) At 0.02 Hz shift.

(d) At 0.03 Hz shift.

(e) At 0.05 Hz shift.

(f) At 0.07 Hz shift.

Fig. 4.8: ICs with different widths obtained at different shifts.

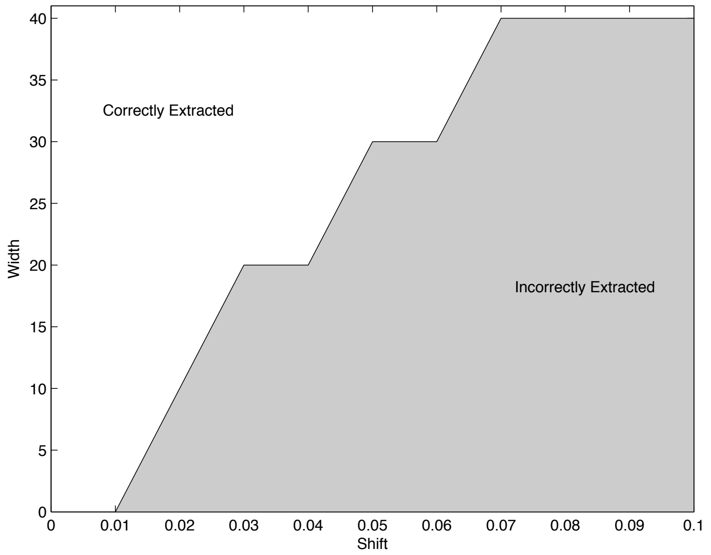


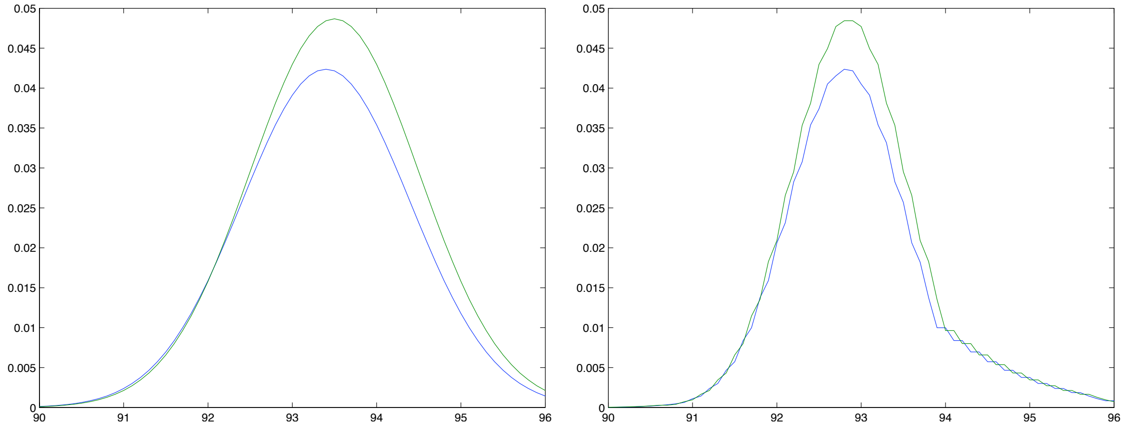
Fig. 4.9: Amount of shift against width of the component incorrectly extracted.

4.3.3.1 Considering a Pre-Aligning Approach

A logical approach to circumvent the problem of frequency shift would be to align the data before applying ICA. As will be discussed later, when dealing with Global distortions, properly pre-aligning data does provide good results. However, when Local distortions are involved, because each component is shifted independently from each other, aligning two spectra with each other is not trivial. Shifting the whole spectrum to align one component results in another component becoming misaligned. If this approach is to be applied correctly, a more sophisticated method is necessary.

Dynamic Time Warping (DTW) is an alignment method first introduced by Sakoe and Chiba in 1978 [92]. It was first proposed as a method to describe how similar two signals of different lengths are by identifying a mapping between the two signals. The mapping it creates can be used to align the two signals, with the focus on ensuring that the local areas of both signals are similar. DTW became very popular in Speech Recognition for aligning vocal recordings [67], and later in Batch Processing where the dynamics can proceed at different rates [66]. DTW is a flexible method and is able to align the local structure of equal-length signals, such as frequency spectra. In this regard, a multivariate version of DTW has been successfully used to align chromatographs [107], however it was concluded that aligning over one dimension (reducing the whole problem to a univariate model) was enough to solve the alignment issue.

To identify the benefits offered by DTW, the technique was applied to the shifted data set introduced in Section 4.2. ICA was then applied to the "re-aligned" data set and identified 91 ICs, compared with 8 without applying DTW.



(a) Zoomed spectra before aligning.

(b) Zoomed spectra after aligning.

Fig. 4.10: Result of aligning spectral data using DTW

Figure 4.10a shows one of the peaks in the spectra of samples 2 and 7 before alignment using DTW and Figure 4.10b after. These figures show that, although DTW did align both peaks, it also modified their shape considerably. ICA identifies all the different shapes of one component throughout the data set as different components, resulting in far more components than expected and, in turn, decreasing the quality of the results. This is the reason why ICA identified 91 components from the processed data.

Another sophisticated aligning procedure is Alignment by Fast Fourier Transform (RAFFT or PAFFT) [112]. This method divides the spectra into an optimal combination of different segments, given by an FFT transform of the signal. Each segment is then aligned locally. However, from the examples shown in the original work [112], the alignment procedure also changes important features of the data.

The reader is encouraged to consider the effects of pre-aligning locally[3] before applying ICA. Local alignment methods artificially distort the spectra, modifying the local shapes of the spectrum. Consequently, ICA identifies each modified local shape as a different component and, as a result, the significance of the extracted components is reduced considerably. Therefore, it is not sensible to locally modify the sampled spectra before processing.

[3] Globally aligning a signal involves modifying its shape uniformly, hence local shapes are maintained.

4.4 Conclusions

The effects of Spectral Shift on popular component extraction algorithms were explored. Spectral Shift was chosen as it is one of the most frequently observed distortions in the Industry. If a distortion as simple as Spectral Shift produces any problems, it can be expected that more complex distortions will do so also.

The component extraction algorithms discussed were Principal Component Analysis (PCA), Independent Component Analysis (ICA) and Alternating Least Squares (ALS), which are the most popular algorithms in use today. In all three cases, the distortion affected the performance of the methods considerably. It is important to note that PCA is used in a wide range of Self-Modelling Curve Resolution (SMCR) methods, not only for identification purposes, but also for estimating the number of components to extract. This means that all SMCR methods that rely on PCA/SVD will be affected by Spectral Shift.

It was shown that if a component suffers from moderate amounts of shift, Independent Component Analysis and, more precisely, the FastICA algorithm, was unable to extract the same components that it does from non-shifted data. No other changes to the data, other than shifting it, was implemented, so it can be concluded that it is the only factor causing the problem.

The results show, however, that FastICA was able to properly identify the ICs in certain circumstances; i.e. when the shift was small, and the component peaks were wide. Hyvärinen has stated that "Actually, and perhaps surprisingly, it turns out that [to solve the ICA problem] it is enough to assume that [the sources] $s1(t)$ and $s2(t)$, **at each time instant** t, are statistically independent."[48]. Meaning that, as in other machine learning and advanced statistical methods, ICA considers each frequency point of a spectrum as if it were a variable, the consistency of which is key to extracting the required information (in this case, the ICs). When shift occurs, however, the information from one variable is passed onto others, resulting in ICA being unable to follow the correct variation of the variables. However, if the shift is small enough and/or the components are wide enough, the variation of neighbouring variables are similar enough for ICA to consider the data as if it would not have suffered from any shift.

The sensitivity that ICA has towards shift illustrates the importance of developing an algorithm that is insensitive to Spectral Shift, as well as other types of distortion. In [46] the authors created a variation of the Independent Component Analysis algorithm that contained shift-invariant features, which were from what the authors referred to as a *feature subspace*. This means that the actual sources were not extracted. Furthermore, the invariancy was measured from responses of the specific feature subspaces, meaning that the resulting shift-invariance was very application-specific. Still, the authors provided insight into how Spectral Shift can be simulated, which is relevant to this thesis.

In this chapter, it was also shown that the approach of pre-aligning the data cannot resolve the local shift problem, as it changes the local shape of the spectra which will affect the results obtained using ICA. Hence, a post-processing algorithm is a more sensible approach to make ICA robust against Spectral Shift.

Chapter 5

Spectral Analysis Framework

The proposed methodologies are aimed to be applied in a generic form for Spectral Component Analysis. To this effect, an optimisation approach was employed such that its objective function was flexible in considering different types of spectral distortion. A diagram summarising the objective function for the Spectral Analysis Framework is shown in Figure 5.1.

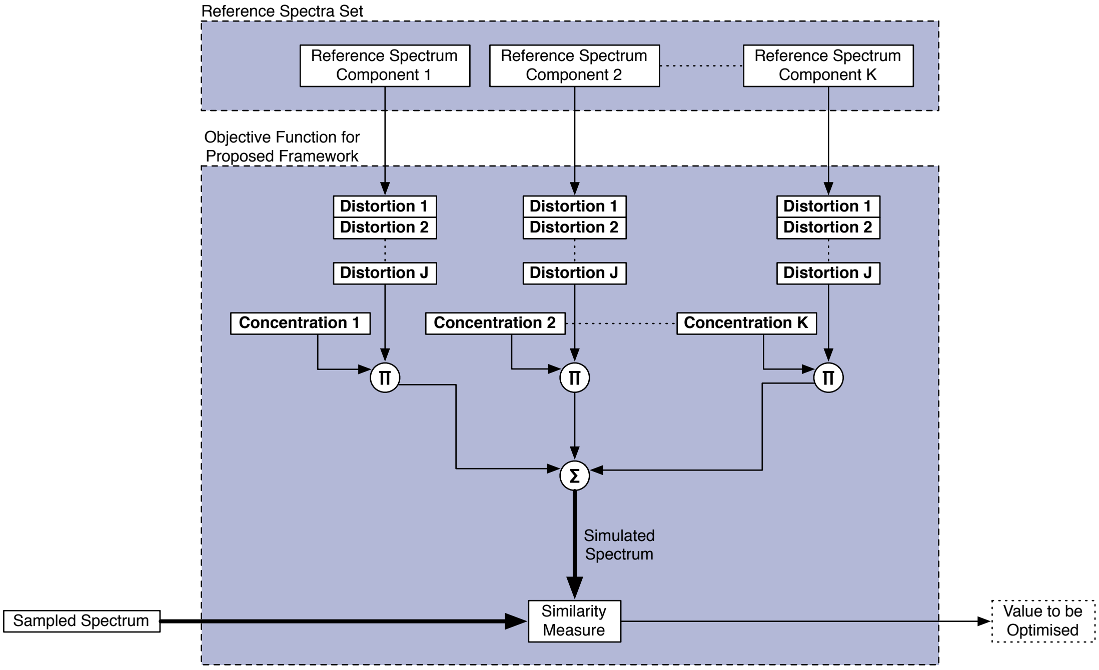


Fig. 5.1: The objective function used by the proposed framework.

As it is shown in Figure 5.1, a set of distortions are simulated over each of the reference spectra. The simulation of each distortion is assumed to be supervised by a plant expert, as the application of specific models need to be congruent with the plant sensors, materials, and suspected distorting factors. In addition, the order in which each distortion is applied is required to be advised by the plant expert, and it is assumed that the effect of each distortion is independent from the rest. Furthermore, the cumulative effects of two or more distortions are considered to be additive.

The distorted reference spectra are then multiplied by their respective concentrations and added together.

The resulting simulated distorted spectrum $\hat{M}$ is then compared with the measured spectrum M. Although there are several methodologies with which such a comparison can be carried out, the Euclidian Space, s, between the $\hat{M}$ and M, described in (5.1), provided good results in this work.

$$s = -\sqrt{\sum_{f}^{F} (M(f) - \hat{M}(f))^2} \tag{5.1}$$

Where $M(f)$ is the energy at frequency location f of M. The maximum value of D is 0. However, the similarity methodology can be modified if necessary.

The optimisation process proceeds to find the distortion values (D) and concentration values (C) that create a $\hat{M}$ that best fits M, such defined in (5.2).

$$[C, D] = \mathbf{pso}(M, S, \mathit{distortion\_models}) \tag{5.2}$$

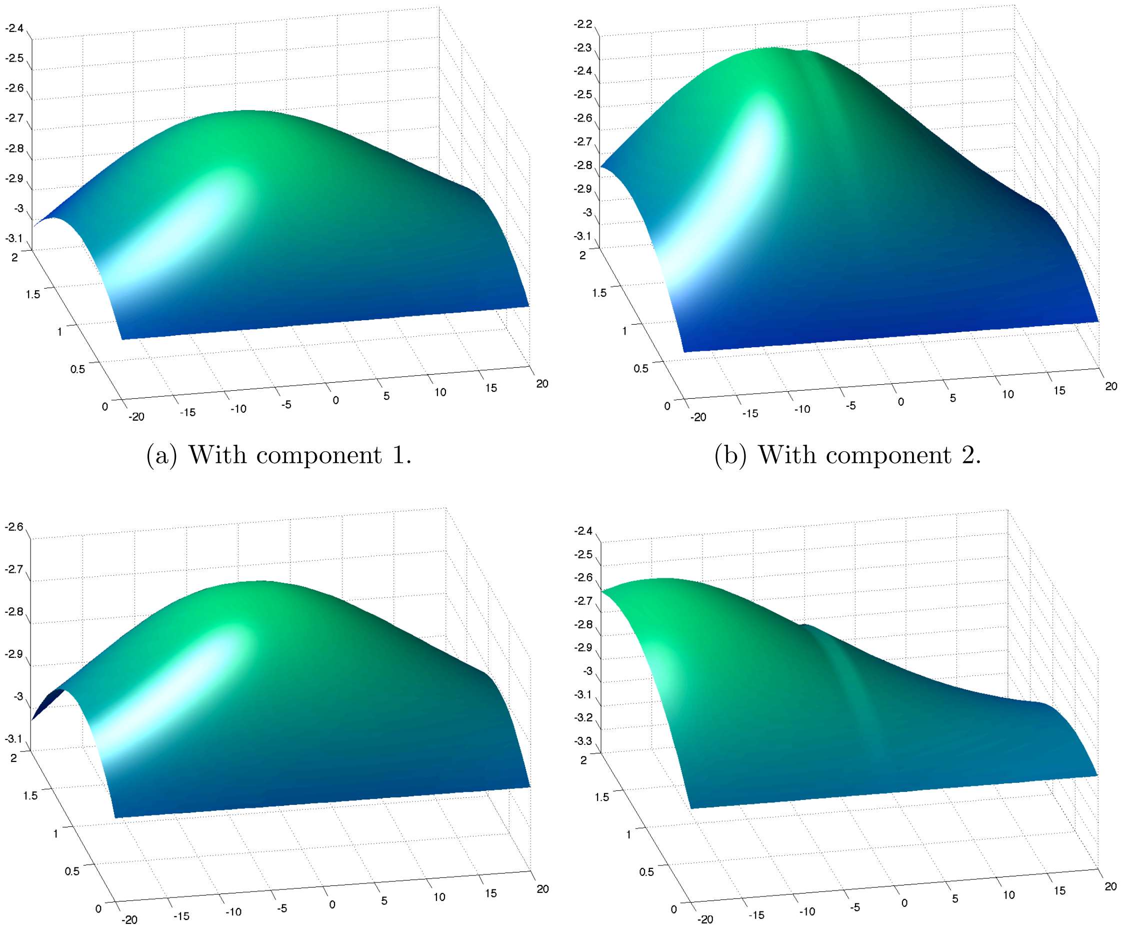
Where **pso** is the optimisation algorithm that is maximising the objective shown in Figure 5.1, which is Particle Swarm Optimisation (PSO); S is the reference spectra set and *distortion_models* is a set of functions with which the suspected distortions are simulated over S.

The objective of the framework is to identify the amount of distortion D and the concentration C of each component.

The supervision of a plant expert is essential for this method to provide congruent results. Every simulated distortion should have a physical meaning related to the sensor being used and the material being inspected. Artificially distorting the reference without such considerations may produce results without any real significance to the process.

5.1 Solution Space

This approach is aimed to be applied in different type of situations, and, therefore, it should be able to incorporate diverse types of information. Thus, the solution space that is desired to be optimised is expected to be unpredictably complex. To confirm this, using the components described in Section 4.2, a sampled spectrum was artificially created that suffered from random Local Shifts. In Figure 5.2, a graphical representation of the solution space for each component in the sampled spectrum are shown, plotted with concentrations ranging from 0 to 2 and shift values from -20 to 20.



(a) With component 1.

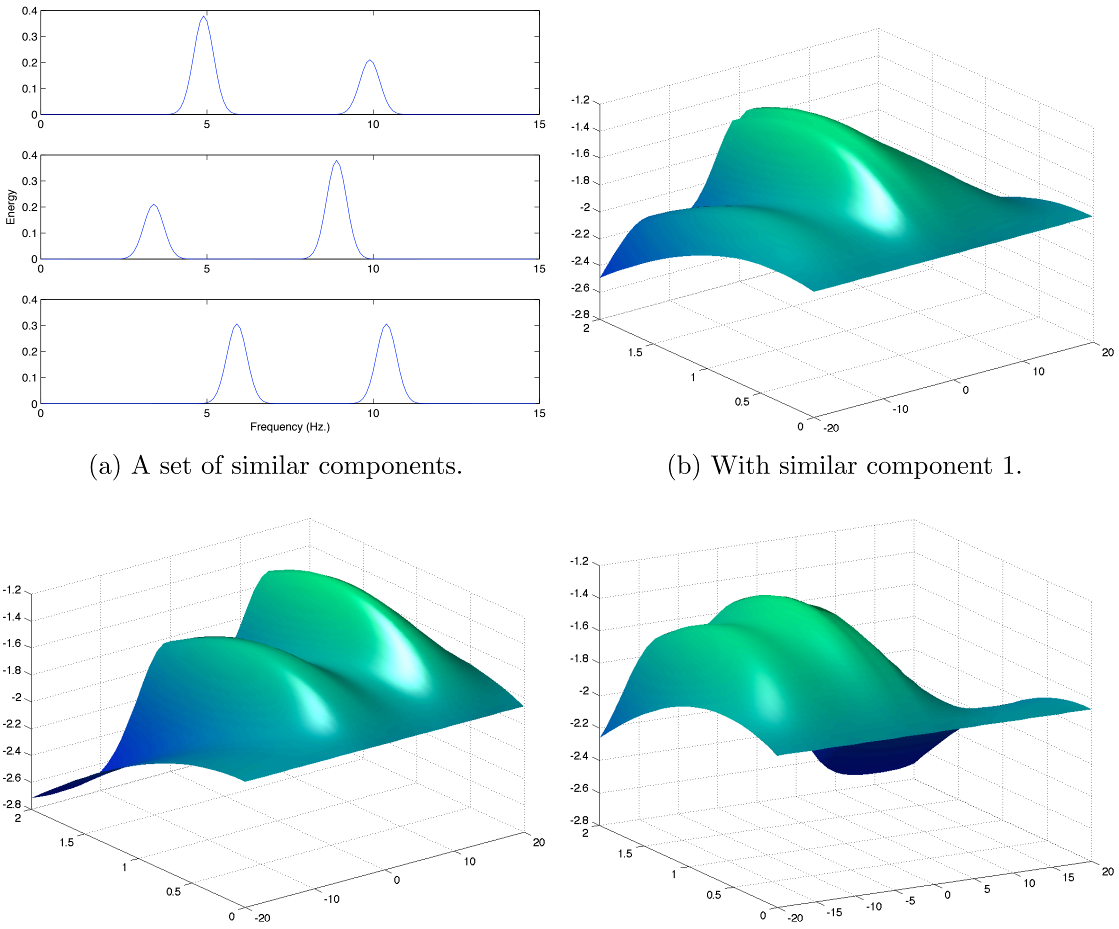
(b) With component 2.

(c) With component 3.

(d) With component 4.

Fig. 5.2: Solution space observed with reference spectra used in Section 5.2.

The solution spaces shown in Figure 5.2 are seen to be close to convex. However, in Components 2 and 4, shown in Figures 5.2b and 5.2d respectively, there is a small hill in the area near the shift value of 0. This corresponds to features that are similar between the components. To expose this issue further, another set of spectra which had very similar features were created, and shown in Figure 5.3a. The solution space that is created when adjusting the shift and concentrations values and comparing it to a random sample are shown for each component in Figures 5.3b, 5.3c, and 5.3d.



(a) A set of similar components.

(b) With similar component 1.

(c) With similar component 2.

(d) With similar component 3.

Fig. 5.3: Solution space observed with similar reference spectra.

As can be seen in Figure 5.3, the solution spaces will vary depending on the spectral shape of the components, and the presence of important local optima is to be expected. Similarly, applying different distortions will create different solution spaces. However, as seen in the tests in Section 3.4, PSO is able to find the global optima in solution spaces much more complex than the ones shown here.

The unpredictability of the solution space, based on the spectral signatures of the components and the different distortions taking place, which, in turn, are expected to be evolving from one sample to another, suggests a need for a Black-Box optimisation algorithm. Hence, an algorithm, such as PSO, is the only sensible approach if a generalised solution is to be developed.

To test the applicability of the proposed framework in different situations an experiment was constructed, described in the following section, in which a mixing plant which suffers from Local Shift is simulated. A second experiment, related to Musical Note Identification, was developed and is described in Appendix A.

5.2 Example: Use of Shifted Spectral Measurements in Feedback Control

This example describes an approach which identifies the concentrations of compounds in a material mixture, product of a simulated mixing plant whose spectral sensors are suffering from spectral shift. It does so by searching for the best linear fit of a reference spectra set and comparing it to the measured spectra, taking account of any shifts that may be present. Both the magnitude of the shift and the concentration of each spectral component can be identified using Particle Swarm Optimisation (PSO).

In Section 5.2.1, the simulation used to demonstrate the proposed approach is described; in Section 5.2.2, the results from a series of experiments are provided; and in Section 5.2.3, conclusions and several areas of future work are discussed.

5.2.1 Simulations

To test the suitability of using PSO in spectral processing rather than other spectral analysis tools such as Classical Least Squares Regression (CLSR) [40, 16], a mixing process under feedback control was simulated. Such process is defined in Figure 5.4. It consisted of a mixer of four ingredients, each with its own feed stream, the flow rate of which were controlled by their respective valve. The product of the mixing process was then monitored by a spectral sensor that provided a sampled spectrum of the material. This sensor suffered from an external factor that presented itself as a Local Spectral Shift in the sampled spectrum.

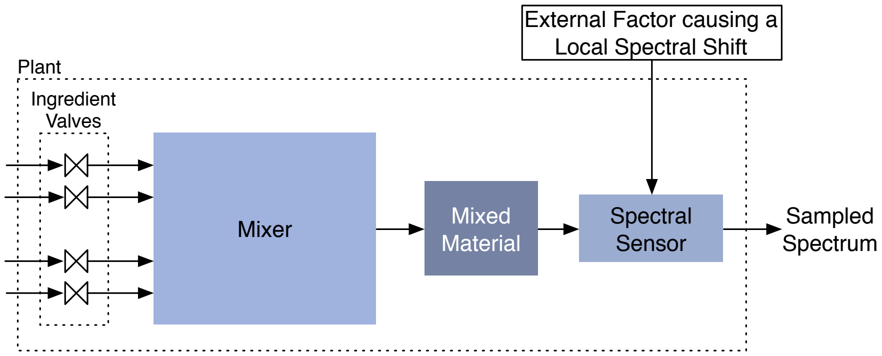


Fig. 5.4: Mixing plant to be simulated.

The objective of the control system was to achieve the desired concentrations of each of the four materials in the mixture by manipulating the valves of the feeds of each ingredient. The concentration of the materials in the mixture was not measured directly and was instead extracted from the spectral measurements. In this study, this was achieved by applying PSO and CLSR. A schematic of the simulated process is provided in Figure 5.5.

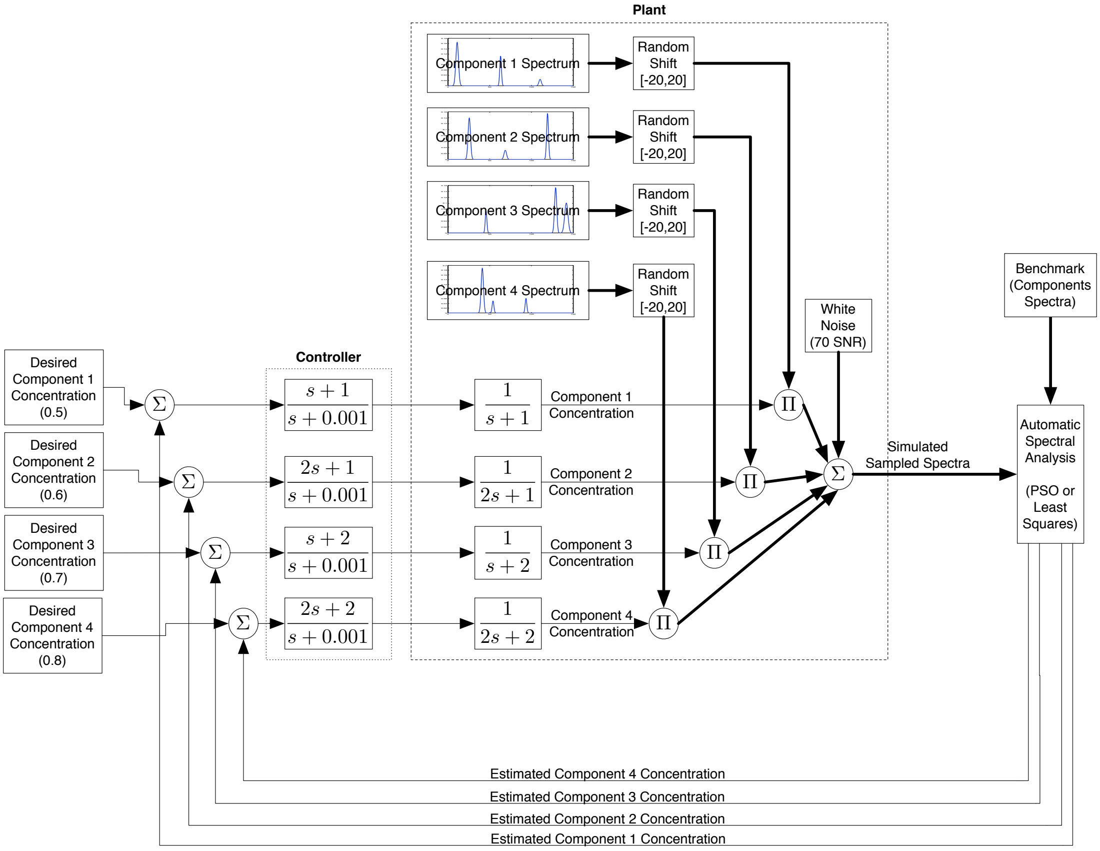

Plant
Component 1 Spectrum
Component 2 Spectrum
Component 3 Spectrum
Component 4 Spectrum
Random Shift [-20,20]
Random Shift [-20,20]
Random Shift [-20,20]
Random Shift [-20,20]
White Noise (70 SNR)
Benchmark (Components Spectra)
Controller
Desired Component 1 Concentration (0.5)
Desired Component 2 Concentration (0.6)
Desired Component 3 Concentration (0.7)
Desired Component 4 Concentration (0.8)
Σ
Σ
Σ
Σ
$\frac{s+1}{s+0.001}$
$\frac{2s+1}{s+0.001}$
$\frac{s+2}{s+0.001}$
$\frac{2s+2}{s+0.001}$
$\frac{1}{s+1}$
$\frac{1}{2s+1}$
$\frac{1}{s+2}$
$\frac{1}{2s+2}$
Component 1 Concentration
Component 2 Concentration
Component 3 Concentration
Component 4 Concentration
Π
Π
Π
Π
Σ
Simulated Sampled Spectra
Automatic Spectral Analysis (PSO or Least Squares)
Estimated Component 4 Concentration
Estimated Component 3 Concentration
Estimated Component 2 Concentration
Estimated Component 1 Concentration

Fig. 5.5: The whole simulated system.

Both methods require a set of reference spectra to work with. The four spectral components introduced in Section 4.2, were used in this study. These spectra are used inside the plant to simulate the measured spectrum of the mixture, where each spectrum is multiplied by their corresponding concentration and mixed with the other components. Finally, white noise is applied (with a signal-to-noise ratio of 70) to simulate measurement noise.

The control system was designed to achieve the desired concentrations of the individual compounds leaving the mixing vessel at specific values in centigrams (cg). These were: 0.5 cg for the first component, 0.6 cg for the second, 0.7 cg for the third, and 0.8 cg for the fourth. The response of the control system when there was no shift in the spectral measurement is shown in Figure 5.6. This figure shows a 20-second simulated response measured from the output of the transfer functions inside the plant. The response in Figure 5.6 shows that every component has reached its desired concentration. This response is referred to as the 'optimal response' of the control system and the one that it is aspired to when using CLSR and PSO to extract the component concentrations.

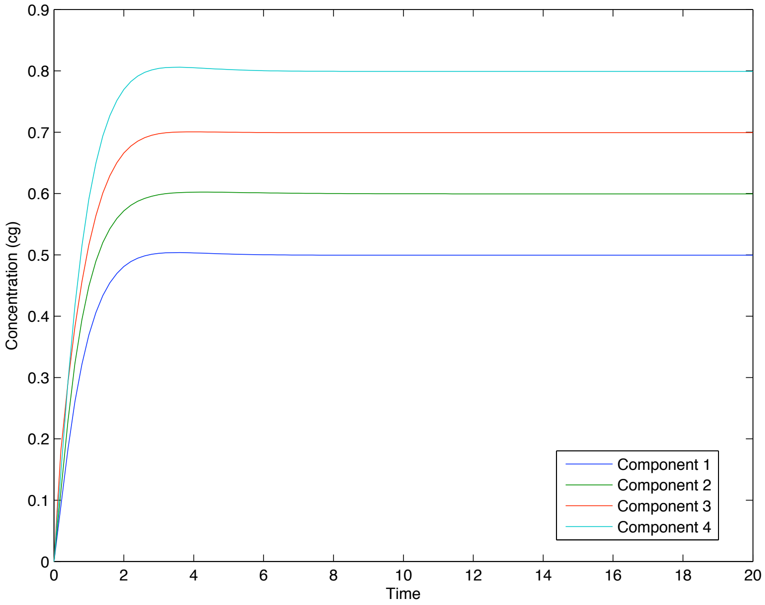


Fig. 5.6: Control response without using spectral data (i.e. no shift).

In the following section, the effect that frequency shift had on the control system is illustrated and the ability of PSO to recover the optimal control performance is investigated.

5.2.2 Results

Following the introduction of the frequency shift, two experiments were conducted. In the first experiment, Classical Least Square Regression (CLSR) was used to resolve the concentrations of each of the components in the mixture directly from the spectral measurement and those concentrations were fed back through the control system. In the second experiment, the concentration of each of the components in the mixture was identified with the aid of PSO, which accounted for the frequency shift.

5.2.2.1 Classical Least Squares Regression

Classical Least Squares (CLS) is a statistical method which can be used to obtain a set of pure-component spectra from a series of samples. These spectra can then be used to estimate their concentration in other samples using CLS Regression (CLSR) [40]. The CLSR method has been reported to be the most commonly used analytical tool for concentration estimation in Fourier Transform Infrared spectrometry (OP/FT-IR) [16], even so that many OP/FT-IR systems are supplied with a CLS analysis package [16].

The CLSR approach requires knowledge of the reference spectra for each of the components in the mixture. This reference, S, is often available and using the spectra obtained from the mixture, D, the concentrations of each of the components can be identified using the following expression:

$$C = DS(S^T S)^{-1} \tag{5.3}$$

The concentration obtained using (5.3) were then fed back through the controller and the resulting concentrations within the mixture are shown in Figure 5.7. This figure shows that the response of the controller has been severely degraded as a result of the shift, when compared to the response of the control system obtained in Section 5.2.1 when there was no shift present.

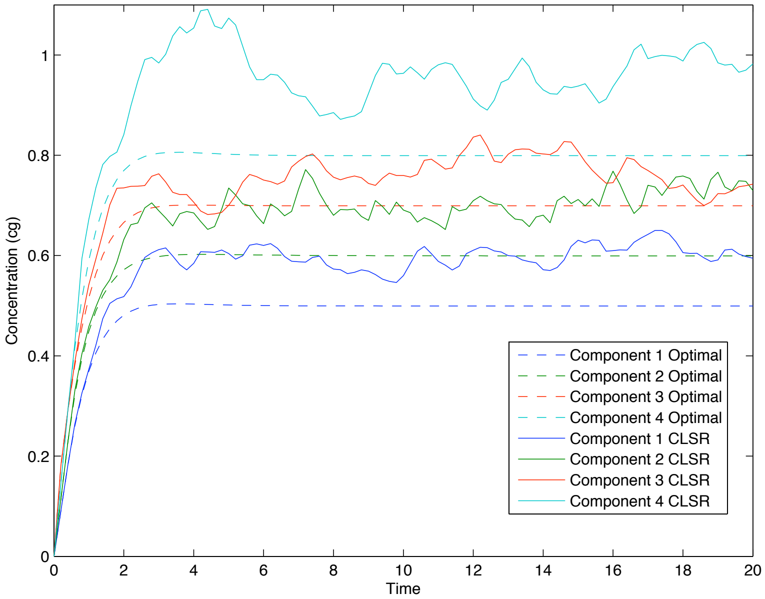


Fig. 5.7: Control response using shifted data and CLSR as spectral analysis tool.

5.2.2.2 Using Particle Swarm Optimisation

In this test PSO was used to estimate the concentrations of the components from the spectrum obtained from the mixture. In the PSO, a swarm of 50 particles was created, with each particle considering 5 nearby particles as neighbours. Each particle had 8 directions to 'fly' in, two per component: one direction dealt with the concentration of the component and the other with its spectral shift. To grade the fitness of each particle, each reference spectrum was shifted and multiplied by its corresponding values derived from the location of the particle. The four modified spectra were then linearly mixed, resulting in a 'test spectrum', which was then compared to the spectrum obtained from the plant by calculating the Mean Square Error (MSE) between them. The MSE is a measure of the 'distance' between the ideal value and the acquired value, calculated by the square root of the addition of the squared differences between them at each frequency point. In effect, it is the Euclidian distance between the two spectra, meaning that a value close to zero is desired.

The value of each concentration direction was limited between 0.01 and 1 cg and its velocity between -0.1 and 0.1; every shift direction was limited between -30 and 30 frequency points and its velocity between -3 and 3. These values were obtained empirically and were found to give the best results in this application.

Time-decreasing inertia was also applied, increasingly reducing the velocities from 0% to 60% of their value in the first 300 iterations, and maintaining that value for the remainder of the search. Convergence to a solution was assumed when 120 iterations passed without a change in the identified global optima. The response of the process when the concentrations obtained using the proposed PSO technique were fed back through the control system are displayed in Figure 5.8.

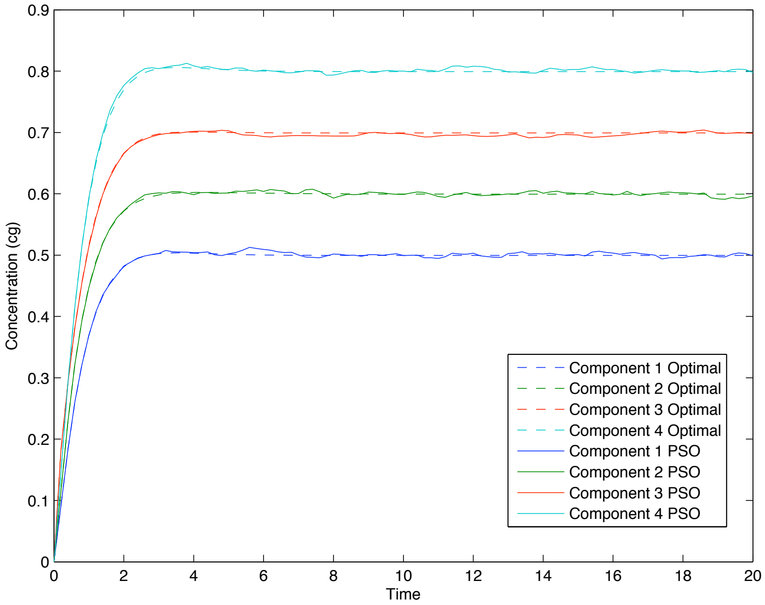


Fig. 5.8: Control response using shifted data and PSO as spectral analysis tool.

Comparing Figures 5.7 and 5.8 shows that the response using the PSO spectral processing technique enables the controller to track the desired response much more closely than was obtained using CLSR. For a more precise comparison, the MSE between each response and the optimal one is provided in the Table 5.1 and it can be seen that PSO outperforms CLSR significantly.

Component	1	2	3	4
MSE (CLSR)	0.9639	1.0171	0.6966	1.6604
MSE (PSO)	0.0339	0.0319	0.0375	0.0373

Table 5.1: Mean Square Error using PSO and CLSR

It is important to note that there are still variations in the control response, as shown with the response of Component 4 in Figure 5.9, which shows a close-up of the period between 7 to 20 seconds of Figure 5.8. However, such variations are still mostly inside the 99% confidence bounds, as it is shown in Figure 5.9, which is an important improvement over using CLSR when Spectral Shift is involved.

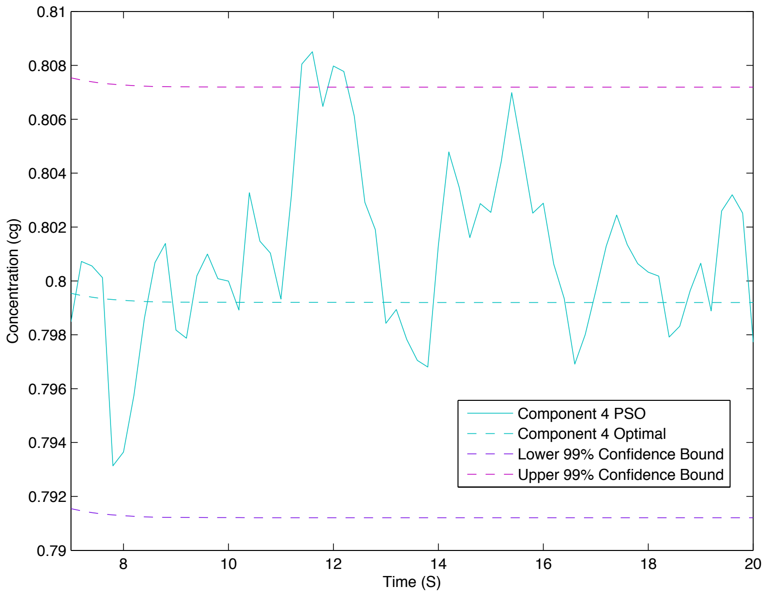


Fig. 5.9: Control response of Component 4 between 7 and 20 seconds.

5.2.3 Conclusions & Discussion on Mixing Plant Example

It has been shown that using common automatic spectral analysis algorithms, such as Classical Least Squares Regression, may lead to poor control performance if these measurements are used directly in a control system. The reason for this is that frequency shift may be present in the measurements and this can have a significant effect on the controller. CLSR assumes the presence of each spectral component in a constant frequency location, a variation of which results in a deviation of the relationship between frequency and spectral intensity. When this relationship changes and nothing is done to compensate for it, the estimated values of the concentrations of each component will become inaccurate.

The Spectral Analysis Framework was able to consider the possibility that the frequency-intensity relationship can change, and compensated for it by searching for the magnitude of shift affecting each component as well as their concentration. This lead to accurate estimates of the concentrations, when compared to those produced by CLSR when shift was introduced into the system. Using the concentrations identified using PSO, the performance of the control system was comparable to that achieved when there was no shift present in the spectral measurements.

5.3 Conclusions & Discussion

A framework was proposed with which an initial generic solution to Spectral Component Analysis could be created. Two experiments from different areas were investigated, and the proposed Framework was able to cope and provide good results in both. It was also commented that it is important to consider that a generic solution for a specific problem may not provide the best efficiency. The proposed framework is meant as an initial gateway into which the knowledge of a process expert can be experimented with, and either confirm or deny the presence of distortions that may be suspected of occurring. It is then possible to use the knowledge, confirmed by this framework, to create a solution much more specific to the problem at hand, improving the efficiency of the process.

For example in Appendix A the Framework was applied for the task of Polyphonic Music Retrieval. It was found that, even though the Framework was successful, another solution could be developed that was more efficient. However, the knowledge used to create this solution was confirmed using the proposed Framework, to the point that some limitations of the Framework were still in place in the more efficient method. Furthermore, the fact that the proposed framework could be used in the Musical arena and provide good results is a testament of its versatility.

An important requirement of the proposed Framework is that of a set of reference spectra. It could be assumed that these are available, but it is possible that this may not be the case. Component extraction methods may be able to obtain a set of reference spectra from sampled measurements. However, as discussed in Chapter 4, current methods are sensitive to spectral distortion, which is to be expected if the proposed Framework is already being considered to be used. In the following chapters, the Framework is extended to the area of component extraction from distorted spectral data.

Chapter 6

Alignment of Generally Distorted Spectra

As discussed in Chapter 5, the Spectral Analysis Framework (SAF) proposed in this thesis is able to analyse spectral data that has been distorted in several ways. However for this analysis to take place, a set of reference spectra is required and in many applications this may not be available. One solution is for Component Extraction methods to be applied to a set of sampled spectra so that a set of components, which the SAF can work with, can be extracted. Unfortunately, as discussed in Chapter 4, current Component Extraction methods are sensitive towards spectral distortion.

As observed in Section 2.2.1, the ways in which a spectrum may be distorted can be categorised in two groups: the Global type, in which the distortion occurs uniformly throughout the spectrum, and the Local type, where the amount of distortion is different for each component. This chapter will describe an algorithm that enhances current Component Extraction only when dealing with Globally distorted data. The topic of Component Extraction from Locally distorted data will be discussed in Chapter 7.

In this chapter, an algorithm is proposed that re-aligns globally distorted data. This re-alignment is shown to enhance the results of current Component Extraction methods. In the examples provided in this chapter, Global Warp and Global Shift distortions are discussed. However, the proposed algorithm was developed so that it could incorporate any type of Global distortion.

6.1 Alignment Using an Optimisation Approach

The objective of the proposed aligning algorithm is to remove or 'cancel' the global distortions taking place in a given spectral data set. In the following explanation, it is assumed that only Global Shift and Global Warp distortions are present.

The alignment method is divided in three phases:

1. **Choosing a reference**. A sample is chosen to be the reference by which every other sample is to be aligned.

2. **Aligning with reference**. Every sample is artificially modified, using a set of pre-defined distortions, such that it best fits the reference sample. The pre-defined set of distortions would be defined by a plant expert with knowledge of the likely distortions.

3. **Rectify reference distortions**. The reference sample may have been distorted itself, therefore the whole set is rectified to compensate.

A summary of the algorithm is provided in Figure 6.1.

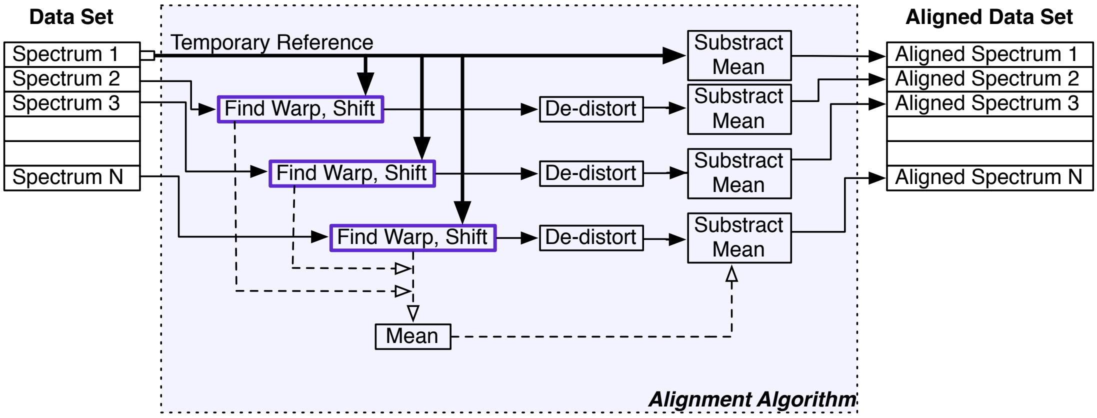


Fig. 6.1: Summary of aligning algorithm with only Warp and Shift being applied.

A detailed description of each step in the algorithm is given in the following paragraphs.

Choosing a Reference for Alignment. A reference sample is chosen randomly. It can be argued that the chosen sample may not be appropriate as a reference, as some components may not be present. Nevertheless, because the spectral samples during the alignment procedure are going to be artificially distorted uniformly, it can be expected, and was observed in tests, that good results are provided if at least one component is present in the chosen sample. It is reasonable to assume that all samples will have at least one component, hence, any sample is an appropriate choice to be used as a reference.

Alignment Procedure. Having defined a reference spectrum, every other sample is then aligned to it. To do this, a set of artificial distortions is defined, which every other sample will be subjected to in an effort to identify the precise nature of the distortion in the data. An optimisation algorithm, in this work Particle Swarm Optimisation, is then applied to determine the magnitude of each distortion that needs to be applied to a sample such that it is aligned with the reference.

The objective function used by the optimisation algorithm is based on the Pearson Correlation coefficient, as only the shape of the spectrum is required to be aligned, without considering their magnitudes. The objective function is given in (6.1).

$$M = \frac{\sum (R - \overline{R})(X_{i_d} - \overline{X_{i_d}})}{\sqrt{\sum (R - \overline{R})^2 \sum (X_{i_d} - \overline{X_{i_d}})^2}} \tag{6.1}$$

where M is the measure of correlation between the reference sample (R) and the artificially distorted sample (X_{i_d}), $\overline{R}$ and $\overline{X_{i_d}}$ are the mean values of R and X_{i_d} respectively, and X_{i_d} is defined by (6.2).

$$X_{i_d} = \textbf{distort\_signal}(X_i, \textit{global\_shift}, \textit{global\_warp}) \tag{6.2}$$

where X_i is the ith sample in the spectral data set X. *global_shift* and *global_warp* are the amounts of Global Shift and Global Warp, respectively, applied to X_i to create X_{i_d}. The **distort_signal** function shifts and warps X_i by the amounts given by *global_shift* and *global_warp*; it is described in detail in Appendix C.

The PSO search aims to find the values of *global_shift* and *global_warp* that create a X_{i_d} that best aligns with R.

Compensating for Reference Distortions. The reference sample R was chosen from a data set that suffered from distortion. Therefore, it is reasonable to assume that R suffers from distortion itself. This means that the whole data set will be biased towards the distortions taking place in R, which may introduce complications later. To compensate for this, it can be assumed that the data set is now aligned, and that a bias is present uniformly throughout the data set. To remove this bias it is necessary to estimate the distortions taking place in R. To do this, the amount that each sample was artificially distorted by to align with R is recorded and, depending on the assumed statistical distribution of distortion values, the amount that R was distorted by can be estimated. For example, if a uniform distribution is assumed, the mean of all the distortion values can be used as an estimate of the distortion taking place in R.

Having estimated the distortion present in R, all the samples of the data set (including R) have this distortion removed, resulting in an aligned data set without bias.

6.1.1 Experiments & Results

To verify the applicability of the proposed aligning algorithm, two reference spectra were randomly created in the same way as described in Section 4.2. These spectra are shown in Figure 6.2. The reference spectra were the basis for creating a data set of 50 spectral samples, each warped and shifted by different amounts uniformly distributed between 0.95 and 1.05 (for warping) and ±20 *fp* (for shifting), shown in Figure 6.3.

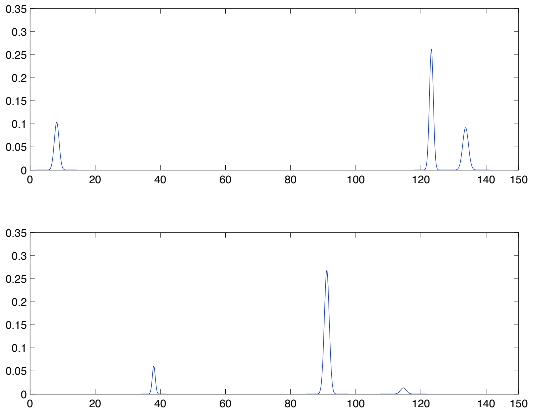


Fig. 6.2: Reference spectra used to create data set in Figure 6.3.

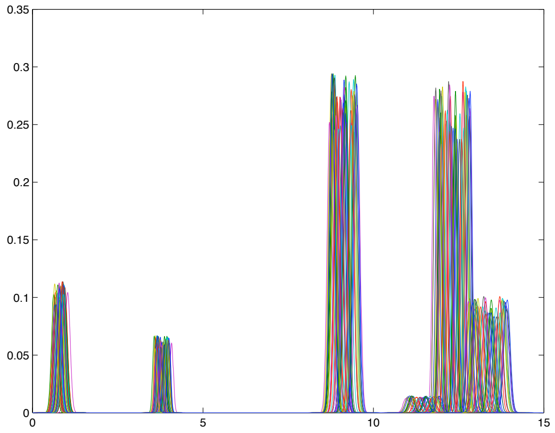


Fig. 6.3: Data set used in experiment.

The alignment procedure was applied, using the first sample as the reference. The aligned data set is shown in Figure 6.4.

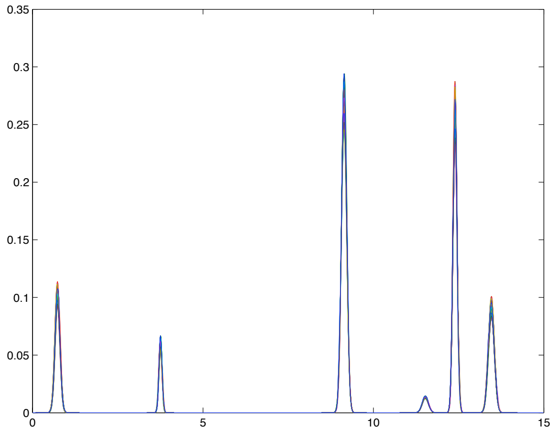
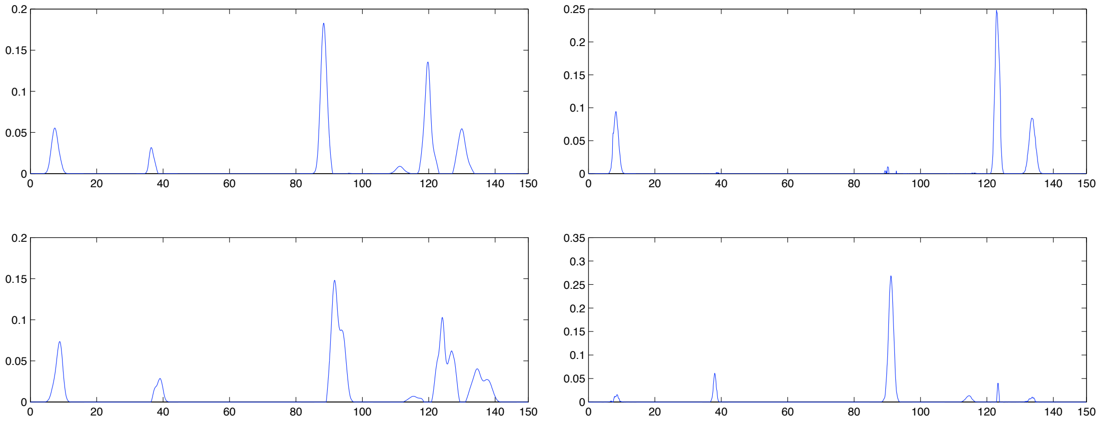


Fig. 6.4: Data set after alignment.

The compensation phase, which assumed a uniform distribution of the distortion values, estimated that the reference sample was shifted by -10.84 *fp*, which can be approximated to -11 *fp*, as the shift can only take integer values, and warped by 1.0158. The real distortion values for the first sample were a shift of -11 *fp* and a warp of 1.0142, indicating that the proposed algorithm has identified the shift well in this data set.

To see if the global alignment method does enhance the performance of component extraction algorithms, Alternating Least Squares was applied to both the original and the aligned data set, and for each case, two components were extracted. The components identified in each data set are shown in Figure 6.5.



(a) Using original data set. (b) Using pre-aligned data set.

Fig. 6.5: Extracted components from original and aligned data set.

As seen in Figure 6.5, the components that ALS extracted from the aligned data set are more similar to the reference spectra in Figure 6.2 than those extracted from the original data set. A correlation measure between the reference spectra and each component is given in Table 6.1. This table shows there is a significant increase in performance when the data set was pre-aligned.

	From original data set	From pre-aligned data set
Comp. 1	0.0345	0.9879
Comp. 2	0.5651	0.9938

Table 6.1: Correlation between extracted components and reference spectra.

6.1.2 Solution Space Observed

The solution space that the PSO search was given to optimise in the above example can be observed in Figure 6.6, where a sample was shifted and warped by different amounts and compared to a reference sample. This figure shows that the space is almost convex and, for this type of problem, PSO has been shown to perform well [29].

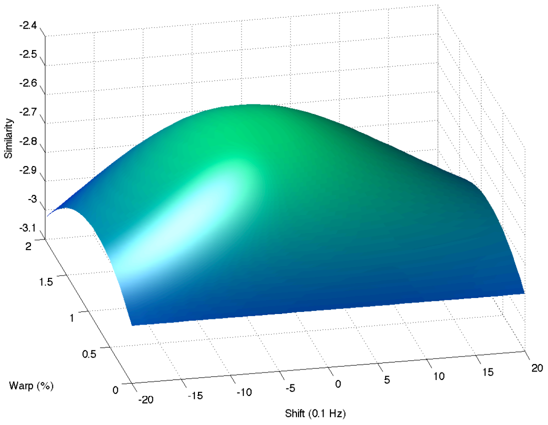


Fig. 6.6: Solution space observed with data in previous experiment.

However, when using other data sets, such as those described in Chapter 5, a solution space as complex as the one shown in Figure 6.7 can be obtained.

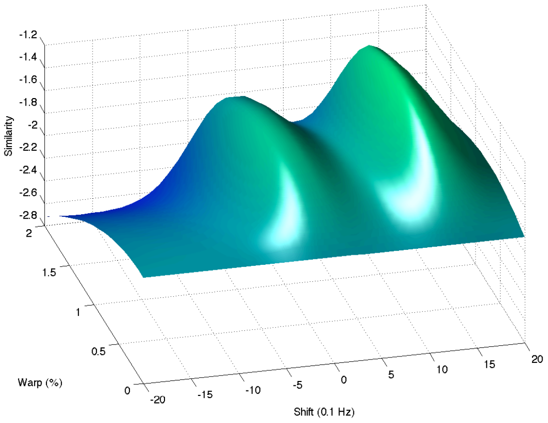


Fig. 6.7: Solution space observed using other spectra (Chapter 5).

The solution space that PSO is required to optimise is defined purely by the shape of the spectra and the types of distortions taking place, making the optimisation problem unpredictable as well as intricate. Fortunately, PSO is able to cope with these complex solution spaces. Given that the shape of the solution space will be unknown for each problem, a Black-Box-Oriented optimisation algorithm, such as PSO, is well-suited to solve it.

6.2 Conclusions

It has been shown that when a spectral data set is distorted in a global manner, a relatively simple algorithm can be used to remove this distortion. This was shown to be important if a component extraction algorithm is to be applied to the data set. A case study was used to show that the application of the pre-alignment algorithm enhanced the ability of ALS to subsequently identify the underlying components in the data set. However, the proposed algorithm assumes that a plant expert is able to identify likely distortions present in the data.

It is important to reiterate that the alignment procedure aims to preserve the overall shape of the spectrum, thus, only artificial distortions that affect the signal globally were applied. The algorithm presented here is not able to re-align spectra that was distorted in a local level. To do that it would be necessary to change the overall shape of the spectrum, which not only contradicts the physical factors of the sensor and the inspected material but, as discussed in Section 4.3.3.1, also degrades the results of any component extraction algorithm. Extracting the components out of locally distorted data is addressed in the following chapter.

Chapter 7

Component Extraction on Locally Distorted Spectra

Global distortions, as seen in the previous chapter, can be resolved beforehand using a pre-aligning approach, where the spectrum is modified in a uniform manner, based on expert knowledge of the system. By doing so, it was shown that analytical tools, such as Alternating Least Squares, were able to extract the components from re-aligned data that was globally distorted.

Unfortunately, local spectral distortions, such as Local Shift or Local Warp, are difficult to remove using a pre-aligning approach without damaging the overall shape of the spectrum. The effects of re-aligning a locally distorted data set before it is analysed have been explored in Section 4.3.3.1, where it was concluded that a post-processing approach should be used to complement established analysis tools, such as Independent Component Analysis. Another alternative would be to create a new component extraction method that makes the fundamental assumption that the data being analysed may be distorted.

In this chapter, two methodologies are presented that are successful in extracting components from locally distorted data. One post-processes the results obtained using ICA when the data is locally shifted, and the other is a novel method that extracts the components after assuming that local distortion is present within the spectral measurements.

7.1 Post-Processing Estimates of Independent Component Analysis

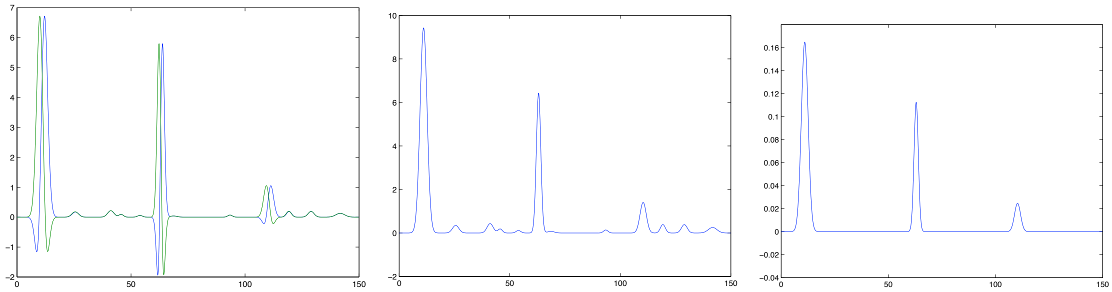
As observed in earlier chapters, when applying ICA to locally shifted spectral data, several more significant components are identified than would be otherwise. By combining these components together, the proposed post-processing method is able to accurately identify the source components in the spectrum. Finding the most appropriate combination of the identified components can be formulated as a non-linear optimisation problem.

Independent Component Analysis was tested with artificial data sets suffering from different degrees of Local Shift in the experiments described in Section 4.3.3. The effect of Local Shift in ICA can be observed in Figure 4.6, where several ‘similar’ components were identified instead of just one. These partial components are different from the sources not only in peak locations, but in peak shapes as well, so they cannot be considered as source estimates by themselves.

7.1.1 Proposed Post-Processing Algorithm

Figure 7.1a shows two related components that were identified when ICA was applied to an artificial data set, suffering from a maximum Local Shift of 1 *fp*, as described in Section 4.2. Figure 7.1b shows the spectra that results from simply adding these two spectra together. Figure 7.1c shows the reference spectra that is most similar to the two spectra identified using ICA. These figures clearly show that by adding the two partial components together, an accurate approximation of the reference spectra is obtained.

The combined component illustrated in Figure 7.1b is referred to in this paper as the Estimated Independent Component (EIC), and is the estimate of the source that the combined ICs are related to. This relatively simple technique provides a feasible solution to identifying the source components in a spectrum affected by frequency shift. However, exhaustive testing has shown that when more partial components are identified, the simple addition of related components does not produce accurate approximation of the source spectra.



(a) ICs found to be related. (b) Combination of the ICs. (c) Source the ICs are related to.

Fig. 7.1: Result of combining related ICs.

The post-processing algorithm proposed here operates in 3 stages:

1. Partial components which are related are identified and grouped.

2. The grouped components are combined to produce an optimal EIC.

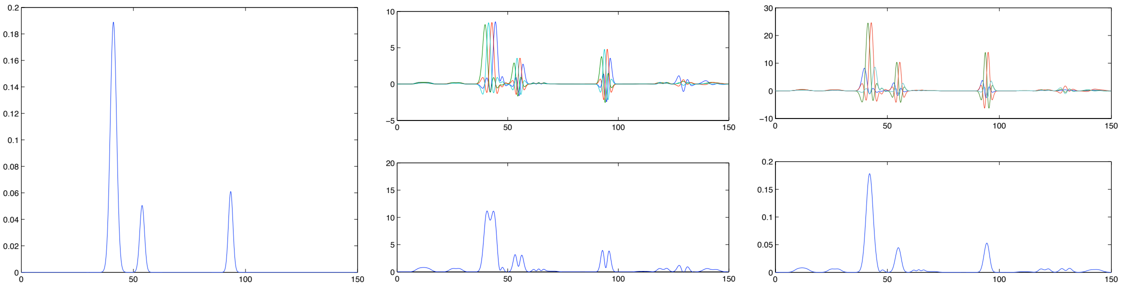
3. Final processing of the EICs is undertaken to remove artefacts from other components.

Each of the these three stages is now described in detail:

Grouping Related Components. Two components that are 'related' to each other will have a high correlation coefficient in an area near the origin in their Normalised Cross-Correlation Vector ($NCCV$). In this example, an area of 40 *fp* ($\sim$ 8 Hz) with a cut-off value for the $NCCV$ of 0.7 gave the best results in terms of finding which components were 'related'. However, further studies suggested that the cut-off value of 0.7 was too strict. A method of finding a balanced cut-off value was found by applying all the values between 0.4 and 1 (with a step size of, say, 0.01) and recording the number of groups of ICs that were obtained for each value[1]. The number of groups most frequently recorded was found to be a good estimate of the number of sources in the data, and any cut-off value that produced this number was appropriate. However, the cut-off value will not always be optimal, and some components may get left out. In such cases, user intervention may be necessary.

[1] Any values lower than 0.4 may give false positives of relation between ICs.

Combining Related Components. When only two components are found to be related, scaling them to be of the same height and adding them often provides a reasonable approximation to a source spectrum, as shown in Figure 7.1. However, when more than two components are found to be related, it is not enough to equalise heights and combine them, as demonstrated in Figure 7.2. In this example, four 'grouped' components are identified and displayed in the upper graph of Figure 7.2b. The resulting EIC displayed in the lower graph of Figure 7.2b has a very different shape to the source, shown in Figure 7.2a.



(a) Source.

(b) ICs from a 2 Hz maximum shift.

(c) Shape-corrected EIC.

Fig. 7.2: Result of PSO search to correct the shape of the linear mix.

To resolve this problem it is necessary to find the optimal combination of the scaling factors for the related components. For example, the upper graph of Figure 7.2c shows the four, re-scaled, grouped components which, when added together, provide the EIC in the lower graph of Figure 7.2c, which accurately describes the source in Figure 7.2a. To find the optimal combination, Particle Swarm Optimisation (PSO) was applied. The optimal EIC was defined as:

$$EIC = \mathbf{shift}(IC_1 + IC_n + \sum_{i=2}^{n-1} a_i IC_i, l) \tag{7.1}$$

and the following function was minimised

$$P = \min(\mathbf{Pearson}(EIC, \mathit{data\_sample})) \tag{7.2}$$

where IC_i is the ith IC in the group, n is the number of ICs inside the group, a_i is the weighting factor of the ith IC, the **Pearson** function is a measure of similarity based on the Pearson product-moment coefficient, and *data_sample* is a randomly-chosen sample spectrum from the data set. PSO aims to find the optimal combination of weighting factors (a_is) for each IC_i. The range of the weights was chosen empirically to be between 0.8 and 3 as in the examples studied, they provided good optimisation speed without losing accuracy. It needs to be considered that the EIC is being compared to a data sample that may be shifted itself. So, to obtain an optimal fit, the EIC is artificially shifted an amount l that PSO also aims to find.

The Pearson coefficient was used as it only takes into account the similarity of the shapes of the spectra, and not their magnitudes. To force only one optimal combination to exist, the first and last ICs of the group (IC_1 and IC_n) remain unchanged throughout the search. If all the ICs vary, different combinations would exist that give the same measure of optimality.

Final De-Correlation. The lower graph in Figure 7.2c shows that when calculated, the EIC may contain artefacts from other components (e.g. the small peaks between 20-35 Hz), implying that the EICs are not completely de-correlated. To reduce this correlation, the following update to every EIC is applied until convergence is reached:

$$EIC_i \leftarrow EIC_i - \left(EIC_j \frac{EIC_i(f_{j_m})}{EIC_j(f_{j_m})} \right) \tag{7.3}$$

where EIC_i is the EIC to be updated; EIC_j is any other EIC; f_{j_m} is the frequency location with the most energy in EIC_j.

7.1.1.1 Estimated Independent Components in Further Analysis

Once identified, the EICs can be used as a reference spectra for further analysis, such as determining the concentrations of the various components in a measured spectrum. This can be achieved by finding a combination of scaling factors and shift values for each EIC, which, when combined, create a spectrum that is the closest in shape to each of the measured spectra. To find such a combination, PSO can be applied to optimise the function, P:

$$P = \min(E) = \min \left(\sqrt{\sum_{f}^{\mathrm{F}} \left(C(f) - \frac{MS(f)}{\|MS\|} \right)^2} \right) \tag{7.4}$$

where C is:

$$C = \sum_{i=1}^{n} (c_i \mathbf{shift}(EIC_i, l_i)) \tag{7.5}$$

meaning that C is the spectrum created when applying the concentrations c_i and shift values l_i to their respective normalised EICs (EIC_i) and adding them; MS is the measured spectrum; $C(f)$ and $MS(f)$ are the energies located at frequency f in both spectra of size F. E is a measure of dissimilarity between C and the normalised measured spectra ($MS/\|MS\|$), based on the Euclidean distance between them, which has been found to give good results with this method. P is the minimum distance, which identifies the optimal combination of concentrations and shift values that best fit MS.

When the optimal combination is found, the estimated concentrations of the normalised version of the sources inside the measured spectra ($\hat{x}_i$) can be calculated by:

$$\hat{x}_i = c_i \|MS\| \tag{7.6}$$

This optimisation approach was applied to the simulated feedback control loop, described in detail in Section 5.2.

The proposed algorithm was tested in different scenarios, described in the next section.

7.1.2 Case Studies

The ability of the proposed post-processing technique to compensate for shift in artificial and real sets of spectral data was investigated. The first case study test the proposed post-process algorithm with artificial data sets that suffer different degrees of Local Shift, and then uses the extracted components as reference spectra to estimate the concentration with a set of artificial test spectra. The following study uses a set of NIR spectra sampled from pharmaceutical tablets, which suffer from slight local shifts, which can be attributed to variations in the sampling and calibrating procedures. The final case study tests the algorithm with a group of NIR measurements of carbonised ice sampled at different temperatures.

7.1.2.1 Case Study 1: Artificial Data Sets

For the following experiments, 20 different data sets were created as described in Section 4.2, meaning that each sampled spectrum was composed of four components. Each data set had a different maximum shift value, which ranged from 1 *fp* (0.1 Hz) to 20 *fp* (2 Hz).

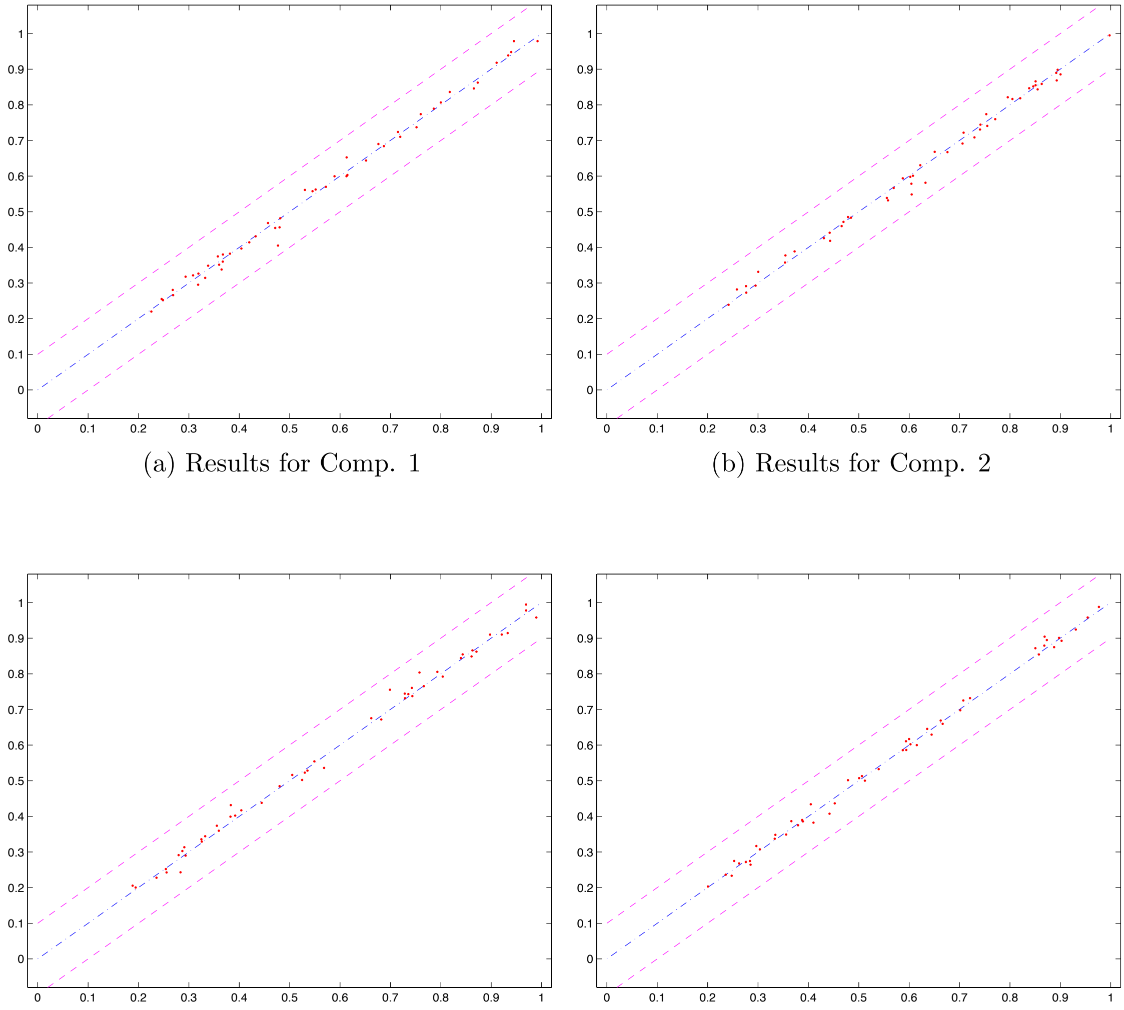
The proposed post-processing technique was applied, and for each data set, four EICs were automatically identified, matching the number of components expected to be extracted. This means that the technique was able to successfully estimate the correct number of components even in cases of severe shift.

The maximum correlation coefficient between the reference sources and their corresponding EICs shifted between -4 and +4 Hz, was used as a measure of similarity. Table 7.1 provides the mean value of the similarity metric for each data set, denoted by their maximum shift value. The bold numbers highlight the most and least similar sets of EICs. These values indicate that the proposed approach successfully identified the four sources even in situations of severe shift.

Set	Corr.	Set	Corr.	Set	Corr.	Set	Corr.	Set	Corr.
0.1	0.9997	0.5	0.9945	0.9	0.9922	1.3	0.9924	1.7	0.9926
0.2	**0.9999**	0.6	0.9924	1.0	0.99924	1.4	0.9935	1.8	0.9917
0.3	0.9899	0.7	0.9930	1.1	**0.9871**	1.5	0.9937	1.9	0.9902
0.4	0.99942	0.8	0.9934	1.2	0.9923	1.6	0.9928	2.0	0.9882

Table 7.1: The average similarity metric of each EIC-set against the reference spectra.

To provide a measure of the accuracy with which PSO can estimate concentrations in measured spectra using these EICs, another data set was generated with a maximum shift value of 1 Hz. In Figure 7.3 and 7.4, the estimated and real concentrations are plotted, using the most and least similar EICs respectively, and it can be seen that the estimated concentrations are well within the 90% confidence bounds. The mean square error being 2.9707E$-$04 and 9.9480E$-$04 for the 0.2 Hz and 1.1 Hz EIC sets respectively. These results highlight the accuracy with which PSO can estimate the concentration of each identified component in the mixtures.



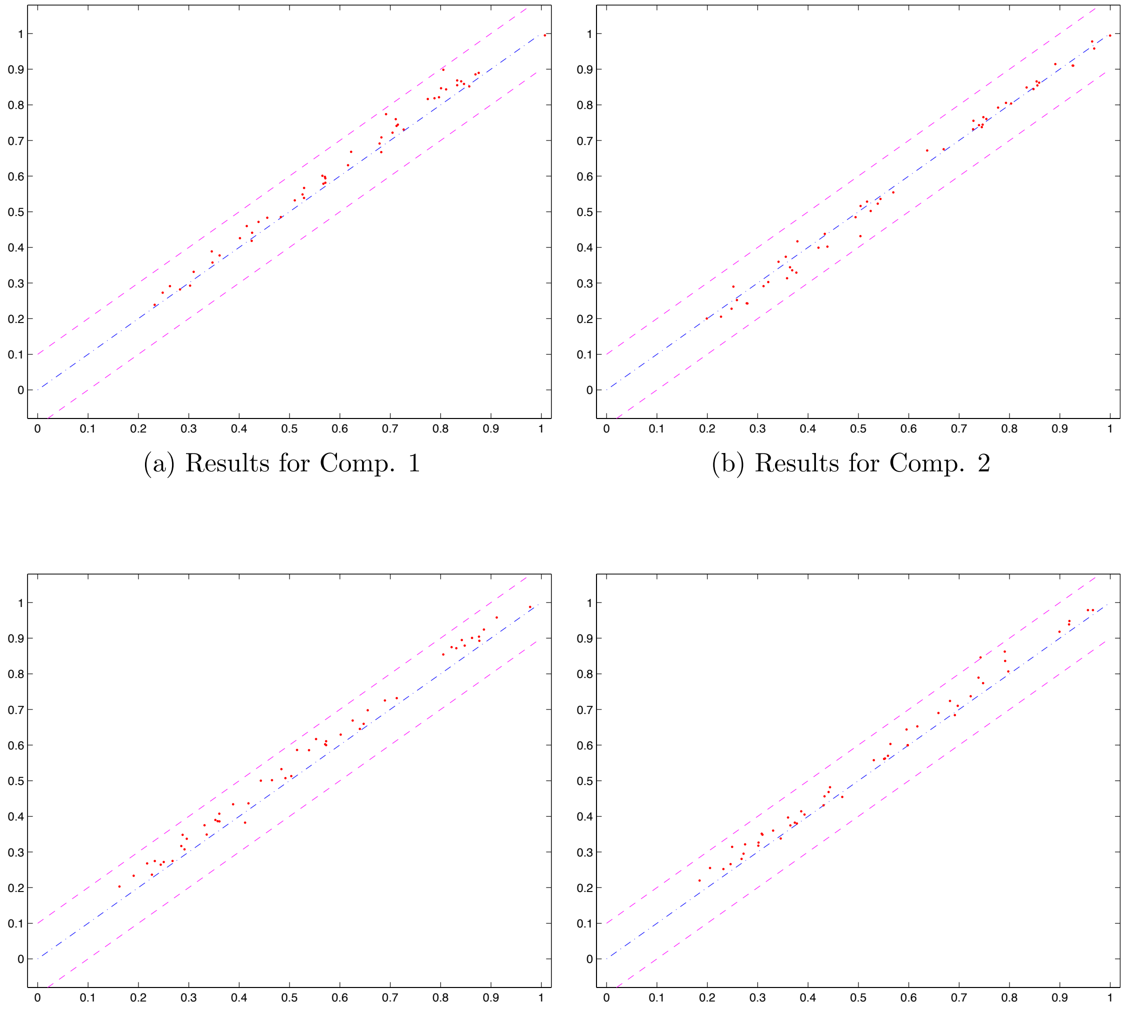
(a) Results for Comp. 1

(b) Results for Comp. 2

(c) Results for Comp. 3

(d) Results for Comp. 4

Fig. 7.3: PSO performance using the most similar EICs (0.2 Hz set), with 90% confidence bounds.



(a) Results for Comp. 1

(b) Results for Comp. 2

(c) Results for Comp. 3

(d) Results for Comp. 4

Fig. 7.4: PSO performance using the least similar EICs (1.1 Hz set), with 90% confidence bounds.

7.1.2.2 Case Study 2: Pharmaceutical Tablets

The data set that was used in this study is part of a publicly available database consisting of 310 NIR spectra, sampled from pharmaceutical tablets of Escitalopram® [28]. It is composed of several batches, differing by the size of the tablet (5, 10, 15 and 20 mg), and each batch has 3 sub-batches that differ in production scale (full scale, pilot scale, and laboratory scale).

The mixture inside the tablets is comprised of an active ingredient and several excipients, such as mycrocrystalline cellulose (dominant), magnesium stearate and talc. In this section, the results using the laboratory-scaled, 5-mg-tablets batch is presented. These results were comparable to those from the other data sets. Figure 7.5 shows the spectra from this batch of measurements.

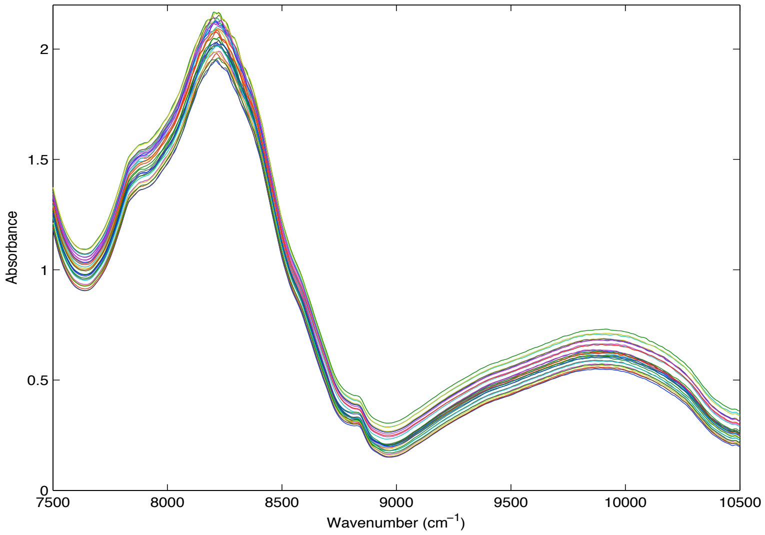


Fig. 7.5: Spectral data of pharmaceutical tablets.

Detailed examination of the spectra in Figure 7.5, shown in Figure 7.6, suggests that there is some frequency shift in the data, specifically in the area between 8050 and 8400 cm^{-1}, as well as the presence of noise.

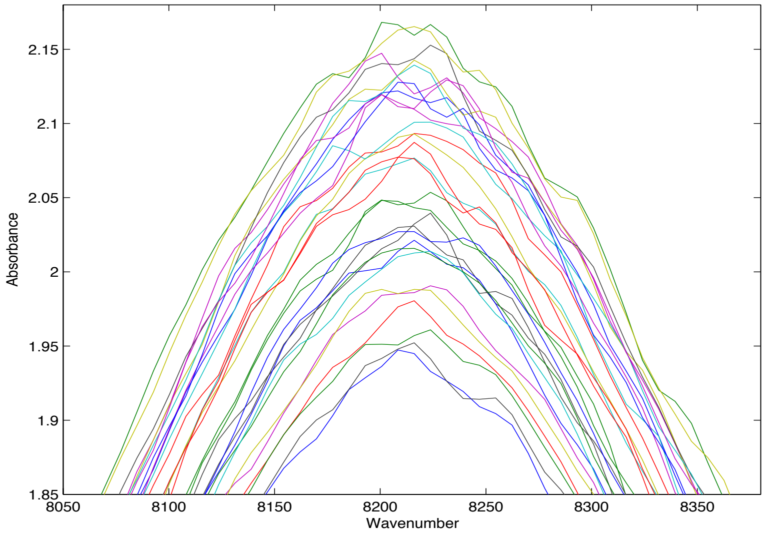


Fig. 7.6: Zoomed version of Figure 7.5, between 8050 and 8400 cm^{-1}.

When ICA was applied to this data, 30 components were identified (shown in Figure 7.7). Considering that there are 30 spectra in the data set, it can be deduced that ICA is unable to reduce the observations to their unique components. This result illustrates the sensitivity of ICA to frequency shift.

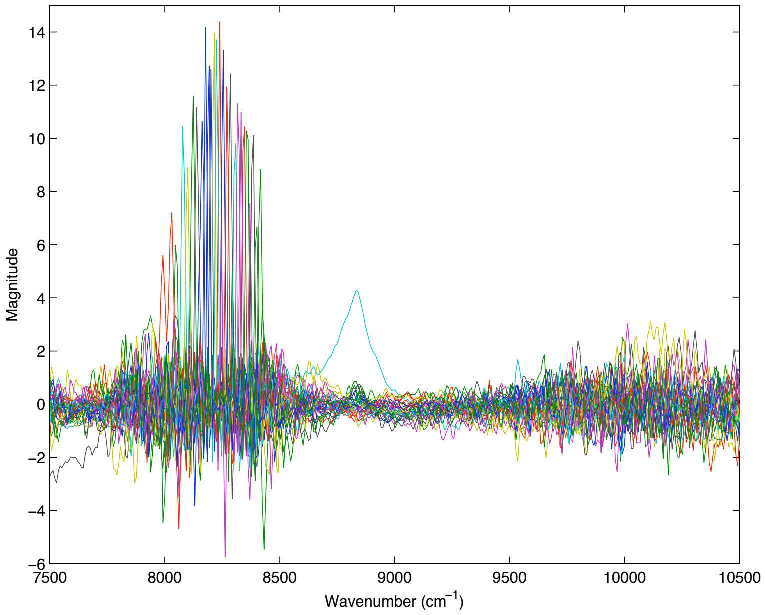
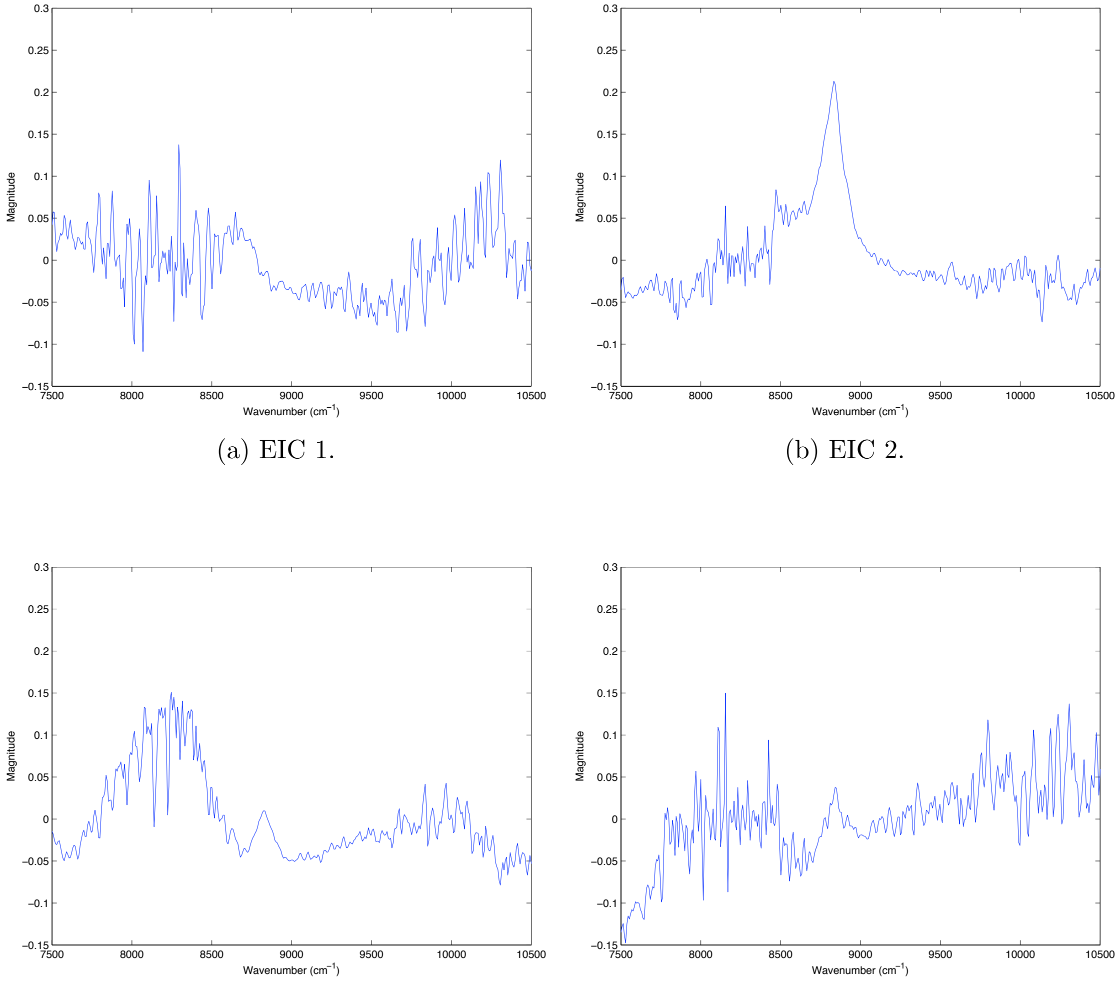


Fig. 7.7: ICs that were found by ICA in spectral data.

The proposed shift compensation algorithm was applied to the data set, and the resulting components are shown in Figure 7.8.

Details provided in [28] suggest that the active ingredient has an important peak near 8830 cm^{-1}, and visual comparison of Figure 7.8b to the reference spectrum provided in [28] shown in Figure 7.9 indicates that the component in Figure 7.8b is similar to the spectral signature of the active ingredient. [28] also states that the dominant excipient, mycrocrystalline cellulose, has a prominent peak near 8200 cm^{-1} [28], suggesting that the component in Figure 7.8c is a good candidate for this material. Information on the other materials in the tablet were not available.

It is important to note that the spectral signature of the active ingredient did not suffer from shift in this data set. However, the severe shift in the other components made it difficult to correctly identify it using ICA alone. This implies that even seemingly irrelevant components (such as excipients in pharmaceutical tablets) may cause difficulties when identifying the active ingredients. The post-processing method was able to circumvent these difficulties, by reducing the amount of components to be inspected, and grouping them appropriately.



(a) EIC 1.

(b) EIC 2.

(c) EIC 3.

(d) EIC 4.

Fig. 7.8: EICs obtained from the results in Figure 7.7.

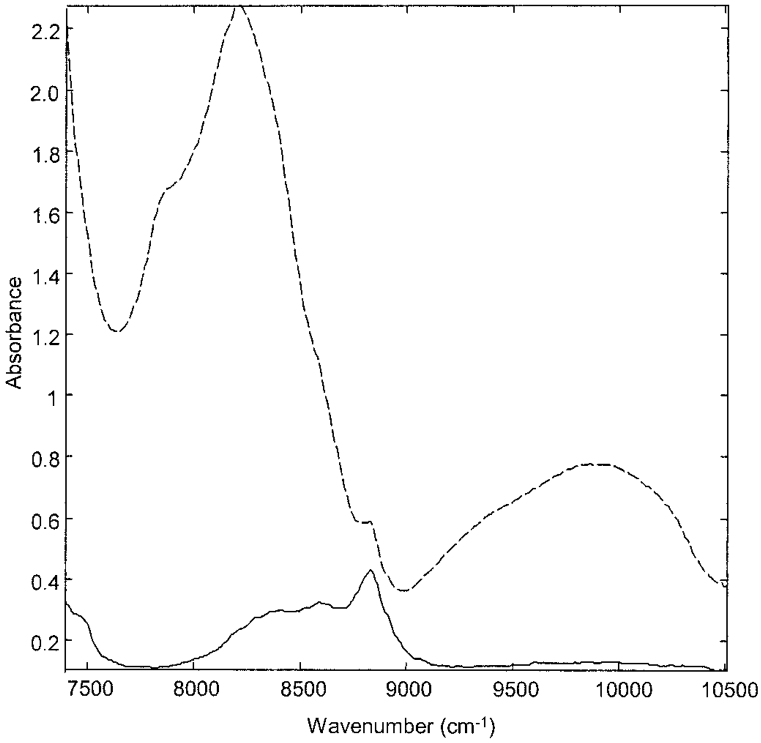


Fig. 7.9: NIR Raw spectra of a 20 mg tablet (dashed line; transmittance spectrum) and the pure active substance (solid line; reflectance spectra). (Scanned image from [28]).

7.1.2.3 Case Study 3: Carbonised Ice Analogs

The data set that was used in this study was a publicly available data set consisting of 9 NIR samples that were measured from samples of carbonised ice (H_2O+CO). These measurements were initially measured in an experiment to observe the effect that heat had on the NIR spectrum [35]. The two most relevant variations that were recorded are highlighted in Figure 7.10. It shows the 9 NIR spectra and highlights, with arrows, the effect that the increase in heat has on the spectra. All the spectra shown in this case study are in cm^{-1}.

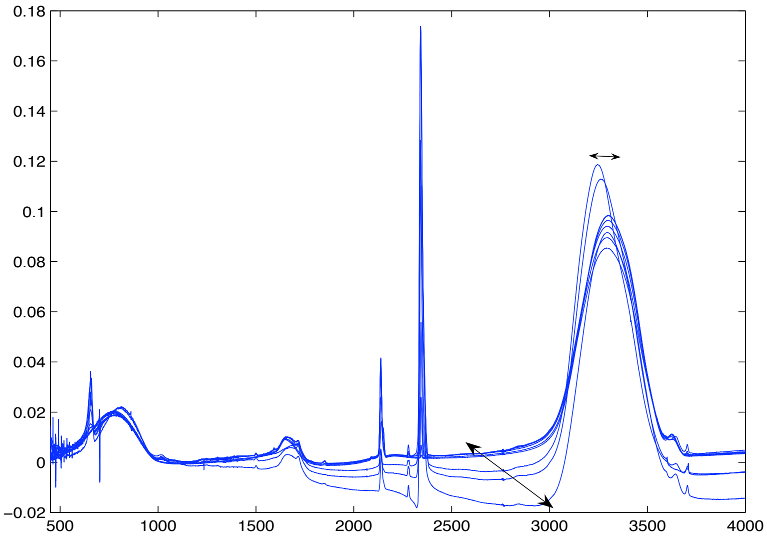


Fig. 7.10: Spectral data of ice UV-radiated in different amounts.

FastICA was initially applied to the 9 spectra and the ICs that were identified are shown in Figure 7.11.

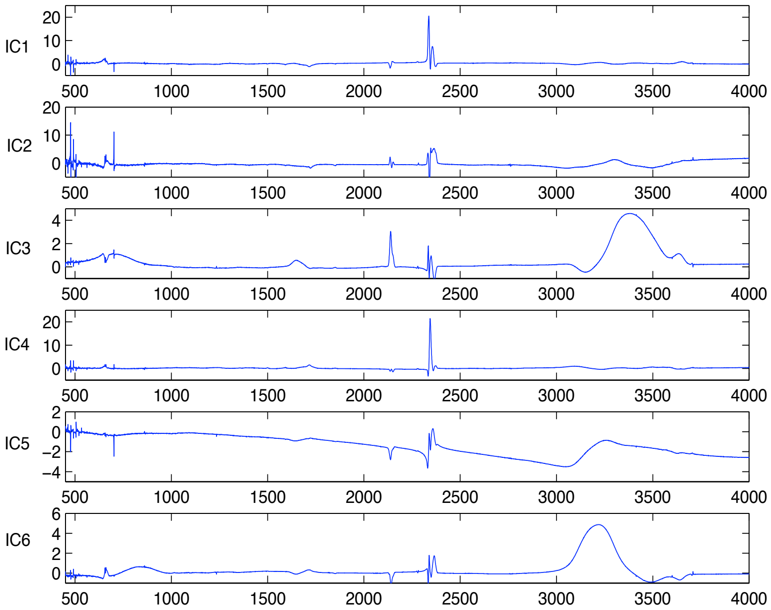


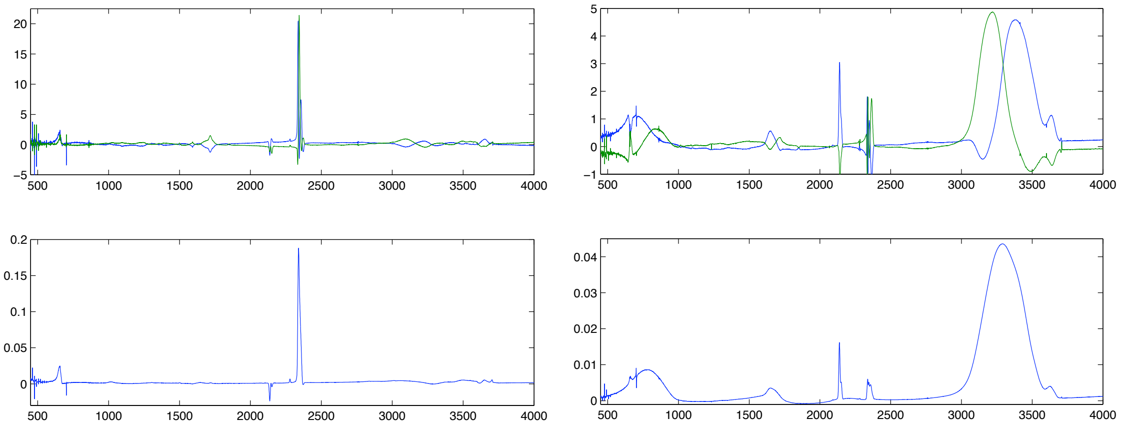
Fig. 7.11: ICs that were found by ICA in ice data.

Four groups were identified using the automated procedure described in Section 7.1.1 and are shown in Table 7.2.

Group	1	2	3	4
Components	IC1,IC4	IC2	IC3,IC6	IC5

Table 7.2: Groups identified in ice data.

Groups 2 and 4 are each comprised of only two components, thus, they can simply be added together to obtain their corresponding EIC. Both groups and their EICs are shown in Figure 7.12.

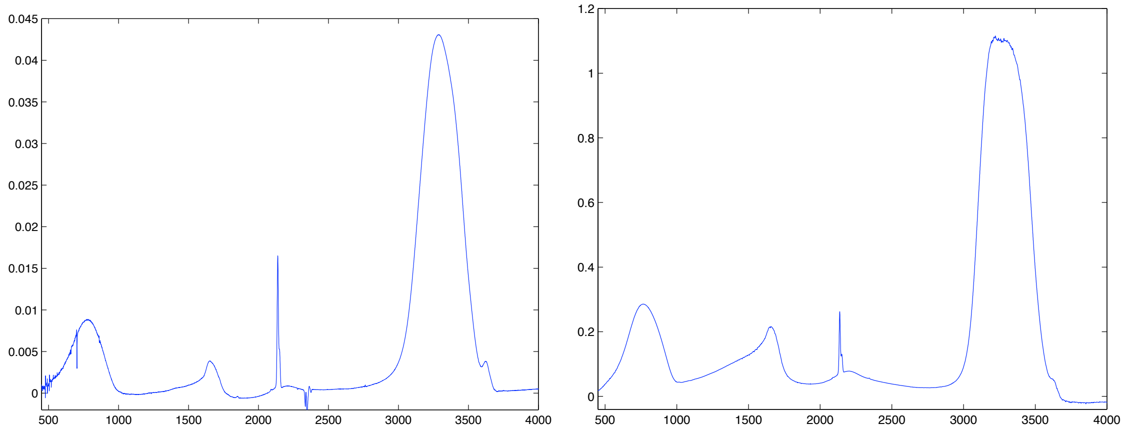


(a) IC1 and IC4 and their combination (EIC1&4). (b) IC3 and IC6 and their combination (EIC3&6).

Fig. 7.12: ICs that were identified as belonging together in groups.

Figure 7.12 shows that there is still some correlation between the two identified EICs. For example, the peak at approximately 2340 cm^{-1} is present in EIC1&4 and EIC3&6. The de-correlation algorithm was applied to these EICs and the resulting EIC3&6 is shown in Figure 7.13a. Figure 7.13b provides the reference spectrum for carbonised ice obtained from The Cosmic Ice Laboratory at NASA [34]. Comparison of Figures 7.13a and 7.13b illustrate that the proposed post-processing algorithm has identified a component with a spectral signature similar to the expected material.

It is important to remember that the process aims to extract the shape of the spectrum, thus, the scale of the resulting spectral signature is not recovered. However, in Figure 7.14 both spectra are superimposed and, although it can be seen that they are not a perfect fit, similarities between the two are clear.



(a) EIC3&6.

(b) Reference spectrum from NASA.

Fig. 7.13: Recovered component and its reference spectrum.

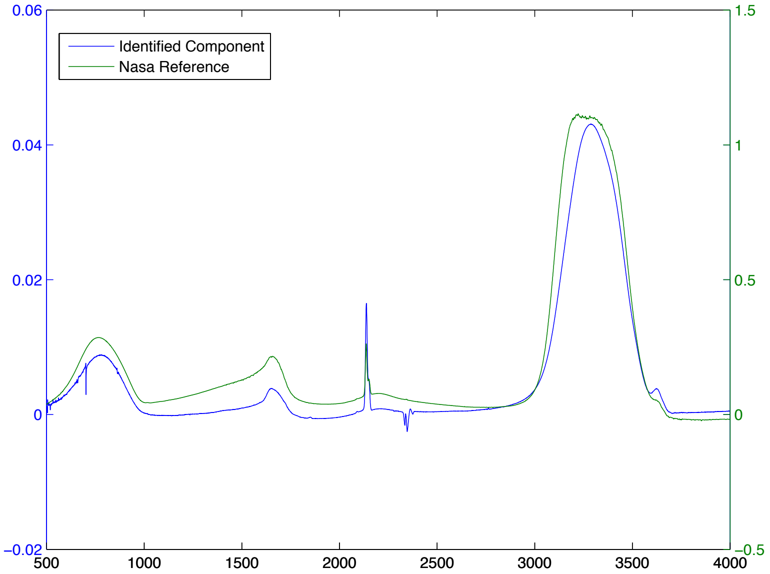


Fig. 7.14: EIC3&6 with NASA reference spectra superimposed.

A numerical representation of their similarity is desired, and, because of the scale difference, a Pearson correlation measure was chosen, as it calculates the similarity between two variables regardless of their scale. The Pearson correlation between the two spectra is of **0.9809**, which means that there is a strong correlation between the identified component and the reference spectrum.

7.1.3 Conclusions on Post-Processing Estimates of Independent Component Analysis

It has been shown that post-processing the results from ICA provides good estimates of the locally distorted component inside the data in both simulated and real data. It was also shown how the grouping mechanism built into the method was able to eliminate the need for the user to inspect the irrelevant components retrieved by ICA.

However, it is important to note that this method is only able to cope with spectra that have suffered solely from Local Shift. If any other distortion is to be addressed, the method would need to be reconsidered. A better solution to this problem would be able to cater for any type of local distortion. To do this, a new component extraction algorithm was developed, and this is described in the following section.

7.2 Blind Source Separation by Sample Subtraction

In this section, a BSS algorithm is proposed that aims to extract the underlying components in a spectral data set by subtracting scaled versions of a sample out of the rest of the set such that the samples are left with only one unique component. Two versions of this algorithm are described in the following sections. The first assumes perfect alignment of the underlying components and shows improvements in speed and performance over other BSS algorithms. The second makes the fundamental assumption that the data is distorted and accounts for it when subtracting samples from each other.

7.2.1 BSS by Subtraction of Non-distorted Data

The objective of the first version of the algorithm is to remove a component out of all the samples in a data set, with the exception of one sample. If the number of the samples being processed is equal to the number of components to be retrieved, then the result of the algorithm would be that the features appearing in one sample would be those of only one component. Therefore, the number of required samples is equal to the number of components being retrieved[2]. Given the additive nature of the components inside the spectra, which is assumed by all component extraction algorithms discussed in this thesis, the process of removing one component is based on the subtraction of a sample from one another.

To further illustrate the subtraction process an example is given. Two simple components, shown in Figure 7.15, were used to create the two simulated sampled spectra shown in Figure 7.16.

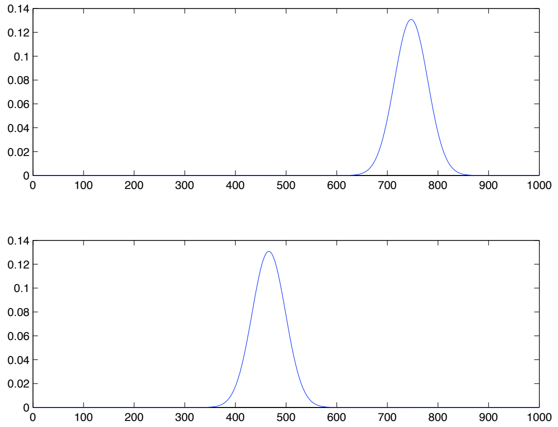


Fig. 7.15: Reference spectra used in example.

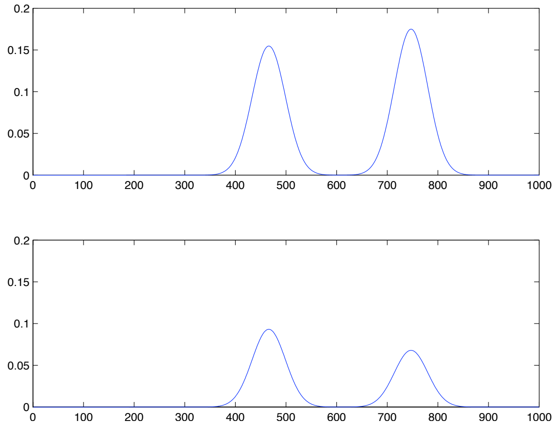


Fig. 7.16: Two simulated spectral samples created using reference from Figure 7.15.

Sample 1, shown in the upper graph of Figure 7.16, is scaled such that its maximum value is equal to the value of that location in Sample 2. In the upper

[2] Methods to work around the issue of estimating the number of components in distorted data with the proposed algorithm are discussed later.

graph of Figure 7.17, the green line is the scaled version of Sample 1 plotted over Sample 2. Subtracting the scaled Sample 1 from Sample 2 results in the spectrum shown in the lower graph of Figure 7.17, and, because the samples do not suffer from any distortion, the second peak has been eliminated from Sample 2, whilst the remaining features remained of the spectra are left intact.

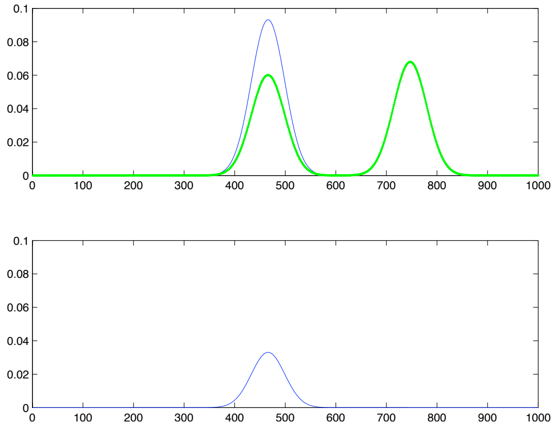


Fig. 7.17: Result of subtracting scaled Sample 1 from Sample 2 without shift.

If this process is then carried out in reverse, scaling Sample 2 with Sample 1 and subtracting, both samples will result in bearing the features of only one component, as shown in Figure 7.18.

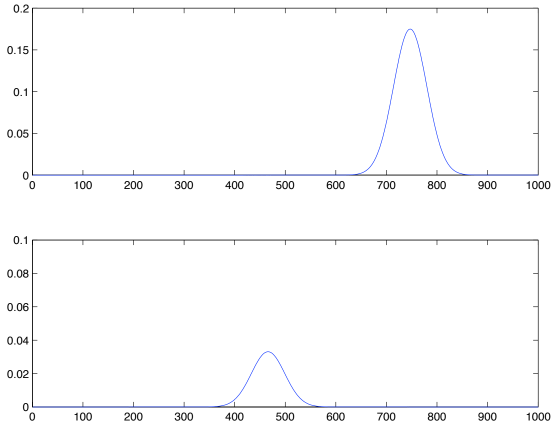


Fig. 7.18: Samples after subtracting scaled Sample 2 from Sample 1 without shift.

Comparison of the shapes of the resulting subtracted components to the reference spectra in Figure 7.15 suggests that they are very similar to the sources. It is necessary to maintain the size of the components during the iterations, but it is relatively easy to do so, as the original spectral magnitudes can be carried forward. Algorithm 1 summarises the method when no distortion is present in the spectra.

Algorithm 1 Simple Blind Source Separation Algorithm.

```
repeat
  for all $\hat{S} \Rightarrow \hat{S}_a$ do
    $\hat{S}_{a_m} = \mathbf{max}(\hat{S}_a, \textit{return index})$
    for all $\hat{S} \neq \hat{S}_a \Rightarrow \hat{S}_b$ do
      $\hat{S}_b \leftarrow \hat{S}_b - \hat{S}_a(\hat{S}_b[\hat{S}_{a_m}]/\hat{S}_a[\hat{S}_{a_m}])$
    end for
  end for
until convergence
```

where $\hat{S}$ is a subgroup of the data set S which has k samples, equal to the amount of components to be retrieved. Convergence is reached when the Mean Square Error is below a pre-specified tolerance, which should be specified to avoid overshooting. For a set of four spectra, each of length 1500 points, a tolerance of 0.001 gave the best results in this study.

7.2.1.1 Comparison with ALS and ICA & Discussion

An un-shifted data set of 100 samples was created using the spectra shown in Figure 4.3 with concentrations ranging from 0.2 to 1.8. The Simple BSS method, as well Independent Component Analysis and Alternating Least Squares, were applied and the time taken for the algorithms to converge was recorded. The performance measure described in Section 4.3.2 was used as the basis for comparison. The results are shown Table 7.3.

Algorithm	Time (sec.)	Performance
Simple BSS	0.0577	1.0000
ICA	0.2239	0.9975
ALS	35.5488	0.9827

Table 7.3: Performance and time comparison between algorithms.

No spectral distortion and no noise was applied to the data set, so a 100% performance rate was achievable. Because Simple BSS also works as a decorrelation algorithm, if the components have no correlation then identification of perfect components is possible. Table 7.3 indicates that all three algorithms were able to extract good estimates of the sources. However, the Simple BSS algorithm did so in a shorter period of time.

The initial aim was not to create a new BSS algorithm, although a variation of this algorithm that is robust against spectral distortion is introduced in Section 7.2.2. This algorithm was first employed in the Final De-Correlation phase of the algorithm proposed in Section 7.1.1. However, the results in Table 7.3 show that other algorithms, such as ICA, may not be utilising all the information they can use from the assumptions made. Thus, such algorithms may be approaching the task in ways that are unnecessarily convoluted, considering that approaches as simple as the one proposed obtained better results in a shorter amount of time.

7.2.2 BSS by Subtraction of Distorted Data

Applying distorted data to the algorithm described in Section 7.2.1 will break the assumption it relies on to remove one component from one sample by subtracting the scaled version of another. However, a variation of this algorithm has been developed that is robust against spectral distortions such as Local Shift.

When a Local Shift is introduced to the example data in Section 7.2.1, two issues arise. First, the peak in which the maximum value of Sample 1 lies needs to be re-scaled and aligned to the one in Sample 2 such that it is completely eliminated from Sample 2, as shown in Figure 7.19. The lower graph shows the result of appropriately aligning and scaling Sample 1 such that the second peak in Sample 2 is eliminated.

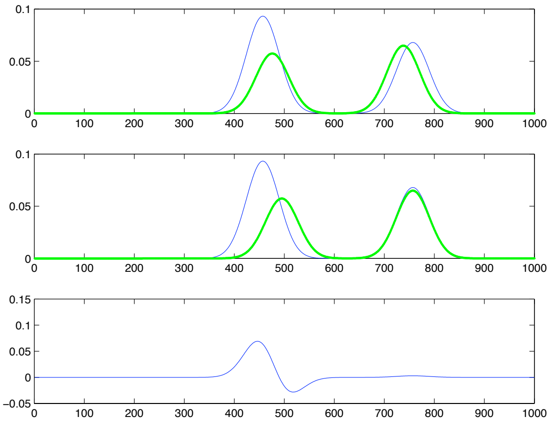


Fig. 7.19: Result of subtracting scaled Sample 1 from Sample 2 with shift.

The second issue can be observed in the lower graph of Figure 7.19, where the shape of the remaining peak is considerably distorted. This can be overcome by adding back to Sample 2 only 'certain parts' of the re-scaled, aligned Sample 1. In the lower graph of Figure 7.20, the remaining peak in Sample 2 is restored by adding back the line in red shown in the upper graph, which is an appropriately chosen section from Sample 1.

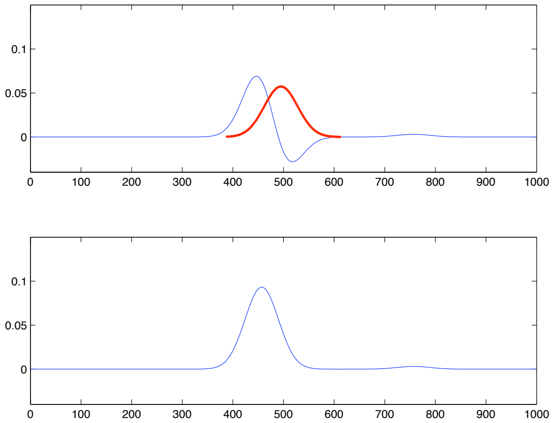


Fig. 7.20: Restoring remaining features of Sample 2 by partially adding back Sample 1.

The alignment phase and the restoring phase were incorporated into the definition of the original algorithm. The pseudo-code given in Algorithm 2 provides a generalised definition of the developed algorithm, in which it is assumed that a subset of randomly chosen samples $\hat{S}$ has been defined.

Algorithm 2 BSS by Substraction.

```
repeat
  for all $\hat{S} \Rightarrow \hat{S}_a$ do
    for all $\hat{S} \neq \hat{S}_a \Rightarrow \hat{S}_b$ do
      $\hat{S}_{a_{distorted}} = \textbf{find\_best\_alignment}(\hat{S}_a, \hat{S}_b)$
      $s_{a_m} = \textbf{max}(\hat{S}_{a_{distorted}}, \textit{return index})$
      $N = \hat{S}_b[s_{a_m}]/\hat{S}_{a_{distorted}}[s_{a_m}]$
      $\hat{S}_b \leftarrow \hat{S}_b - (\hat{S}_{a_{distorted}} N)$
      $I = \textbf{find\_indexes\_to\_repair}(\hat{S}_b)$
      $\hat{S}_b[I] \leftarrow \hat{S}_b[I] + (\hat{S}_{a_{distorted}}[I] N)$
    end for
  end for
until convergence
```

The **find_best_alignment** function finds the best way to temporarily distort $\hat{S}_a$ such that its maximum value best aligns with the features that are near it in $\hat{S}_b$. N is the value which normalises the distorted $\hat{S}_a$ such that its maximum value is equal to the corresponding location in $\hat{S}_b$. The **find_indexes_to_repair** function locates the indexes I of $\hat{S}_b$ that need repairing, and, hence, defines the sections of $\hat{S}_{a_{distorted}}$ that need to be added back to $\hat{S}_b$. An explanation of both of these functions, as well as several observations made from the implementation, are now discussed.

Finding Best Alignment. The process of finding the best distorted alignment involves estimating the amount of global distortion needed to be artificially applied to $\hat{S}_a$ such that the area near its maximum value when being subtracted from $\hat{S}_b$ is close to zero. Meaning that $\hat{S}_a$ is artificially globally distorted and subtracted from $\hat{S}_b$ to create a temporary spectrum D, from which the values near the maximum value of $\hat{S}_a$ are gathered. This is summarised in (7.7).

$$D = \hat{S}_b - \textbf{distort\_signal}(\hat{S}_a, \mathit{distort\_measures}) \tag{7.7}$$

An optimisation algorithm, as well as a brute-force approach, can be applied to find the optimal set of *distort_measures*. However, it was found that when dealing with two or more spectral distortions at the same time, a brute-force approach was too time consuming. The use of an optimisation algorithm, such as Particle Swarm Optimisation, provided faster results.

Depending on the type of distortions, the objective function used by the optimisation algorithms may differ. For example, when observing local shifts, the objective function in (7.8) sufficed.

$$M = -\textbf{abs}(D[s_{a_m}]) \tag{7.8}$$

where

$$s_{a_m} = \textbf{max}(\hat{S}_{a_{shifted}}, \mathit{return\ index}) \tag{7.9}$$

and

$$\hat{S}_{a_{shifted}} = \textbf{distort\_signal}(\hat{S}_a, \mathit{shift}). \tag{7.10}$$

The **max** function, with the *return index* flag, returns the index of the maximum value of a spectrum, and M is the measure of optimality.

When dealing with a local shift coupled with a global warp, the objective function in (7.11) provided good results.

$$M = -\sum_{i=s_{a_m}-range}^{s_{a_m}+range} iD[i]^2 \tag{7.11}$$

where *range* is a pre-defined number of points to the left and right of s_{a_m}, the location of the maximum value of $\hat{S}_{a_{shifted\_and\_warped}}$ defined by (7.12).

$$\hat{S}_{a_{shifted\_and\_warped}} = \mathbf{distort\_signal}(\hat{S}_a, \mathit{shift}, \mathit{warp}) \tag{7.12}$$

Each value in D is weighted such that the values in the right side of s_{a_m} are given more weight. The reason for this is that, because of the nature of the warp distortion, the differences between the features are more predominant and more easily identifiable to the right side of the maximum value.

It is important to note that the objective function needs to be defined by using appropriate knowledge of the distortion taking place. However, being able to consider several distortions at the same time is an important quality of the proposed algorithm, as knowledge from a plant expert can be incorporated in a relatively easy manner.

Finding Locations to Repair. The process of finding which frequency locations require modification after the subtraction has taken place is relatively simple. It was found that if the value of a frequency location is outside a pre-defined range, then it is necessary for it to be repaired. This range does need to be tuned for the specific spectral signals being used, but a value of 1% of the full frequency spread was found to be suitable for both limits (one negative and the other positive). If the peaks have sharp edges, having a negative limit close to zero and increasing the positive limit were found to smoothen the peaks. However, overshooting may result in slow convergence as well as thin features being created.

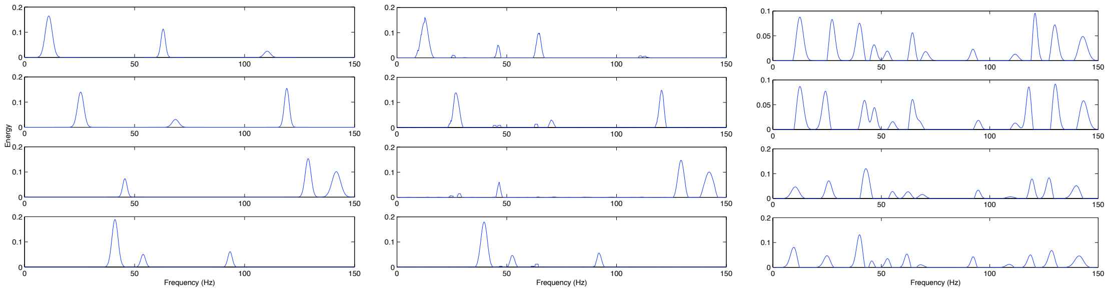
Post-Filtering. If there is a small amount of overlap or noise in the spectra, all the locations to be repaired may be difficult to identify. Therefore, the samples will have an increasing amount of small features, as more locations are not identified. However, because of the small nature of these features, they can be filtered out relatively easy. A moving average window filter with a window of length of 1% of the range of the spectrum was used in this work to remove the small features. This filter was found to remove a large quantity of these features, without significantly affecting the structure of the identified component features.

Number of Samples to Use. The number of samples in the subset $\hat{S}$ must be the same as the number of components to be extracted, k. As in any other Curve Resolution Method, k can be estimated using several techniques. However, these techniques have already been shown to be fragile to spectral distortion. If using locally shifted data, however, the methodology applied in the Post-Processing technique described in the Section 7.1.1 can provide a good estimate of the number of components in the data. It has also been observed that the number of components can be estimated if the BSS by Subtraction algorithm described here is applied repeatedly, with an increasing amount of samples every time, until the extracted components begin repeating themselves.

7.2.2.1 Experiments & Results

Two data sets of 100 samples each were created using the methods described in Section 4.2. One data set suffered from local shift with a range of ±20 *fg* (±2 Hz), and the other suffered from both the same local shift distortion as well as a global warp that ranged between 95% to 1.05%. The concentrations of the components ranged from 0.5 to 1.5 in both data sets.

Alternating Least Squares and the proposed technique, BSS by Subtraction (SubBSS), were applied to both data sets, both assuming that the correct numbers of components was known *a-priori*. Four randomly chosen samples were used by BSS by Subtraction, whilst the whole data set was used for ALS. The results for the first data set are shown in Figure 7.21 and for the second data set in Figure 7.22.

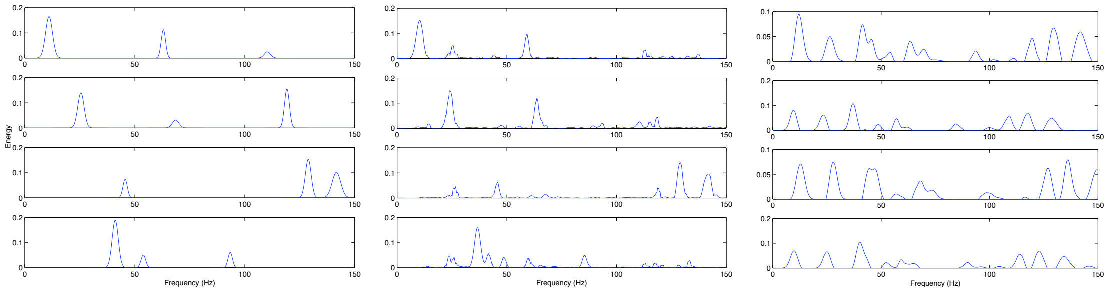


(a) Benchmark used.

(b) Components extracted by SubBSS

(c) Components extracted by ALS

Fig. 7.21: Components extracted from data set suffering from local shift.



(a) Benchmark used.

(b) Components extracted by SubBSS

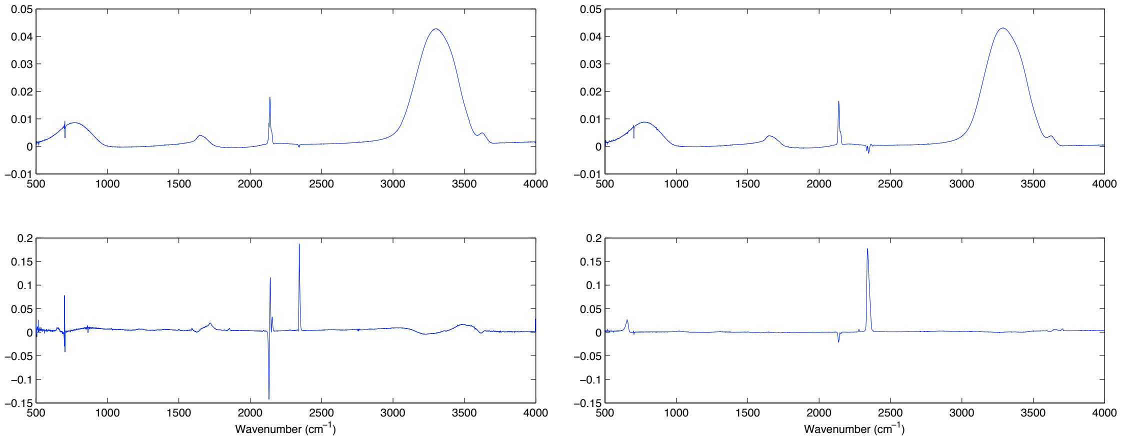
(c) Components extracted by ALS

Fig. 7.22: Components extracted from data set suffering from local shift and global warp.

Figures 7.21 and 7.22 show that Alternating Least Squares was not able to extract the correct components. In fact, in both cases, ALS reached 2000 iterations without convergence. Both Figures show that the BSS by Subtraction technique was able to estimate four components similar to the benchmark. As can be observed, when two distortions occur at the same time, the accuracy of the technique is reduced. However, the extracted components are still similar to the actual source components and significantly better than those extracted by ALS.

7.2.2.2 Case Study: Ice Analogs

The spectral data used in the case study described in Section 7.1.2.3 was used with the developed method of BSS by Substraction. Using rank analysis, such as Singular Value Decomposition, to estimate the number of components was not possible, as the data suffers from severe shift. To work around this, the method was applied repeatedly, using an increasing number of samples, until the extracted components began repeating themselves. Using this methodology, it was concluded that there were two components in the spectra, which are shown in Figure 7.23a. For comparison, the two predominant EICs extracted using the PSO-Based Post-Processing technique described earlier are show in Figure 7.23b.



(a) Using BSS by Subtraction.

(b) Using PSO-Based Post-Processing Technique

Fig. 7.23: Extracted components from ice data, described in Section 7.1.2.3.

As can be seen from Figure 7.23, the first component is similar when using both methods. The second component is also similar when using both methods; the major differences are the negative peak near 2100 cm^{-1} and a number of small features between 3000 and 3700 cm^{-1}. Hence, both methods provide similar results with this data set. However, only Local Shift was present; if the data set would have presented other types of distortions, the PSO-Based Technique would not be able to extract the components.

7.2.3 Conclusions on Blind Source Separation by Subtraction

The new method described here has great potential for resolving the problem of component extraction from locally distorted spectra. It was shown that it was able to cope with two different types of distortions taking place at the same time, and the fact that more distortions can be considered bodes well for its generalised application.

However, it is important to note that this method is a variation of the decorrelation process of the post-processing method described in Section 7.1.1. Hence, the components to be extracted are assumed to be non-correlated. If there is correlation, then shared features will appear in only one of the components.

7.3 Review of Usage of Developed Algorithms

A brief review of the limitations of the methods is provided here.

The post-processing technique relies on the results of Independent Component Analysis, specifically the FastICA implementation, to recover the components. It was developed this way because of the observed behaviour of ICA when locally shifted data is used. If another type of distortion is taking place in the data, such as warp, the effects it has on ICA will need to be explored and, if the results are found to be repairable, the post-processing technique would need to be modified accordingly. Furthermore, for every type of distortion possible, ICA would need to be explored and corrected accordingly, resulting in a high investment of time.

The second method was implemented to avoid this problem. The BSS by Subtraction technique can be modified in a relatively simple manner to include different types of distortion if need be, as there is no other algorithm that it relies on. It is realistic to assume that the more types of distortions the data suffers from, the more difficult it would be for the algorithm to estimate an appropriate set of components. In addition, an important aspect of the post-processing technique that the second method does not share is that the number of components to be extracted is assumed to be known *a-priori*. Techniques that can be used to work around this issue have been discussed, but the task of estimating the number of components inside distorted spectral data remains an open problem.

Chapter 8

Conclusions

Spectral data can be used to extract important information from an inspected material or product, and is relevant in many academic areas as well as Industry. For example, medical spectral signals can be used to find the presence of disease in a patient, the composition of Pharmaceutical tablets from spectral measurements can be analysed to estimate component concentrations and temperature, and speech and face recognition can be enhanced with the use of spectral information obtained from audio and video.

This thesis has highlighted that sampled spectral data can be very sensitive towards changes in the environment. Temperature deviations, the presence of foreign materials, light emanations, etc. have been shown to distort spectral data, which degrades the ability of spectral analysis tools to extract meaningful information from the data.

In this thesis, a Spectral Analysis Framework was proposed which was able to extract information from spectral data suffering from a pre-defined set of distortions. If a set of reference spectra is available, the framework simulates the effect of these distortions in the reference data set, such that their combination best fits an analysed spectral sample. A Black-Box optimisation approach was adopted, with which the framework was able to consider any combination of distortions occurring simultaneously. As a result, the effects of each distortion can be estimated by looking at spectral data alone, as well as provide the user with the desired information (such as component concentrations estimates). This is important to a plant monitor, as the information the framework provides can be used to identify if it is necessary to calibrate the spectral sensor.

The proposed framework was tested in a simulated mixing plant whose spectral sensors were sampling shifted data. The plant initially used Classical Least Squares Regression (CLSR) to estimate the material concentrations of a product from spectral data, and a simple controller was able to provide a good system response. However, when the spectral data suffered from shift, CLSR did not provide good estimates, and the system response degraded considerably. The Spectral Analysis Framework proposed in this thesis was used to replace CLSR for estimating concentrations. By simulating a Shift distortion, the framework provided good estimates of both the component concentrations and the shift suffered. As a result, the system response was improved and was very close to the response when no shift was present. Furthermore, the amount of shift suffered by every sample was provided by the framework throughout the experiment, thus, the plant monitor could use this information to calibrate the spectral sensor in an upcoming scheduled plant maintenance. In the mean time, the framework could continue its work without requiring to stop the plant, minimising the plant downtime.

The proposed framework is able to estimate the concentrations of underlying components in a composite spectrum, given a set of reference spectra of the components, which meets the first objective of this research project. However, a set of reference spectra may not always be available. To circumvent this issue, a set of algorithms were developed to extract the underlying components out of a distorted spectral data set, following the same optimisation approach of the framework. It was found that component extraction methods are affected differently depending of the type of spectral distortion encountered. In this regard, two groups of types of distortion were defined. One group, referred to as Global Distortions, were categorised as those that modify the spectrum uniformly. The other group, the Local Distortions, are those that affect each component by a different amount.

For data that suffered from a Global type of distortion, a method was developed that aimed to auto-align a spectral data set given a set of pre-defined Global distortions. This alignment method was shown to considerably improve the performance of Alternating Least Squares with simulated data that suffered simultaneously from Global Shift and Global Warp. Given the nature of the distortions assumed to be taking place, other component extraction methods will benefit from this method.

For Local types of distortion, two methods were developed. The first was able to correct the results provided by Independent Component Analysis when the data set suffered from Local Shift. This method was tested using simulated data sets that suffered from shift with good results. It was tested with two laboratory data sets as well. The first was a public data set sampled from pharmaceutical tablets in which one component suffered from severe shift and noise. ICA was able to identify one component (which was the one not suffering from any shift), the rest of the identified components suffered heavily from *component division.* The developed method was able to correctly group the 'partial components' provided by ICA and combined them into a comprehensive spectral component data set that was congruent with the information available concerning the rest of the underlying components. The second was a public data set sampled from carbonised ice analogs at different temperatures, that suffered from Local Shift. This method was able to extract the underlying components that corresponded to the expected spectral signatures.

The first of the Local Distortion methods only considered the Local Shift distortion. A second, more flexible algorithm was developed that could consider more types of Local Distortions. A set of pre-defined Local Distortions are required, as well as knowing *a-priori* the number of underlying components to extract. It was tested with simulated data suffering from simultaneous Local Shift and Local Warp, and outperformed Alternating Least Squares. In addition, it was successful in extracting the same components from the carbonised ice analogs data set as the first method.

These three methods, in conjunction, are able to extract the underlying components out of spectral data set suffering from any type of distortion. The extracted components were shown as being able to be used as a reference spectral data set in the Spectral Analysis Framework, meeting the second and final objective of this research project.

8.1 Future Work

Two areas in which the developed algorithms could be improved were encountered during this work. In the following sections, the issues of computation time reduction and spectral distortion identification are discussed.

8.1.1 Computation Time Reduction

It has been shown that analysing distorted spectral data can be carried out as an optimisation problem. This approach introduces great flexibility in the variety of spectral shapes and types of distortions that can be considered. However, approaching it as an optimisation task comes with an important trade-off, which is a *long computation time*. Depending on the size of the spectra being analysed, the number of components to extract, and the number and types of distortions to consider, the proposed algorithm can take up to several minutes to complete. This reduces significantly the types of processes in which the framework could be used as a real-time monitoring tool.

One approach to reduce computation time is to use a faster computer to carry out the optimisation task. The algorithms, with the same data set, were deployed in two different machines: first, in a machine with a 1.7 GHz processor, and then in another with a 2.5 GHz processor. The decrease in computation time was expected to be near the ratio of the speeds of both machines, however, it was much higher. Hence, the main improvements to the speed of the optimisation process will be carried out by making the optimisation algorithm more efficient, rather than by upgrading hardware. Although, the latter is always welcome.

An important reason for the long computation time is the nature of the defined problem, where it is assumed that no information is known about it, i.e. it is a Black-Box problem. This means that the algorithm has to 'discover' the shape of the solution space for every analysed sample. However, speed improvements have been made by considering information specific to the process. In the simulated mixing plant described in Chapter 5, it was observed that an important amount of time was invested in the exploratory phase of the optimisation. It has been noted that Spectral Shift is usually the result of changes to ambient factors, such as temperature, pressure, and humidity, which are slow-varying. Hence, it can be assumed that the amount of distortion occurring in sample k is similar to that occurring in sample $k + 1$. Given this heuristic, the optimisation search can be initialised to start near the values found in the previous sample, and to therefore jump directly to the exploitation phase of the search. Speed improvements were shown, although not to the point of being applicable in plants with sampling times of a few seconds. However, this improvement does show potential, thus, it is of interest to further investigate the inclusion of process heuristics to enhance the optimisation task.

8.1.2 Spectral Distortion Identification

An important aspect of analysing distorted data is defining the possible distortions occurring in the data. Before being able to artificially distort a spectrum to best fit another, the manners in which to modify the spectrum are required to be defined by an expert of the field. To this effect, any artificial modification carried out in a spectrum should have a physical meaning relevant to the process.

Initially, it would be of interest to create an extensive database of known distortions, given a specific spectral sensor, material, and external factors. However, it is reasonable to consider the possibility of a plant supervisor not being able to preemptively detect any spectral distortion taking place. Such circumstances will benefit from a tool that would be able to *discover* what type of distortions are occurring in a spectral data set, relying on a pre-defined database of possible distortions. The creation of such a database would be a difficult challenge, but, if met, would greatly benefit this project, as well as the fields that are involved in the analysis of spectral data.

Bibliography

[1] A. A. Ali. Fm spectrum of video signals. *Proceedings of the IEEE*, 70(3):306–307, 1982.

[2] J. Backus. *The Acoustical Foundations of Music*, chapter 8, pages 152–153. W. W. Norton & Company, Inc., 1977.

[3] K. A. Baggerly, J. S. Morris, and K. R. Coombes. Reproducibility of seldi-tof protein patterns in serum: comparing datasets from different experiments. *Bioinformatics*, 20(5):777–785, Mar 2004.

[4] D. Barash. *Genetic algorithms applied to nonlinear and complex domains.* PhD thesis, University of California, June 1999. Chair-Ann E. Orel.

[5] M. P. Bernstein, D. P. Cruikshank, and S. A. Sandford. Near-infrared laboratory spectra of solid h2o/co2 and ch3oh/co2 ice mixtures. *Icarus*, 179(2):527–534, December 2005.

[6] M. W. Berry, M. Browne, A. N. Langville, V. P. Pauca, and R. J. Plemmons. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173, 2007.

[7] T. Bianchi and F. Argenti. Effects of a carrier frequency offset on wavelet transceivers. *Proceedings of the Sixth Baiona Workshop on Signal Processing in Communications*, pages 265–270, September 2003.

[8] J. C. Brown. Musical fundamental frequency tracking using a pattern recognition method. *The Journal of the Acoustical Society of America*, 92(3):1394–1402, September 1992.

[9] M. Castanys, M. J. Soneira, and R. Perez-Pueyo. Automatic identification of artistic pigments by raman spectroscopy using fuzzy logic and principal component analysis authors: M. castanys,m. j. soneira, r. perez-pueyo. *Laser Chemistry*, pages 11–19, October 2006.

[10] V. Černý. Thermodynamical approach to the traveling salesman problem: An efficient simulation algorithm. *Journal of Optimization Theory and Applications*, 45(1):41–51, January 1985.

[11] L. Cetto. Alignment of mass spectromotry data. United States Patent: 7365311, April 2008.

[12] A. K. Chan. *Wavelet toolware: software for wavelet training.* San Diego, Calif. ; London : Academic Press, 1998.

[13] C.-T. Chen. *Digital Signal Processing: Spectral Computation and Filter Design.* Oxford University Press, 2001.

[14] Z.-P. Chen and J. Morris. Improving the linearity of spectroscopic data subjected to fluctuations in external variables by the extended loading space standardization. *The Analyst*, 133:914–922, 2008.

[15] Z.-P. Chen, J. Morris, and E. Martin. Correction of temperature-induced spectral variations by loading space standardization. *Analytical Chemistry*, 77(5):1376 – 1384, March 2005.

[16] J. W. Childers, W. J. Phillips, E. L. Thompson, D. B. Harris, D. A. Kirchgessner, D. F. Natschke, and M. Clayton. Comparison of an innovative nonlinear algorithm to classical least-squares for analyzing open-path fourier transform infrared spectra collected at a concentrated swine production facility. *Applied Spectroscopy*, 56(3):325–336, March 2002.

[17] C. K. Chui. *An introduction to wavelets.* Boston ; London : Academic Press, 1992.

[18] M. Clerc and J. Kennedy. The particle swarm - explosion, stability, and convergence in a multidimensional complex space. *Evolutionary Computation, IEEE Transactions on*, 6(1):58–73, Feb 2002.

[19] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, April 1994.

[20] K. R. Coombes, J. Herbert A. Fritsche, C. Clarke, J. neng Chen, K. A. Baggerly, J. S. Morris, L. chun Xiao, M.-C. Hung, and H. M. Kuerer. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clinical Chemistry*, 49:1615–1623, October 2003.

[21] M. A. Czarnecki, Y. Ozaki, M. Suzuki, and M. Iwahashi. Quantitative study of the dissociation of dimeric cis-9, cis-12-octadecadienoic acid in pure liquid by the ft-ir liquid film technique. *Applied Spectroscopy*, 47(12):2157–2161, 1993.

[22] A. de Juan and R. Tauler. Multivariate curve resolution (mcr) from 2000: Progress in concepts and applications. *Critical Reviews in Analytical Chemistry*, 36(3-4):163–176, December 2006.

[23] M. de Veij, A. Deneckere, P. Vandenabeele, D. de Kaste, and L. Moens. Detection of counterfeit viagra with raman spectroscopy. *Journal of Pharmaceutical and Biomedical Analysis*, 46:303–309, 2008.

[24] D. J. DeFatta, J. G. Lucas, and W. S. Hodgkiss. *Digital Signal Processing: A System Design Approach*. John Wiley & Sons, 1988.

[25] S. Doraisamy and S. M. Rüger. An approach towards a polyphonic music retrieval system. *Proceedings of the 2nd Annual ISMIR*, pages 187–193, October 2001.

[26] M. Dorfler. Time-frequency analysis for music signals: A mathematical approach. *Journal of New Music Research*, 30(1):3–12, 2001.

[27] J. S. Downie. Music information retrieval. *Annual Review of Information Science and Technology*, 37:295–340, 2003.

[28] M. Dyrby, S. Engelsen, and L. Nørgaard. Chemometric quantitation of the active substance (containing c=n) in a pharmaceutical tablet using near-infrared (nir) transmittance and nir ft-raman spectra. *Applied Spectroscopy*, 56(5):579–585, May 2002.

[29] R. C. Eberhart and Y. Shi. Comparing inertia weights and constriction factors in particle swarm optimization. In *Proceedings of the 2000 Congress on Evolutionary Computation*, volume 1, pages 84–88 vol.1, 2000.

[30] S. L. Eix, S. A. Schlueter, and A. Anderson. Raman and infrared spectra of solid dibromodifluoromethane. *Journal of Raman Spectroscopy*, 23:495–499, 1992.

[31] N. Erk. Determination of active ingredients in the pharmaceutical formulations containing hydroclorothiazide and its binary mixtures with benazepril hydrochloride, triamterene and cilazapril by ratio spectra derivative spectrophotometry and vierordt method. *Journal of Pharmaceutical and Biomedical Analysis*, 20(1-2):155–167, June 1999.

[32] W. P. Findlay and D. E. Bugay. Utilization of fourier transform-raman spectroscopy for the study of pharmaceutical crystal forms. *Journal of Pharmaceutical and Biomedical Analysis*, 16(6):921–930, February 1998.

[33] T. Fujishima. Realtime chord recognition of musical sound: a system using common lisp music. *Proceedings of the International Computer Music Conference*, pages 464–467, 1999.

[34] P. A. Gerakines and Moore. Spectral database, the cosmic ice laboratory . nasa.

[35] P. A. Gerakines, W. A. Schutte, and P. Ehrenfreund. Ultraviolet processing of interstellar ice analogs. *Astronomy and Astrophysics*, 313:289–305, 1996.

[36] Y. Ghebremeskel, J. Fields, and A. Carton. The use of near infrared (nir) spectroscopy to study specific interactions in polymer blends. *Journal of Polymer Science*, 32:383–386, 1994.

[37] P. Giudici. *Applied Data Mining: Statistical Methods for Business and Industry.* Wiley, 2003.

[38] L. Guanter, V. Estelles, and J. Moreno. Spectral calibration and atmospheric correction of ultra-fine spectral and spatial resolution remote sensing data. application to casi-1500 data. *Remote Sensing of Environment*, 109:54–65, 2007.

[39] L. Guanter, R. Richter, and J. Moreno. Spectral calibration of hyperspectral imagery using atmospheric absorption features. *Applied Optics*, 45(10):2360–2370, 2006.

[40] D. M. Haaland and R. G. Easterling. Improved sensitivity of infrared spectroscopy by the application of least squares methods. *Applied Spectroscopy*, 34(10):539–548, September 1980.

[41] K. H. Hazen, M. A. Arnold, and G. W. Small. Temperature-insensitive near-infrared spectroscopic measurement of glucose in aqueous solutions. *Applied Spectroscopy*, 48(4):477–483, 1994.

[42] S. Hongtao, D. D. Feng, and Z. Rong-chun. Face recognition using multi-feature and radial basis function network. *ACM International Conference Proceeding Series*, 16:51–57, 2003.

[43] H. Hoshino, S. Tajima, and T. Tuschiya. The effect of the temperature on the mass spectra of aliphatic primary alcohols and 1-alekenes. i. *Bulletin of the Chemical Society of Japan*, 46:3043–3048, 1973.

[44] A. Hyvärinen. New approximations of differential entropy for independent component analysis and projection pursuit. *Neural Computation*, 9(7):1483–1492, 1997.

[45] A. Hyvärinen. Fast and robust fixed-point algorithms for independent component analysis. *IEEE Transactions on Neural Networks*, 10(3):626–634, 1999.

[46] A. Hyvärinen and P. Hoyer. Emergence of phase and shift invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, 12(7):1705–1720, 2000.

[47] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis.* New York ; Chichester : Wiley, 2001.

[48] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 13(4-5):411–430, May-June 2000.

[49] S. ichi Kamemaru, H. Itoh, and J. ichi Yano. Character recognition by feature extraction using cross-correlation signals from a matched filter. *Optical Engineering*, 32:26–32, 1993.

[50] K. Ichige, M. Imai, and H. Arai. Fastica-based blind signal separation and its applications to radio surveillance. *14th Workshop on Statistical Signal Processing*, pages 546–550, August 2007.

[51] A. Inoue, K. Kojima, Y. Taniguchi, and K. Suzuki. Effects of temperature and pressure on the near-infrared spectra of hod in d20. *Journal of Solution Chemistry*, 16(9):727–734, 1987.

[52] T. I. Ivanov, M. O. Vieitez, C. A. de Lange, and W. Ubachs. Frequency calibration of b 1&sigma;+u&ndash;x 1&sigma;+g (6,0) lyman transitions in h2 for comparison with quasar data. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 41(3):035702 (4pp), 2008.

[53] T. Iwata, J. Koshoubu, C. Jin, and Y. Okubo. Temperature dependence of the mid-infrared oh spectral band in liquid water. *Applied Spectroscopy*, 51(9):1269–1275, 1997.

[54] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, May 2005.

[55] J.-H. Jiang, Y. Liang, and Y. Ozaki. Principles and methodologies in self-modeling curve resolution. *Chemometrics and Intelligent Laboratory Systems*, 71(1):1–12, April 2004.

[56] K. A. D. Jong. *An analysis of the behavior of a class of genetic adaptive systems.* PhD thesis, University of Michigan, Ann Arbor, MI, USA, 1975.

[57] C. Jutten and J. Herault. Blind separation of sources, part i: An adaptive algorithm based on neuromimetic architecture. *Signal Processing*, 24(1):1–10, July 1991.

[58] V. Kachel, O. Kempski, J. Peters, and F. Schödel. A method for calibration of flow cytometric wavelength shift fluorescence measurements. *Cytometry*, 11:913–915, 1990.

[59] E. J. Karjalainen. The spectrum reconstruction problem : Use of alternating regression for unexpected spectral components in two-dimensional spectroscopies. *Chemometrics and Intelligent Laboratory Systems*, 7(1-2):31 – 38, 1989.

[60] J. Kennedy and R. Eberhart. Particle swarm optimization. *Proceeding of the IEEE International Conference on Neural Networks*, IV:1942–1948, 1995.

[61] A. P. Klapuri. Automatic music transcription as we know it today. *Journal of New Music Research*, 33(3):269 – 282, September 2004.

[62] B. S. Krongold, K. Ramchandran, and D. L. Jones. Frequency-shift-invariant orthonormal wavelet packet representations. *Proceedings of the 1997 International Conference on Image Processing (ICIP '97)*, 1:2579–2582, October 1997.

[63] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, October 1999.

[64] K. Lee and M. Slaney. Automatic chord recognition from audio using a supervised hmm trained with audio-from-symbolic data. *Proceedings of the 1st ACM workshop on Audio and music computing multimedia*, pages 11 – 20, October 2006.

[65] A. Lerch. On the requirement of automatic tuning frequency estimation. *Proceedings of the 7th International Conference on Music Information Retrieval*, pages 212–215, October 2006.

[66] Y. Li, C. L. Wen, Z. Xie, and X. H. Xu. Synchronization of batch trajectory based on multi-scale dynamic time warping. *Proceedings of the Second International Conference in Machine Learning and Cybernetics*, pages 2403–2408, November 2003.

[67] C.-Y. Lin, J. S. R. Jang, and M.-Y. Hsu. An automatic singing voice rectifier design. *Proceedings of the 11th ACM International Conference on Multimedia*, pages 267 – 270, 2003.

[68] J. Lin. Near-ir calibration transfer between different temperatures. *Applied Spectroscopy*, 52(12):1591–1596, 1998.

[69] J. Liua, S. A. Billings, Z. Q. Zhu, and J. Shen. Enhanced frequency analysis using wavelets. *International Journal of Control*, 75(15):1145–1158, October 2002.

[70] B. Logan. Mel frequency cepstral coefficients for music modeling. *Proceedings of the Int. Symposium on Music Information Retrieval*, pages 1–11, October 2000.

[71] N. C. Maddage, C. Xu, M. S. Kankanhalli, and X. Shao. Content-based music structure analysis with applications to music semantics understanding. *Proceedings of the 12th Annual ACM international Conference on Multimedia*, pages 112–119, October 2004.

[72] N. C. Maddage, C. Xu, C.-H. Lee, and M. Kankanhalli. Statistical analysis of musical instruments. *IEEE Pacific Rim Conference on Multimedia*, pages 581–588, 2002.

[73] H. Malika, A. Khokhar, and R. Ansari. Improved watermark detection for spread-spectrum based watermarking using independent component analysis. *Proceedings of the 5th ACM workshop on Digital rights management*, pages 102–111, 2005.

[74] P. Masri, A. Bateman, and N. Canagarajah. A review of time–frequency representations, with application to soundymusic analysis–resynthesis. *Organized Sound*, 2(3):192–205, 1997.

[75] M. Maurya, R. Rengaswamy, and V. Venkatasubramanian. Fault diagnosis by qualitative trend analysis of the principal components. *Chemical Engineering Research and Design*, 83(A9):1122–1132, September 2005.

[76] M. Meissner, M. Schmuker, and G. Schneider. Optimized particle swarm optimization (opso) and its application to artificial neural network training. *BMC Bioinformatics*, 7:125, 2006.

[77] S. Molau, M. Pitz, R. Schluter, and H. Ney. Computing mel-frequency cepstral coefficients on the power spectrum. *Proc. Int. Conf. on Acoustic, Speech and Signal Processing*, 1:73–76, June 2001.

[78] D. F. Morrison. *Applied Linear Statistical Methods.* Prentice Hall, 1983.

[79] R. Nagarajan and M. Upreti. Correlation statistics for cdna microarray image analysis. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 3(3):232–238, July 2006.

[80] U. S. D. of Commerce. *Manual of Regulations and Procedures for Federal Radio Frequency Management*. US Government Printing Office, May 2003.

[81] T. O'Haver. Peak finding and measurement.

[82] M. Oltean. Evolving evolutionary algorithms for function optimization. In K. C. (et al), editor, *The 7th Joint Conference on Information Sciences*, volume 1, pages 295–298, North Carolina, Sept. 2003. Association for Intelligent Machinery.

[83] D. Ozdemir and R. Williams. Multi-instrument calibration with genetic regression in uv-visible spectroscopy. *Applied Spectroscopy*, 53(2):210–217, 1999.

[84] J. T. M. Pearce, T. J. Athersuch, T. M. D. Ebbels, J. C. Lindon, J. K. Nicholson, and H. C. Keun. Robust algorithms for automated chemical shift calibration of 1d 1h nmr spectra of blood serum. *Analytical Chemistry*, 80(18):7158–7162, August 2008.

[85] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[86] M. Preghenella, G. Pezzotti, and C. Migliaresi. Comparative raman spectroscopic analysis of orientation in fibers and regenerated films of bombyx mori silk fibroin. *Journal of Raman Spectroscopy*, 38(5):522–536, May 2007.

[87] C. Raphael. Automatic transcription of piano music. *Proc. Int. Symposium on Music Information Retrieval*, pages 15–19, October 2002.

[88] C. Rascon, B. Lennox, and O. Marjanovic. Using lagged spectral data in feedback control using particle swarm optimisation. *Proceedings of the UKACC Control 2008 Conference*, September 2008.

[89] H. Ressom, R. Varghese, D. Saha, E. Orvisky, L. Goldman, E. Petricoin, T. Conrads, T. Veenstra, M. Abdel-Hamid, C. Loffredo, and R. Goldman. Particle swarm optimization for analysis of mass spectral serum profiles. *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 431–438, 2005.

[90] S. Roberts and R. Everson. *Independent component analysis : principles and practice.* Cambridge University Press, 2001.

[91] S. J. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach.* Prentice Hall, 2 edition, December 2002.

[92] H. Sakoe and S. Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(1):43–49, February 1978.

[93] J. D. Schaffer, R. A. Caruana, L. J. Eshelman, and R. Das. A study of control parameters affecting online performance of genetic algorithms for function optimization. In *Proceedings of the third international conference on Genetic algorithms*, pages 51–60, San Francisco, CA, USA, 1989. Morgan Kaufmann Publishers Inc.

[94] Y.-W. Shang and Y.-H. Qiu. A note on the extended rosenbrock function. *Evolutionary Computation*, 14(1):119–126, 2006. PMID: 16536893.

[95] G. Shetty, C. Kendall, N. Shepherd, N. Stone, and H. Bar. Raman spectroscopy: Elucidation of biochemical changes in carcinogenesis of oesophagus. *British Journal of Cancer*, 94:1460–1464, April 2006.

[96] Y. Shi and R. C. Eberhart. Parameter selection in particle swarm optimization. *Proceedings of the 7th International Conference on Evolutionary Programming VII*, pages 591–600, 1998.

[97] P. Smaragdis and J. C. Brown. Non-negative matrix factorization for polyphonic music transcription. *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 177–180, 2003.

[98] S. W. Smith. *The Scientist and Engineer's Guide to Digital Signal Processing.* California Technical Publishing, 2 edition, 1999.

[99] J. V. Stone. *Independent component analysis : a tutorial introduction.* Cambridge, Mass. ; London : MIT Press, 2004.

[100] R. Storn. Designing nonstandard filters with differential evolution. *Signal Processing Magazine, IEEE*, 22(1):103–106, Jan. 2005.

[101] R. Storn and K. Price. Minimizing the real functions of the icec'96 contest by differential evolution. pages 842–844, May 1996.

[102] R. Storn and K. Price. Differential evolution –a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11(4):341–359, 12 1997/12/01/.

[103] R. Szostak and S. Mazurek. Quantitative determination of acetylsalicylic acid and acetaminophen in tablets by ft-raman spectroscopy. *The Analyst*, 127:144–148, 2002.

[104] Y. Tie and M. Sahin. Separation of spinal cord motor signals using the fastica method. *Neural Engineering*, 2:90–96, 2005.

[105] A. Törn and A. Zilinskas. *Global Optimization*, volume 350 of *Lecture Notes in Computer Science*. Springer-Verlak, 1989.

[106] J. H. G. M. van Geffen and R. F. van Oss. Wavelength calibration of spectra measured by the global ozone monitoring experiment by use of a high-resolution reference spectrum. *Applied Optics*, 42(15):2739–2753, 2003.

[107] J. Vial, H. Noçairi, P. Sassiat, S. Mallipatu, G. Cognon, D. Thiébaut, B. Teillet, and D. N. Rutledge. Combination of dynamic time warping and multivariate analysis for the comparison of comprehensive two-dimensional gas chromatograms: Application to plant extracts. *Journal of Chromatography A*, 1216(14):2866 – 2872, 2009. 32nd International Symposium on Capillary Chromatography and 5th GCxGC Symposium.

[108] F. Vogt and K. Booksh. Influence of wavelength-shifted calibration spectra on multivariate calibration models. *Applied Spectroscopy*, 58(5):624–635, 2004.

[109] F. Vogt and B. Mizaikoff. Dynamic determination of the dimension of pca calibration models using f-statistics. *Journal of Chemometrics*, 17(6):346 – 357, August 2003.

[110] Y. Wang and M. Gu. Application of comprehensive calibration to mass spectral peak analysis and molecular screening. United States Patent 7451052, November 2008.

[111] D. H. Wolpert and W. G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, 1:67–82, 1997.

[112] J. W. H. Wong, C. Durante, and H. M. Cartwright. Application of fast fourier transform cross-correlation for the alignment of large chromatographic and spectral datasets. *Analytical Chemistry*, 77(17):5655–5661, August 2005.

[113] C.-P. Wu, P.-C. Su, and C.-C. J. Kuo. Robust audio watermarking for copyright protection. *SPIE 44th Annual Meeting, Advanced Signal Processing Algorithms, Architectures, and Implementations*, 3807:387–397, July 1999.

[114] Z. Ye and G. Auner. Principal component analysis approach for biomedical sample identification. *IEEE International Conference on Systems, Man and Cybernetics*, 2:1348–1353, 2004.

[115] Z. Ye, G. Auner, and P. Manda. Raman spectra calibration, extraction and neural network based training for sample identification. *Proceedings of the International Joint Conference on Neural Networks*, 1:622–626, 2003.

[116] L. X. Yu, R. Lionberger, A. Raw, R. D'Costa, H. Wu, and A. Hussain. Applications of process analytical technology to crystallization processes. *Center for Drug Evaluation and Research, Food and Drug Administration*, pages 1–46, August 2003.

[117] M. Zuppa, C. Distante, K. C. Persaud, and P. Siciliano. Recovery of drifting sensor responses by means of dwt analysis. *Sensors & Actuators: B. Chemical*, 120(2):411–416, January 2007.

Appendix A

Music Analysis

Musical Information Retrieval (MIR) has attracted significant attention in the scientific community. Many of the identified research areas in this field are very challenging, such as the automatic identification of specific songs and content. These problems are generic across science disciplines and successful solutions will benefit many other applications.

In this chapter, a brief literature review will be given about the use of musical spectra for Musical Analysis, followed by the application of the framework proposed in this thesis for note identification. Then, the description will be provided of a specialised algorithm developed to analyse musical spectra after having used the framework to confirm the type of distortion taking place.

A.1 Background on Musical Information Retrieval

Many areas of this field are being tackled [27], but only two of them are of interest to this project. The first is Bibliographical [27], where information such as the name of the artist, name of the song, and year of creation are retrieved from the song. The field of Song Watermarking could potentially meet this task. It is currently being used for digitally signing a piece of music to enforce copyright law [73, 113]. The digital signature is encoded inside the recording as a low energy frequency component, such as white noise, which is retrieved by the use of Independent Component Analysis. The approach could be generalised to code the name of the artist and song. The digital signature needs to be hidden in specific points in the recording, as not to be easily extracted out of the recording illegally. Although it has been noted that this procedure is robust against time and frequency scaling, the authors themselves have admitted that the tests were carried out with shift rates below 10%, whilst higher rates are definitely possible.

The second is Harmonic [27], where the specific notes being played at any point in the recording are retrieved. Identifying a specific note and its time location is crucial for any algorithm involving harmony identification, key detection, automatic music transcription, song structure identification, etc. [71].

One approach to note identification is to assume a perfectly-tuned instrument, as with the Mel-Frequency Cepstral Coefficients method [72] or when applying music-oriented filter banks, such as the Constant-Q Trasnform [8]. The main frequency in which the recorded instrument is tuned can be assumed to be known *a priori*, as with the Pitch Class Profiles developed by Dr. Fujishima [33]. However, if the tuning of the instrument deviates by amounts as small as 1%, these approaches begin to fail [8]. It has been stated that it is important to check the tuning of the system as "frequency shifts can occur during the process of sampling" [8].

Another approach, shown in [87, 64], is to train using every type of chord variation[1], instead of relying on note identification. However, the tuning issue is still relevant, unless the training includes all of the possible chords with all of the possible combinations of de-tuned notes. In fact, in [64], the errors obtained were blamed on the need for more types of chord implementations during the training stage. Despite this, the authors did not propose an estimation of the size of the training data set sufficient for an error-free estimation, which obviously would be of impractical dimensions.

The main reason for the 'perfectly-tuned instrument' assumption of the described algorithms is the very nature of their approach: the frequency location of the notes are pre-defined. However, the musician may not be aware of how far from exact tuning the instrument is at any time, or even if the instrument is de-tuned at all. In addition, many instruments are made of wood or metal alloys, which are flexible materials whose shape can change depending on the room temperature or humidity. As a result, the tuning of the instrument is not constant throughout a live performance.

Methods to address the deviation from precise tuning have been proposed by

[1] A chord is a group of notes played at the same time.

automatically detecting the tuning or *key* of the instrument [65, 72]. However, a string instrument, such as a guitar, has several strings that may each have been tuned to a different base-frequency, because of human error or outside influences. This means that even if an accurate tuning-base-frequency or key detection algorithm was developed, it would be an ‘average’ tuning-base-frequency or key of all the strings which may cause detection problems in the future.

It is necessary, then, to identify each note as an individual component in the recording, a process known as Polyphonic Music Retrieval [61, 25]. Unfortunately, this technique uses methods that rely on the specific nature of the note spectral signature, making them non-generalisable. Non-Negative Matrix Factorisation has also been tried [97], but it bases itself upon detecting ‘unique events’, not spectral components. This means that if two notes are always played together in the recording, those two notes will be considered the same ‘event’. In the same way, a de-tuned version of a note would be considered another ‘unique event’. Still, it is important to mention that, of all the algorithms described, this has been the closest one to achieve the goal of Polyphony.

It is essential to note that component extraction is of little relevance in the Music arena. This statement may seem dubious, as there is a great amount of information that can be retrieved from the spectrum of a note, such as its base-frequency and overtone magnitudes. However, the spectral signature of all possible notes are identical warped versions of each other. As will be seen later in this work, this fact provides a clear heuristic with which the aforementioned information can be extracted without the need to extract the spectral signatures of the notes. This means that the information extraction algorithms used in the Music field are simpler and faster than the ones used in other fields, solely because the spectral signatures of all possible musical components are already known.

Nonetheless, the Music field does not have the definitive answer to the problem in question. The methods used in this field are too specific and non-generalisable. Even so, because musical data is relatively easy to obtain[2], the Music field will be seen in this work as a testbed for algorithms, rather than an area of interest for application.

[2] The author of this work is a music hobbyist, as is his supervisor.

A.2 Using the Framework for Note Identification

As described in the last section, there is little interest in spectral signature extraction and identification in the field of Musical Note Identification. The spectral signature of a musical note is a good identifier, but, as will be discussed further in this section, there are other features that are simpler and more powerful to identifiers of a note. However, it is important to note that the approach proposed in this thesis, which can be of great relevance in the fields of Pharmaceutical Industry and Biomedicine, was founded and developed using musical data, because such data is prevalent and frequently contains spectral distortions. Although not studied here in great detail, there is the possibility of using the described approach for Polyphonic Music Retrieval [61, 25].

A.2.1 Introduction

Identifying the musical notes played during a musical recording is an important aspect of Music Information Retrieval. A musical note can be identified in several ways, one of which is its frequency spectrum, an example of which is shown in the upper graph in Figure A.1. The first, left-most peak is located at the *base-frequency* of the note, the essential identifier of a musical note used to derive the location of every other peak in the spectrum, which are known as *overtones*. The frequency at which the second harmonic is located is double the base-frequency, triple in the third, and so forth.

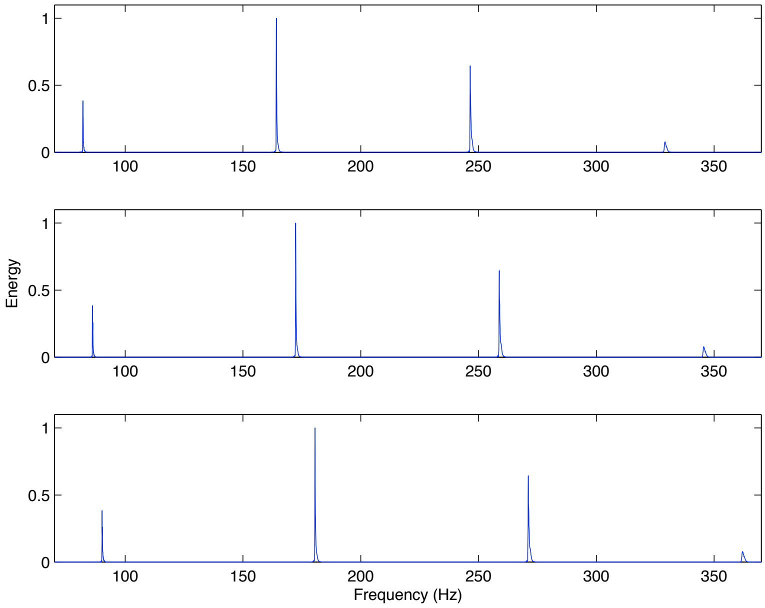


Fig. A.1: Example of spectrum stretch/shrink in de-tuned musical note.

An important problem faced during the identification procedures is the sensitivity of the instruments to variations in the tuning during the recording. Such inconsistencies can be caused by a change in temperature near the instrument being recorded, or by human error when tuning. The distortion that takes place in the spectrum when de-tuning occurs is essentially a Local Warp. In the middle and lower graph in Figure A.1, the spectra of two notes de-tuned by varying amounts are shown. It can be seen how the base-frequency varied by a small amount, but the later peaks are shifted by an exponentially growing amount.

The Spectral Analysis Framework proposed in this thesis can be used to identify the notes played in a recording of a de-tuned instrument by simulating the Local Warp distortion as described in Appendix C. As for the reference spectral data set, because of the simple nature of a musical spectrum, it can be manufactured by applying (A.1).

$$B_n(w, f, H, a) = \begin{cases} 1; w = hf, h = 1, \dots, H \\ 0; w \neq hf \end{cases} \tag{A.1}$$

B_n is a simulated spectrum of a component n that is part of a benchmark B; H is the number of harmonics in the benchmark spectrum (which includes the one in its base-frequency). An example of the reference spectrum created for the

E2 note, with five harmonics, is shown in Figure A.2.

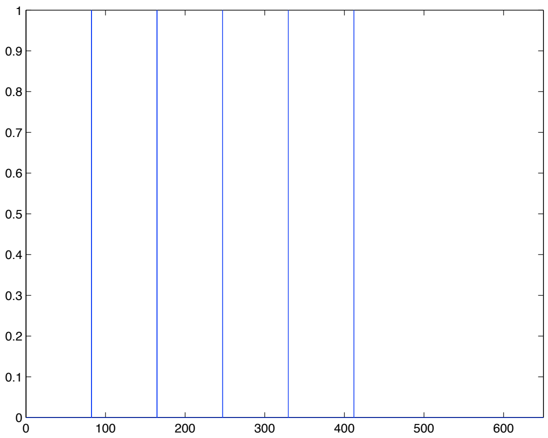


Fig. A.2: Reference spectrum of an E2 Note located at 82.407 Hz.

A set of reference musical spectra can be created using the documented base-frequencies for each musical note, a brief list of which can be found in Appendix B. Such a set will be used as a database from which the optimisation algorithm can obtain an optimal subset of notes that best resembles the recorded spectrum, as well as the values of warp and contribution for each.

There are many notes that a musical spectrum may have, but, if only one person is playing the instrument, the number of notes that can be played at a specific time is reduced. A flute can only produce one note at a time, six with a guitar, and up to ten with a piano. Meaning that it is not necessary to look for all possible musical notes at each given time in the recording. A preliminary exploration of which notes are ‘possibly’ being played is conducted to simplify the solution space, carried out by either PSO or a brute force methodology. A minimum and maximum value for a warp distortion are defined (0.98 and 1.02 achieved the best in these experiments), and a warp value was found where the Pearson correlation coefficient of the note against the recorded spectrum was the highest. If a value of 0 was found to be the highest, then this note was filtered out from the main search process.

The main PSO search is then carried out by looking for an optimal combination of contribution and warp values for the musical notes in the resulting subset. The warp and contribution values that were found in the preliminary stage are used as starting points in the main search, to reduce the number of search iterations. The values were modified stochastically by a small factor between 0.99 and 1.01, to avoid the particles starting in the same point.

To test the approach, three chords played by a clean electric guitar were recorded and their frequency spectrum was obtained. Each chord consisted of two notes: E2 and D3; G2 and B2; E2 and C3. A preliminary search was conducted and the relevant notes for each recording were identified. A PSO search was then conducted to find which of the relevant notes were present in the recording. The relevant notes that were found and their contribution values are shown in Table A.1.

Recording 1						
Relevant notes	**E2**	G2	A2	B2	**D3**	
Contribution	**0.4**	0	0	0	**0.1**	
Recording 2						
Relevant notes	E2	F#2	**G2**	**B2**	D3	
Contribution	0	0	**0.4**	**0.2**	0	
Recording 3						
Relevant notes	**E2**	F2	G2	A2	B2	**C3**
Contribution	**0.4**	0	0	0	0	**0.11**

Table A.1: Contribution values of the relevant notes of each recording.

The notes in bold are the only ones that were found in the main PSO search to have important contribution in the recorded musical spectra, and are the notes that were expected to be identified for each recording.

It could be argued that the subset of notes found in the preliminary search are the notes present in the musical spectra and that the main search is unnecessary. However, because of the nature of the musical spectrum, which comprises peaks at frequency locations of constant ratios, such features are also contained in the spectra of other notes, commonly called *harmonic notes*. This means that the contribution of one note results in a partial contribution of its harmonic notes, and only by searching for the presence of both the original and the probably non-present harmonic notes that a proper identification can be carried out.

The base-frequencies of harmonic notes are located at specific ratios of the base-frequency of the original note. A note located at $\frac{4}{3}$ of a note base-frequency is called a *perfect fourth*, and at $\frac{3}{2}$ results in a *perfect fifth*. Ratios calculated by 2^n, where n is any integer other than 0, result in the same note but in a different *octave*. For instance, the G3 note is located at 196 Hz; it is the G note located in the third octave[3]. G2 is located at half the G3 base-frequency ($2^{-1} * 196 = 98$ Hz), G4 at double ($2^1 * 196 = 392$ Hz), and G5 at quadruple ($2^2 * 196 = 784$ Hz).

A.2.2 Limitations of the Proposed Framework in Music Applications

The Framework proposed in this thesis has an important set of limitations in the Musical arena. The main objective of Polyphonic Music Retrieval is to identify the presence of a note in a recording. However, using the frequency spectrum for such a task is unnecessarily complex. The musical spectra of two different notes are similar between each other; in fact, it can be shown that the frequency spectra of two different notes are warped versions of each other. In addition, notes are identified by relying on the information located in thin peaks of the spectrum, a text-book definition of a sparse representation, which, when processed, results in unnecessary time handling non-relevant information.

Furthermore, the partial overlapping of harmonic notes is an issue dealt with to a limited extent. Components in the Musical arena are expected to have important overlapping features, which can be resolved by applying the approach twice, the first acting as a filter for the second.

It is important to note that the objective of the proposed Framework is to provide a generic solution to spectral analysis problems, although it may not be the most efficient one. However, the proposed Framework confirmed the assumptions of the spectral distortions and how they affected the measured spectra. With this knowledge, other techniques can be implemented that are specific to the problem, and be much more efficient. In the following section, a detailed description of a technique that fulfilled the objectives of Polyphonic Music Retrieval is given, that was created with the knowledge confirmed by the proposed Framework.

A.2.3 Conclusions

To test the feasibility of the proposed framework in other areas that apply spectral analysis, it was used to identify the notes being played in de-tuned musical recordings. How well a musical instrument is tuned can be influenced

[3] The numbering of octaves are calculated by the piano keyboard, which starts at A0 and ends on C8.

by changes in the environment or human error. When a note is out of tune, its spectral signature suffers from Warp and many current methods in Polyphonic Music Retrieval require to consider it. The framework was able to identify the notes being played in a recording of de-tuned electric guitar, by considering a database of known musical spectral signatures as its reference spectral data set, and the simulation of a Warp distortion. Although the framework should not be considered as an efficient method for Polyphonic Music Retrieval, the fact that it was able to achieve its goal is a testament of its versatility.

A.3 Polyphonic Music Retrieval by Peak-Finding

The spectrum of a note can be represented by the locations and heights of its peaks, known as *overtones*. For example, an E2 note, with its base-frequency at 82.407 Hz, can be simply identified by the values presented in Table A.2, which is much simpler than considering the full spectrum displayed in Figure A.2.

	Base-Frequency	First Overtone	Second Overtone	Third Overtone	...
Location	82.407 Hz	164.814 Hz	247.221 Hz	329.628 Hz	...
Magnitude	1	1	1	1	...

Table A.2: Contribution values of the relevant notes of each recording.

A composite musical spectra can be represented in a similar way, by finding the locations of the base-frequencies of the notes present in the recording. To do this, first, all the peaks in the spectrum need to localised, which can be achieved by the peak-finding algorithm implemented by Tom O'Haver in 1995 (and revised in 1996) [81]. It locates the peaks in a spectrum by searching for zero-crossings in a smoothened, derived version of the measured spectra. A musical spectrum comprised of notes E2, A2, and D3, such as the one shown in Figure A.3, can be represented by the values in Table A.3.

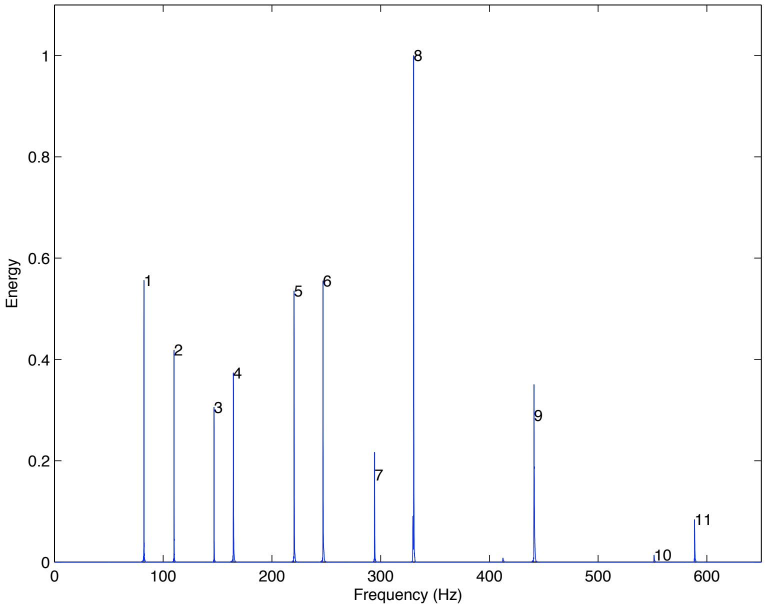


Fig. A.3: Composite spectrum of musical recording of notes E2, A2 and D3.

Peak #	Location (Hz)	Magnitude
1	82.2215	0.5561
2	109.9792	0.4185
3	146.8211	0.3058
4	164.5271	0.3735
5	220.3789	0.5350
6	246.9589	0.5551
7	294.3993	0.1722
8	330.3160	1.0000
9	441.0522	0.2893
10	551.4519	0.0137
11	588.7144	0.0836

Table A.3: Peak representation of spectrum shown in Figure A.3.

The objective of the proposed Peak-Finding-Based Note Identification algorithm can be achieved by calculating a matrix that explains which peaks belong to the same note. From this matrix, the number of notes being played in the recording can be calculated, as well as the base-frequency of each, retrieved by the frequency location of the leftmost peak of each identified note. Each note can then be identified by the closest reference base-frequency found in Appendix B.

From the peak locations presented in Table A.3, it is possible to deduce which peaks belong to the same note by observing which are close to being an integer number of times apart from each other. A matrix can be formed such that each row contains the locations of the peaks divided by the location value of one peak. The relation of one peak to another can be judged by how close the result of the division is to an integer. The implementation is summarised in (A.2).

$$M = 1 - \left| \begin{bmatrix} \frac{p_1}{p_1} & \frac{p_2}{p_1} & \cdots & \frac{p_n}{p_1} \\ \frac{p_1}{p_2} & \frac{p_2}{p_2} & \cdots & \frac{p_n}{p_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{p_1}{p_n} & \frac{p_2}{p_n} & \cdots & \frac{p_n}{p_n} \end{bmatrix} - \begin{bmatrix} rd\left(\frac{p_1}{p_1}\right) & rd\left(\frac{p_2}{p_1}\right) & \cdots & rd\left(\frac{p_n}{p_1}\right) \\ rd\left(\frac{p_1}{p_2}\right) & rd\left(\frac{p_2}{p_2}\right) & \cdots & rd\left(\frac{p_n}{p_2}\right) \\ \vdots & \vdots & \ddots & \vdots \\ rd\left(\frac{p_1}{p_n}\right) & rd\left(\frac{p_2}{p_n}\right) & \cdots & rd\left(\frac{p_n}{p_n}\right) \end{bmatrix} \right| \tag{A.2}$$

where p_i is the frequency location of the ith peak; the rd function is a rounding function; the $|\cdot|$ operator represents the absolute value of the operation; and M is a matrix that holds the information of which peak belongs to the same note as another. The row number represents one peak, and the column number represents another; it has been found that if the cell that represents both peaks has a value higher than 0.97, then both peaks belong to the same note. The values in M can then be converted to either 1, if the original value was higher than 0.97, or to 0, otherwise.

The rows that have only one cell with a value of 1 are then deleted from M, to filter out outlying peaks. Rows that have all of their values present in another are also removed, as they are redundant. The resulting matrix has the same amount of rows as notes in the frequency spectrum.

The resulting matrix calculated with the data from Table A.3 is given in Table A.4.

Peak #	1	2	3	4	5	6	7	8	9	10	11	Base-Freq. (Hz)	Closest Ref.
Note 1	*1*	0	0	1	0	1	0	1	0	0	0	82.2215	**E2**
Note 2	0	*1*	0	0	1	0	0	1	1	1	0	109.9792	**A2**
Note 3	0	0	*1*	0	0	0	1	0	1	0	1	146.8211	**D3**

Table A.4: Peaks belonging to the same note in spectrum shown in Figure A.3.

Table A.3 indicates that Note #1 has its base-frequency in the location of Peak #1 (82.2215 Hz), and the closest reference to this is E2 (82.41 Hz) . The base-frequency of Note #2 is at Peak #2 (109.9792 Hz), with A2 being the closest (110 Hz). And Peak #3 represents the base-frequency (146.8211 Hz) of Note #3, having D3 as its closest reference (146.83 Hz).

The algorithm described here was tested against 45 different guitar recordings, ranging from 1 to 4 notes being played at the same time, with different stroking and different hand positions. Table A.5 shows the result from these tests.

#	Notes Expected	Notes Identified	#	Notes Expected	Notes Identified	#	Notes Expected	Notes Identified
1	E2	E2	16	E2	E2	31	E2	E2
2	A2	A2	17	A2	A2	32	A2	A2
3	E2,A2	E2,A2	18	D3	D3	33	D3	D3
4	D3	D3	19	E2,A2,D3	E2,A2,D3	34	F#3	F#3
5	E2,D3	E2,D3	20	G2	G2	35	G3	G3
6	A2	A2	21	D3	D3	36	E2,A2, D3,G3	E2,A2, D3,G3
7	E3	E3	22	G3	G3	37	B2	B2
8	A2,E3	A2,E3	23	G2,D3,G3	**G2,D3**	38	G#3	G#3
9	G2	G2	24	A2	A2	39	E2,B2, E3,G#3	**E2,B2, G#3**
10	B2	B2	25	E3	E3	40	B3	B3
11	G2,B2	G2,B2	26	A2,E3,G3	A2,E3,G3	41	B2,E3, G#3,B3	**B2,E3, G#3**
12	C3	C3	27	A3	A3	42	E4	E4
13	E2,C3	E2,C3	28	E4	E4	43	E3	E3
14	C#3	C#3	29	G4	G4	44	D3,G3, B3,E4	D3,G3, B3,E4
15	E2,C#3	E2,C#3	30	D3,A3,E4, G4	D3,A3,E4, G4	45	E2,B2, F#3,G#3	E2,B2, F#3,G#3

Table A.5: Notes identified by algorithm. Bold numbers represent mis-identified notes.

From Table A.5, it can be seen that the algorithm was able to identify correctly nearly all the notes in the 45 recordings, except in three cases. All three incorrectly identified examples involved the same note being played but at different octaves at the same time. Because of the way that the reference spectra is built and the nature of the linear model that is assumed, a note that is played in two or more octaves at the same time (such as E2 and E3) will be identified as only one note. Solving this particular problem remains open in the field of Music Information Retrieval. It has been addressed by using a high resolution frequency spectrum and assuming that the peaks of the notes in different octaves do not overlap [71], but the performance of this approach was still not acceptable.

It is important to note that the mean time in which the note identification process took place was 68.8 ms per recording, which, compared favourable to the mean time of $\sim$ 6 minutes for the optimisation approach. This represents a significant improvement.

Appendix B

Musical Note Frequencies

In Table B.1 the standard base-frequencies (in Hz) for notes ranging from A0 to C8 are shown. These values were obtained from [2], and correspond for the notes in the tempered scale, which is very popular in both professional and amateur settings. There are other types of scales, which differ in the ratios between the harmonics, but it was decided to use this scale for its popularity and the fact that the base-frequency deviations from one scale to another deviate by a relatively small amount.

All the notes that can be played in a piano are shown. As a frame of reference for guitarists, the E2 note is the lowest note of a guitar in standard tuning, and E5 is the note in the 12th fret of the highest string.

It is important to note the relationships between harmonics. A4 is double the frequency of A3, which is exactly one octave below it. It has 2/3 the frequency of E5, its perfect fifth, and it 3/4 the frequency of D5, its perfect fourth.

Note	Base-frequency	Note	Base-frequency	Note	Base-frequency
A0	27.5	D#3/Eb3	155.56	A5	880
A#0/Bb0	29.14	E3	164.81	A#5/Bb5	932.33
B0	30.87	F3	174.61	B5	987.77
C1	32.7	F#3/Gb3	185	C6	1046.5
C#1/Db1	34.65	G3	196	C#6/Db6	1108.73
D1	36.71	G#3/Ab3	207.65	D6	1174.66
D#1/Eb1	38.89	A3	220	D#6/Eb6	1244.51
E1	41.2	A#3/Bb3	233.08	E6	1318.51
F1	43.65	B3	246.94	F6	1396.91
F#1/Gb1	46.25	C4	261.63	F#6/Gb6	1479.98
G1	49	C#4/Db4	277.18	G6	1567.98
G#1/Ab1	51.91	D4	293.67	G#6/Ab6	1661.22
A1	55	D#4/Eb4	311.13	A6	1760
A#1/Bb1	58.27	E4	329.63	A#6/Bb6	1864.66
B1	61.74	F4	349.23	B6	1975.53
C2	65.41	F#4/Gb4	369.99	C7	2093
C#2/Db2	69.3	G4	392	C#7/Db7	2217.46
D2	73.42	G#4/Ab4	415.31	D7	2349.32
D#2/Eb2	77.78	A4	440	D#7/Eb7	2489.02
E2	82.41	A#4/Bb4	466.16	E7	2637.02
F2	87.31	B4	493.88	F7	2793.83
F#2/Gb2	92.5	C5	523.25	F#7/Gb7	2959.96
G2	98	C#5/Db5	554.37	G7	3135.96
G#2/Ab2	103.83	D5	587.33	G#7/Ab7	3322.44
A2	110	D#5/Eb5	622.25	A7	3520
A#2/Bb2	116.54	E5	659.26	A#7/Bb7	3729.31
B2	123.47	F5	698.46	B7	3951.07
C3	130.81	F#5/Gb5	739.99	C8	4186.01
C#3/Db3	138.59	G5	783.99		
D3	146.83	G#5/Ab5	830.61		

Table B.1: Notes in tempered scale with corresponding base-frequencies in Hz.

Appendix C

Algorithm to Artificially Distort a Signal

Artificial warping and shifting was a frequent task during the development of this project. An algorithm to artificially distort a given signal was therefore implemented and, although its original use was intended towards frequency spectra, it can be used with any signal, even if it is not continuous. Its objective is to modify a signal in a way that its overall shape is maintained, but with severe local distortions. A secondary objective is to generalize the ways that a signal can be distorted. When a sensor is not calibrated, the sampled signal may be a distorted version of the expected one by more ways than just a linear shift.

The types of distortions implemented here are the ones described in the rest of the work, and were chosen because they are the ones that have been found the most prevalent in the reviewed literature as well as being observed in experiments. However, because of the way that the algorithm is programmed, any other type of modification could be added on if necessary.

- **Shift**. A linear displacement of the signal towards the left or right of the horizontal axis.
- **Warp**. A non-linear shift that stretches or shrinks the signal horizontally.

The following sections will describe the way that these distortions were developed, in a conceptual manner. These implementations are at the heart of many of the developed algorithms, so it is important for them to be simple and efficient to avoid bottle necks in the overall code. To this effect, the distortions are implemented in a very small amount of lines using structures already built into MatLab. However, only explaining how this algorithm was developed in MatLab may lead to it being unavailable in other platforms, such is the reason of describing the algorithm only abstractly in the following sections and leaving specific-language-efficiency concerns up to the programmer.

C.1 Shift

Shift, as described before, is a linear displacement towards the left of right in the horizontal axis. To implement a shifted signal, the problem is divided in two. If the shift is towards the left, the algorithm cuts a 'chunk' off at the start of the signal of a size of the wanted shift. If the shift is towards the right, the algorithm creates an all-zero signal which is the size of the wanted shift, and then concatenates it with the original signal.

Algorithm 3 Pseudo-code for the shift distortion.

```
$S \leftarrow signal\_to\_be\_shifted$
$\hat{S} \leftarrow resulting\_signal$
**if** $shift > 0$ **then**
    $\hat{S} = \textbf{concatenate}(\textbf{arrayofzeros}(shift), S)$
**else**
    $\hat{S} = S[shift : end]$
**end if**
```

It is to be noted that this distortion will produce a signal of different length than the original, however in Section C.3 it is discussed how this issue is dealt with, as the warp distortion (described in C.2) also produces this.

C.2 Warp

Warping involves uncommon array handling and the procedures for shrinking and for stretching differ. However, it resulted to be a very simplified version of Dynamic Time Warping (DTW), but done in reverse. DTW gives a measure of similarity between two signals, given that one is a warped version of the other [66]. DTW basically puts the two signals in two different axis and tries to find the best path to map the points between the signals; the length of this path is the measure of similarity.

This modification does what can be called as an Inverse DTW (IDTW): its objective is to find the warped version of one signal given a proportion of its length. It does this by obtaining from that proportion the length of the warped version of the signal, and then obtaining the best path between the two signals, which is based on the difference of lengths between the original signal and its warped version.

For example, a known original signal (signal A) has a length of 100 sampled points and the user wants a *stretched* version of it with a proportion of 1.01 of its length. The unknown warped signal (signal B) is going to have a length of $100 * 1.01 = 101$ points; a difference of 1 point. To create B, it will be necessary to divide A into 2 chunks of equal length and add that point between them. The value that this new point will have is obtained by averaging the values of the points beside it[1] (the end point of chunk 1 and the start point of chunk 2). If the proportion would have been 1.1, B would have had a length of $100 * 1.1 = 110$ points; a difference of 10 points. To create B in this case, it is necessary to divide A in 11 chunks of equal length and add 1 point between the chunks.

In the case of *shrinking* signal A, the user now wants signal B to have a proportion of 0.99 of the length of A. B is now going to have a length of $100*0.99 = 99$ points; a difference of -1 point. To create B, it will be necessary to divide A into two chunks and remove the point that is at the end of the first chunk. If the proportion would have been 0.9, B would have had a length of $100 * 0.9 = 90$ points; a difference of -10 points. To create B in this case, it will be necessary to divide A in 11 chunks and remove the point at the end of each chunk, except the last one.

It is important to note that the warp proportion, for this algorithm, can only go from 0.5 to 2; this means that the smallest warped version of a signal that this algorithm can calculate is going to be half the length of the original, and the largest is going to be double. The reason why it can only go down to 0.5 is that the minimum length of a chunk is of 2 points when shrinking; this is because every chunk (except the last one) will have its end point removed and, if a chunk is permitted to have a length of only 1 point, its resulting length after warping will be of zero. The same conundrum applies to the opposite, when stretching,

[1] This is to ‘smooth out’ the newly added areas.

as every chunk has the minimum size of 1 point, the maximum amount of chunks that can be produced is equal to the length of the signal.

The pseudo-code for this distortion is given in Algorithm 4. However, it is very inefficient in its use of *for* loops, which are very CPU-intensive. If using MatLab, this can be overcome by taking advantage of its efficient array handling. Every chunk is identified by an index, calculated the same way as for the chunks. If the signal is being shrinked, it is only necessary to remove the points at the indexes. If it is being stretched, an array of means between the chunks is created, and, with clever indexing, both the original and the new arrays are fitted onto another array. Refer to the MatLab source code presented in Table C.1, where it is shown how this was implemented.

Algorithm 4 Pseudo-code for the warp modification, in a conceptual manner.

```
$S \leftarrow signal\_to\_be\_warped$
$\hat{S} \leftarrow resulting\_signal$
$N = |\textbf{round}(\textbf{length}(S) * warp - \textbf{length}(S))| + 1$
$G = \textbf{split}(S, N)$
if $warp > 1$ then
  for $i = 1 : (N - 1)$ do
    $\hat{S}.\textbf{push}(G[i])$
    $\hat{S}.\textbf{push}(\textbf{mean}(G[i][end], G[i+1][start]))$
  end for
else if $warp < 1$ then
  for $i = 1 : (N - 1)$ do
    $\hat{S}.\textbf{push}(G[i][start : end - 1])$
  end for
end if
$\hat{S}.\textbf{push}(G[N])$
```

C.3 Enforcing the Original Length

The shift and warp distortions change the length of the signal, so it is necessary to restore the length of the original signal after the fact. To do this, a comparison between the two signals is made. If the distorted version is bigger than the original, the extra points are removed; if it is smaller, it is zero-padded. The pseudo-code for this enforcement is given in Algorithm 5.

Algorithm 5 Pseudo-code for enforcing the original length.

```
$S \leftarrow original\_signal$
$\hat{S} \leftarrow signal\_to\_be\_corrected$
$\hat{\hat{S}} \leftarrow resulting\_signal$
if length($S$) > length($\hat{S}$) then
    $\hat{S}$.push(arrayofzeros(length($S$) − length($\hat{S}$)))
    $\hat{\hat{S}} = \hat{S}$
else
    $\hat{\hat{S}} = \hat{S}[start : (end - (\mathbf{length}(\hat{S}) - \mathbf{length}(S)))]$
end if
```

C.4 MatLab Implementation

```
function y = distort_signal(x, shift, warp, base, mult)

[mx nx] = size(x);

%warp section
if warp > 1    %stretching
    add_points = round(nx*warp - nx);     %# of added points
    ny = nx+add_points;                   %length of warped input
    in_betw = nx/(add_points+1);          %# of points btw the added ones
    y = zeros(1,ny);
    seq_to_dir = 1:ny;
    seq_to_takeout = round(in_betw+1:in_betw+1:ny);
    seq_to_dir(seq_to_takeout) = [];
    y(seq_to_dir) = x;
    y(seq_to_takeout) = mean([y(seq_to_takeout-1);y(seq_to_takeout+1)]);
elseif warp < 1    %shrinking
    rem_points = round(nx - nx*warp);    %# of removed points
    ny = nx-rem_points;                  %length of warped input
    in_betw = ny/(rem_points+1);         %# of points btw the removed ones
    seq_to_dir = 1:nx;
    seq_to_takeout = round(in_betw+1:in_betw+1:nx);
    seq_to_dir(seq_to_takeout) = [];
    y = x(seq_to_dir);
else
    y = x;
end

%shift section
if shift > 0
    y = [zeros(1,round(shift)) y];
elseif shift < 0
    y = y(round(shift):end);
end

%forcing the same length as the original one
[my ny] = size(y);
if ny > nx
    y = y(1:nx);
elseif ny < nx;
    y = [y zeros(1,nx-ny)];
end
```

Table C.1: Source code of distort algorithm in MatLab.