Automated Baseline Estimation for Analytical Signals

by

Rukhshinda Jabeen

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

at

Dalhousie University
Halifax, Nova Scotia
August 2013

Table of Contents

List of Tables

List of Figures

Abstract

The interfacing of analytical measurement instrumentation to small computers for the purpose of on-line data acquisition has now become standard practice in the modern laboratories. An important aspect of digital data acquisition is the possibility of performing post-run data analysis and signal processing. There is a large number of pre-processing methods available including noise reduction, peak resolution, baseline removal and complex signal decomposition prior to further data analysis. Baseline artefacts often interfere with the interpretation and quantitation of analytical data by signal distortion and can complicate the data analysis. Hence, baseline subtraction methods have become an important pre-processing tool.

During the last decade, many baseline estimation methods have been proposed, but many of these approaches are either only useful for specific kinds of analytical signals or require the adjustment of many parameters. This complicates the selection of an appropriate approach for each kind of chemical signal and the optimization of multiple parameters itself is not an easy task. In this work, an asymmetric least squares (ALS) approach is used with truncated and augmented Fourier basis functions to provide a universal basis space for baseline approximation for diverse analytical signals. The proposed method does not require extensive parameter adjustment or prior baseline information. The basis set used to model the baselines includes a Fourier series truncated to low frequency sines and cosines (consistent with the number of channels) which is then augmented with lower frequencies. The number of basis functions employed depends mainly on the frequency characteristics of the baseline, which is the only parameter adjustment required for baseline estimation. The weighting factor for the asymmetric least squares in this case is dependent mainly on the level of the noise. The adjustment of these two parameters can be easily performed by visual inspection of results.

To estimate and eliminate the baseline from the analytical signals, a novel algorithm, called Truncated Fourier Asymmetric Least Squares (TFALS) was successfully developed and optimized. It does not require baseline representative signals or extensive parameter adjustments. The method is described only with parameters optimization using simulated signals. The results with simulated and experimental data sets having different baseline artefacts show that TFALS is a versatile, effective and easy-to-use baseline removal method.

List of Abbreviations Used

airPLS	Adaptive Iteratively Reweighted Penalized Least Squares
ALS	Asymmetric Least Squares
ALSS	Asymmetric Least Squares
AWP	Adaptive Wavelet Transform
DFT	Discrete Fourier Transform
DWT	Discrete Wavelet Transform
ESI	Electro Spray Ionization
FBS	Fourier Basis Set
FT	Fourier Transform
FT-IR	Fourier Transform Infrared
FT-MS	Fourier Transform Mass Spectrometry
FT-NMR	Fourier Transform Nuclear Magnetic Resonance
GC	Gas Chromatography
IWT	Inverse Wavelet Transform
LC	Liquid Chromatography
MALDI	Matrix-Assisted Laser Desorption Ionization
NAE	Negative Absolute Error
NMR	Nuclear Magnetic Resonance
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
PLS	Penalized Least Squares
RMSE	Root Mean Squared Error
SVD	Singular Value Decomposition
TFALS	Truncated Fourier Asymmetric Least Squares
TFB	Truncated Fourier Basis
WPT	Wavelet Packet Transform

Acknowledgments

Completing this thesis was never an endeavour I could have taken at my own. Had I not enjoyed the support and help of many extra-ordinary people, this undertaking would be far from where it is now. There are number of people I would really like to thank here.

My research supervisor Peter Wentzell was an ever encouraging influence for this research. It was his continuous support, patience and guidance over the past three years, that made this thesis possible. I would like to express my sincere and heartfelt gratitude to him for all his time and support.

Club Wentzell: a group of friendly, talented and helpful people, whom I am privileged to work with. Over the past years I have interacted with the many wonderful people in this group, I would like to express special thanks to Siyuan Hou, Joe Boutilier and Bjorn Wielens.

Special thanks to Michelle Everist and Fabiola Manhas to share their data. Many thanks to my friend Abir Lafsay for her moral support in every step of my life during past years.

It would be unjust not to mention my family for always being supportive. My parents, though separated by seas, have always had me in their hearts and prayers. My joyful daughters and wonderful husband, who made my tiring days full of wonder and helped me morally and financially.

Chapter 1
Introduction

1.1 Overview

The success or failure of any data analysis method mainly relies on the acquisition of signals consistent with assumptions made by the underlying model. For any analytical technique, the data must be adequately representative of the analyte or system to be examined. All analytical instruments provide indirect measurements of the chemical or physical characteristics of an analyte, for example the amount of radiation absorbed, the number of gas phase ions produced, or the amount of current produced during charge transfer. These characteristics are measured by a detector and commonly converted to a signal response which is a function of some ordinal variable such as wavelength, mass-to-charge ratio, time or voltage (i.e. a vector or first-order measurement). The response vector not only provides signals characteristic of the analyte(s), but may also include some unwanted artifacts, for example signals arising from the matrix or solvent and stray radiation. These artifacts manifest themselves as broad, featureless profiles in the ordinal domain (low frequency components in the Fourier domain) as well as sharper, more distinct peaks (interferences) and are convolved with higher frequency oscillations of lower magnitude associated with instrumental noise. All of these can alter the ideal analyte response and complicate the data analysis. Of these artifacts, estimation and elimination of low frequency components is relatively more important for modeling and quantitation due to their relatively high amplitude and overlap with the signal of interest. At the same time, such features are more complicated to identify and remove due to the

variable shape of the resulting signal along the ordinal variable. It is these artifacts that are the focus of this thesis.

1.2 Motivation

As noted above, analytical experiments that are designed to measure first-order signals often result in data that are superimposed with some slowly varying non-specific signals, complicating the accurate determination of analytical parameters. These slowly varying components are commonly known as the baseline and sometimes referred to as background in the literature [1, 2] and are illustrated in Figure 1.1. The origin and shape of the background depends on the specific measurement approach (see next section for details), but often it is slowly varying signal with broad curvature, whereas analyte peaks usually have relatively sharp and narrow features that are superimposed on broad baselines.

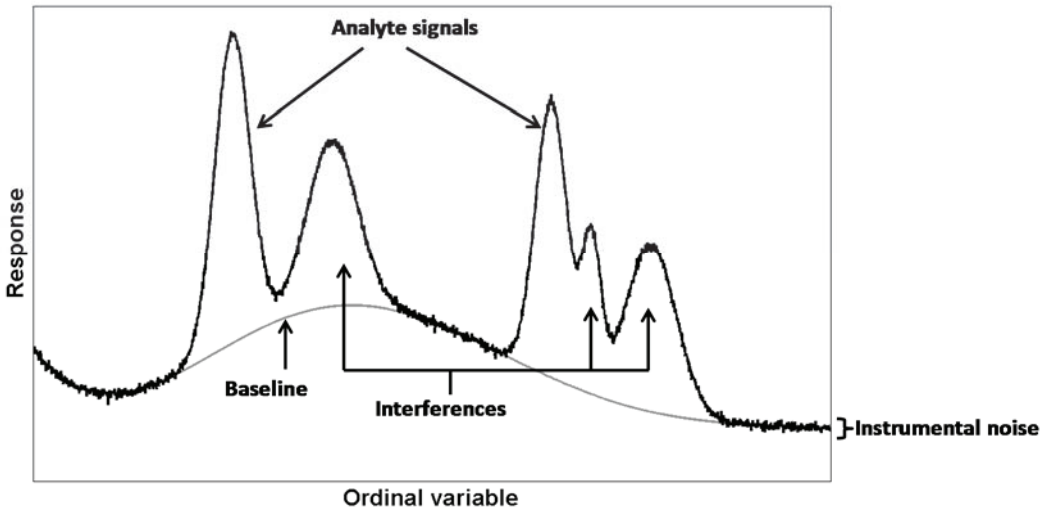

Figure 1.1 Typical components of a first-order analytical measurement.

After signal acquisition, data analysis is the common approach in analytical chemistry for peak location identification, integration or modeling. The presence of baseline usually has relatively less effect on the determination of peak location, but it can lead to significant complications in quantitative analysis. Therefore, baseline estimation and subtraction is an important component of data analysis.

Usually experiments are designed to minimize the effects of these non-specific baseline components. For example a "blank" may be subtracted from the signal of sample analyte. However, this is not possible in cases where the blank is unavailable or sample dependent, as for biological samples where the matrix is variable. In addition, this experimental elimination step often does not completely remove the baseline. Consequently, many approaches exist in the literature, from hardware modifications to post-processing of the signal to overcome this problem [3-6].

Although some baseline problems can be minimized by adjusting instrumental parameters [3, 4] or exploiting digital filtering [7] and over-sampling in Fourier transform instrumentation [8, 9], post-processing data treatment would be a more general idea for baseline correction, since finding the causes of artifacts and ways to eliminate them can be a complicated, time-consuming and a non-generalized approach. Many post-processing approaches have been employed for baseline estimation in the literature. These approaches include model-free approaches [10-44], curve fitting [45-52], removal of low frequency components in the frequency domain [53-63] and manual and automated baseline estimation in time domain [64], [28-35]. Many of these approaches are successfully able to remove the low frequency baseline artifacts, but it has been acknowledged that every approach has its limitations and strengths [5, 6]. Some

approaches are useful for specific analytical signals, which means they work for certain baseline shapes that can be associated with particular analytical signals. Other approaches are good for multiple kinds of signals but have many parameters to adjust and these parameters depend on analyte response (signal-to noise-ratio, S/N).

This thesis report describes a novel approach for baseline estimation that avoids some of the drawbacks of existing methods stated above. The work focuses on the development and optimization of the proposed approach, including quantitative comparisons with two other existing methods through simulated data sets. Applications are also included for a variety of experimental signal vectors to provide a qualitative visual evaluation for real situations where the true baseline is not known.

1.3 Sources of Baselines in Analytical Measurements

The position, shape and relative intensity of analyte signals are the basic characteristics of interest in any signal vectors to be measured for further analysis, including the structural or functional characterization of molecules and determiation of analytes. The determination of these characteristics can be confounded by the presence of a slowly varying baseline. In this section, some baseline problems associated with specific analytical techniques will be discussed. Due to the wide range of analytical measurements, this discussion cannot be comprehensive, so the emphasis will be on spectroscopic measurements because of their widespread usage. A few other techniques are mentioned as well, however.

In spectroscopic techniques, a slowly varying baseline signal can arise from the radiation source, solvent or matrix contributions, instrumental measurement effects or a

combination of these. A few common sources of baselines in spectroscopic measurements are indicated in the literature [1, 2, 65] as follows.

1) Elastic scattering of the radiation source by the sample cell, solvent or optical components.
2) Radiation reflection by the sample cell, optical components or cell compartment.
3) Raman scattering by the solvent and its components.
4) Fluorescence from the solvent and dissolved particles.
5) Luminescence from the cell wall or optical components.
6) Stray radiation due to leakage of room light.
7) Radiation source (flame, light source, plasma etc.)

Fourier transform (FT) spectroscopy has become a very common and relatively standard measurement technique for certain types of analytical instrumentation (infrared (IR), Raman, Nuclear Magnetic Resonance (NMR)) due to its fast data acquisition, fewer optical elements and better response in comparison to dispersive systems [2]. In the dispersive instruments (referred as grating or scanning spectrometers) source energy is sent through both a sample and a reference path. In these systems, the intensity over a narrow region around each point in the spectrum can be determined by slowly moving the grating. Then, the separates the wavelengths of light in the spectral range and directs each wavelength individually to the detector or alternatively a continuous array of detectors is used to measure all the regions simultaneously. On the other hand, in FT instruments, all wavelength channels are measured simultaneously (multiplex system), so the time required to collect all the data to form complete spectrum is often drastically reduced. Since all of the frequencies are collected at the same time without dispersion,

the wavelengths and their signal intensities are overlapped and are then deconvolved by Fourier analysis to form a spectrum. A limitation of FT instruments is the single beam configuration that requires a background correction following a separate background run, where possible [1].

In Fourier transform infrared (FT-IR) spectroscopy, the IR beam passes through not only the sample but also through a length of air. Since air contains two major IR active molecules, CO_2 and H_2O, absorption from these two molecules is present in every resultant spectrum. The other source of background is the variable intensity of the infrared source; the lamp usually has maximum intensity somewhere in the middle of the spectrum and gradually diminishes towards either end. Due to the uneven intensity of light source through the entire spectral range, the FT-IR spectrum typically exhibits a slowly varying curvature [2].

In NMR spectroscopy, the presence of a baseline can be the result of a bad choice of measurement parameters, corruption or non-linearity in the instrument condition (such as non-linear filter-phase response), an intense solvent signal relative to analyte (called a dynamic range problem), a dead time problem in the pulsed NMR, or the discrete nature of Fourier transform in FT-NMR [8, 66]. Dead time is the time interval between the centre of the excitation pulse and the point where the first sample of the time domain signal is obtained. The dead time in the time domain appears as missing sample points in frequency domain and the more missing sample points, the more severe the baseline distortion appears in the spectrum. Otting et al [67] demonstrated that a source of baseline in FT-NMR is the incorrect use of the Fourier transform algorithm, the so called

'first data point problem', and sampling delay. These baseline distortions in spectra can cause incorrect peak integration and peak picking. Oversampling and digital signal processing have been implemented in modern spectrometric hardware to minimize the baseline [68], but some broad signals also come from the sample itself.

Raman spectroscopic signals are also known for background signals. These background signals come from either non-source induced emission processes (e.g. black body radiation, room light, etc.) or fluorescence from sources other than the analyte (e.g. solvent, optics, matrix) [69]. Raman spectroscopy has been extensively applied in a variety of research areas in recent years; however background is a major obstacle in the successful implementation. The non-source emissions can be eliminated by instrumental optimization, but the sample fluorescence is almost unavoidable. It is often observed as a broad band background signal that is superimposed on the signal of analyte. This background signal can be an order of magnitude or greater than the Raman signals and consequently dominates the spectrum, resulting in very small Raman peaks in the spectrum. Therefore, the subtraction of background is essential to extract reliable analytical information from the Raman signals.

In addition to spectroscopy, other widely used analytical techniques also exhibit signals characterized by broad baselines. In chromatography, baseline variations can be caused by changes in the mobile phase composition that affect properties of the eluent, such as refractive index or absorptivity (optical detection in liquid chromatography (LC)), electrical properties (conductivity detection ion chromatography) or thermal conductivity in gas chromatography (GC). Other sources, such as pump noise or column bleed can also play a role.

In mass spectrometry, a variety of factors can lead to baseline variations, depending on the ion source (e.g. matrix-assisted laser desorption ionization (MALDI), electrospray ionization (ESI)) and type of mass spectrometer. For example the chemical noise (referred as baseline in the mass spectroscopic literature [70]) or the unwanted interferences due to matrix or chemical impurities in the sample, is totally dependent on the type of ionization source (e.g., MALDI, ESI). Typically matrices are a low molecular weight organic acid that is mixed in large molar excess compared to the sample (protein or peptide). The primary role of the matrix material is to absorb laser energy and transfer the energy to the sample to ionize the sample molecules without fragmentation. During this process, matrix material vaporises and the gaseous ions are sent to the electric field for ion separation. The interaction of matrix materials with the sample gradually decreases its contribution due to molecular ionization and consumption of matrix material, giving relatively higher baseline contribution at low m/z ratio regions. Consequently, baseline is a mass-to-charge dependent offset on which information-bearing spectral components are superimposed, complicating the quantitation and also affecting peak detection and sample-to-sample comparison [62, 65]. When mass spectrometry is used as a detection method for chromatography, baseline variations can arise in both dimensions.

Finally, other techniques where baseline can be a problem include X-ray fluorescence, activation analysis and various electrochemical methods, but these will not be discussed in detail here.

1.4 Baseline Removal Methods

The literature on baseline estimation and elimination is widely dispersed among many fields of research including analytical chemistry and chemometrics [see for example 22, 71, 72], nuclear physics [73], X-ray spectroscopy [74, 75], NMR [8-10, 16], Raman Spectroscopy [38, 50], and bioinformatics [13, 76, 77]. Early publications on baseline correction were mainly based on hardware modifications [3, 4]. The first few computer-based approaches were published in late 1960's for automated baseline removal [19, 76]. Later, the development and use of computation and automated approaches were accelerated through 1970's and 1980's due to the increasing availability of computers [45, 67, 80-85]. During the last few decades, due to extensive advances in the field of analytical instrumentation and quantitative research, many baseline estimation approaches have been published [6 -101]. These approaches range from manual baseline methods to semi-automated and automated approaches. To avoid an extensive detailed discussion of individual approaches published in the literature, brief descriptions of some commonly used general approaches will be presented within this section, along their advantages and drawbacks. The most common methods are based on the following approaches: the adaptive Kalman filter [87-89], window-based approaches [10-22], smoothing and derivative methods [23-31], polynomial or spline baseline modeling [45-52,92-101], frequency domain filtering using Fourier or wavelet transforms [53-66] and asymmetric least squares methods [32-44].

1.4.1 Notations

In this section and the remainder of the thesis, various equations are presented to represent baselines and their manipulation through models. Because analyte signals and baselines are represented as vectors, equations often involve vectors and matrices as well as scalars. Unless otherwise stated, the standard notations will be used as follows: Vectors are represented by bold lower case letters and matrices by bold uppercase letters. Scalars will be given in italics (upper or lower case). The transpose of vectors or matrices is indicated with a superscript ' T ' and generally, unless otherwise indicated, column vectors can be assumed. Matrix inverses are indicated with a ' -1 ' superscript.

1.4.2 Adaptive Kalman Filtering

The application of adaptive Kalman filter to chemistry was proposed by Rutan and Brown [84, 85]. While its more successful application has been in multi-component calibration and curve resolution [86, 87], Gerow and Rutan used this filter for background elimination in 1986 [88]. They used derivatives to reduce the relative magnitude of low frequency deviations in conjunction with the adaptive Kalman filter for modeling. Then, in 1988, they used factor analysis to model the spectral profiles of the fluorescence background and the adaptive Kalman filter to calculate the weighting factors for each of the abstract background components [89]. The adaptive Kalman filter is a modification of Kalman filter, proposed by R. E. Kalman for digital data processing in engineering and orbital mechanics in 1960 [90]. The Kalman filter in its simple form is a recursive, linear least squares filter. In a recursive manner, it processes the data points one at a time and obtains the best estimates of the model parameters (e.g. slope, intercept, mean) using each

new measurement. The general Kalman model can be described by the following equations;

$$\mathbf{x}(k) = \mathbf{F}(k, k-1) \cdot \mathbf{x}(k-1) + \mathbf{w}(k) \qquad \text{(State Model)} \qquad (1.1)$$

$$z(k) = \mathbf{h}^{\mathrm{T}}(k) \cdot \mathbf{x}(k) + v(k) \qquad \text{(Measurement Model)} \qquad (1.2)$$

In the state model, $\mathbf{x}(k)$ is the $(n \times 1)$ state vector at iteration k, which in this case consists of the weighting factors for each of n background components. Here k corresponds to the wavelength channel of the fluorescence spectrum. The $(n \times n)$ matrix $\mathbf{F}$ describes how the state vector is supposed to change at each iteration, but in this case is set to the identity matrix, $\mathbf{I}$, since the baseline estimate is fixed. The $(n \times 1)$ vector $\mathbf{w}(k)$ is the noise in the state vector and in this case is set to zero. The measurement model describes how the state vector $\mathbf{x}(k)$ gives rise to the observed measurement at iteration k, $z(k)$. In this example, $z(k)$ is the observed fluorescence at wavelength channel k and the $(n \times 1)$ vector $\mathbf{h}(k)$ represents the normalized spectrum of each background component at that wavelength. The scalar quantity $v(k)$ represents the measurement error in the fluorescence measurement at wavelength channel k.

Based on these equations and additional definitions for the system noise covariance ($\mathbf{Q}(k) = 0$ in this case) and measurement variance ($\mathbf{R}(k)$), the algorithm describes how the estimates of the background components $(\mathbf{x})$ are updated at each iteration. This algorithm is described in the literature [30] and will not be reproduced here. In essence, the Kalman filter fits the background or baseline profiles to each spectrum, but the "adaptive" algorithm allows the filter to be turned off when analyte

components are present by inflating estimates of measurement noise in those regions, effectively performing a weighted least squares fit. This permits the background to be estimated in the presence of analyte peaks. This approach has strong parallels to the asymmetric least squares approach discussed later.

The major drawback of the adaptive Kalman filter is that it requires an accurate set of basis functions to consistently describe the background profiles, which means that representative background profiles (spectra) must available and valid for analyte spectra. In addition, the recursive implementation is cumbersome and it has largely been displaced by other methods.

1.4.3 Window-based Approaches

Commonly used window-based approaches detect noise regions and construct the baseline curve by interpolating identified noise regions directly or by using curve fitting (e.g., polynomial, Gaussian, cubic spline). In these approaches, median filtering [10-12], iterative thresholding [13-20] or statistical entropy [21, 22] methods are commonly employed.

In 1995, Friedrichs used median filtering of signals in moving windows of 32 points each for baseline approximation and then applied a Gaussian smoothing function to remove any discontinuities [10]. Kourkoumelis used a Fourier transform to reduce the high frequency components in the signal and then identified each convex set as a window to find the median and locate baseline points [13]. Coombes [15] and Golotvin [16] used specific window sizes to identify the noise points by comparing the intensity range of small neighborhood with some defined threshold and eliminate peak regions to find the

baseline regions. Krishnan [21] and Phillips [22] used the maximum entropy approach for LC-MS and FT-IR spectra to locate the baseline points. In the former case, the baseline of the total ion chromatogram was reduced by including only mass channels with low entropy (less than a threshold ξ), thereby removing channels consisting only of noise. In the FT-IR application, signal, baseline and noise components were separated by maximizing the entropy of the noise and employing information about signal line shape.

All of these window-based methods rely heavily on robust noise region identification and therefore may not achieve optimal baseline correction in many cases. Threshold approaches are based on the detection of noise points by comparing the intensity range of a small neighborhood with the standard deviation of noise regions or comparing the statistical entropy parameter. It has been observed that these methods occasionally identify the low signal points in some cases as noise [13]. Noise standard deviation estimation is also theoretically biased to be smaller than the true value in a statistical view, and leads to additional inaccuracy in detection of noise data points [13]. Another drawback of these methods is that they are often limited to specific signal types (e.g. MS, NMR) and data acquisition parameters (e.g. sampling interval) and therefore may not be widely applicable.

1.4.4 Smoothing and Derivative Methods

Derivatives are also widely used to remove baseline artifacts from the analytical signals [23-25]. The simplest form of a first derivative calculation is where each sample point is subtracted from its neighboring point. When performed on the entire signal, the first derivative removes the signal which is same between the two points (variables),

which in turn attenuates the low frequency baseline. In principle, the first-derivative should completely remove a constant baseline, whereas the second derivative should remove linear baseline components. Since derivatives de-emphasize low frequency baseline and emphasize high frequency noise, Savitzky-Golay derivative filters are often used to simultaneously smooth the data. Thus, a properly designed derivative filter represents a balance between low and high frequency attenuation, requiring adjustment of filter parameters optimum for a particular signal. However, the biggest drawback of signal differentiation is the change in signal shape, which may be undesirable. For example, the first derivative of a signal peak will give positive and negative peaks bracketing the original, which may have implications for interpolation, integration, or other operations. Therefore, derivative methods can be a good choice for peak identification or data smoothing and noise removal, but for baseline correction it changes the peak shape at both ends and makes the signal interpretation difficult for quantitation, since derivatives introduce some features at the edges of peaks.

Savitzky-Golay digital filters [91], or polynomial least squares filters are widely known as a noise filters and are considered as standard for data smoothing [26]. The general characteristic of smoothing filters is that the lower frequencies, generally associated with chemical signals, are passed after filtration, while the high frequency components uniquely associated with the noise are attenuated. The design of Savitzky-Golay filter requires the selection of the widow size, the order of polynomial and the order of derivative. The simplest Savitzky-Golay filter is the moving average filter (zero-order, zeroth derivative) which averages points in a symmetric window around the filtered points. In baseline removal applications, smoothing filters such as this are

applied with wide windows to remove both signals and noise with frequency components higher than that of the baseline. However, the filtered signal is only a starting point for baseline estimation and further refinement is generally necessary, making such methods more complex and signal dependent. For example Schulze et al [27, 28] used small and large window zero-order Savitzky-Golay filters in conjunction with peak stripping to attenuate the high frequency noise and the signal of interest and isolate the low frequency baseline. The selection of window size and the order of polynomial are highly dependent on signal-to-noise ratio and shape of the baseline. Other moving average approaches fall into in the same category, since the average of past data points [29], neighboring data points [30], or only two alternate points [31] are taken. These approaches are usually based on the moving average smoother or linear interpolation of a moving average of a certain window size.

1.4.5 Polynomial or Spline Baseline Modeling

Polynomial fitting [45-52] and spline smoothing [92-101] are essential parts of the literature in the baseline estimation area. The simplest version of polynomial fitting employs a set of basis functions in the ordinal variables: $x^0, x^1, x^2, x^3, ..., x^n$. For a signal vector y , the model for polynomial fitting can be written as,

$$y(x) = a_n x^n + a_{n-1} x^{n-1} .... + a_1 x + a_0 + e \qquad (1.3)$$

where a represents the coefficients to be determined, n is the order or degree of polynomial, and e represents the measurement error.

A spline is a smooth polynomial function which is piecewise-defined and holds a high degree of smoothness at the connection of two polynomial pieces. The splines are generated on the basis of a sequence of chosen points called nodes. Usually a third-order polynomial (cubic spline), a third-order Hermite polynomial (piecewise cubic spline) or a 2.5 degree polynomial (spline 2.5) are used for modeling baselines. Splines fulfill the four main criteria:

1. Neighboring polynomials should have joint nodes, $f_i(x_i) = f_{i+1}(x_i)$
2. Neighboring polynomials must have same slope at joint nodes (continuous first derivative), $f_i^{'}(x_i) = f_{i+1}^{'}(x_i)$
3. Neighboring polynomials must have same curvature at their joining points (continuous second derivative), $f_i^{''}(x_i) = f_{i+1}^{''}(x_i)$
4. The start of the first and end of the last polynomial of a curve do not have joint points, so the slope is assumed to be zero. $f_1^{'}(x_i) = f_{n-1}^{'}(x_i) = 0$

Many different approaches have been proposed using polynomial fitting or spline smoothing; ranging from low-order [48, 49, 101] and multiple polynomial fitting [50, 51] to linear [93-96] and cubic [97-99] spline smoothing in the time or frequency domain [100]. Polynomials and splines have been used as a baseline estimation tool through linear interpolation [60, 68, 93-96], least squares fitting [45, 92, 101] and iterative threshold based fitting [50]. Usually, baseline point identification is done prior to the function fitting. Many different approaches are used to identify baseline reference points, including manual point selection [94, 99], and derivatives for baseline identification [46, 101]. Manual fitting is not as effective and fast, since it is totally dependent on user's

experience and understanding of the data. For automated approaches, the performance declines for lower signal-to-noise ratio and signal-to-background ratio environments [77]. Polynomial and spline fitting can provide fairly good baseline estimation for specific analytical signals, but cannot be used as a generalized approach for multiple analytical techniques. Moreover, an appropriate choice of the degree of the polynomial is quite crucial to reliable baseline estimation.

1.4.6 Frequency Domain Filtering using Fourier or Wavelet Transforms

Fourier and wavelet transforms are two powerful tools in signal analysis in different areas of study. Many approaches exist in the literature for baseline correction based on these methods [9, 53-57], and these will be reviewed briefly here. Baseline correction based on the discrete Fourier transform first represents the signal using sine and cosine functions, transforming the signal into the frequency domain [102].

For a discretely and uniformly sampled signal, $y(t)$, with N points, the Fourier decomposition can be represented in a number of equivalent ways, as shown below.

$$\begin{aligned}
\mathbf{y}(t) &= \sum_{n=0}^{N-1} D_n e^{-i2\pi n f_s t/N} \\
\mathbf{y}(t) &= \sum_{n=0}^{N/2}\left[A_n \cos\left(\frac{2\pi n f_s t}{N}\right) + B_n \sin\left(\frac{2\pi n f_s t}{N}\right)\right] \\
\mathbf{y}(t) &= \sum_{n=0}^{N/2}\left[C_n \cos\left(\frac{2\pi n f_s t}{N} + \varphi_n\right)\right]
\end{aligned} \tag{1.4}$$

In these equations, A_n, B_n, C_n and D_n represent coefficients in the Fourier series. Focusing on the last equation, the Fourier transform represents the amplitude (C_n) and phase (φ_n) as a function of the frequency of the sinusoid (nf_s/N). Although the independent

variable, t, traditionally represents time (giving rise to the "frequency" interpretation) any ordinal variable also works. The quantity f_s is the sampling frequency, equal to $1/\Delta t$, where Δt is the interval between measurements.

Baseline removal with the Fourier transform (FT) works in a manner similar to derivative filtering. Low frequency coefficients associated with the baseline are removed or adjusted and the inverse FT is performed. High frequency (noise) components can also be removed. In practice, this approach can lead to large distortions (like wide-window smoothing) and is usually only used for signals originating in the Fourier domain [9].

The discrete wavelet transform (DWT) is based on principles similar to the FT, but has been more widely applied for baseline removal. The decomposition with the DWT employs a special function called a mother wavelet ψ, the most popular and useful forms being the Daubechies, Coifelt and Symmlet families of wavelets [63]. The transformation decomposes the signal into two new vectors at each level, referred to as the approximation and the detail. If the original signal has N elements, the approximation vector at the first level contains $N/2$ low frequency coefficients (the result of a low pass filter), whereas the detail vector contains $N/2$ coefficients corresponding to the high frequency components (the result of a high pass filter). The wavelet transform allows multilevel decomposition by transformation of every level of the approximation coefficients into a new lower level of detail and approximation coefficients. An illustration of wavelet decomposition is presented in Figure 1.2.

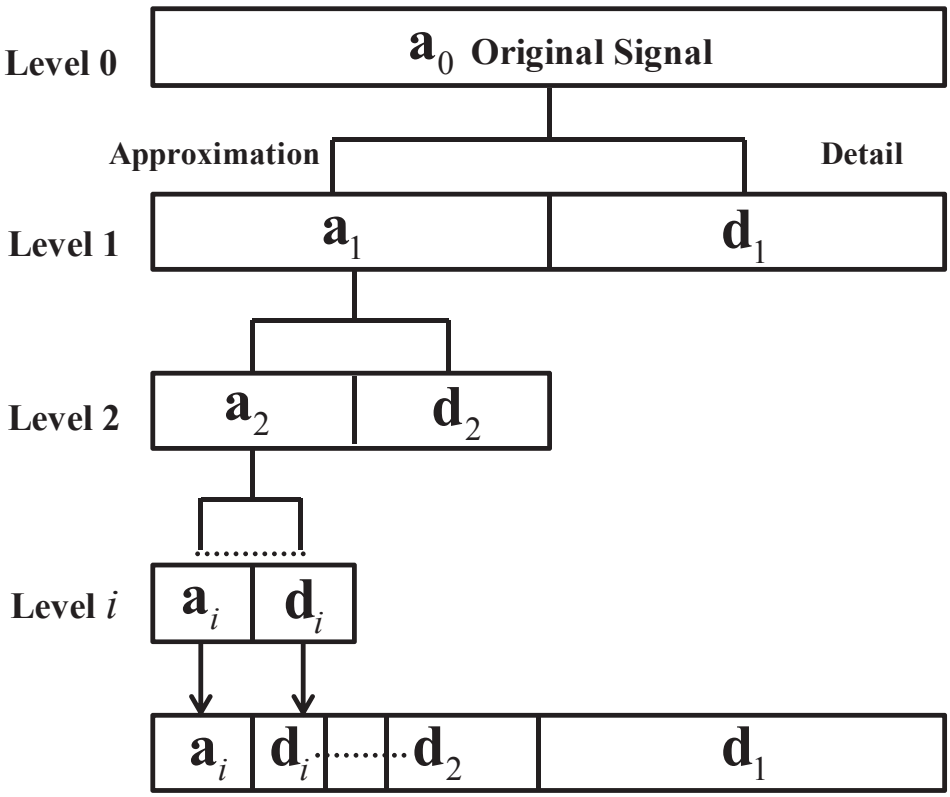

Figure 1.2 Schematic illustration of wavelet decomposition to obtain detail and approximation coefficients at each resolution level.

There are several variants of this basic approach, such as the wavelet packet transform (WPT) and adaptive wavelet transform (AWT). When used for noise removal, the assumption is that the noise dominates in the detail (high frequency) components and these are compared to some threshold and removed (hard thresholding) or adjusted (self thresholding) before applying the inverse wavelet transform (IWT) to regenerate the signal. For baseline removal several strategies can be employed. One is to remove the low level approximation (low frequency) vector associated with the baseline [55-59] before applying the IWT, although this may result in distortion. Another approach is to use the wavelet decomposition to identify peaks or baseline regions and then apply other methods (polynomial, spline, etc.) to estimate the baseline [38, 61-63].

While wavelet transforms have been shown to be useful for baseline removal, their largest drawback is the large number of parameters that have to optimized for a given system being studied. These include the mother wavelet to be used, the number of levels of approximation and the type of thresholding. These methods are sensitive to the baseline and the noise, and the requirement for optimization is an impediment to general applications.

1.4.7 Asymmetric Least Squares Methods

In recent years, baseline removal by methods based on asymmetric least squares (ALS) [32-42] and other closely related approaches [43, 44] have become popular, as evidenced by the large number of citations in a relatively short period of time. Such methods are also most closely related to the technique described in this work so, while a brief description is given here, more details are presented in Chapter 2.

ALS estimation was first introduced in 1987 by Newey and Powell in Econometrics [103], and then Boelens et al used the same approach for baseline estimation in chromatography [32]. The basic principle behind ALS is to carry out a weighted least squares fit of experimental data, y, to a function intended to model the baseline, where the baseline estimates are given by z. The weighted sum of squared residuals is given by

$$S = \sum_i w_i \left(y_i - z_i \right)^2 \tag{1.6}$$

To permit the function z to model the baseline but not the positive peaks, the weights are assigned to largely ignore positive deviations from the model by assigning them according to the following equation.

$$w_i = \begin{cases} p & \text{if } y_i > z_i \\ 1-p & \text{if } y_i \le z_i \end{cases} \tag{1.7}$$

Here p is an asymmetry parameter between 0 and 1, such that $p = 0.5$ is equivalent to ordinary least squares and $p < 0.5$ will decrease the importance of positive residuals. The model parameters for the baseline, with new weights assigned at each iteration are determined iteratively. The baseline function, z , can be any appropriate function, but in the initial application [32], a linear combination of background vectors was used.

One problem with this original method was the requirement for an explicit baseline model. To solve this, Eilers later combined asymmetric least squares with a Whittaker smoother [33], originally introduced by Whittaker [104] in 1922 and also described in the chemical literature by Liang et al in 1999 [105]. The combination of the Whittaker smoother with asymmetric least squares, also called penalized least squares (PLS), was first described in the appendix of a paper on chromatographic time warping in 2004 [106]. A more complete description is in a draft manuscript that is available online [33] but, interestingly, this has never been published. Later in this thesis, this specific baseline correction method, [33] will be termed as asymmetric least squares smoothing (ALSS) instead of the generalized term PLS.

PLS attempts to achieve a balanced combination of two conflicting goals: fidelity to the data (remaining close to the baseline) and smoothness. The approximation result requires optimal choice of the 'order' of smoothing difference, q , penalizing factor, λ and the asymmetric weight, p . It minimizes the penalized least squares function, which is the sum of weighted squared residuals and a penalized sum of differentials.

$$S = \sum_i w_i \left(y_i - z_i \right)^2 + \lambda \sum_i \left(\Delta^q z_i \right)^2 \tag{1.8}$$

In this equation y represents the data, w the asymmetric weight (previously described), and z is the fitted baseline, which is penalized with the roughness penalty λ (positive).

The positive and negative residuals get different weights. The smaller weights (near zero) for positive residuals (peak regions) and larger weights for negative residuals (nonpeak regions) will tend to follow the baseline. Therefore, starting with all weights equal to 1 and iteratively changing the asymmetric weights with appropriate values of p and λ will converge to the baseline estimate.

The penalty factor, λ, basically is a balancing factor between the fit and smoothness. The larger the value of λ, the greater the contribution of smoothing in the minimization function and the smoother the baseline approximation will be. For very small values of λ, the minimization function relies more on the regression function. An extremely high value of lambda leads to a least squares fit to a polynomial of $(q-1)$ degree. Therefore, the choice of the optimum value of λ depends on the signal-to-noise ratio and becomes crucial. Similarly, the choice of difference order, q, (or differential) also depends on the signal-to-noise ratio.

Since its introduction, PLS and variants of it have been employed in a variety of applications for baseline estimation and removal [32-40]. Recently, a mixture model has been proposed [41, 42], where the baseline is modeled as a smooth curve using 'n' cubic B-splines and a discrete roughness penalty.

Asymmetric least squares based approaches either require baseline representative signals or functions for baseline modeling, or the adjustment of parameters that are dependent on the characteristics of the analytical signal. There is a need for a baseline approximation approach which would neither require representative baseline data (signals) nor extensive parameter tweaking dependent on signal characteristics.

1.5 Thesis Outline

From the preceding sections of this chapter, it is apparent that many approaches have been developed and used to minimize or remove the slowly varying baseline artifacts from analytical signals. The objective of this work was to develop an algorithm to remove baseline artifacts from a wide variety of analytical signals in an automated manner without the need for representative baseline signals or human intervention during baseline estimation, and requiring minimal parameter optimization. The proposed method consists of two steps. First, basis functions are calculated using a truncated Fourier series along with some augmentations. In the second step, asymmetric weighted least squares is used with this basis set to estimate the baseline and reconstruct the baseline removed signal. The study will cover the development, optimization and comparison of proposed algorithm with other commonly used approaches.

The thesis is divided into five chapters. Chapter 1 has provided an introduction to the problem of baselines in analytical signals, the sources of baseline components, and a review of current baseline estimation methods, outlining the general approaches and their advantages and drawbacks. The fundamental purpose of this chapter has been to motivate the work and put the objectives into perspective.

Chapter 2 describes the theoretical principles at each step in the development of proposed method. The choice of the optimal basis set will also be demonstrated with the help of simulated data sets. Finally, the Truncated Fourier Asymmetric Least Squares (TFALS) algorithm is presented. An overview of the TFALS algorithm is presented in

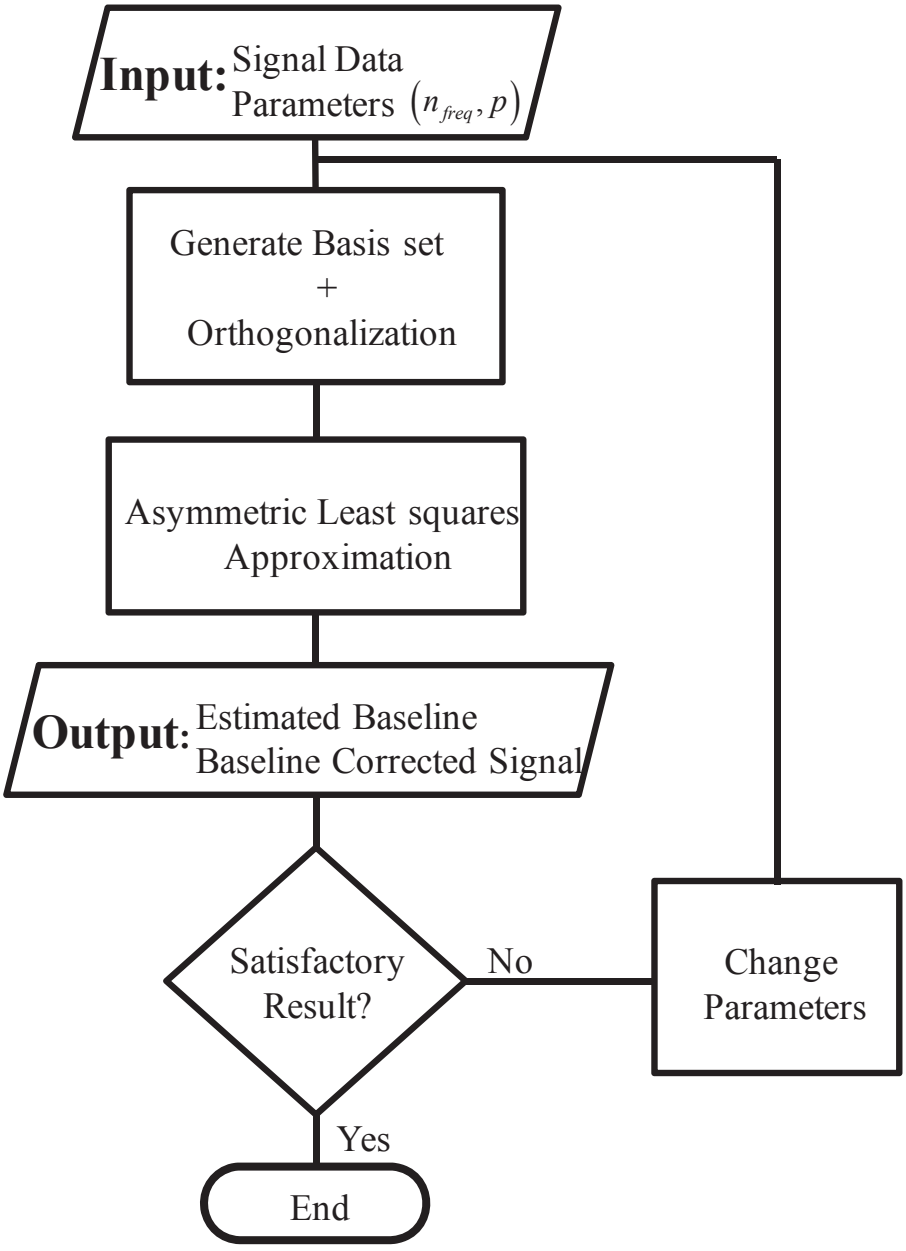

Figure 1.3 Visual representation of the proposed method for the research in this thesis.

Figure 1.3. TFALS requires only two adjustable parameters to be specified in the baseline estimation and removal procedure: the number of frequencies, n_{freq} and asymmetric weighting parameter, p .

In Chapter 3, a comprehensive study is carried out to investigate the dependency of the baseline estimation on these adjustable parameters. It is demonstrated here that choice of number of basis functions influences the shape or frequency of the baseline estimated, a characteristic that depends on the analytical technique, type of sample or analytical conditions. On the other hand, the optimal value of the asymmetric weight is not crucially dependent on analyte signal, but is more of a function of signal noise.

Chapter 4 will provide a detailed description of the results of applying the TFALS algorithm. This chapter consists of two parts. The first part focuses on the performance assessment of TFALS compared to two other approaches using simulation studies, whereas second part demonstrates the successful application of TFALS to experimental data sets in a variety of analytical signals, and a visual comparison of TFALS with a method from the literature for few experimental data signals will also be presented. Finally, in Chapter 5, conclusions are drawn and some recommendations for future work are provided.

CHAPTER 2
Baseline Correction Algorithm

2.1 Introduction

It has been noted that almost all of the approaches in the literature have obtained satisfactory results for baseline removal within specific limitations and drawbacks, as documented in Chapter 1. The approaches employed for baseline correction typically require representative signals for the baseline and/or the intervention of the analyst to select baseline regions or to adjust parameters (e.g. fit parameters, wavelet basis). It is often not possible to obtain representative baseline signals, as in the case of biological samples, where the matrix cannot be separated from the analyte. Even when such signals can be obtained, their acquisition represents an added inconvenience for methods that require them. Likewise the selection of baseline regions and tuning of parameters is a time consuming process and requires the expertise of the analyst to achieve satisfactory results. The objective of this work was to develop and study a novel approach that avoids or minimizes these requirements for baseline correction. This chapter will focus on the theoretical development of a proposed approach for baseline estimation of analytical signals.

The strategy employed in this work is based on the principles of asymmetric least squares, which has become quite popular in recent years for effective baseline removal. A drawback of this method is the need for representative baseline signals or mathematical functions to act as basis functions. Eilers [33] developed a method based on the

Whittaker smoother to avoid explicit basis functions, but this requires the adjustment of three parameters and may not be ideal for all situations, especially at signal edges. The use of an explicit functional form for the baseline has an advantage of having well-understood constraints, and later Eilers employed a combination of B-spline as basis function [41, 42]. However, a natural choice for the baseline basis functions are the sinusoidal components of a truncated Fourier series, since the baseline naturally contains the lowest frequency signal components. This was the approach used in the method developed here, although it was not without its challenges, as described in the sections that follows.

This chapter is divided into six remaining sections. The first of these focuses on the theory of the continuous and discrete Fourier transform (DFT), with a description of some of the limitations of the DFT and ways to minimize those problems. Section 2.3 covers a description and exploration of usage of Fourier series (as a basis set) as an option for the baseline estimation. Section 2.4 describes the linear transformation used to orthogonalize the non-orthonormal basis set. Section 2.5 provides a simulation study to compare and optimize the Fourier-based basis sets, with and without augmentation, to approximate the slowly varying baseline and Section 2.6 focuses on the asymmetric least squares regression approach to estimate the best basis set fit to the baseline corrupted signal. Finally, a summary is provided in Section 2.7.

2.2 Fourier Transforms

The Fourier transform (FT) is an essential signal processing tool in modern data acquisition and processing. The FT has become an integral part of many analytical

instruments (e.g. FT-IR, FT-MS) where it is necessary to perform domain transformations, and it has also become a standard method for post-acquisition signal processing (e.g. smoothing, deconvolution). The Fourier transform can be applied to both continuous and discrete functions, but the latter dominate in analytical instrumental applications where signals are typically digitally sampled at fixed intervals [107]. Therefore focus of this section will be on the discrete Fourier transform (DFT).

The general definition of the continuous Fourier transform, $\mathcal{F}$, and its inverse, $\mathcal{F}^{-1}$, are as given below,

$$H(f) = \int_{-\infty}^{\infty} h(t) e^{-i2\pi ft} dt = \mathcal{F}\{h(t)\} \tag{2.1}$$

$$h(t) = \int_{-\infty}^{\infty} H(f) e^{i2\pi ft} df = \mathcal{F}^{-1}\{H(f)\} \tag{2.2}$$

Here $H(f)$ represents the Fourier transform and $h(t)$ the signal. The FT operator $\mathcal{F}$ generates the Fourier spectrum from the signal and the inverse FT operator $\mathcal{F}^{-1}$ restores the signal from the spectrum. Typically, the signal is defined as a function of the variable time, t, and the spectrum as a function of the variable frequency, f. However, these designations are arbitrary and the independent variable in the signal domain does not need to be associated with time. The FT and its inverse should be regarded as complementary domain transformations that transform the same information between two alternate representations, arbitrarily referred to as the time (signal) and frequency (spectral) domain. Note that the FT and its inverse are symmetric operators, the only changes being the independent variable and the sign of the exponent.

In practice, the measurement of a signal usually gives us a finite number of data measurements at discrete intervals. Consequently, the FT reproduces the signal by the addition of finite number of sinusoids at defined frequency intervals with variable amplitudes and phases. Therefore, the infinite integral is replaced by a finite sum from $-N/2$ to $N/2-1$, where N is the number of sample points [108]. The Fourier transform calculated in this way is called the discrete Fourier transform (DFT).

The spectrum calculated from the discrete signal sampling through the DFT is actually an approximation of the continuous Fourier transform of the function underlying the data at the sampled points. The spectrum obtained is

$$H(f) = \sum_{j=-N/2}^{N/2-1} h(j\Delta t) e^{-i2\pi fj\Delta t} \tag{2.3}$$

Where $h(j\Delta t)$ represents the signal at $t = j\Delta t$ and Δt is the sampling interval. The signal is arbitrarily assigned time points from $(-N\Delta t/2)$ to $(N/2-1)\Delta t$ for a period of $T = N\Delta t$. $H(f)$ represents the Fourier transform (spectrum) at frequencies of $[-N/2, ..., N/2+1]\cdot\Delta f$, where $\Delta f = 1/T$.

When time domain signal, $h(t)$, contains N points (where N is an even number) the FT (spectral domain) also contain N complex coefficients. Through Euler's relationship,

$$e^{-i\theta} = \cos\theta - i\sin\theta \tag{2.4}$$

These coefficients can be thought of as the coefficients for a series of sinusoids at specific frequencies, varying from zero (sometimes referred to as DC, by analogy with electrical

direct current) to the Nyquist frequency, $f_N = f_s/2$, where f_s is the sampling frequency, $1/\Delta t$. This combination of sinusoids can reproduce the signal points exactly. One way to present the results of the FT is to plot the real and imaginary parts of the coefficients separately as a function of f, referred to as real and imaginary spectra. More commonly, the combination of sines and cosines is transformed into a combination of sines and phase angles, and the FT is plotted as an amplitude and phase spectrum, where the amplitude gives the total contribution at each frequency. This is calculated as

$$amp\left(H\left(f\right)\right) = \sqrt{\left[real\left(H\left(f\right)\right)\right]^2 + \left[imag\left(H\left(f\right)\right)\right]^2} \tag{2.5}$$

Usually, the amplitude spectrum is of more interest than the phase spectrum. To illustrate this, consider the example in Figure 2.1(a), which shows a signal sampled at 32 time points at 1 second intervals (Note that the $t = 0,...,31$ on the time axis, although for the purposes of equation 2.3, it is transformed to $-16,...,15$).

The amplitude spectrum produced by the FT is shown in Figure 2.1(b) and shows contributions at frequencies ranging from $-0.5\,\text{Hz}$ to $0.5\,\text{Hz}$. Because the original time signal is real (not complex), the amplitude spectrum is symmetric for positive and negative frequencies and in most applications only the positive half is shown. The FT produces 32 complex coefficients for frequencies of -0.5 to $+0.46875\,\text{Hz}$. For the figure, an additional point has been included at $f = 0.5\text{Hz}$ by symmetry, so 33 coefficients are actually displayed.

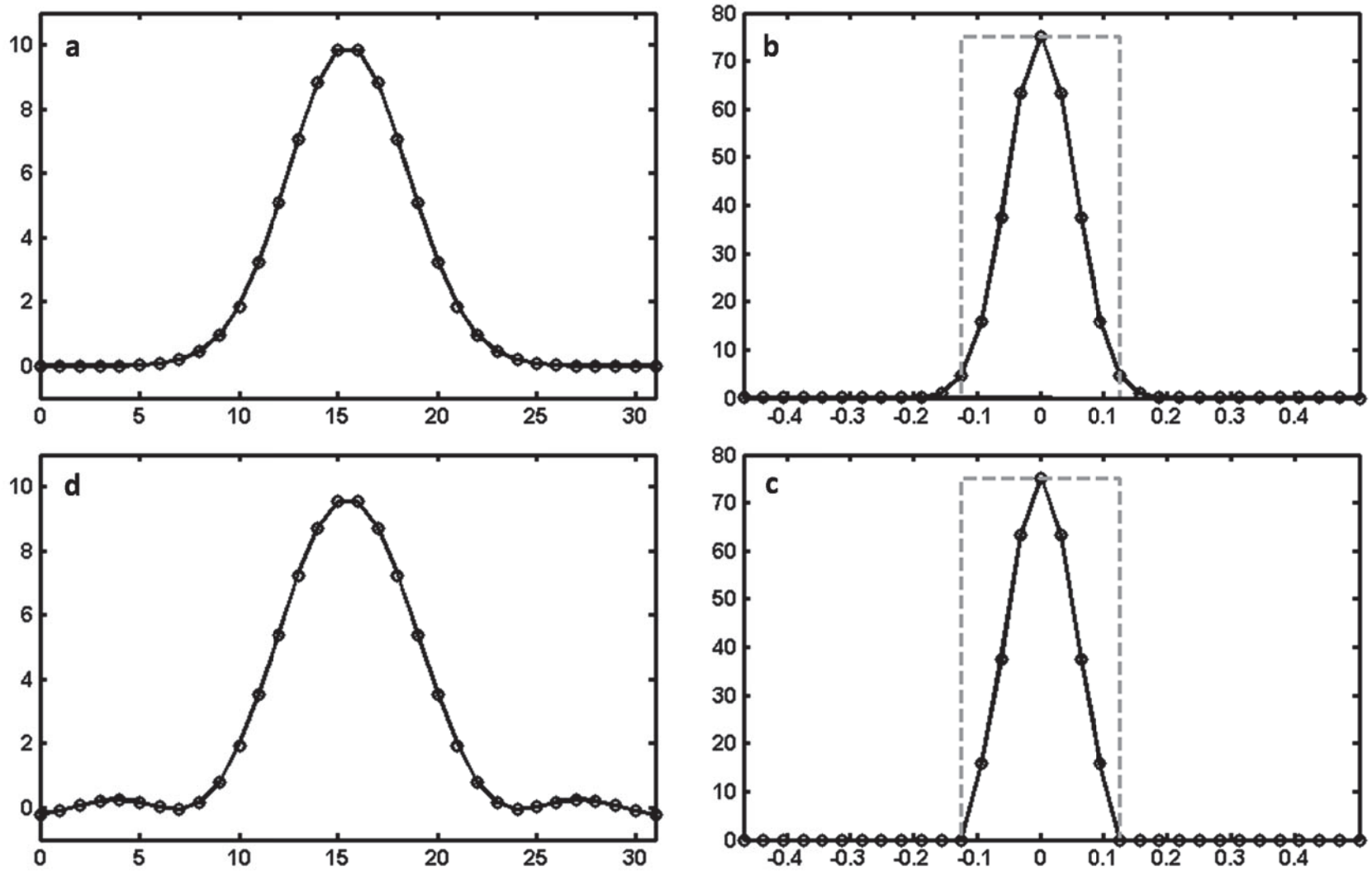

Figure 2.1 **(a)** A Gaussian signal sampled at 32 time points. **(b)** Amplitude spectrum of Gaussian signal and boxcar truncation function. **(c)** Amplitude spectrum of Gaussian signal after boxcar truncation. **(d)** Time domain Gaussian signal after truncation.

Since the DFT is only an approximation of a continuous signal with a finite number of data points, it cannot provide the same amount of information as an infinite signal, which leads to some unavoidable consequences compared to the infinite case. One of these is referred to as aliasing, where frequencies in the continuous signal that are higher than the Nyquist frequency are folded back into lower frequencies within the range of the DFT. Another consequence, that is more important in the current context, results from truncation of an infinite series and is referred as spectral leakage [109]. This is discussed in the next section.

2.2.1 Truncation Effect (Spectral Leakage)

When the DFT is applied for signal smoothing, the strategy is to remove high frequency spectral components and perform the inverse FT to generate the smooth signal. The rational for this is that signal information is mainly in the low frequency components, whereas white noise is distributed throughout the spectrum, so removing high frequency components should improve the signal-to-noise ratio.

One way to do this is to apply boxcar function, $\Pi(f)$, to the real and imaginary parts of the FFT. The boxcar function is a function that is unity at low frequencies and zero at high frequencies, the boundary being defined by a cut-off frequency, f_c .

$$\Pi(f) = \begin{cases} 1, & |f| < f_c \\ 0, & |f| \geq f_c \end{cases} \tag{2.6}$$

Multiplication of the DFT by the boxcar function in the frequency domain is equivalent to convolution of the original signal with the inverse FT of the boxcar in the time domain.

$$
\begin{aligned}
h'(t) &= \mathcal{F}^{-1}\{\Pi(f)\cdot H(f)\} \\
h'(t) &= \mathcal{F}^{-1}\{\Pi(f)\} * \mathcal{F}^{-1}\{H(f)\} \\
h'(t) &= \mathcal{F}^{-1}\{\Pi(f)\} * h(t)
\end{aligned}
\tag{2.7}
$$

Here, $h(t)$ indicates the smoothed signal in the time domain and " $*$ " indicates convolution.

The inverse Fourier transform of the boxcar function is

$$
\mathcal{F}^{-1}\{\Pi(f)\} = \int_{-f_c}^{f_c} e^{i2\pi ft} df \tag{2.8}
$$

$$
\mathcal{F}^{-1}\{\Pi(f)\} = 2f_c \frac{\sin(2\pi f_c t)}{2\pi f_c t} = 2f_c \,\mathrm{sinc}(2\pi f_c t) \tag{2.9}
$$

Hence, the effect of truncation on the Fourier transforms results in the convolution of the true signal with a sinc, or cardinal sine function:

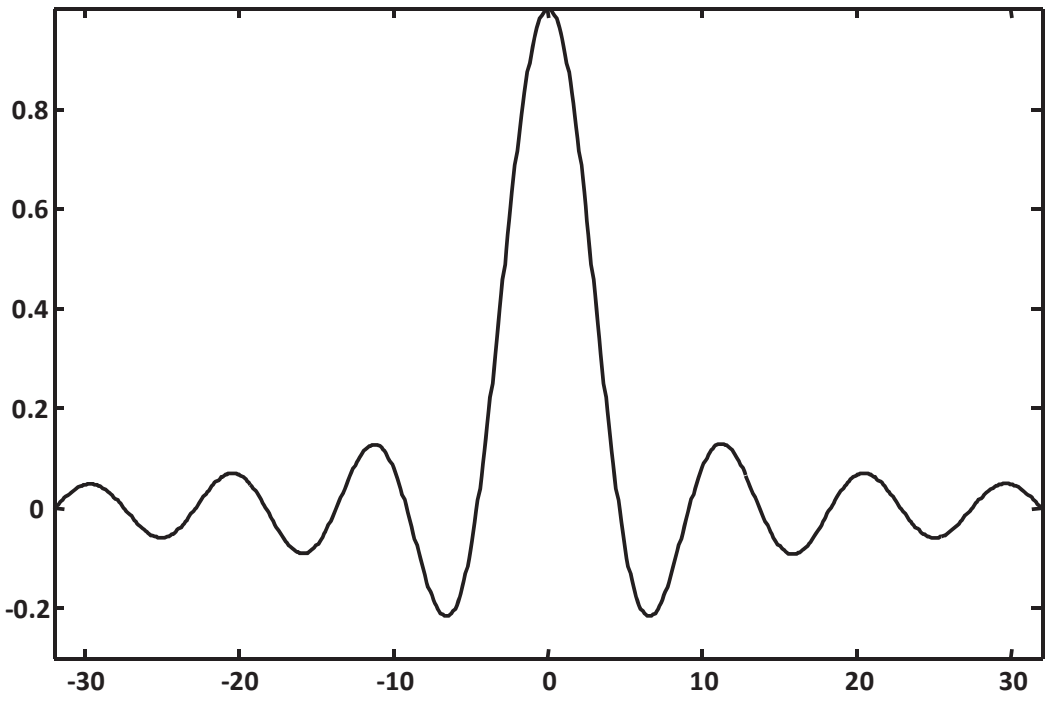

Figure 2.2 An example of the sinc function for $f_c = 7/64$.

$$h'(t) = 2f_c \,\mathrm{sinc}(2\pi f_c t) * h(t) \tag{2.10}$$

The sinc function can be regarded as a damped sinusoid centered on $t = 0$. An example of this function for $f_c = 7/64$ and $t = -32$ to 32 is shown in Figure 2.2.

The effect of this truncation in the frequency domain is a convolution of the time domain signal with the sinc function, resulting in distortions, or "wiggles", in the appearance of the signal [110]. This is illustrated in Figure 2.1, where a boxcar function, indicated by the dashed lines, has been applied to the DFT in Figure 2.1(b), resulting in the truncated DFT in Figure 2.1(c). The inverse FT of this, shown in Figure 2.1(d), clearly shows the distortions introduced by the truncation.

There is an inverse relationship between the width of the truncation function and its inverse FT [110], since time scale expansion corresponds to frequency scale compression

$$h(kt) = \frac{1}{k} H\left(\frac{f}{k}\right) \tag{2.11}$$

Here, k is a real constant greater than zero. Hence, if the truncation function is increased in length then the sinc function approaches an impulse. The more closely the sinc function approaches an impulse, the less ripple effect will be introduced [110]. Conversely, when the cut-off frequency is reduced, the extent of distortions becomes greater. This is relevant when the DFT is used for baseline estimation, since f_c is adjusted even lower to remove signal and high frequency noise, in principle leaving only the baseline. In practice, however, the truncation leads to a severe distortion of the estimated baseline which is not useful (see next section). For smoothing applications,

alternative windows (e.g. Hann, Bartlett or Welch) can be used [111], but these have limited utility in baseline removal.

2.3 Fourier Basis Set for Baseline Estimation

The Fourier series is an expansion of a periodic function using a sum of sines and cosines. It uses the orthogonality relationships of the sine and cosine functions to build an orthogonal basis set. This orthogonal set of functions can be used to approximate a given function which is sampled at discrete intervals. For a signal vector of length N, a Fourier series of finite length can be written as:

$$f(t) = a_o + a_1 \cos\left(\frac{2\pi f_s t}{N}\right) + b_1 \sin\left(\frac{2\pi f_s t}{N}\right) + a_2 \cos\left(\frac{2\pi \cdot 2 \cdot f_s t}{N}\right) + b_2 \sin\left(\frac{2\pi \cdot 2 \cdot f_s t}{N}\right) + ... \\ + a_n \cos\left(\frac{2\pi \cdot (N/2) \cdot f_s t}{N}\right) + b_n \sin\left(\frac{2\pi \cdot (N/2) \cdot f_s t}{N}\right) \tag{2.12}$$

This can be written as in general form as:

$$f(t) = \sum_{n=0}^{N/2} \left[a_n \cos\left(\frac{2\pi n f_s t}{N}\right) + b_n \sin\left(\frac{2\pi n f_s t}{N}\right) \right] \tag{2.13}$$

where f_s is the sampling frequency, N is the number of points sampled, and a_n and b_n are the Fourier coefficients.

The series given above will reproduce the measurements in the signal vector exactly since it is simply another representation of the inverse FT. However, the premise of this work is that the low frequency sinusoidal functions in this series can be used to model the baseline of the signal in conjunction with asymmetric least squares. This is

analogous to the Fourier smoothing described in the previous section, except that the FT is not directly involved, ultimately allowing asymmetric least squares to be employed. The first requirement, however, is that the truncated Fourier series be able to fit the baseline in the absence of signal peaks. Unfortunately, least squares fitting gives rise to the same spectral leakage problem as in Fourier smoothing, as will be illustrated with the following example.

Figure 2.3(a) shows a typical baseline signal (open circles) of 32 time points (a small number was chosen for illustration). Figure 2.3(b) shows the amplitude spectrum of the baseline (positive frequencies only) and Figure 2.3(c) shows the same spectrum after applying a boxcar function with a cut-off frequency of 0.15Hz. The solid black line in Figure 2.3(a) shows the reconstructed baseline after applying the low pass filter and performing the inverse FT. Although the reconstructed signal follows the baseline, the presence of "wiggles" due to spectral leakage is evident, especially at the edges. Based on the discussion in the previous section, this is expected.

Another way to fit the baseline is to fit it using sinusoids with the frequencies as were employed in the DFT. Specifically, with five frequency components the fit equation is

$$\begin{aligned} y_{est} = a_0 &+ a_1 \cos\left(\frac{2\pi f_s t}{N}\right) + b_1 \sin\left(\frac{2\pi f_s t}{N}\right) + a_2 \cos\left(\frac{2\times 2\pi f_s t}{N}\right) \\ &+ b_2 \sin\left(\frac{2\times 2\pi f_s t}{N}\right) + \ldots + a_4 \cos\left(\frac{4\times 2\pi f_s t}{N}\right) + b_4 \sin\left(\frac{4\times 2\pi f_s t}{N}\right) \end{aligned} \tag{2.14}$$

where $f_s = 1\text{Hz},\ N = 32,\ t = 0,\ldots 31\text{s}$ and the a's and b's are the coefficients fit by ordinary least squares. These basis functions are shown in Figure 2.3(d), and the fit is

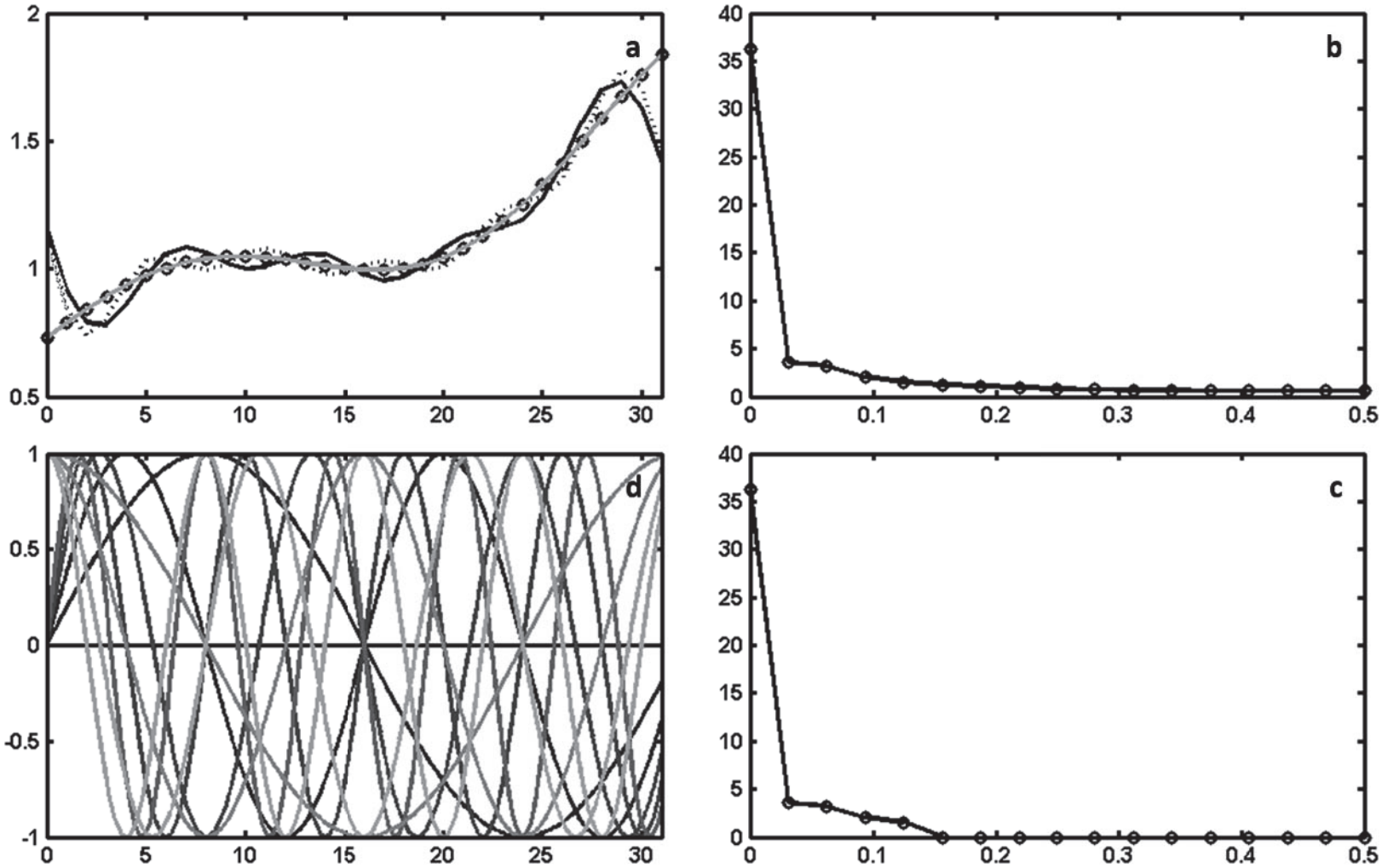

Figure 2.3 Illustration of spectral leakage in baseline fitting using a low pass Fourier filter (▬), Fourier basis (▪▪▪▪) and Fourier basis with augmentations (▬). **(a)** Baseline points (open circle) and various baseline fits **(b)** FT of baseline. **(c)** Truncated FT. **(d)** Fourier basis set.

shown by the dotted black line in Figure 2.3(a). The fit to the baseline is not the same as for Fourier smoothing, since the least squares method uses different criteria, but the two are very similar in their characteristics, including the spectral leakage that continues to be a problem in this approach, and the baseline fit is inadequate for modelling. For the proposed method to work, a better fit to the pure baseline is needed.

One way to improve the fit to the baseline would be to extend the basis set to include higher frequencies, but this would then allow the model to fit higher frequency signal peaks, which is undesirable. Another alternative that was investigated was to extend the basis set to lower frequencies. To test this, two additional frequencies were added to the model used in Figure 2.3, one at one-half the lowest frequency and one at one-quarter of the lowest frequency. Thus the fitted equation is now

$$y_{est} = a_0 + a_{0.25}\cos\left(\frac{2\pi\cdot(0.25)f_s t}{N}\right) + b_{0.25}\sin\left(\frac{2\pi\cdot(0.25)f_s t}{N}\right) + a_{0.5}\cos\left(\frac{2\pi\cdot(0.5)f_s t}{N}\right) + b_{0.5}\sin\left(\frac{2\pi\cdot(0.5)f_s t}{N}\right) + \sum_{n=1}^{4} a_n\cos\left(\frac{2\pi n f_s t}{N}\right) + \sum_{n=1}^{4} b_n\sin\left(\frac{2\pi n f_s t}{N}\right) \quad (2.15)$$

The fit using this extended basis set is shown by the solid gray line in Figure 2.3(d). A very good fit is obtained, indicating that the extension of the truncated Fourier series to lower frequencies is a viable way to model slowly varying baseline.

The challenge at this point is to determine the optimum basis set or optimum frequencies to accurately model the baseline while excluding components of the analytical signals. This involves two questions: (1) How many terms to include in the truncated Fourier series and (2) how to expand the basis set at low frequencies (what frequencies, how many). These questions are investigated in the section that follows.

First, however, some mathematical complications in the implementation of the augmented Fourier series are discussed in the next section.

2.4 Orthonormal Basis and Transformations

The fitting of a model to a baseline signal by ordinary least squares or asymmetric least squares requires the solution of a system of linear equations involving the inversion of an appropriate matrix. If the basis functions evaluated at N time points are given by an $N \times p$ matrix $\mathbf{X}$, the estimate of the $p \times 1$ vector of model coefficients, $\hat{\boldsymbol{\beta}}$, is given by the following equation.

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{y} \tag{2.16}$$

or $$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^{\mathbf{T}}\mathbf{W}\mathbf{y} \tag{2.17}$$

Here, the first equation applies to ordinary least squares and the second to asymmetric least squares, where $\mathbf{y}$ is the $N \times 1$ vector of measurements and $\mathbf{W}$ represents a weighting matrix. Both equations require the term in the parenthesis to be invertible (non-singular) which in turn requires that $N \geq p$ and $\mathbf{X}$ is non-singular.

The truncated Fourier series proposed in the previous section forms an orthogonal basis set that requires only normalization to have orthonormal basis set. A subset $\{\mathbf{v_1}, \mathbf{v_2}, ..., \mathbf{v_r}\}$ of a vector space $\mathbf{V}$ is called orthonormal if $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ when $i \neq j$, that is the vectors are mutually perpendicular, and they are unit length $\mathbf{v}_i \cdot \mathbf{v}_i = 1$ [112]. A simple example of an orthonormal basis is the standard basis $\mathbf{e}_i$ for R^n. The vector $\mathbf{e}_i$ is the vector with all zeros except for 1 at each i_{th} coordinate. A rotation or flip through the

origin transforms an orthonormal set to another orthonormal set. These are the transformations which preserve the inner product and are called orthogonal transformations.

If the matrix of basis functions, $\mathbf{X}$, has orthonormal columns, as it is in the case of the normalized truncated Fourier series, it not only ensures that $\mathbf{X}$ is non-singular, but also simplifies the least squares solution, since $\left(\mathbf{X}^{\mathrm{T}}\mathbf{X}\right)=\mathbf{I}$, the identity matrix. Unfortunately, when the truncated Fourier series is augmented with lower frequency components, the columns of $\mathbf{X}$ are no longer orthonormal. Moreover, in certain cases that depend on the number of terms in the truncated series and the frequencies used in the augmentation, $\mathbf{X}$ becomes singular. This means that the columns of $\mathbf{X}$ are no longer linearly independent and the least squares equations cannot be solved. For instance, in the example presented earlier which used $n=0,1,2,3,4$ augmented by $n=1/4,1/2$, $\mathbf{X}$ is still non-singular with the further augmentation of $n=1/8,3/8$ but then becomes singular with the further addition of $n=1/16$. This singularity means that one or more of the columns of $\mathbf{X}$ is a linear combination of the others to computational precision. Because it is difficult to predict which combinations of truncated and augmented Fourier basis sets will lead to singularities when testing for optimal basis functions, an alternative strategy was sought through principal component analysis (PCA).

2.4.1 PCA for Linear Transformation

Principal component analysis (PCA), also referred to as singular value decomposition (SVD), is a widely used tool in different areas of study. It has been used for dimensionality reduction in multivariate regression and calibration and in exploratory

data analysis, among other applications. The basic idea is simple but it has been re-invented many times and documented in different fields of research over the years [113-119]. The method itself was first introduced in physics by Cauchy [115] in 1829 then formulated by Pearson [116] in 1901 in what is probably the most famous paper. In 1930, Hotelling transformed PCA to its current form [117] in psychometrics. It was popularized in chemistry by Edmund Malinowski [118] as factor analysis; however, it has also been claimed that the earliest non-specific reference to PCA was given by Adcock in the chemical literature in 1878 [119].

Mathematically, PCA is a linear transformation that transforms the data into a new orthogonal coordinate system in such a way that the greatest variance in the data is accounted for by the first coordinate (first principal component) and the next greatest variance by the second coordinate and so on. The PCA basis vectors, referred to as principal components, eigenvectors or loadings, have two essential properties: all components are of unit length and they are all orthogonal to each other. In other words, they form orthonormal basis set. Visual inspection for orthogonality and unit length is feasible for up to three dimensional data sets but for higher dimensions these conditions can be satisfied mathematically. The dot product properties can solve this problem easily. Mathematically, these properties can be described as

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{x}_i &= 1 \\ \mathbf{x}_i \cdot \mathbf{x}_j &= 0 \end{aligned} \tag{2.18}$$

In the current context, PCA is used to solve the problem of singularity in the matrix of basis functions, $\mathbf{X}$. When augmentation leads to such singularity, direct calculation of

the coefficients is not possible. Although singularity implies that there is redundant information in the basis vectors, it is difficult to select individual basis functions for removal to solve the singularity problem. Instead, the basis vectors in $\mathbf{X}$ are subjected to principal components analysis, resulting in a set of orthonormal basis vectors $\mathbf{V}$. From the perspective of baseline fitting, $\mathbf{V}$ contains the same information as $\mathbf{X}$, only it is organized into a set of orthonormal basis vectors. The nature of PCA means that the last columns of $\mathbf{V}$ can be eliminated (without loss of information) until it is no longer rank deficient (singular). For example, if $\mathbf{X}$ has 19 columns but has a mathematical rank of 18 (the largest number of columns that are linearly independent), the removal of the last column of $\mathbf{V}$ will give rise to a set of basis vectors that contain the same information but are non-singular. It is important to note that the original augmented Fourier series is still being fit to the baseline, but this approach circumvents some of the computational problems.

2.5 Truncated Fourier Basis Set Selection

It has been noted in Section 2.2 that the truncation function returns a convolution of a signal with the sinc function, which gives some artifacts in the signal. It has also been noted that these effects can never be removed completely but can be minimized by changing some parameters. These artifacts can also presumably appear in the approximation of a signal or part of a signal, using discrete Fourier series of a finite length as a basis. In approximating the slowly varying baselines of a signal of finite length using truncated Fourier basis sets, a truncation effect appears in the estimated baseline in terms of some artifacts, as noted in Figure 2.3. To minimize the truncation

artifacts and get a better approximation of low frequency baselines, the truncated Fourier basis sets Fourier sets were augmented with additional low frequencies and tested as candidate basis functions for baseline approximation and tested using three different baseline functions (linear, exponential, sinusoidal). Six different groups of augmented basis sets were tested along with the truncated Fourier basis set (TFB). Comparative approximation results for each basis set group are presented and described briefly, followed by the conclusive summary. The optimization of ALS parameters for the selected basis set is presented in Chapter 3.

2.5.1 Experimental

It has been shown that a Fourier basis set (FBS) when truncated to low frequency components, i.e. a truncated Fourier basis (TFB), and augmented with additional low frequency components can provide a good estimation of a slowly varying baseline. The issue that remains is the number of augmented frequencies and their values. In addition to provide a good fit to the baseline using the augmented frequencies with the TFB, it is also desirable for model to have as few terms as is reasonable to adhere to principles of parsimony and simplicity. To this end, seven different sets, or groups, of basis functions were generated, as described below.

The first group of functions (designated as group I) was simply the truncated Fourier basis set (TFB). The basis functions can be defined as

$$y_i(x) = \cos\left(\frac{i+1}{2}.\frac{2\pi(x-1)}{N}\right) \qquad \text{for } i = \text{odd} \tag{2.19}$$

$$y_i(x) = \sin\left(\frac{i}{2}.\frac{2\pi(x-1)}{N}\right) \qquad \text{for } i = \text{even} \tag{2.20}$$

Here i is an integer that ranges from 0 to $2n_{\max}$ and x is an integer that ranges from 1 to N where N is the number of channels. Thus the TFB includes the frequencies 0 (DC), f_1 (the lowest non-zero frequency in the FT, or the sampling frequency), $2f_1$, $3f_1$ etc. up to $n_{\max}f_1$. The value of $n_{\max}$ determines the truncation point and is limited to $N/2$. In this study, $n_{\max}$ was set to the integer nearest to $N/40$ (5% of the Nyquist frequency), which was found to be more than adequate. Thus the number of frequency terms (including DC) for Group I was $n_{freq} = n_{\max} + 1$.

Since Group I was not augmented, it was not expected to perform well, but was included for reference. As indicated in Table 2.1, the second group, designated as Group II (hi) (for "half interval") included frequencies of $1.5f_1$, $2.5f_1$, etc. up to $(n_{\max} - 0.5)f_1$ in addition to the TFB, or in other words, $n = 0,\ 1,\ 1.5,\ 2,\ 2.5,\ldots n_{\max}$. for a given value of n, the corresponding basis functions were

$$\cos\left(\frac{n \times 2\pi(x-1)}{N}\right) \quad \text{and} \quad \sin\left(\frac{n \times 2\pi(x-1)}{N}\right) \tag{2.21}$$

Note that only the cosine term is used for $n = 0$, so the total number of basis functions is

Table 2.1 Basis set groups with their arrangements.

GROUP#	n
I	[0,1,2,... $n_{\max}$]
II(hi)	[0,1,1.5,2,2.5,... $n_{\max}$]
III(h)	[0,0.5,1,2,... $n_{\max}$]
IV(q)	[0,0.25,0.5,1,2,... $n_{\max}$]
V(oe)	[0,0.125,0.25,0.375,0.5,...,1,2,.. $n_{\max}$]
VI(ot)	[0.1,0.2,..1,2,.. $n_{\max}$]
VII(os)	[0,0.0625,0.125,0.1875,0.25,...,1,2,.. $n_{\max}$]

$4n_{\max} - 1$.

Since augmentation for slowly varying baselines is likely to be more effective at frequencies less than f_1, the third group, designated as Group III(h) (for "half") augmented the TFB (Group I) with a single frequency at $f = 0.5f_1$ $(n = 0.5)$, adding two more basis functions. This frequency corresponds to a sinusoid with a half-cycle over the range of the signal.

Additional groups were created as indicate in Table 2.1. Group IV (q) ("quarter") added a frequency of $0.25f_1$ to Group III, and Group V (oe) ("one eighth") added seven additional frequencies between 0 and f_1 at an interval of $f_1/8$, as $f_1/8$, $f_1/4$, $3f_1/8$, $f_1/2$, $5f_1/8$,..., $7f_1/8$ to Group I. Group VI (ot) ("one tenth") include nine additional frequencies at uniform intervals of $0.1f_1$ between 0 and f_1. Finally, Group VII (os) ("one sixteenth") added fifteen low frequencies at an interval of $f_1/16$ between 0 and f_1 as $f_1/16$, $f_1/8$, $3f_1/16$,..., $15/f_1$ to the TFB.

Although other frequency combinations could be incorporated, these groups were considered to give a broad representation of the types of combinations that could be useful for modeling slowly varying baselines.

To evaluate the performance of all seven basis set groups for baseline approximation, each group was tested individually for individual simulated data set and the root-mean-squared errors (RMSE) of estimation were also calculated and tabulated in Section 2.5.2 for comparison.

2.5.1.1. Computational Aspects

All data processing was carried out using programs written by the author in Matlab®2010b (MathWorks, Natick, MA) under Windows 7 Professional 2009© on a 2.10 GHz processor with 2.00 Gb of memory.

2.5.1.2. Data Simulations

Three simulated data sets were used to evaluate the baseline estimation ability of each group of basis sets. These data sets were intended to determine the fidelity of the baseline estimation using each basis set group individually and compare the estimation errors to obtain the best basis set group for baseline approximation. Neither peaks nor noise was added to these data signals since the only purpose of this study was to compare the approximations of individual baseline functions.

Data Set 1(a) consisted of a vector of 2000 points and Data Set 1(b) consisted of a vector of 200 points. A linear function with a slope of 0.05 for Data Set 1(a) and 0.5 for Data Set 1(b) and intercept of zero was used as the baseline function.

$$\mathbf{y}_i = m\mathbf{x} \tag{2.22}$$

Here m is the slope and x_i ranges from 1 to 2000 or 1 to 200 in steps of unity. The value of slope was chosen to give maximum amplitude of 100.

Data Set 2(a) consisted of a vector of 2000 points and Data Set 2(b) consisted of a vector of 200 points. Ten exponential functions were generated as a baseline with ten randomly generated with decay rates between 9×10^{-4} and 0.05 using

$$\mathbf{y}_2 = 100e^{-k\mathbf{x}} \tag{2.23}$$

Here k determines the decay rate of exponential function, corresponding to half-lives between about 14 and 770 channels. The baselines generated in this manner for Data Set 2(a) are shown in Figure 2.4a. For Data Set 2(b) the baselines were truncated at channel length 200.

Data Set 3 also consisted of a vector of 2000 points and a vector of 200 points. Ten sinusoidal signals were used to simulate the baselines using random angular frequencies between 4×10^{-5} and 0.05 random phase angles (between $-\pi$ and $+\pi$), employed in equation 2.24.

$$\mathbf{y}_i = 100.\sin(\omega t + \varphi) \tag{2.24}$$

The sinusoidal functions had maximum amplitude of 100 and for Data Set 3(a) (2000 points) correspond to about 0.13 to 1.6 cycles, while for Data Set 3(b) (200 points) correspond to a range of about 0.013 to 0.16 cycles over the range of the signal. The baseline functions generated are shown in Figure 2.4b. For Data Set 3(b), the signals were truncated at 200 points.

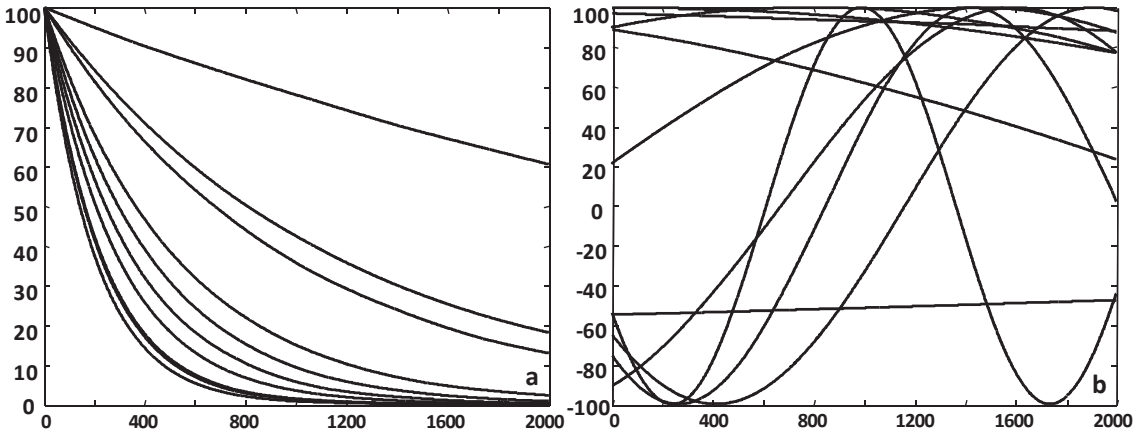

Figure 2.4 (a) exponential baseline functions for Data Set 2a and **(b)** sinusoidal functions baseline functions for data set 3a.

2.5.2 Results and Discussion

All three sets of baselines were tested with the seven different groups of designed basis functions. The groups of basis functions described in Section 2.5.1 were truncated to the first 5 frequencies ($n_{\max} = 5$) for data sets having 200 channels and 50 frequencies ($n_{\max} = 50$) for data sets with 2000 channels. Estimated and actual baselines were then compared using the root-mean-squared error (RMSE) for each simulated baseline vector, calculated as given below.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(b_i^o - \hat{b}_i\right)^2}{N}} \tag{2.25}$$

Here $\mathbf{b}_i^o$ is the actual baseline, $\hat{\mathbf{b}}_i$ is the estimated baseline at that channel, and N is the number of channels. The RMSE of estimation is tabulated for each simulated baseline vector using all seven groups of basis functions for comparison.

Table 2.2(a) shows the RMSE values with each basis set group for the linear baseline at a channel length of 2000 (2000 sample points) and Table 2.2 (b) presents the results for a channel length of 200. It can be clearly deduced from Table 2.2 that the truncated Fourier basis (TFB, Group I) gave a very high residual for both channel lengths, which means that this set is unable to give a good approximation for linear baselines. Group II (one extra set of sine and cosine before every truncated Fourier set) gave a relatively lower RMSE values which in turn seemed to be a better choice for linear baseline approximation. Group III (only one extra pair of sine and cosine terms at half intervals in TFB) gave relatively lower RMSE values than the TFB (Group I) for approximation of linear functions but seems not to be an appropriate choice, especially

for smaller channel lengths. On the other hand, group IV, V , VI and VII gave very small RMSE values, which means that these groups provide a very good approximation for linear functions, even for smaller channel lengths.

Table 2.2(a) Comparison of basis set groups results with linear baseline, channel length 2000.

Slope	I	II (hi)	III (h)	IV (q)	V (oe)	VI (ot)	VII (os)
5.00E-02	3.16E+00	7.54E-14	9.21E-05	7.39E-14	1.11E-13	5.48E-14	7.81E-14

Table 2.2(b) Comparison of basis set groups results with linear baseline, channel length 200.

Slope	I	II (hi)	III (h)	IV (q)	V (oe)	VI (ot)	VII (os)
5.00E-01	9.57E+00	1.80E-07	2.30E-02	4.14E-08	1.14E-09	2.73E-14	3.62E-14

To compare the results for exponential baselines, the root mean squared errors of approximation with each basis set group are tabulated with ten increasing decay rates in Table 2.3(a) for a channel length of 2000 and in 2.3(b) for a channel length of 200.

Table 2.3(a) Comparison of basis set groups results with exponential baseline, channel length 2000.

Decay rate	I	II (hi)	III (h)	IV (q)	V (oe)	VI (ot)	VII (os)
2.50E-03	1.25E+00	7.55E-08	3.72E-05	2.55E-10	1.93E-13	8.49E-14	8.38E-14
8.58E-03	2.60E+00	2.08E-07	9.81E-05	1.34E-09	1.17E-13	9.83E-14	6.16E-14
1.03E-02	2.76E+00	2.48E-07	1.14E-04	1.93E-09	8.89E-14	1.38E-13	1.13E-13
1.90E-02	3.10E+00	5.65E-07	2.22E-04	9.49E-09	1.02E-13	1.06E-13	1.40E-13
2.35E-02	3.14E+00	8.41E-07	2.96E-04	1.84E-08	1.96E-13	2.07E-13	1.51E-13
2.85E-02	3.16E+00	1.28E-06	3.93E-04	3.46E-08	4.19E-13	4.26E-13	3.35E-13
3.38E-02	3.17E+00	2.00E-06	5.17E-04	6.27E-08	9.06E-13	9.35E-13	7.59E-13
4.29E-02	3.17E+00	4.12E-06	7.80E-04	1.50E-07	2.98E-12	3.08E-12	2.58E-12
4.42E-02	3.17E+00	4.55E-06	8.22E-04	1.67E-07	3.48E-12	3.60E-12	3.02E-12
4.91E-02	3.17E+00	6.57E-06	9.93E-04	2.48E-07	6.05E-12	6.26E-12	5.33E-12

Table 2.3(b) Comparison of basis set groups results with exponential baseline, channel length 200.

Decay rate	I	II (hi)	III (h)	IV (q)	V (oe)	VI (ot)	VII (os)
2.50E-03	4.67E-01	8.77E-09	1.12E-03	5.34E-07	5.37E-14	4.01E-14	5.76E-14
8.58E-03	1.51E+00	2.84E-08	3.63E-03	1.75E-06	4.05E-14	9.18E-14	5.74E-14
1.03E-02	1.77E+00	3.35E-08	4.27E-03	2.07E-06	4.31E-14	5.27E-14	5.39E-14
1.90E-02	3.03E+00	5.78E-08	7.37E-03	3.71E-06	3.63E-14	4.10E-14	3.69E-14
2.35E-02	3.59E+00	6.91E-08	8.80E-03	4.56E-06	6.64E-14	3.51E-14	5.40E-14
2.85E-02	4.15E+00	8.09E-08	1.03E-02	5.53E-06	3.21E-14	7.72E-14	6.50E-14
3.38E-02	4.69E+00	9.28E-08	1.18E-02	6.64E-06	3.49E-14	5.60E-14	3.64E-14
4.29E-02	5.50E+00	1.12E-07	1.42E-02	8.76E-06	3.10E-14	2.96E-14	3.62E-14
4.42E-02	5.60E+00	1.15E-07	1.45E-02	9.09E-06	4.41E-14	3.12E-14	4.18E-14
4.91E-02	5.97E+00	1.25E-07	1.57E-02	1.04E-05	6.68E-14	3.59E-14	5.04E-14

The TFB (Group I) again gives a very high RMSE and could not provide a reasonable approximation for the exponential baselines at either of the two channel lengths. Group III again seems to work better than the TFB, giving smaller RMSE values, but the results deteriorate for the smaller channel length. It can be seen that RMSE values increase with increasing decay rate in each of the two channel length cases for all the basis groups under study, likely due to the increased difficulty in fitting the sharper decays with the lower frequencies in the truncated series. Overall groups II, IV, V, VI and VII are again able to provide a good approximation for exponential baselines for both channel lengths.

The RMSE of approximation for sinusoidal baselines with each basis set groups are tabulated for ten increasing frequencies in Table 2.4(a) at a channel length of 2000 and in 2.4(b) at a channel length of 200.

Table 2.4(a) Comparison of basis set groups results with sinusoidal baseline, channel length 2000.

Frequencies	I	II (hi)	III (h)	IV (q)	V (oe)	VI (ot)	VII (os)
4.59E-05	1.91E-01	1.12E-08	5.55E-06	3.44E-11	1.39E-13	1.41E-13	1.76E-13
1.22E-04	4.52E-01	2.51E-08	1.25E-05	6.24E-11	1.27E-13	1.79E-13	1.06E-13
3.40E-04	7.85E-01	4.23E-08	2.11E-05	9.22E-11	9.30E-14	1.09E-13	1.45E-13
4.90E-04	6.39E-01	3.74E-08	1.85E-05	1.12E-10	6.24E-14	7.25E-14	8.77E-14
5.67E-04	1.15E-01	2.03E-09	5.68E-07	1.30E-11	8.45E-14	1.33E-13	7.29E-14
9.35E-04	4.40E+00	1.59E-07	8.31E-05	2.08E-10	8.01E-14	1.34E-13	9.82E-14
1.70E-03	8.44E-02	2.30E-08	1.33E-05	2.26E-09	9.07E-13	9.13E-13	5.22E-13
2.10E-03	4.30E-01	1.65E-08	9.85E-06	3.77E-10	8.86E-14	1.24E-13	1.07E-13
2.70E-03	2.70E+00	1.76E-07	1.54E-04	1.03E-08	2.21E-13	1.03E-13	1.07E-13
4.20E-03	2.66E+00	1.33E-07	6.72E-05	1.98E-10	8.16E-14	7.80E-14	9.35E-14

Table 2.4(b) Comparison of RMSE values for basis set groups results with sinusoidal baselines of varying frequencies at channel length 200.

Frequencies	I	II (hi)	III (h)	IV (q)	V (oe)	VI (ot)	VII (os)
4.59E-05	3.47E-02	6.51E-10	8.33E-05	3.96E-08	4.69E-14	3.94E-14	4.66E-14
1.22E-04	5.01E-01	9.40E-09	1.20E-03	5.71E-07	3.72E-14	5.83E-14	5.17E-14
3.40E-04	8.23E-01	1.54E-08	1.97E-03	9.36E-07	1.26E-14	1.21E-14	1.21E-14
4.90E-04	2.30E-01	4.31E-09	5.51E-04	2.62E-07	1.27E-14	1.24E-14	1.23E-14
5.67E-04	2.53E+00	4.68E-08	5.99E-03	2.72E-06	3.89E-14	4.15E-14	5.21E-14
9.35E-04	1.24E+00	2.32E-08	2.96E-03	1.39E-06	3.98E-14	8.71E-14	4.69E-14
1.70E-03	7.31E+00	1.26E-07	1.63E-02	5.54E-06	2.08E-14	1.86E-14	2.11E-14
2.10E-03	3.70E+00	6.80E-08	8.72E-03	3.85E-06	3.02E-14	2.55E-14	1.98E-14
2.70E-03	5.05E+00	9.16E-08	1.18E-02	4.94E-06	8.69E-15	1.47E-14	9.55E-15
4.20E-03	4.03E-01	7.56E-09	9.67E-04	4.58E-07	3.20E-14	5.74E-14	6.00E-14

It can be observed from the RMSE results, once again, that the TFB (Group I) was unable to approximate these low frequency sinusoids, Group III again gives much smaller RMSE values than Group I, making it an option for approximation of low frequency sinusoids but probably would not be an optimal choice for smaller channel lengths. Overall, among all three types of baseline functions used here, Groups I and III seems to provide better results for sinusoid baselines than exponential and linear baselines perhaps because the Fourier series itself is a combination of sines and cosines.

For each of the two channel lengths, Groups II, IV, V, VI and VII again provide very small RMSE values, indicating that these groups are able to provide a very good approximation for the sinusoidal baselines, and this is consistent with other data sets examined.

2.5.3 Conclusions

It has been noted that the truncated Fourier basis without augmentation is not an appropriate choice to approximate and eliminate the slowly varying baseline functions from the simulated analytical signals. The reason for its poor performance is the frequency difference between the first non-zero Fourier basis frequency (f_1) and the slowly varying baseline. It has been noted in the above simulation study that including just one additional frequency at $0.5f_1$ set immensely reduced the RMSE values in comparison to the TFB, which supports the above statement.

In comparing the results for all seven groups for all three baseline functions at each of two channel length cases, Groups II, IV, V, VI and VII always seemed to provide a very good approximation, giving very low RMSE values in the baseline function approximation. To keep the computation time at a minimum while still providing a good approximation, the choice of the smallest basis set would seem to be more appropriate. Group IV appears to be the best choice in that regard in comparison to the next three groups (V, VI, VII), since Group IV uses a smaller number of augmented basis vectors and also provides reasonably low RMSE values in all of the six simulated data sets studied. Group II uses more basis sets at higher frequencies and therefore is not as appropriate for estimation of baseline artifacts. In addition, Group IV also provided a

relatively better approximation for all three baseline functions than Group II, which give relatively higher RMSE values. These two groups only compete for exponential baselines at a smaller channel length and high decay rate.

Therefore, group IV with the augmentation of two pairs of sine and cosine between a DC and the first frequency of TFB was selected the best choice for the approximation of low frequency baseline functions and used in the implementation of the algorithm in subsequent discussions.

2.6 Asymmetric Least Squares

The ability to model baseline profiles accurately is only one requirement of a baseline removal algorithm, since it also must be constrained so that it does not model or remove analytical signals. In this work, the analytical signals are assumed to occur at higher frequencies and will not be fit well by the truncated and augmented Fourier basis set. However, this is not sufficient, since only baseline regions must be included in the fit. For this purpose, asymmetric least squares (ALS) is used. Whereas ordinary least squares assumes a symmetric distribution of residuals around the model, ALS an asymmetric distribution in which positive residuals can be much larger than the negative ones, since the negative residuals from the baseline contain only instrumental noise, but positive residuals are the result of noise and analytical signal.

Asymmetric least squares or expectile estimation was first introduced by Newey and Powell in 1987 [120] as a least squares alternative to regression quantiles and described as easily computable and more efficient than quantile regression, which was introduced by Koenker and Bassett in 1978 [121]. Later, Eilers (2004) used this idea in

baseline estimation [32, 33]. Asymmetric least squares regression is a weighted generalization of ordinary least squares regression. A baseline function, also referred to as an expectile curve, is obtained by iteratively re-weighted least squares.

In ordinary least squares (OLS) the trend is estimated by minimizing the sum of squared residuals

$$SSR_{OLS} = \sum_{i=1}^{n} (y_i - \mu_i)^2 \tag{2.26}$$

where n is the sample size, y_i is the response variable and μ_i is the expected value based on current model parameters. In Figure 2.5 (a) simulated analytical data are presented along with the trial baseline obtained using OLS with a truncated/augmented Fourier series with $n = 1/4, 1/2, 1, 2, 3, \ldots$. Note that, while the basis functions are insufficient to model the signal peaks, OLS tries to include them, resulting in an entirely unsatisfactory baseline model.

Analytical signals can be regarded as having two main regions, areas with signals of analytical interest (peaks) and those without peaks. Hence we have more interest in the upper and lower boundaries than in the mean trend. To a first approximation, the area of the signal higher than the expected line in Figure 2.5(a) could be regarded as an upper boundary containing analytical information and the area lower than the expected line as lower boundary containing baseline information. Asymmetrically weighted least squares, which is a weighted generalization of ordinary least squares, would be an appropriate choice here to estimate this boundary. Thus, the baseline can be estimated by minimising

$$SSR_{ALS} = \sum_{n=1}^{n} w_i \left(y_i - \mu_i \right)^2 \tag{2.27}$$

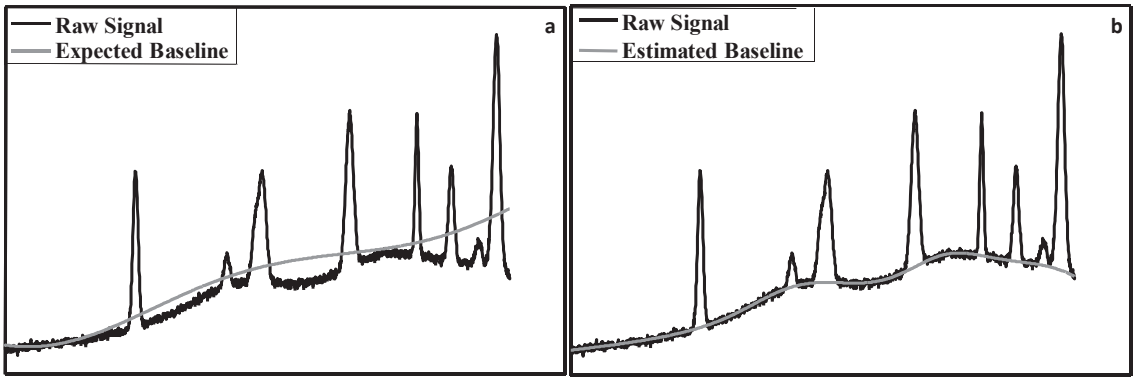

Figure 2.5 (a) Raw signal with expected baseline estimated by OLS.**(b)** Raw signal with estimated baseline estimated by ALS.

The weights w_i are assigned as

$$w_i = \begin{cases} p & if\ y_i > \mu_i \\ 1-p & if\ y_i \le \mu_i \end{cases} \tag{2.28}$$

where y_i is the response variable and μ_i is the expectile or expected baseline value for asymmetry parameter p with $0 < p < 1$. The final resultant baseline is presented in Figure 2.5 (b). The user usually has to choose the asymmetry parameter p, which gives very small weight for the area where an analyte peak is present and a weight of $1-p$ for negative residuals in areas where peaks are absent. The final model is obtained iteratively, solving least squares problem and re-evaluating the weights at each step until there is no further change in weights. With the set of basis functions $\mathbf{B}$ $(n_{channel} \times n_{basis})$, the expected baseline $\boldsymbol{\mu}$ $(n_{channel} \times 1)$ is calculated as

$$\boldsymbol{\mu} = \mathbf{Bq} \tag{2.29}$$

$$\mathbf{q} = \left(\mathbf{B}^{\mathrm{T}}\mathbf{W}\mathbf{B}\right)^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{W}\mathbf{Y} \tag{2.30}$$

Where $\mathbf{W}$ is an $\left(n_{channel} \times n_{channel}\right)$ diagonal matrix of weights. The basis set matrix is calculated as described in the previous section based on the truncation and augmentation of the Fourier series chosen. The number of basis functions, n_b, is calculated from the number of frequency , n_{freq} parameter that is selected by the user. Starting from unit weight, the regression vector (2.23) and consequently expected baseline is calculated (2.22). With these results, new weights (2.21) are determined for each data point and a new expected baseline is estimated in each iteration until the weights do not change anymore. Ordinary least square is a special case of ALS with asymmetry parameter $p = 0.5$.

2.7 Summary

Asymmetric least squares is a convenient and effective method for baseline estimation, but optimal performance requires a function of appropriate form to model the baseline without modeling signal peaks. A truncated Fourier series, retaining only low frequencies consistent with the baseline, is a natural choice for this function but leads to artifacts that are a consequence of the truncation. These artifacts can be mitigated if we introduce more basis functions with lower frequencies, eliminating the ripple effect from the baseline approximation result. Since the baseline of a signal is comprised of very low frequency components, augmentation with frequencies between DC and the first non-zero frequency Fourier component was able to provide a good approximation of low frequency baseline components with no ripple effect.

The proposed method is referred as Truncated Fourier Asymmetric Least Squares (TFALS). TFALS requires the adjustment of two parameters to obtain the optimal baseline estimation. The relationship of these parameters with the attributes of analytical signals is studied and discussed in next chapter, followed by the application of TFALS to simulated and experimental data in Chapter 4. The application of the TFALS to experimental data will better demonstrate the appropriateness of TFALS basis set to estimate a true baseline.

CHAPTER 3
Parameter Optimization for Baseline Results

3.1 Introduction

In recent years, the most promising methods for baseline estimation have been based on the asymmetric least squares (ALS) approach, which has been used in conjunction with baseline representative signals as a basis or with a Whittaker smoother. In this work, a truncated Fourier series augmented with low frequency components is used to model slowly varying baselines in conjunction with ALS fitting. This approach will be referred to as the Truncated Fourier Asymmetric Least Squares (TFALS) method.

TFALS assumes that the baseline is a combination of low frequency components. To model these slowly varying baseline components of the analytical signal vector, a set of basis functions is first generated using a Fourier series heavily truncated to the lower frequencies along with the augmentation of two additional lower frequency components and a DC component. The total number of frequency components (n_{freq}) used to model an individual baseline (each frequency resulting in both sine and cosine terms) is chosen by the user (number of frequencies). The total number of basis functions generated, n_{basis}, is $2n_{freq}-1$ since only one term is generated for the DC component. After normalization of the basis functions, singular value decomposition (SVD) is applied to obtain the orthonormal basis set that establishes a non-singular matrix of basis functions. ALS regression is then applied to get the coefficient vector (regression vector) by iteratively changing the weights of each variable point, giving weight near zero(p)

around peak areas and $(1-p)$ otherwise. Finally, the product of the basis functions matrix and the coefficient vector provide the estimated baseline vector.

TFALS requires the adjustment of two parameters (n_{freq} and p) for the estimation and elimination of slowly varying baseline components from the analytical signals. It is necessary to understand the relationship between the parameters and the characteristics of the baseline and signal to choose the optimal values for a specific baseline and obtain a baseline corrected signal. The motivation to study this relationship was to provide a good explanation and a better understanding of the choice of parameters for a specific analytical signal based on its attributes, since an incorrect choice of parameters could lead to suboptimal baseline correction.

This chapter begins with a complete description of the TFALS algorithm and its adjustable parameters (number of frequencies, n_{freq}, and the asymmetric weight, p) and follows with a study of the relationship of these parameters with the analytical signal attributes in the context of baseline estimation. The studies incorporated a variety of simulated data sets designed to examine and validate the parameter dependencies. These included the asymmetric weight dependency on signal-to-noise ratio and the relationship of both parameters with characteristics of the analyte peak (peak height, peak width and peak location). The effect of both adjustable parameters on baseline estimation for various signal attributes was evaluated by error estimation in baseline approximation. The estimated errors were then plotted against individual signal attribute values (e.g. different baseline amplitudes or different peak heights) and the TFALS parameter values.

Two and three dimensional plots of results are included to visualize the individual signal attribute effect on TFALS parameters.

3.2 Truncated Fourier Asymmetric Least Squares Algorithm

Consider that the measured signals are collected in a raw data matrix $\mathbf{X}_{ab}\left(n_{chan} \times n_s\right)$ having a number of related samples, n_s, measured at a number of equally spaced channels, n_{chan}. These channels can be any ordinal variable (e.g. time, wavelength), assuming each sample vector $\mathbf{x}_{ab}\left(n_{chan} \times 1\right)$ is the sum of the baseline $\mathbf{x}_b$ and the analyte signal $\mathbf{x}_a$ with some noise components.

$$\mathbf{x}_{ab} = \mathbf{x}_a + \mathbf{x}_b \tag{3.1}$$

The subscript 'ab' represents the raw spectrum (analyte with the baseline), 'a' represents the quantities for analyte only, and 'b' represents the quantities for baseline only.

Assuming that the baseline vector is a linear combination of predefined functions, it is modeled with the user-specified number of orthonormal basis functions that form the matrix $\mathbf{B}\ \left(n_{chan} \times n_{basis}\right)$.

$$\mathbf{x}_{ab} = \mathbf{x}_a + \mathbf{Bq} \tag{3.2}$$

Various functions can form the basis set used in $\mathbf{B}$ (e.g. baseline signals, polynomial), but in this work, the truncated and augmented Fourier series is used in this capacity. In principle, the estimated regression vector, $\mathbf{q}\ (1 \times n_{basis})$, becomes the Fourier coefficients

for the baseline estimate. In practice, however, singularity restrictions may require a redefinition of the basis functions actually used, as discussed in Chapter 2.

3.2.1 Truncated and Augmented Fourier Basis Set

The TFALS algorithm requires user specification of two parameters, the number of frequency terms to use in the truncated/augmented Fourier series, n_{freq}, and the asymmetric weight, p. The first of these relates to the basis functions used, while the second relates to the ALS fitting. It is the choice of n_{freq} and its relationship to the basis functions that is discussed here.

Given the $n_{nchan} \times 1$ signal vector $\mathbf{x}_{ab}$, the corresponding frequencies for the Fourier transformed signal will be designed as $\left[f_0, f_1, f_2, ..., f_{n_{chan}/2} \right]$, where $f_0 = 0$ (DC), f_1 is the lowest non-zero frequency component, $f_2 = 2f_1$, and $f_{nchan/2}$ corresponds to the Nyquist frequency. As discussed in Chapter 2, the augmented frequency set used to generate the basis functions is $\left[0,\ 0.25f_1,\ 0.5f_1, f_1,\ 2f_1, ...,\ \left(n_{freq} - 3 \right) f_1 \right]$. Therefore, the simplest baseline $\left(n_{freq} = 1 \right)$ would consist of only a baseline offset (DC) component. The addition of increasingly higher frequencies allows the modeling of more complex baselines.

If the matrix of basis functions is given by $\mathbf{B}\left(n_{chan} \times n_{basis} \right)$ then the jth column vector of $\mathbf{B}$, designated as $\mathbf{b}_j$ represents the jth basis function. The number of basis functions, n_{basis}, is related to the number of frequency components, n_{freq}, by

$$n_{basis} = 2n_{freq} - 1 \tag{3.3}$$

since there are sine and cosine terms for each frequency except for zero. If b_{ij} represents the ith element $(i = 1....n_{chan})$ of basis function $\mathbf{b}_j$, the basis functions for the TFALS method are defined as follows.

$$b_{i1} = 1 \qquad (f = 0) \tag{3.4}$$

$$b_{i2} = \sin\left(\frac{0.5\pi(i-1)}{n_{chan}}\right) \qquad (f = 0.25f_1) \tag{3.5}$$

$$b_{i3} = \cos\left(\frac{0.5\pi(i-1)}{n_{chan}}\right) \qquad (f = 0.25f_1) \tag{3.6}$$

$$b_{i4} = \sin\left(\frac{\pi(i-1)}{n_{chan}}\right) \qquad (f = 0.5f_1) \tag{3.7}$$

$$b_{i5} = \cos\left(\frac{\pi(i-1)}{n_{chan}}\right) \qquad (f = 0.5f_1) \tag{3.8}$$

For $j > 5$, the basis function is defined as follows.

$$b_{ij} = \sin\left(\frac{2\pi(j-4)(i-1)}{2n_{chan}}\right) \qquad \left(j = \text{even}, f = \frac{j-4}{2}f_1\right) \tag{3.9}$$

$$b_{ij} = \cos\left(\frac{2\pi(j-5)(i-1)}{2n_{chan}}\right) \qquad \left(j\text{=odd}, f = \frac{j-5}{2}f_1\right) \tag{3.10}$$

Note that the maximum value for j is $n_{basis} = 2n_{freq} - 1$.

This raises the question of how many frequency components to include when optimizing the TFALS algorithm for baseline removal. The lower limit, $n_{freq} = 1$, corresponds to the simplest case of a constant offset. The addition of more frequencies allows more variation in the baseline signal to be modeled but also runs the risk of

modeling analytical signal components (peaks). In principle, the upper limit approaches the Nyquist frequency $\left(n_{chan} \cdot f_1/2\right)$, but this would clearly be too far.

To obtain a more reasonable upper limit for optimization purposes, the following argument is used. The upper limit depends on the width of analytical peaks relative to the overall signal. Narrow peaks on an extended, highly variable baseline would tolerate/require higher frequency baseline components than broader peaks over a shorter window. If we assume that a "narrow" peak is a Gaussian peak with a standard deviation of $\sigma_t = 2$ points in the measurement domain, then we can write,

$$g(i) = h_t e^{-(n_c - i)^2/2\sigma_t^2} \tag{3.11}$$

where $g(i)$ is the signal magnitude as a function of the index, $i(i = 1...n_{chan})$, h_t is the maximum height of the Gaussian in the time domain, and n_c is the index at the centre of the peak. It is known that a Gaussian signal in the time domain produces a Gaussian shaped amplitude spectrum in the frequency domain described by,

$$G(\omega) = h_\omega e^{-\omega^2\sigma_t^2/2} \tag{3.12}$$

where $\omega = 2\pi f$, $G(\omega)$ is the amplitude at ω and h_ω is the amplitude at $\omega = 0$. The Gaussian in the frequency domain is centered at $f = 0$ (symmetric for positive and negative frequencies) and its width is inversely proportional to the width of the Gaussian in the time domain; i.e. broader peaks in the time domain lead to narrow peaks in the frequency domain and vice-versa. If we let j represent the index of the Fourier amplitude in the frequency domain, then we can write

$$\omega = 2\pi f = 2\pi f_1 j = \frac{2\pi f_s j}{n_{chan}} \tag{3.13}$$

where f_1 is the first non-zero frequency and f_s is the sampling frequency. The result was obtained using $f_1 = (2f_N / n_{chan})$ where f_N is the Nyquist (maximum) frequency given by $f_N = f_s/2$. Arbitrarily taking $\Delta t = 1$ (units don't matter), $f_s = 1$ and we can write

$$G(j) = h_\omega e^{-4\pi j^2 \sigma_t^2 / 2n_{chan}^2} = h_\omega e^{-j^2/2\sigma_f^2} \tag{3.14}$$

In this equation, σ_t is the standard deviation of the Gaussian in the frequency domain expressed in units of the index in that domain. Equating the exponents gives.

$$\frac{-4\pi j^2 \sigma_t^2}{2n_{chan}^2} = \frac{-j^2}{2\sigma_f^2} \tag{3.15}$$

which lead to

$$\sigma_f = \frac{n_{chan}}{2\pi\sigma_t} \tag{3.16}$$

Thus the analyte peak will be represented a Gaussian with a standard deviation of σ_f channels in the frequency domain.

In modeling the baseline, we want to minimize the analyte signal components included. We will arbitrarily set a limit of 10% of the amplitude spectrum (low frequencies), for a Gaussian peak, this corresponds to $z = 0.1257$. Taking $\sigma_t = 2$, this leads to

$$j_{\max} = \frac{0.1257 \cdot n_{chan}}{2\pi \cdot 2} = 0.0100 n_{chan} \tag{3.17}$$

Therefore, for the purposes of the proposed algorithm, the maximum number of frequencies included in the maximum number augmented basis set is

$$n_{freq}(\max) = 3 + round(n_{chan}/100) \tag{3.18}$$

In practice, it has been found that the upper limit usually does not need to extend this high, since most baselines are slowly varying. Generally, 4 or 5 frequency terms are sufficient to model most baselines, but this provides a rationale for setting an upper limit in the optimization process.

As a final step in the generation of basis functions each basis vector is normalized to unit length by dividing each column vector by its Euclidean norm:

$$\mathbf{b}_j^{norm} = \frac{\mathbf{b}_j}{\left\|\mathbf{b}_j\right\|} \tag{3.19}$$

While this does not change the shape of the vectors, it may improve mathematical stability.

3.2.2 Orthonormal Basis Set

As noted in Chapter 2, because of the augmentation of sine and cosine components in the Fourier series with lower frequencies that extend beyond the orthogonality interval (2π), the basis set matrix generated here is no longer orthogonal, which may lead to a singularity problem in the least squares estimation. To obtain a non-singular basis set matrix and simplify the least squares problem in the next step, singular value decomposition (SVD) it is used to orthogonalize the normalized basis set matrix.

SVD represents the matrix $\mathbf{B}$ $(n_{chan} \times n_{basis})$ as the product of three matrices: the left singular matrix, $\mathbf{U}$ $(n_{chan} \times n_{basis})$ (orthonormal), the diagonal matrix of singular values, $\mathbf{S}$ $(n_{chan} \times n_{basis})$, and the right singular matrix, $\mathbf{V}$ $(n_{chan} \times n_{basis})$.

$$\mathbf{B} = \mathbf{USV}^{\mathbf{T}} \tag{3.20}$$

If $\mathbf{B}$ is singular, the rank of $\mathbf{B}$ can be determined as the point at which the singular values in $\mathbf{S}$ go to zero to working precision. If this is the case, the matrix $\mathbf{U}$ is truncated to the first p columns, where p is the rank of $\mathbf{B}$, resulting in the truncated matrix $\mathbf{U}_p\left(n_{chan} \times p\right)$. If $\mathbf{B}$ is not singular, then $\mathbf{U}_p = \mathbf{U}$. The matrix $\mathbf{U}_p$ is then used as the matrix of basis functions for asymmetric least squares.

3.2.3 Asymmetric Least Squares Approximation

Having transformed the basis functions in the augmented Fourier series, $\mathbf{B}$, to the alternate space in $\mathbf{U}_p$, the task now is to use asymmetric least squares in conjunction with these basis vectors to estimate the baseline. The model is

$$\mathbf{x}_b = \mathbf{U}_p\mathbf{q} + \mathbf{e} \tag{3.21}$$

Here, $\mathbf{x}_b$ is the baseline component of the signal $\left(n_{chan} \times 1\right)$, $\mathbf{q}$ is the vector of basis function coefficients $\left(p \times 1\right)$, and $\mathbf{e}$ is an error vector $\left(n_{chan} \times 1\right)$.

The regression vector $\mathbf{q}$ in Equation 2.30 describes the shape and intensity changes of analyte signal due to the baseline contribution. This shape and intensity vector is calculated by asymmetric least squares regression, assuming a very small weight p for the area where analyte peaks are present and $\left(1 - p\right)$ in baseline regions. Initially, weights are set to one for all channel points. The regression vector is initially estimated by

$$\mathbf{q} = \left(\mathbf{U}_p^T\mathbf{U}_p\right)^{-1}\mathbf{U}_p^T\mathbf{x}_{\mathbf{ab}} \tag{3.22}$$

Note that this equation uses $\mathbf{x}_{ab}$, the vector of measurements (analytical signal+ baseline) rather than the baseline vector, $\mathbf{x}_b$, since the later is, of course, unavailable. The

objective is to use weighted regression to fit only the baseline regions, but of course these are not known *a priori*. The assumption is that analyte signals will have only positive residuals, as positive residuals are given a much smaller weight in an iterative process. Following the initial fit, the residuals are calculated by

$$\mathbf{r}_1 = \mathbf{x}_{ab} - \mathbf{U}_p \hat{\mathbf{q}}_1 \tag{3.23}$$

For the second iteration, the weight for measurement i ($i = 1...n_{chan}$) is assigned to p (small) if the corresponding residual is positive, or to $(1-p)$ otherwise. These weights are placed on the diagonal of a weight matrix, $\mathbf{W}$ ($n_{chan} \times n_{chan}$) and used for weighted regression in the second step

$$\hat{\mathbf{q}}_2 = \left(\mathbf{U}_p^T \mathbf{W}_2 \mathbf{U}_p\right)^{-1} \mathbf{U}_p^T \mathbf{W}_2 \mathbf{x}_{ab} \tag{3.24}$$

Here the "2" indicates the coefficients and weights at the second iteration. Residuals and weights are recalculated and the processes is repeated based on the following equations

$$\mathbf{r}_j = \mathbf{x}_{ab} - \mathbf{U}_p \mathbf{q}_j \tag{3.25}$$

$$w_{ij} = \begin{cases} p & \text{if } r_{ij} > 0 \\ (1-p) & \text{if } r_{ij} \le 0 \end{cases} \tag{3.26}$$

$$\mathbf{W}_j = diag\left(\mathbf{w}_j\right) \tag{3.27}$$

$$\hat{\mathbf{q}}_{j+1} = \left(\mathbf{U}_p^T \mathbf{W}_j \mathbf{U}_p\right)^{-1} \mathbf{U}_p^T \mathbf{W}_j \mathbf{x}_{ab} \tag{3.28}$$

This process is continued until convergence when no further changes in the weights occur. At that point, the baseline is

$$\mathbf{x}_b = \mathbf{U}_p \hat{\mathbf{q}} \tag{3.29}$$

and the corrected analyte signal is obtained by subtraction

$$\hat{\mathbf{x}}_a = \mathbf{x}_{ab} - \hat{\mathbf{x}}_b \tag{3.30}$$

3.3 Parameter Optimization of TFALS

For the purpose of validation of the proposed algorithm and parameter optimization from the statistical perspective, several simulation studies were done to check the two parameter dependency on analytical signals.

3.3.1 Experimental

3.3.1.1 *Computational Aspects*

All data processing was carried out using programs written by the author in MatLab® 2010b (MathWorks, Natick, MA) under Windows 7 Professional 2009© on a 2.10 GHz processor with 2.00 GB of memory.

3.3.1.2 *Data Simulations*

Four simulated data sets were generated in order to examine the effect of the TFALS parameter values (n_{freq}, p) on baseline estimation in the presence of signals. Each data set considers the individual attributes of analytical signals and was generated separately.

Data set 1 was intended to determine the effect of peak parameters on baseline estimation for a single analytical peak in the presence of a baseline. The nominal signal to be modeled consisted of a vector of 4000 points with a Gaussian signal (height, $h = 50$;

standard deviation, $\sigma = 50$) centered at index 2000. This was superimposed on a sigmoidal baseline generated with the function

$$y(x) = \frac{A}{1 + e^{-\left(\frac{x - x_o}{w}\right)}} \tag{3.31}$$

where x correspond to the element index (1 to 4000), x_o is 2000, A is the amplitude (100), and w determines the slope of the sigmoid and was set to 500. Normally distributed random noise was added to the signal with a standard deviation of 1.

Three subsets of data were generated to test the effect of peak height, peak width, and peak location. For Data Set 1a, the peak height was varied between 10 and 100 in steps to 10. For Data Set 1b, the nominal peak height and location were used, but peak

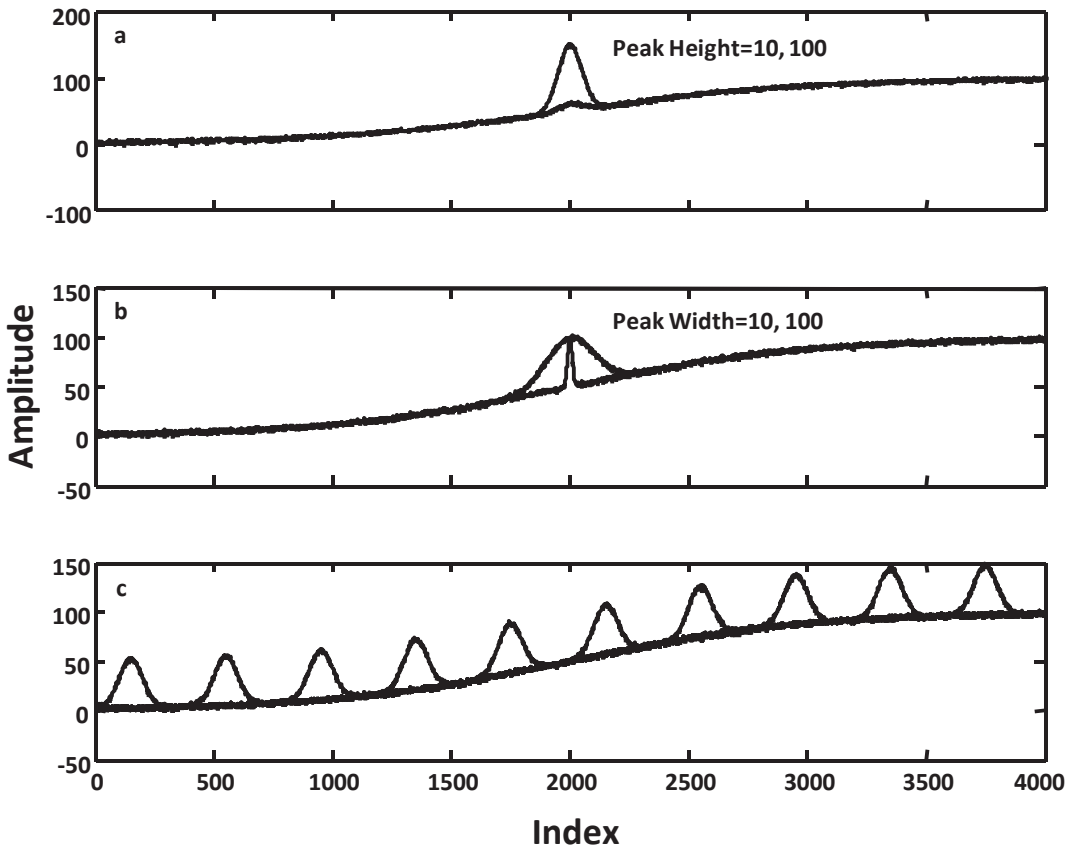

Figure 3.1 Subsets of Data set 1 showing ten differ the range of **(a)** peak heights, **(b)** peak widths and **(c)** peak locations with one sigmoid

width was varied between 10 and 100 in steps of 10. For Data Set 1c, the nominal peak height and width were used, but the peak location was varied between 150 and 3750 in steps of 400. Figure 3.1 represents this simulated data set. Figure 3.1(a) represents Data Set 1a, 3.1(b) represents Data Set 1b, and 3.1(c) represents Data Set 1c. These data subsets were used to examine the dependency of analyte peak height estimation on both of the adjustable TFALS parameters (number of frequency terms and the asymmetric weight).

Data Set 2 was generated to determine the relationship between TFALS asymmetry parameter, p, and the signal-to-noise ratio. The nominal signal to be modeled consisted of a vector of 2000 points with a Gaussian signal (height, $h = 500$, standard deviation,

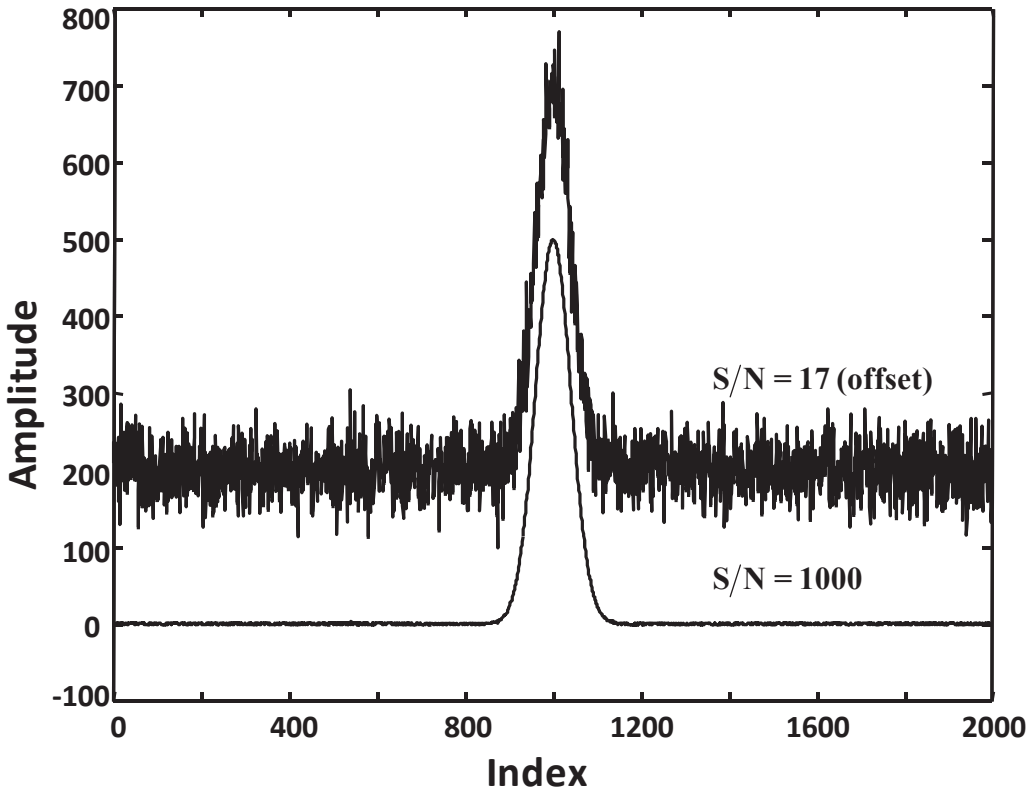

Figure 3.2 Data set 2 with high (lower) and low (upper) signal-to-noise ratios (other signal vectors not shown).

$\sigma = 40$) centered at index 1000. Twenty different normally distributed random noise sequences were added with the standard deviations ranging from 0.5 to 1 in steps of 0.1, 1 to 5 in steps to 0.5, and 6 to 30 in steps of 5. No baseline was added to this signal. The limiting cases of this data set ($S/N = 1000$ and $S/N = 17$) are represented in Figure 3.2.

Data Set 3 was simulated to study the relationship between the two TFALS parameters (number of frequencies, n_{freq}, and asymmetry factor, p). The nominal signal to be modeled consisted of a vector of 2000 points with five Gaussian signals at specified locations (200, 550, 900 1300 and 1750), heights (585, 243, 279, 522 and 315) and standard deviations (21 , 12, 13, 19 and 14). Normally distributed random noise was added to the signals with standard deviations of 4 and 40.

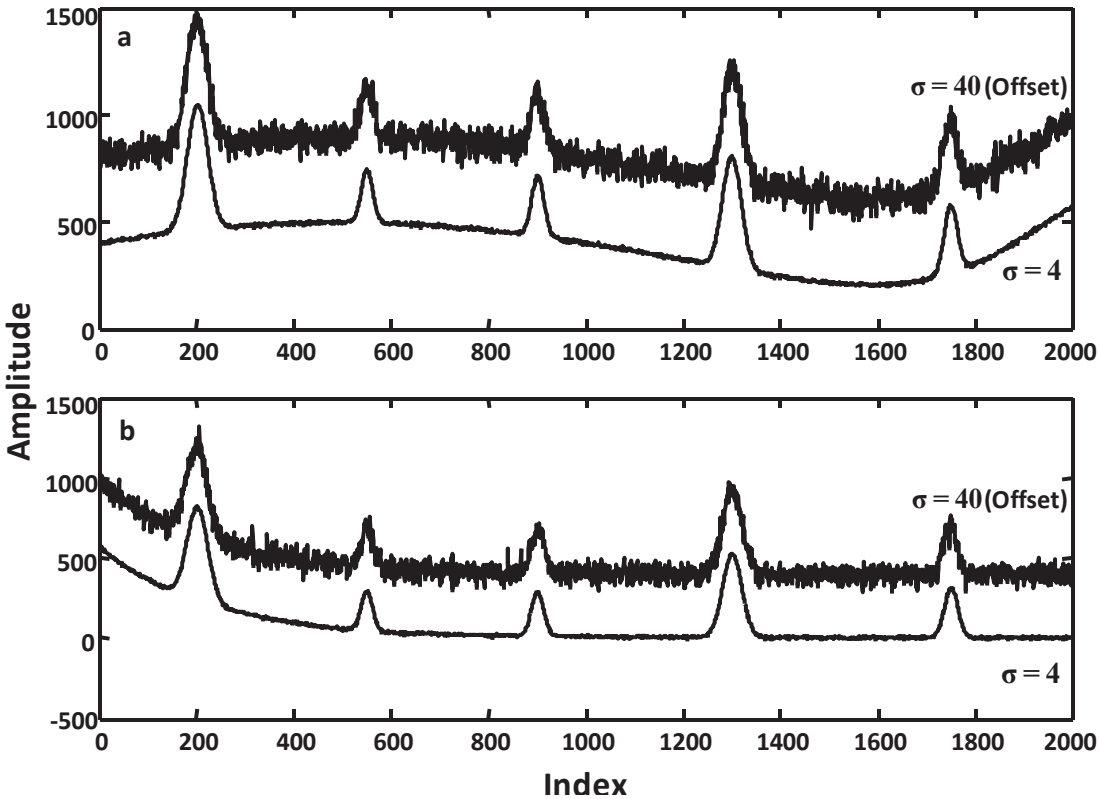

Figure 3.3 Data Set 3, simulated signals with **(a)** Gaussian baseline having low and high signal-to-noise ratios and **(b)** exponential baseline having low and high signal-to-noise ratios.

Two sub-sets of data were generated to test the relationship between the two adjustable parameters in presence of different baseline features. For Data Set 3a, a combination of Gaussians (heights, $h = 500, 700$, standard deviations, $\sigma = 750, 250$) centered at 500 and 2200, was added to the signals as a baseline. For Data Set 3b, an exponential baseline, Ae^{-hx} with a decay rate of $k = 0.0045$ and an amplitude of $A = 573$ was added. The signals with Gaussian baselines are represented in Figure 3.3(a) and the signals with exponential baselines are shown in Figure 3.3(b).

3.3.2 Results and Discussion

3.3.2.1 Effect of TFALS Parameters on Peak Height Estimation

To test the effect of the number of frequencies, n_{freq}, and asymmetry parameter, p, on the estimation of analyte peak heights, three different peak parameters were tested individually: peak height, peak width and peak location. The level of random noise and the shape of the sigmoidal baseline were kept the same for all simulated signal vectors in this test study and only the peak parameters were changed.

Ten different peak heights, widths and locations were used to test the dependence of each of these peak attributes on the two adjustable parameters $\left(n_{freq}, p\right)$ of TFALS. The relationship of each of the three peak attributes with the two adjustable parameters was studied and will be presented separately. To test and visualize the relationships, surface and contour plots were used along with two dimensional line plots for each case. For better visualization of estimation errors in these graphs, 'negative absolute errors' in peak

height estimation were used to measure the estimation error in each case. The negative absolute error was calculated by

$$E_{nabs} = PH_{true} - PH_{estimated} \tag{3.32}$$

In this equation PH_{true} is the true height of the analyte peak measured at the maximum in the absence of noise and baseline, while $PH_{estimated}$ is the peak height of the analyte measured after subtraction of the baseline estimated by TFALS from the analyte signal (in the presence of noise). This definition was chosen so that the sign indicates the deviation of the estimated baseline: a negative sign indicates that the estimated baseline is too low, while a positive sign indicates that it is too high.

3.3.2.1.1 Effect of Peak Height on Baseline Estimation

To examine the relationship between the peak height and the two adjustable parameters of TFALS, ten simulated signal vectors with identical levels of noise and baseline were used. Each signal vector had a single peak with same position and width, but with different peak heights, as previously described for Data set 1a.

Figure 3.4 shows a surface and contour plot of negative absolute errors in peak height estimation for ten different peak heights as a function of the number of frequencies used to model the baseline. The asymmetry parameter, p, was set to 0.021 for these calculations. Figure 3.5 shows the same information in two-dimensional plot to convey the quantitative information more clearly. The general trend is for the errors to increase from more negative to more positive as both the peak height and number of frequencies increases. Ideally, one would like to have all of the errors as close to zero as possible.

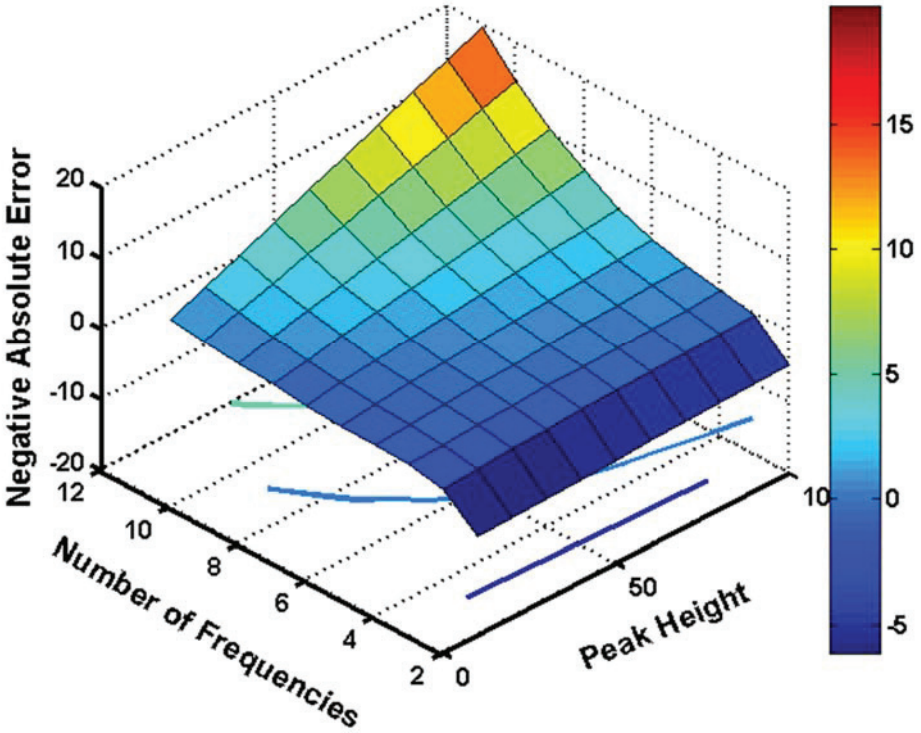

Figure 3.4 Surface and contour plots of estimation errors in peak heights for signals with different peak heights using different number of frequencies.

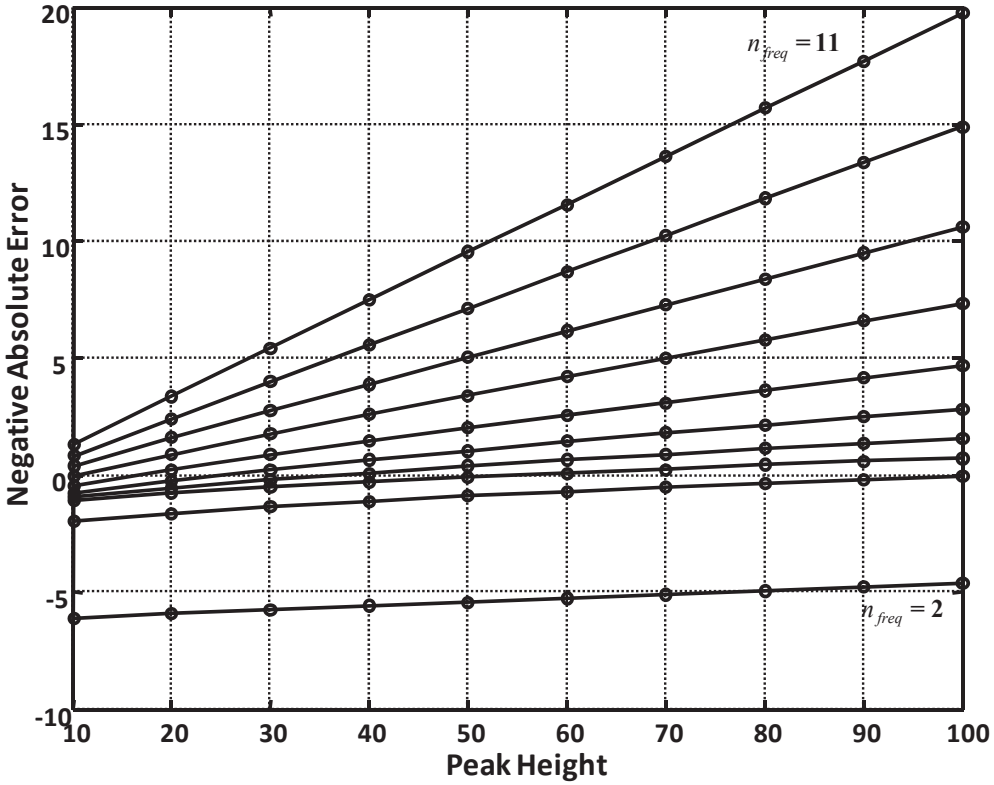

Figure 3.5 Negative absolute errors using number of frequencies from 2 to 11 at different peak heights.

The largest absolute errors are positive and occur when both the peak height and number of basis frequencies is high.

The behavior observed in Figure 3.4 and 3.5 is the result of the interaction of several factors that are a consequence of the asymmetric least squares fitting. In general, because positive weights are much smaller than negative weights, when ALS is applied to a pure baseline signal, it will tend to align itself with the negative excursions of the noise, as shown in Figure 3.6

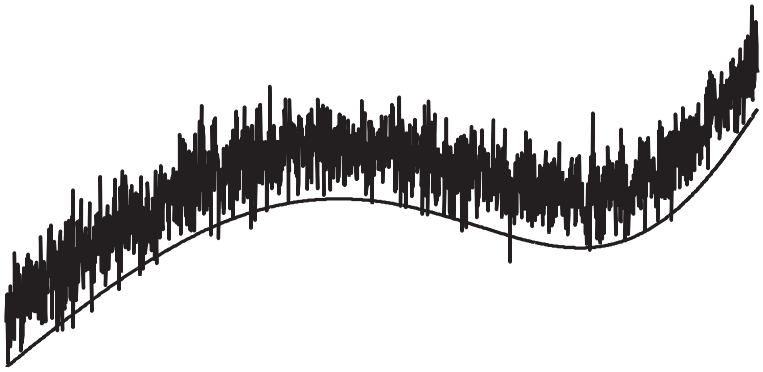

Figure 3.6 Fitting of a noisy baseline by ALS.

As more and more analytical signal components are added, the positive deviations become larger and start to have a greater effect, consequently pulling the estimated baseline higher until they are balanced by more negative deviations. Because higher (or wider) peaks increase this effect, the baseline shifts higher in those cases. The magnitude of the changes depend on the asymmetric weighting factor as well (discussed later). This explains why, for example at $n_{freq} = 4$, the baseline starts with a negative error for small peaks that turns into a positive error for larger peaks. Also note that the magnitude of the errors in on the same order as the noise ($\sigma = 1$) in this case.

The characteristics of ALS also relate to the number of frequencies, n_{freq}, and the ability to fit the baseline and the analytical signal. Recall that n_{freq} is directly related to the number of basis functions used. When the number of frequencies is small ($n_{freq} = 2$) the basis functions are inadequate to model the baseline and relatively large errors result. As the number of basis functions is increased ($n_{freq} = 3$ to 6), the baseline is adequately modeled and smaller errors result. However, when the number of basis functions is increased even further ($n_{freq} = 7$ to 11) the higher frequency components begin to model the positive deviations of the analytical peaks to improve the ALS objective function, resulting in larger positive errors in the baseline (overfitting). As already noted, this effect will be larger for larger peaks.

For the present example, optimum performance across the range of peaks appears to occur for $n_{freq} = 4$, although other values are also acceptable. However, this is not universal and will depend especially on the shape of the baseline and peaks, as well as the number of peaks, asymmetry factor, and signal-to-noise ratio.

One could also argue in this case that $n_{freq} = 9$ would be a good choice, since the relative error (ca. 10% of the peak height) is constant across the range. While this may be true, it is more reliable in the presence of multiple large and small signals to evaluate the baseline estimation relative to the absolute noise level than the signal. This will become evident with the example presented at the end of this section.

Figure 3.7 shows a surface and contour plots of negative absolute errors in peak height estimation for the ten different peak heights as a function of ten different values of

the asymmetry parameter, p , (ranging from 0.001 to 0.046 with an interval of 0.005) used to model the baseline by TFALS. For this plot, the number of frequencies used was $n_{freq} = 4$. Figure 3.8 presents a line plot of the same information in Figure 3.7. It is observed in this plot that each error estimation line starts with negative errors in the baseline and these become more positive with the increment in peak heights. It is also apparent that the signals with higher peaks require a relatively smaller asymmetry parameter value and the smaller peak heights require a higher asymmetry parameter value. For a peak height of 100, for example, a value of $p = 0.011$ gives an error near zero, while a peak height of 30 requires $p = 0.046$ for an error near zero.

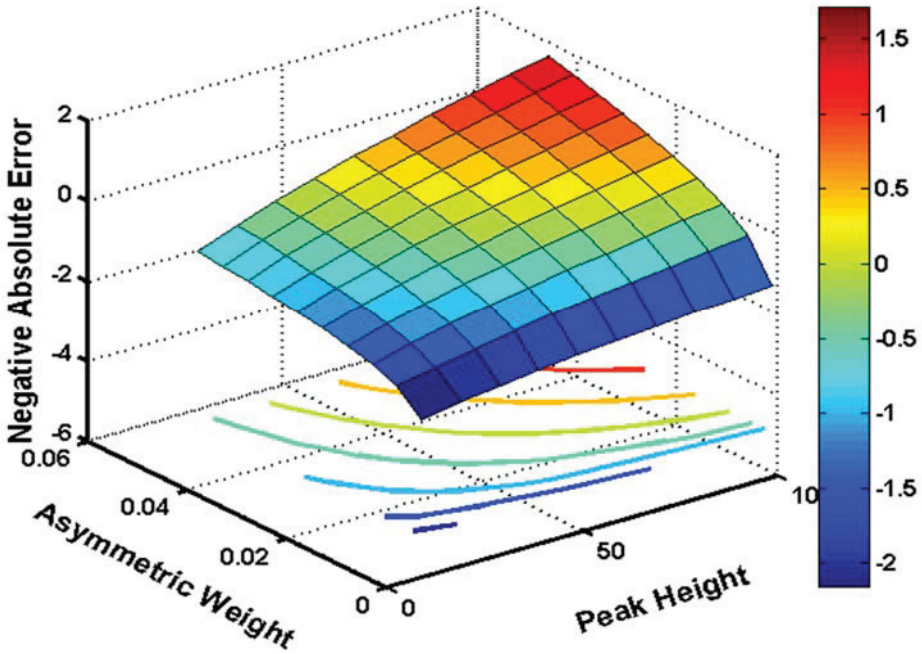

Figure 3.7 Surface and contour plots of estimation errors in peak heights for signals with different peak heights using different asymmetry parameters.

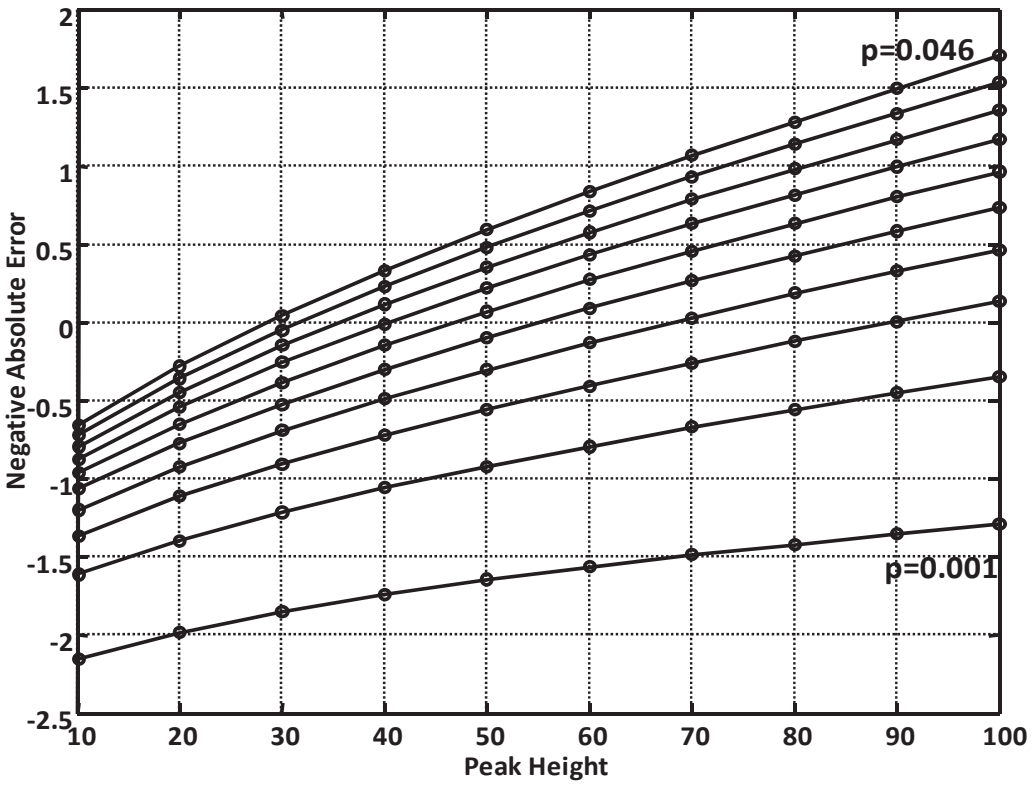

Figure 3.8 Negative absolute errors using different asymmetry parameter values at different peak heights.

The reason for the observed interaction between the asymmetry parameter and the peak height is as previously described. For a pure baseline, the estimation by ALS will generally be low and on the order of the measurement noise. As positive analyte signal components are added, this estimate will be pulled up to improve the objective function, gradually reaching and surpassing the true baseline. A larger asymmetry factor will cause this to happen faster for small peaks, but will result in more positive errors for large peaks. Conversely, a small value of p will make large positive errors less likely for large peaks, but increase the negative errors for small peaks.

The choice of an optimum value for p is difficult because it depends on the number and size of analyte peaks. For the present example, a value of $p = 0.026$ (fifth from the top) achieves a reasonable baseline. However, two points are worth noting. First, the range of errors introduced by changing the choice of p is much smaller than those observed for the choice of n_{freq}, meaning that the selection of this parameter is much less critical than choosing the number of frequencies. Second, since the errors are bounded on the low end by the absence of any signal, but not bounded on the high end, it is safer, when in doubt, to choose a smaller value of p.

Figure 3.9 shows the practical consequences of the TFALS parameters on the simulated signals. Figure 3.9a shows a small signal (height =10) with the true baseline (green) along with baselines estimated with $n_{freq} = 4$ (blue) and $n_{freq} = 11$ (red), both estimated with $p = 0.021$. It is clear that both baseline estimates are similar. In contrast, Figure 3.9b shows the same results with a larger peak (height=100). Here it is clear that the larger number of basis functions ($n_{freq} = 11$) overfits the peak and leads to poor baseline estimation. This effect of peak modeling increases and becomes visually apparent with the increase in peak heights but is also present for small peaks. Therefore, the choice of optimal number of frequencies, n_{freq}, is not dependent on the peak heights but rather the shape of the baseline, and the wrong choice of n_{freq} could lead to incorrect baseline estimation and peak stripping in the signals.

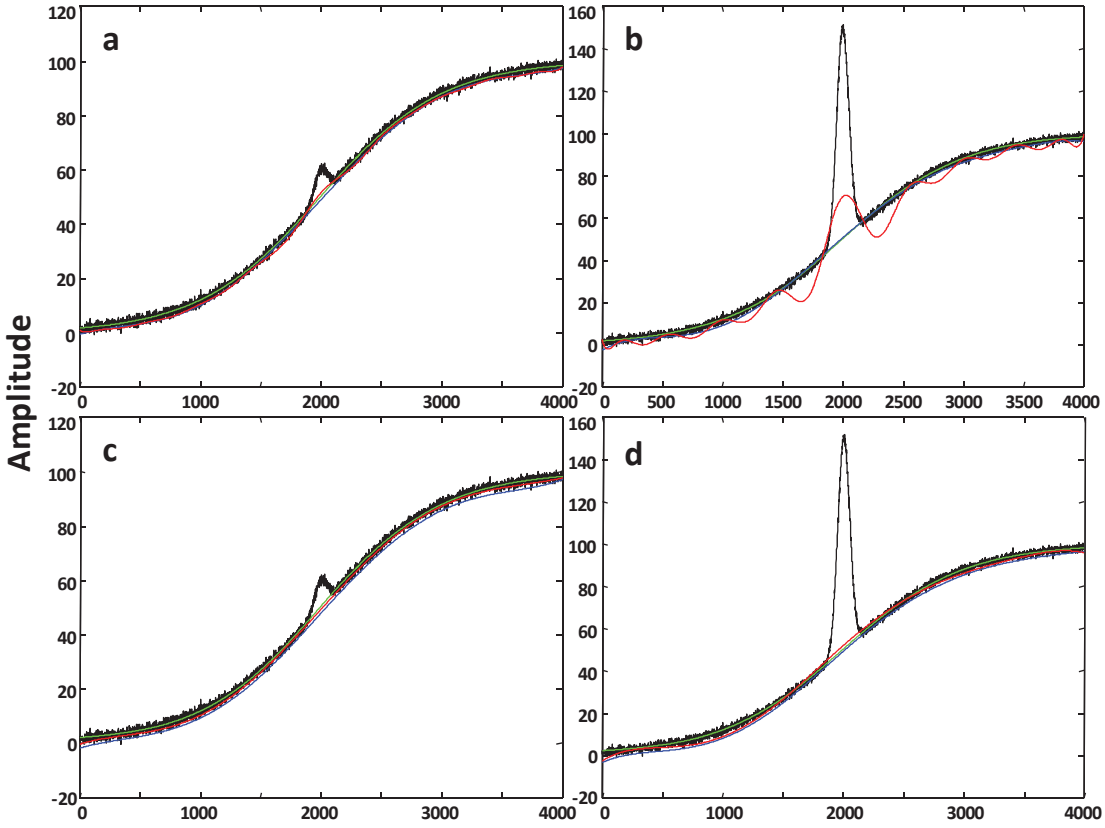

Figure 3.9 Signal vectors (▬) for smallest peak (a, c) and largest peak (b, d) along with true (▬) and estimated baselines. Figure (a) and (b) show the effect of using $n_{freq} = 4$ (▬) and $n_{freq} = 11$ (▬). Figure (c) and (d) show the effect of $p = 0.001$ (blue) and $p = 0.046$ (red).

Figure 3.9c shows the effect of the asymmetry parameter, p, on the small peaks with values of $p = 0.001$ (blue) and $p = 0.046$ (red) compared to the baseline (green) using $n_{freq} = 4$. A similar plot for the large peak is shown in Figure 3.9d. It can be clearly observed from the two plots that, for the small peak, the red line (representing $p = 0.046$) is relatively closer to the true baseline, but both values would provide an acceptable result, whereas for large peak an asymmetry parameter between the two values presented would be a better choice. Again, however, both estimated baselines would be acceptable. It is clearly seen from both baseline estimation plots that the difference in

peak height estimation using two different asymmetry values is not crucially different. Hence, the choice of optimal asymmetry parameter is relatively robust and independent of peak height.

3.3.2.1.2 Effect of Peak Width on Baseline Estimation

For testing the TFALS parameters' relationship to peak width, ten simulated signal vectors with identical levels of noise and baseline and different peak widths were used and tested individually as in the previous section.

Figure 3.10 shows a surface and contour plots of negative absolute errors in peak height estimations for ten different peak widths as a function of number frequencies used

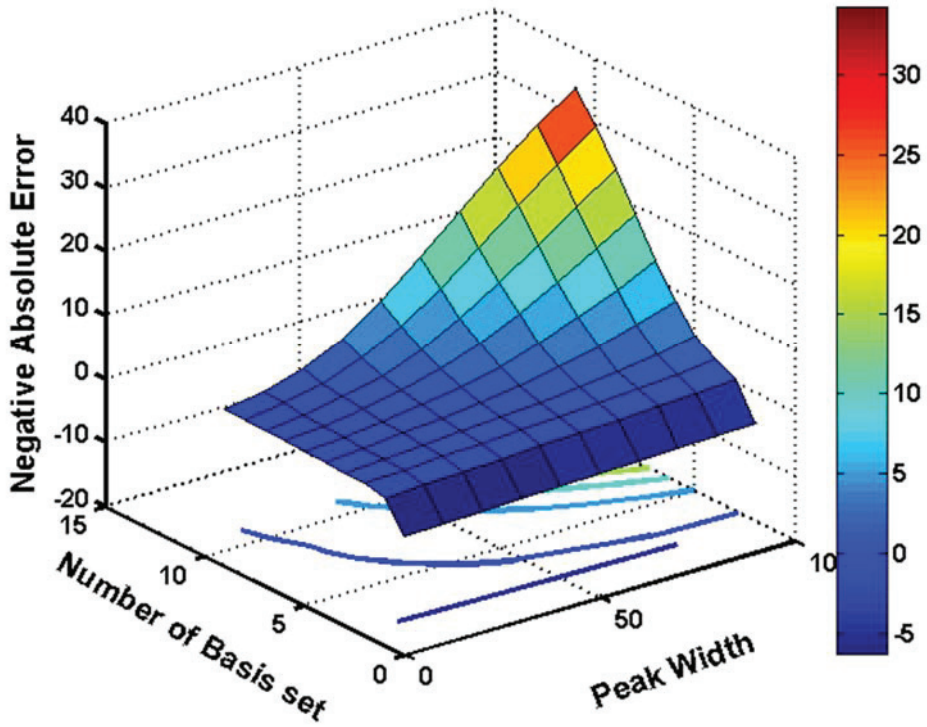

Figure 3.10 Surface and contour plots of estimation errors in peak heights for signals with different peak widths using different numbers of frequencies.

to model the baseline. The value of p was set to 0.021 for these calculations. It is apparent that the estimation errors are minimal at the same number of frequencies used to model the data for all ten peak widths as in the previous section, which probably is the optimal number for that specific shape of baseline; however the estimated error seems relatively higher at higher peak widths, specifically using higher number of frequencies.

Figure 3.11 shows a line plot with the same information continued in Figure 3.10. For this specific data set, the third line from the bottom ($n_{freq} = 4$) seems to be an optimal choice, having small slope near the zero line. It is not surprising that this is consistent with the previous section since the baseline and noise are the same.

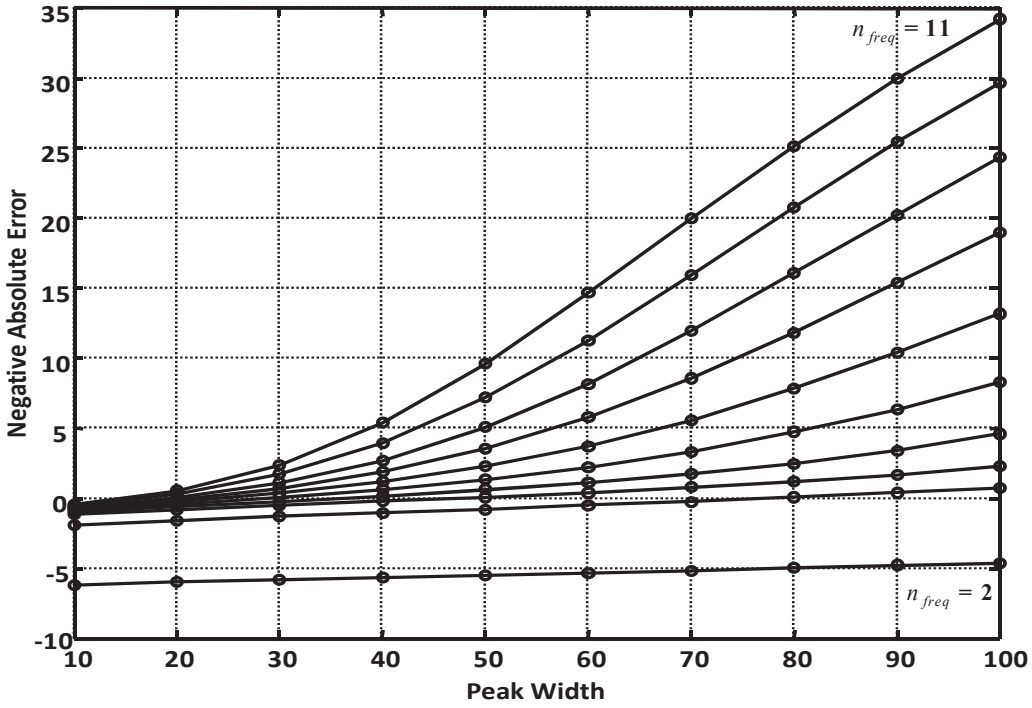

Figure 3.11 Negative absolute errors using number of frequencies from 2 to 11 at different peak widths.

The interpretation of these figures is similar to those in the previous section. For a given set of basis functions (n_{freq}), increasing the peak width increases the amount of positive values. For narrow peaks, the effect of increasing the number of frequencies is relatively unimportant, since the basis functions do not contain sufficiently high frequencies to model the analyte peak. For wider peaks, however, this ability to fit the analyte peaks becomes important at lower and lower frequencies, resulting in larger positive errors (higher baselines) for wider peaks.

Figure 3.12 presents surface and contour plots of negative absolute errors in peak height estimation for ten different peak widths as a function of the asymmetry parameter, p, values (ranging from 0.001 to 0.046 with an interval of 0.005) used to model the

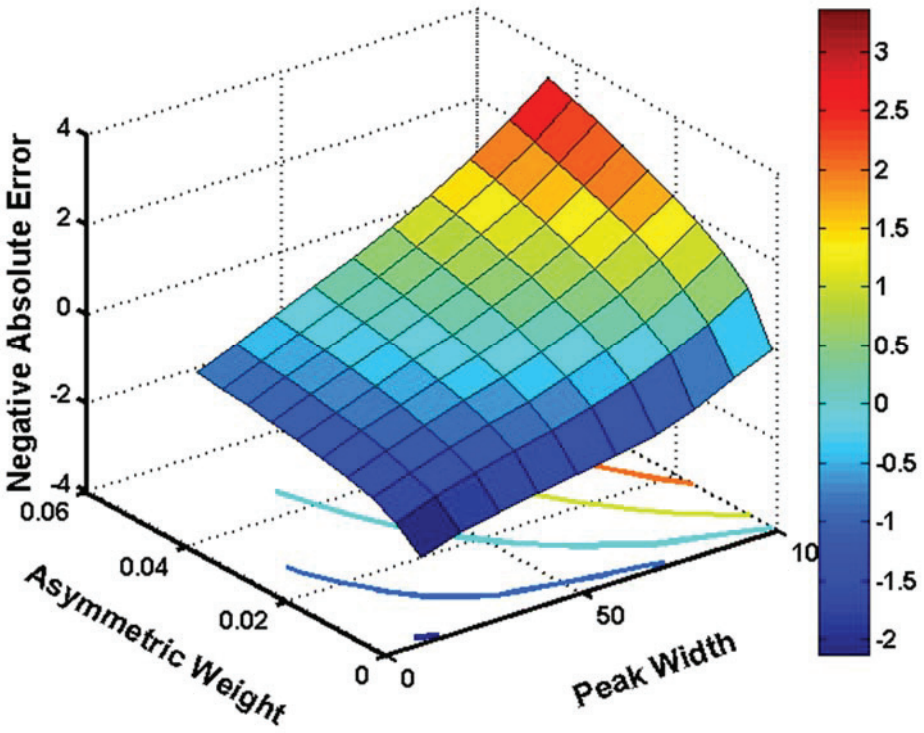

Figure 3.12 Surface and contour plots of estimation errors in peak heights for signals with different peak widths using asymmetry parameters.

baseline with $n_{freq} = 4$. It is apparent from the surface and the contour plots that the modeling of small peaks required relatively higher asymmetric weight, whereas for wider peaks, a smaller asymmetry parameter provides a good peak height estimation.

Figure 3.13 presents a line plot of these results for greater clarity. It is observed from this plot that each error estimation line starts from a negative value and gradually increases with the increment in peak widths. It is also apparent that for peak widths close to 100, smaller asymmetry parameter value (0.001) provides zero absolute error and the largest asymmetry parameter value (0.046)used in this study gave zero error for signal having peak width of 30.

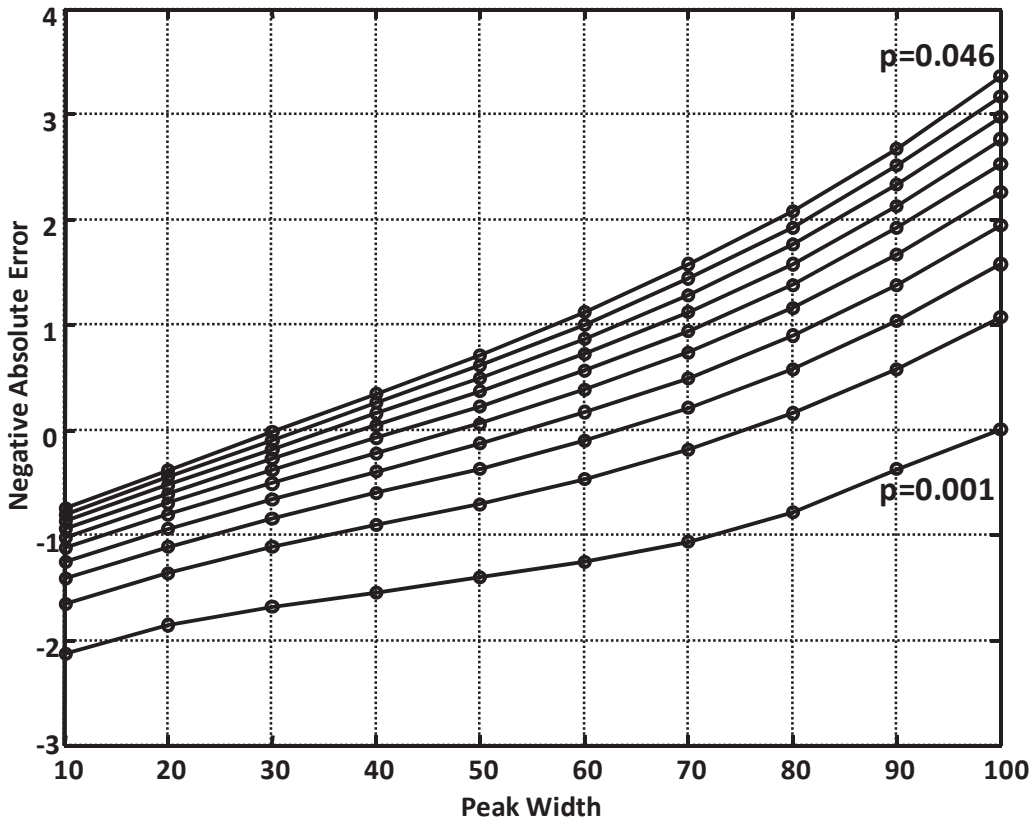

Figure 3.13 Negative absolute errors using different asymmetry parameter values at different peak widths.

The decrease in the optimum asymmetry factor with increasing peak width is analogous to the effects observed for peak height. Wider peaks, like higher peaks, increase the positive signal regions of the objective function higher, thereby pulling the estimated baseline higher. The asymmetry factor balances this effect against the natural tendency of the baseline estimates (in the absence of analyte peaks) to be low, so the optimum will depend on the nature of the analyte signal. As in the previous section, however, the range of errors for this parameter is not very large, so baseline estimation is not critically dependent on its value.

Figure 3.14 represents the results of the simulated signals using different parameters of TFALS for the baseline removal. Figure 3.14a shows the raw signal vector having a narrow peak (peak width=10) with true baseline (green) along with estimated baseline using $n_{freq} = 4$ (blue) and $n_{freq} = 11$ (red) and both estimated with $p = 0.021$. It is clear that both estimates are similar for small peak width. In contrast, Figure 3.14b shows the same results with a wider peak (peak width=100). For wider peak, it is clear that the larger number of basis function ($n_{freq} = 11$) over-fits the peak and provides poor baseline estimation. This effect of peak modeling is also present for narrow peak but is extensively apparent with the increment in peak width. Therefore the choice of optimal number of frequencies, n_{freq}, is not relatively dependent on the peak widths, however the wrong choice of n_{freq} could leads to an incorrect baseline estimation and peak stripping, this effect becomes critical in the signals having wider peaks particularly.

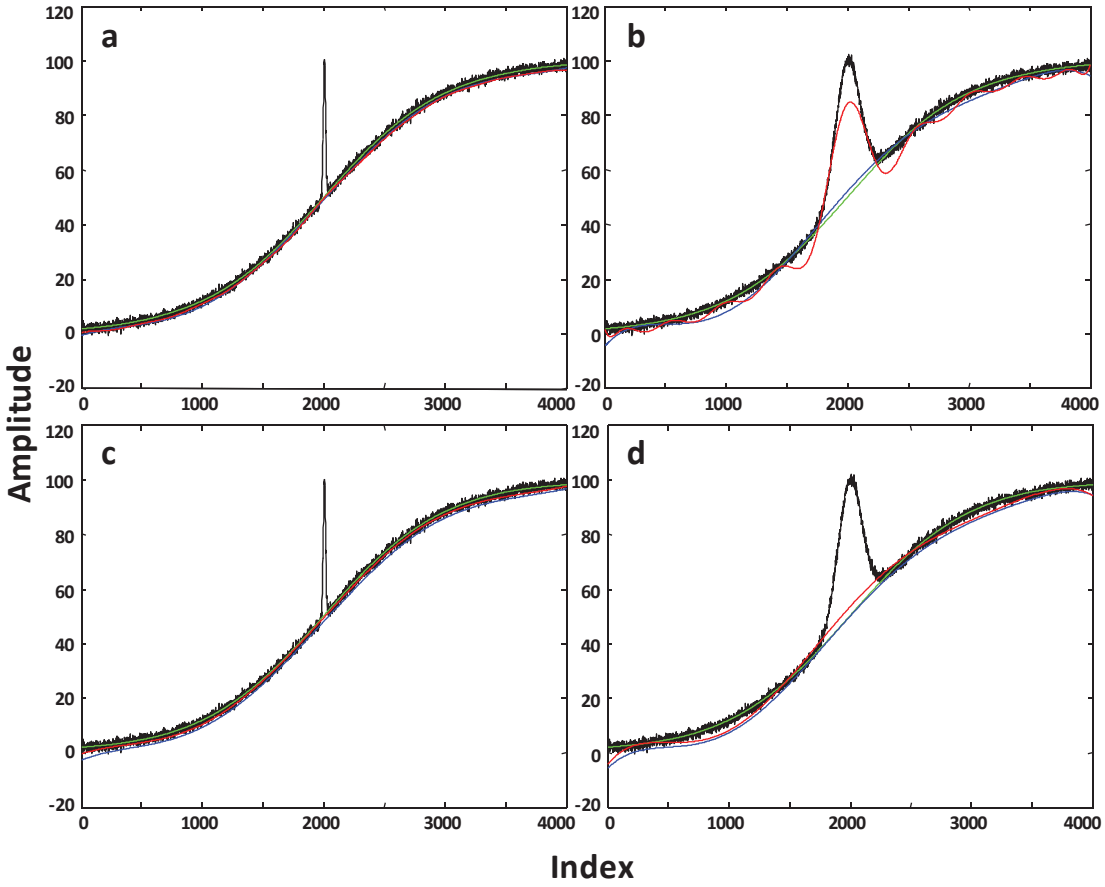

Figure 3.14 Signal vectors (——) for narrow peak (a, c) and widest peak (b, d) along with true (——) and estimated baselines. Figure (a) and (b) show the effect of using $n_{freq} = 4$ (——) and $n_{freq} = 11$ (——). Figure (c) and (d) show the effect of $p = 0.001$ (——) and $p = 0.046$ (——).

Figure 3.14c represents the effect of asymmetry parameter, p, on the narrow peak with the asymmetry values $p = 0.001$ (blue) and $p = 0.046$ (red) compared to the true baseline (green), using $n_{freq} = 4$. Figure 3.14d shows a similar plot for the wider peak. By comparing the two plots, it can be clearly observed that for the narrow peak red line $(p = 0.046)$ is relatively closer to the true baseline but the blue line $(p = 0.001)$ still seems to provide an acceptable result, whereas for wider peak, the blue line ($p = 0.001$) seems to provide a very good peak height estimation and red line ($p = 0.046$) seems not

an ideal choice. It is clearly seen from both of these plots that the difference in peak height estimation using two different asymmetry values is not crucially different. Therefore, the choice of optimal asymmetry parameter is relatively robust and independent of peak widths.

3.3.2.1.3 Effect of Peak Position on Baseline Estimation

To test the TFALS parameters relationship with peak location, ten simulated signal vectors with identical levels of noise and baseline and different peak locations were used and tested individually as in the previous sections.

Figure 3.15 shows surface and contour plots of negative absolute errors in peak

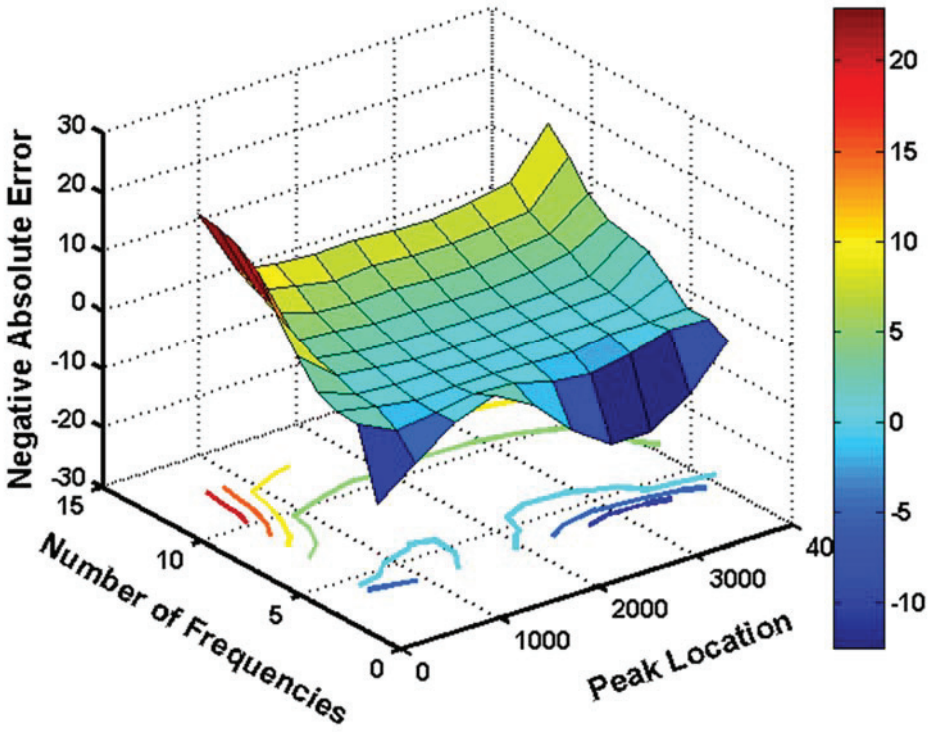

Figure 3.15 Surface and contour plots of estimation errors in peak heights for signals with different peak locations using different numbers of frequencies.

height estimations for ten different peak locations as a function of the number frequencies used to model the baseline with $p = 0.021$. It appears on initial inspection that the optimum number of frequencies is the same in this case as for the previous sections, which is not surprising given the baseline and peak shapes are the same. However the estimated error seems relatively higher for the peaks located at both ends, and this effect seems more obvious using a higher number of frequencies, n_{freq}, to model the baseline via TFALS method.

Figure 3.16 presents the same results as a line plot. It is apparent that the peaks at both ends of the signal range give relatively higher estimation errors using individual n_{freq} values. For this specific data set, the third line from the bottom seems to be an

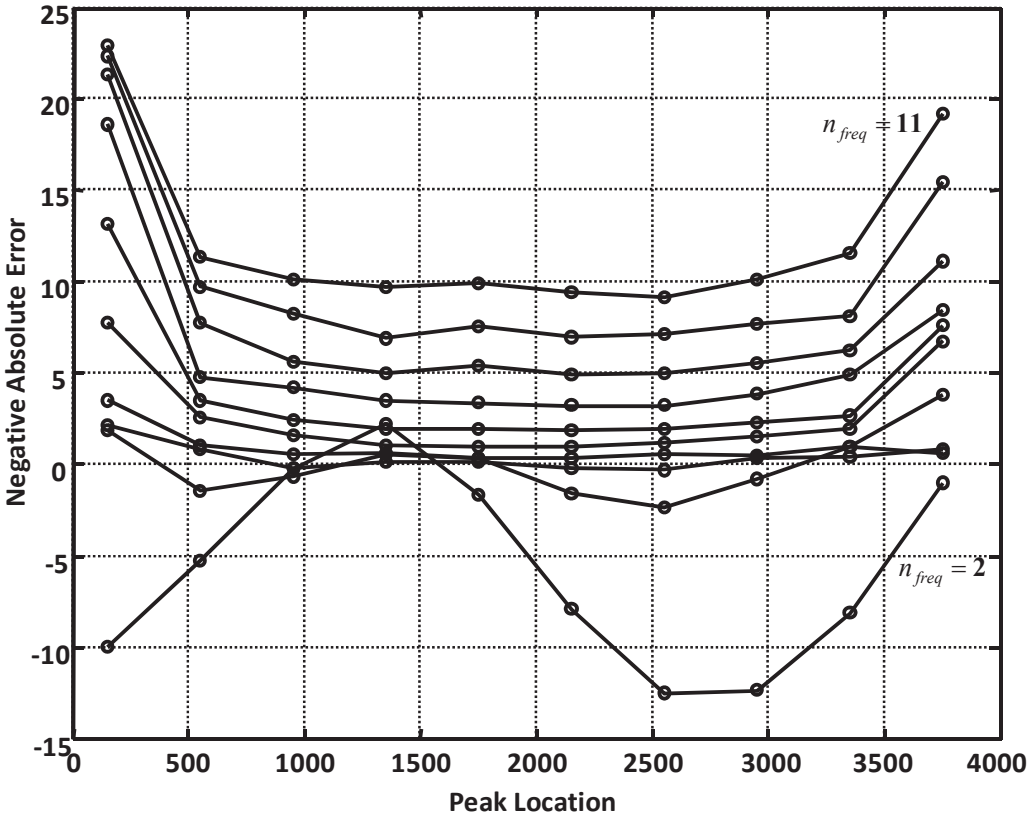

Figure 3.16 Negative absolute errors using number of frequencies from 2 to 11 at different peak locations.

optimal choice, having errors nearest to the zero line across the range, representing the $n_{freq} = 4$ as in the previous sections.

The pattern of behavior in this case is somewhat different from those in the previous sections in that there is no constant trend. Ideally, we would expect that all of the lines should be flat, since the same peak shape is being used in each case and only the location is changing. This is generally the case for peaks toward the middle of the range only, as expected, there is an increasing positive error in the baseline as the number of frequencies is increased and the algorithm attempts to fit the analyte peak. These errors are generally larger for peaks near the ends. The reason for this is the sinusoidal nature of the basis functions. When high frequency components in the baseline model adapt to fit the peaks, this is usually accompanied by negative-going oscillations on either side of the peak that penalizes the ALS objective function and reduces the extent of analyte peak fitting. At the limits of the range, however, one side of the oscillations is missing, so the baseline is permitted to extend higher into the peak, giving larger positive errors.

Another interesting feature of Figure 3.16 is the oscillatory nature of the errors when the number of frequencies is insufficient to model the baseline ($n_{freq} = 2,3$). In these cases, the baseline is under-fit, so the fitted baseline will make wide sweeps above and below the true baseline. Small errors will result if a peak is near one of those crossing points, but otherwise the errors will be large.

Figure 3.17 presents a surface and contour plots of negative absolute errors in the peak height estimation for ten different peak locations as a function of asymmetry parameter, p, values (ranging from 0.001 to 0.046 with an interval of 0.005) used to

model the baseline with $n_{freq} = 4$. It is apparent from the surface and the contour plots that the estimation error changes abruptly with the peak locations and there is no specific relationship between the peak locations and asymmetry parameter, p.

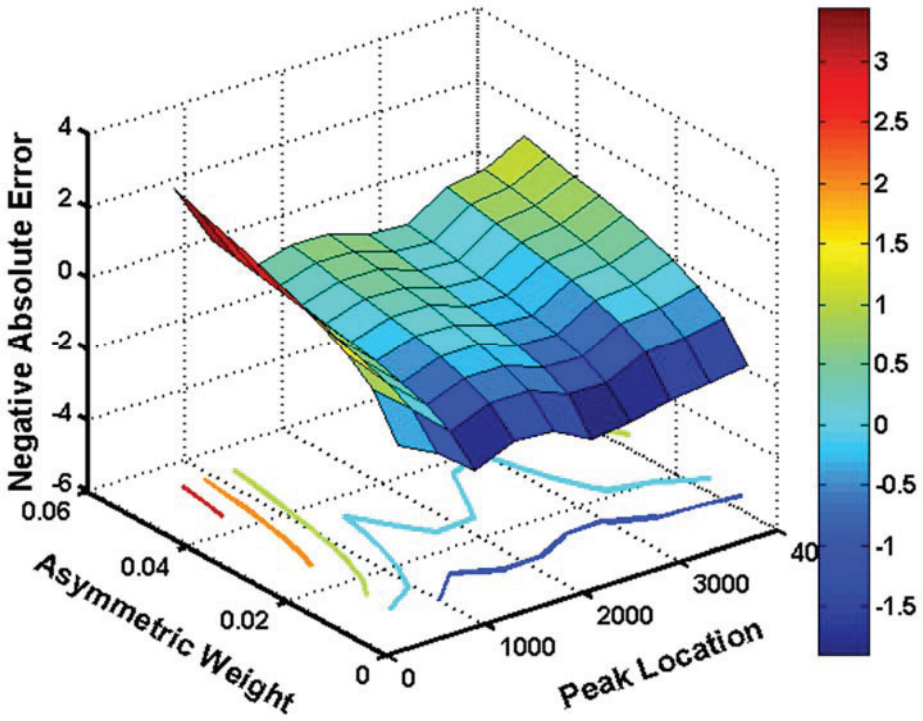

Figure 3.17 Surface and contour plots of estimation errors in peak heights for signals with different peak locations using different asymmetry parameter values.

Figure 3.18 shows a line plot of the results in Figure 3.17. It is observed from this plot that each error estimation line gives a relatively more positive estimation error for the peaks closer to the ends. The behavior observed in this plot is closely related to that observed for Figure 3.16. For peak positions near the middle, the baseline error behaves as expected. For small values of p, baseline estimates are low, and these shift upwards as the asymmetry parameter is increased. The same trend is observed for peaks near the

ends, but they start at a more positive value. This is because , as before, the terminal positions allow the sinusoidal terms more freedom in modeling the analyte peak, since

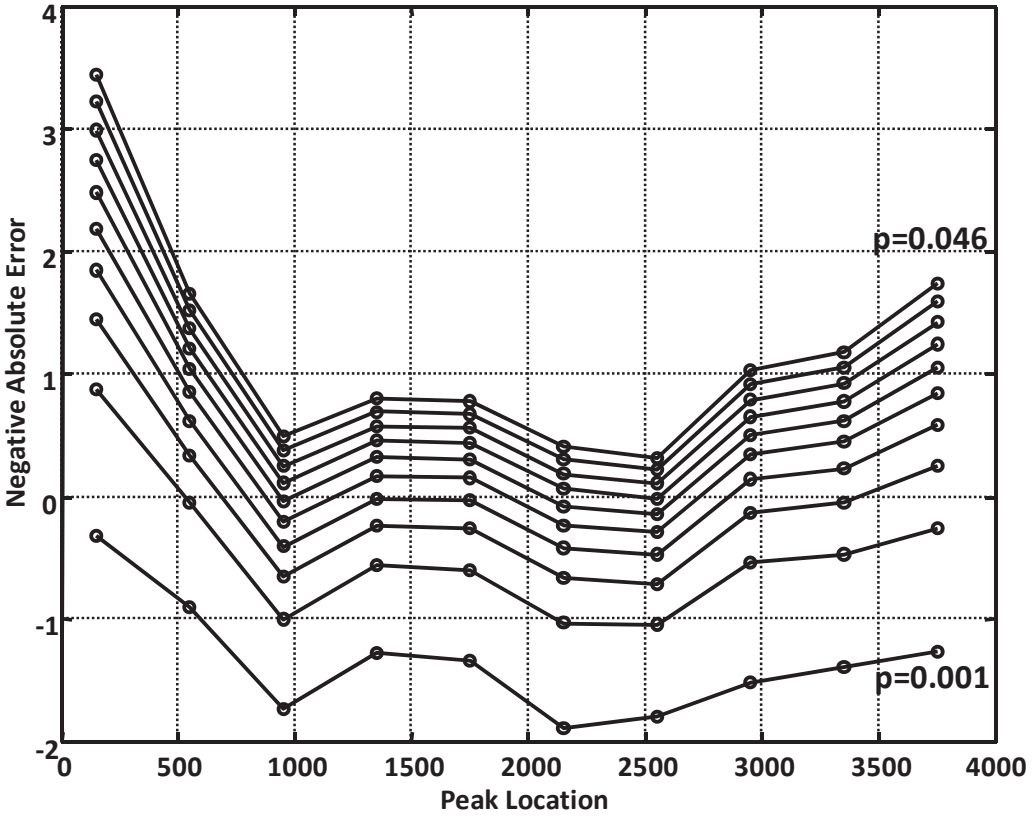

Figure 3.18 Negative absolute errors using different asymmetry parameter values at different peak locations.

they are not constrained to fitting a baseline on two sides. Therefore, the estimated baseline will tend to go higher in these regions. Despite this pattern of behavior, however, it should once again be noted that the magnitude of the effect introduced by the asymmetry parameter is quite small.

Figure 3.19 presents some results to support the observations. In each subplot, the raw data are shown together with the true baseline. Figure 3.19a – c illustrate the effect of the number of frequencies selected, with the calculated baseline for $n_{freq} = 4$ shown

in blue and $n_{freq} = 11$ in red ($p = 0.021$). It is clear in all three subplots that the blue line fits the true baseline fairly well, but the red line uses excessive number of functions and attempts to fit the analyte peak to reduce the objective function. In doing so, negative oscillations on either side of the peak, as in Figure 3.19b, generate positive residuals that limit the extent of fitting that can occur. However, at the end locations, shown in Figure 3.19a and c, one side of these oscillations is removed, allowing the baseline to extend higher into the peak. This is the reason that a greater positive error is observed when the peaks are located at the end.

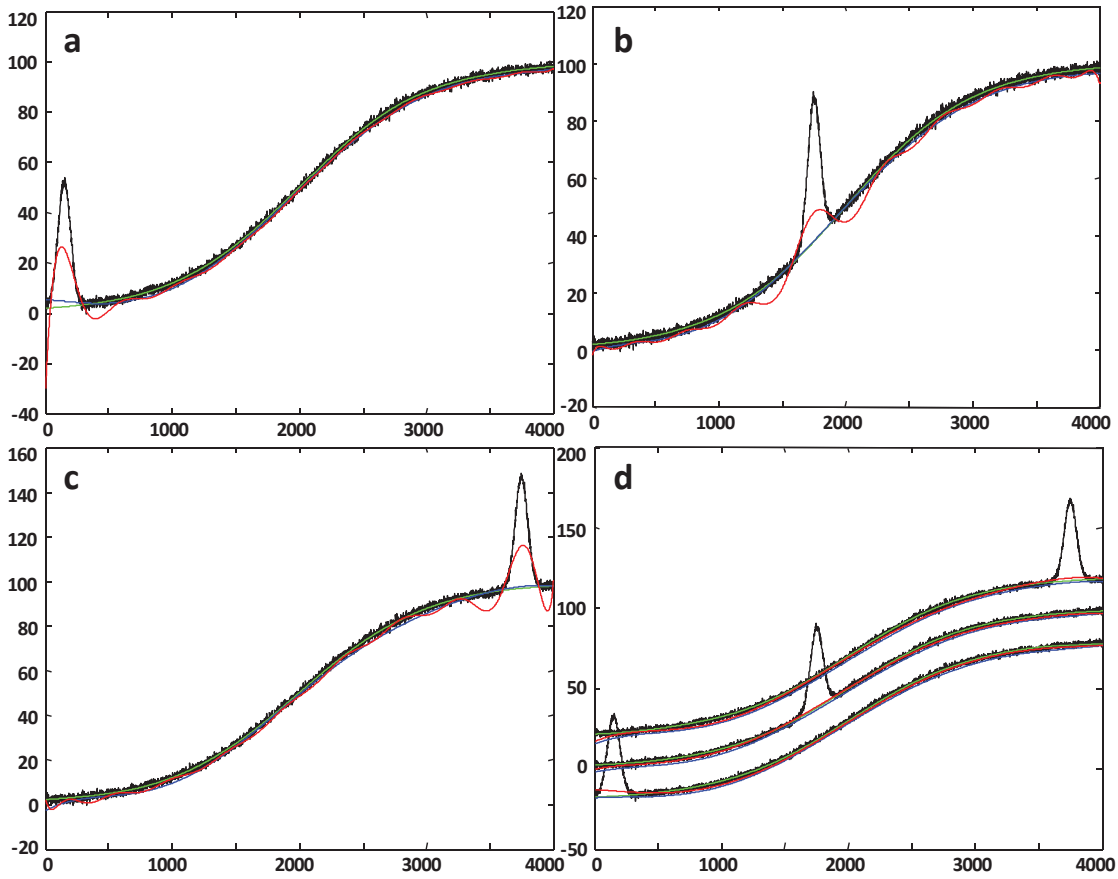

Figure 3.19 Baseline fit for different peak positions. **(a) - (c)** show the effect of the number of frequencies with $n_{freq} = 4$ (blue) and $n_{freq} = 11$ (red) compared to the true baseline (green). **(d)** shows the effect of the asymmetry parameter with $p = 0.001$ (blue) and $p = 0.046$ (red) compared to the true baseline (green).

Figure 3.19d illustrates the effect of the asymmetry parameter at the three locations, with $p = 0.001$ (blue) and $p = 0.046$ (red) for $n_{freq} = 4$. As discussed previously, increasing p raises the baseline, but the effects are larger near the ends of the signal range, at both extremes, however, the baseline estimates are still reasonable.

3.3.2.2 Asymmetric Weight Relationship to Signal-to-Noise Ratio

To examine the relationship of the signal-to-noise ratio to the asymmetric weighting parameter, twenty signals were generated with different signal-to-noise ratios. The baseline for all signals was kept at zero and only one peak with the same height $(h = 500)$, standard deviation ($\sigma = 40$) and location (centered at 1000 points) was used. Only the noise levels were changed to get different vectors having different signal-to-noise ratio. Baselines for each of the fifteen signals were estimated with 40 different asymmetry parameters, p. The values of p were used between 0.005 to 0.02 with an interval of 0.001, and between 0.02 to 0.068 with an interval of 0.002. In applying TFALS, the number of frequencies used was $n_{freq} = 1$, since no baseline function was added.

Figure 3.20 shows a plot of the "Negative Absolute Error" (NAE) as a function of the S/N ratio (reciprocal of relative standard deviation) and asymmetry parameter. As observed previously, increasing p raises the baseline from its initial point (positive of negative errors) to more positive values. For low S/N, the larger values of p give better results, but the effects are very small. For high S/N, however a large value of p causes a positive baseline error of about 10 times the measurement noise and this will increase

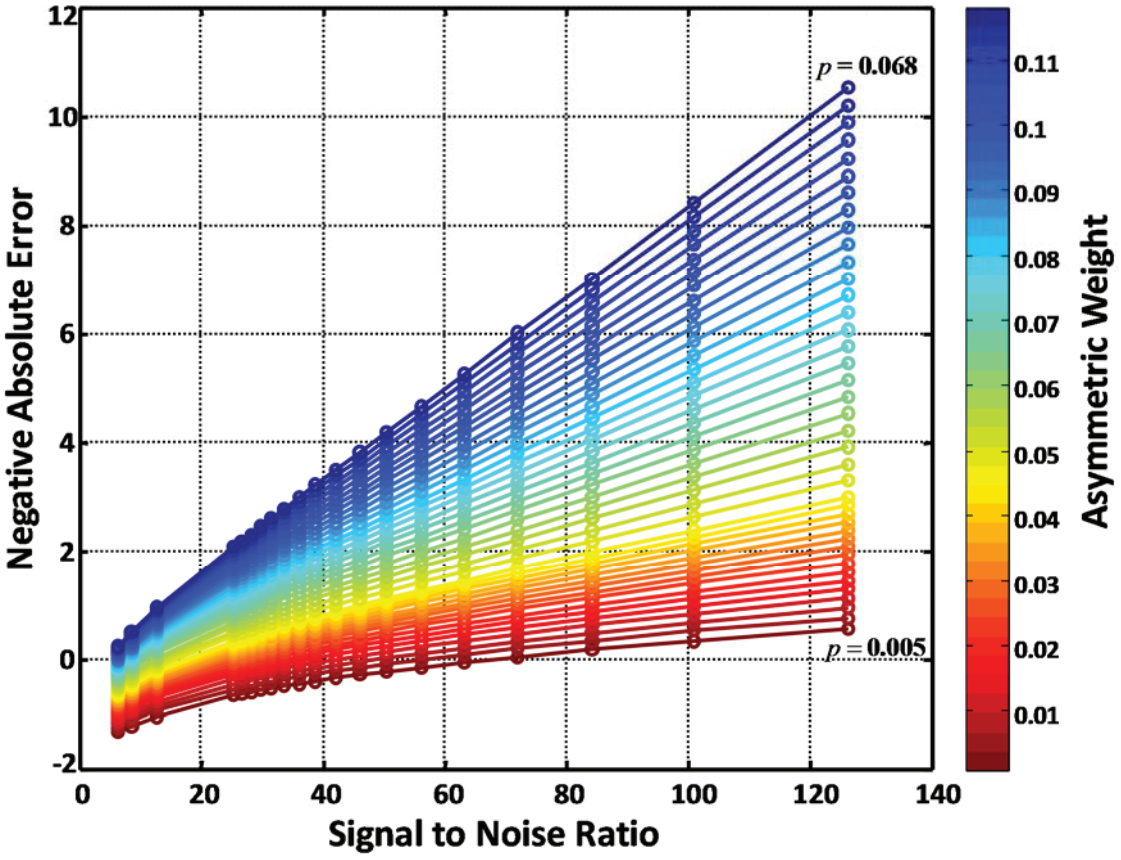

Figure 3.20 Results of asymmetric weight dependency on signal to noise ratio.

linearly with the S/N. Therefore in general, it is advisable, when uncertain, to choose a small value of p. While this is more likely to produce a baseline that is underestimated over the range of the measurements, it will tend to keep the absolute errors more constant relative to the baseline noise.

3.3.2.3 Relationship of TFALS Parameters

So far, the studies carried out have varied the two adjustable TFALS parameters independently, fixing one and examining the behavior of the other. To test the interdependence of the two parameters on each other, two sets of signals were generated

with the same Gaussian baseline and low and high signal-to-noise ratios, and two sets of signals were generated with the same exponential baseline and low and high signal-to-noise ratios. The same set of five Gaussian analyte peaks (height=585, 243, 279, 522, 315; standard deviation=21, 11, 12, 20, 14; location=200, 550, 900, 1300, 1750) were used for each signal vector. The noise standard deviations were $\sigma = 4$ (high S/N) and $\sigma = 40$ (low S/N).

Figure 3.21 shows the results of simulation for the Gaussian baseline and a low S/N. In this case, the results are plotted as the RMS error (RMSE) in the baseline as a function of the number of frequencies, n_{freq} and the asymmetry parameter, p, where the RMSE is defined as,

$$RMSE(\text{baseline}) = \sqrt{\frac{\sum_{i=1}^{N}\left(b_{est} - b_{true}\right)_i^2}{N}} \tag{3.33}$$

Here $\left(b_{est} - b_{true}\right)_i$ is the difference between the estimated and the true baseline at channel i. The RMSE is a measure of the "average" difference between the estimated and the

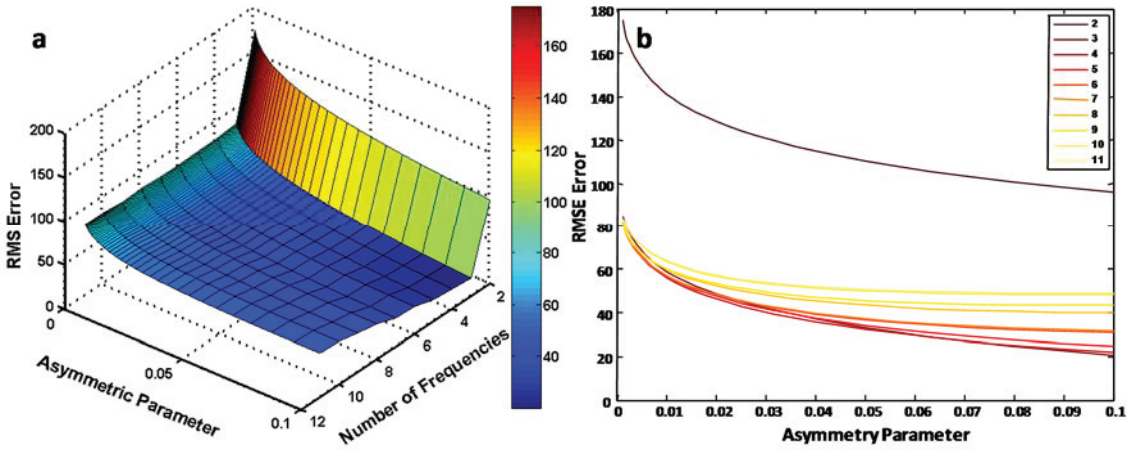

Figure 3.21 **(a)** Surface and **(b)** line plot of RMS error in baseline estimation for signal with Gaussian baseline and low signal-to-noise ratio using different values of TFALS parameters.

true baseline and ideally has a value of zero, but this would not be the case even in the absence of analyte peaks and noise, since the baseline functions cannot model the baseline exactly.

Several points are worth noting regarding Figure 3.21. First, although there is a relationship between the two parameters, the interaction is small and they are largely optimized independently. Second, the dependence on asymmetry parameter is relatively small, while the errors introduced by an incorrect number of frequencies are larger, especially when an insufficient number of frequencies has been employed. A third point worth noting is that, when a sufficient number of frequencies is used ($n_{freq} = 3$ to 5) a minimum RMSE has not been achieved even when $p = 0.1$, which is larger than the values that have been used thus far. This is contrasting however, because the current figure measures the agreement with the true baseline, whereas previous studies considered the errors in peak height. When the asymmetry parameter is low, the estimated baseline will tend to "hug" the negative transients of the noise. This gives an error on the order of the noise and is consistent with the results shown here, where an RMSE of about 50 at $p = 0.01$ corresponds roughly to the noise level of 40 as p is increased, the baseline will rise closer to the true value, but at the expense of errors in the peak heights as previously shown. Therefore, a perfect match with the baseline is not necessarily desirable. Figure 3.22 shows the same results for the Gaussian baseline in the high S/N ratio situation, and similar observations can be made. As before, a higher dependence on number of frequencies than on the asymmetry parameter is observed, and the intersection between the parameters is small. It should also be noted that the RMS errors in the baseline are smaller and pass through a minimum within the range of

asymmetry parameters used (unlike the previous case). These observations are consistent with the lower noise level, which places the "bottom" of the noise closer to the true

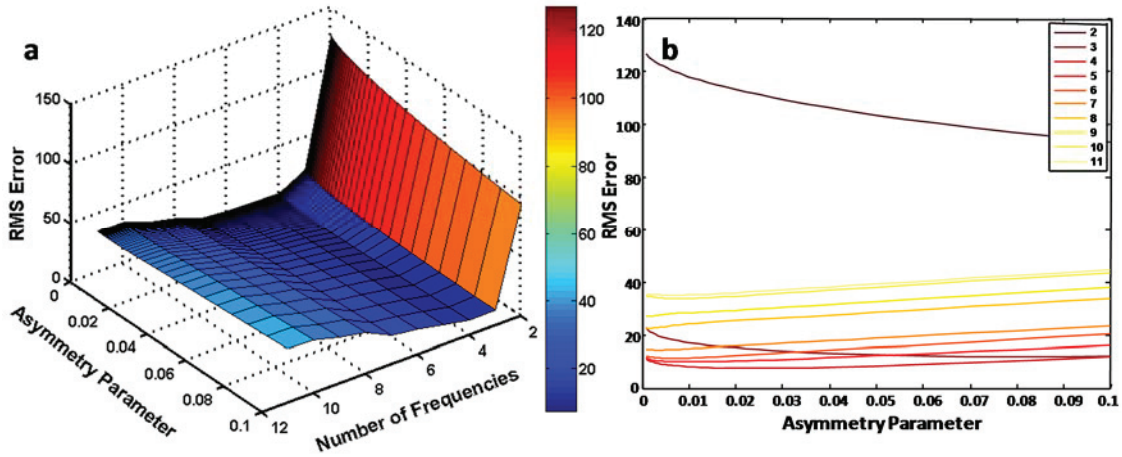

Figure 3.22 **(a)** Surface and **(b)** line plot of RMS error in baseline estimation for signal with Gaussian baseline and high signal-to-noise ratio using different values of TFALS parameters.

baseline to start and also requires less "adjustment" by the asymmetry factor to move it even closer. The lowest level of RMSE also approaches the noise level ($\sigma = 4$). However, even if the noise level becomes very small, there is a limit beyond which the RMSE cannot reach, as it is determined by the lack of fit to the true baseline.

Figure 3.23 and Figure 3.24 show results analogous to those in Figures 3.21 and 3.22, but using an exponential instead of a Gaussian baseline. The interpretation of these results is consistent with those presented earlier and shows that the general behavior is not tied to the type of baseline, although the optimum number of frequencies may vary with the baseline characteristics.

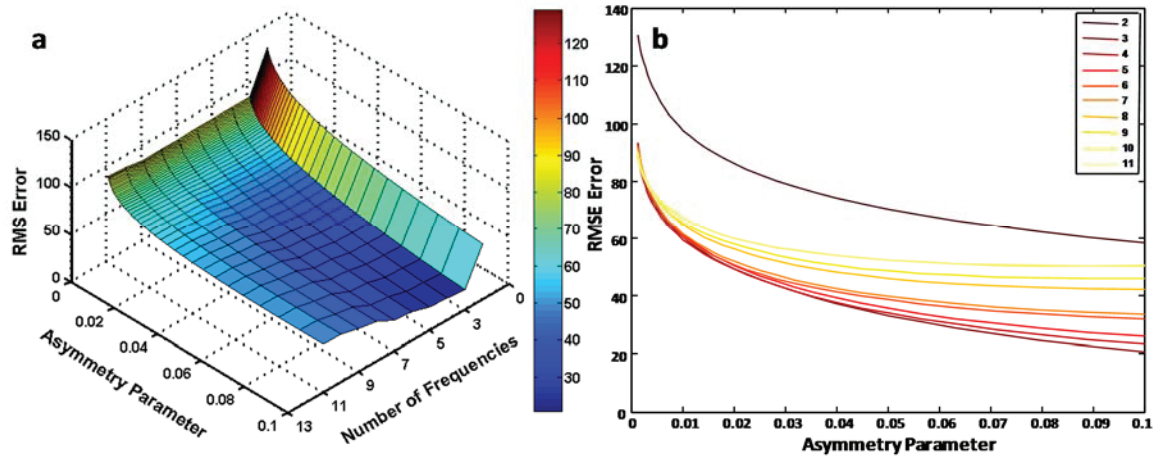

Figure 3.23 **(a)** Surface and **(b)** line plot of RMS error in baseline estimation for signal with exponential baseline and low signal-to-noise ratio using different values of TFALS parameters.

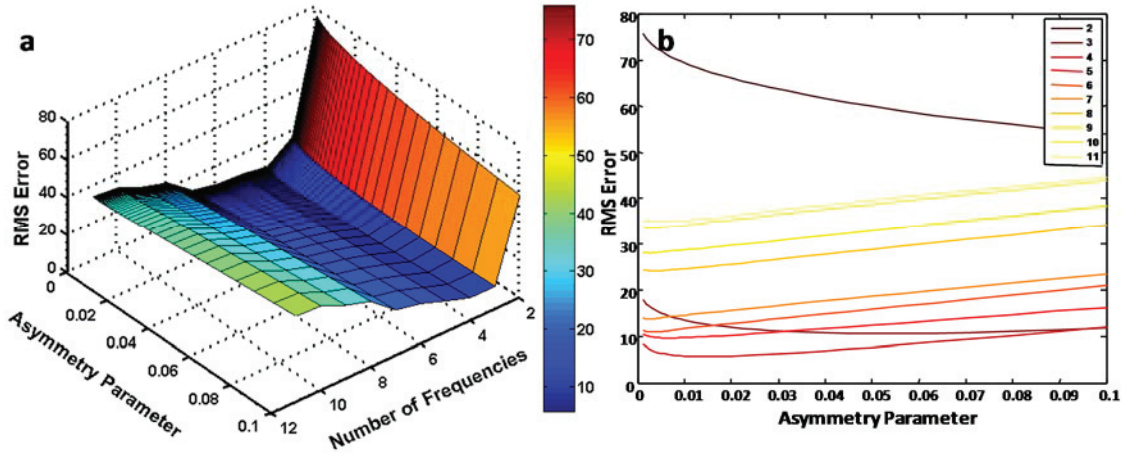

Figure 3.24 **(a)** Surface and **(b)** line plot of RMS error in baseline estimation for signal with exponential baseline and high signal-to-noise ratio using different values of TFALS parameters.

It is clear from these studies that the two parameters are not highly dependent on each other. While there is always an optimal value for each parameter for a particular data set to obtain the best baseline corrected signal, the choice of the asymmetry factor is not very crucial, although it does depend to some extent on signal-to-noise ratio. The

choice of the number of frequencies is most closely connected to the nature of the baseline

3.4 Conclusions

The TFALS method uses only two adjustable parameters to estimate low frequency baselines and these two parameters are not highly dependent on signal attributes. The first parameter, the number of Fourier frequencies used, is completely dependent on the nature of the baseline. It has been shown from the results presented here that the selection of number of frequencies is largely independent of the analyte signal. Therefore, slight changes in the amplitude of baseline or signals of the analyte, which can be commonly seen in replicate data signals or signals of similar analytical samples taken at the same analytical conditions, would provide the same optimal results with the same number of frequencies. Further results of experimental data will be provided in Section 4.4.1 to test this hypothesis.

The results presented in this study also demonstrated that the second adjustable parameter, the asymmetry parameter is also independent of the analyte signal itself although it has a small dependence on signal-to-noise ratio. It has been noted that the signals with lower signal-to-noise ratio require relatively higher asymmetric weight for better estimation, whereas data with higher signal-to-noise ratio can normally be estimated with a very small asymmetric weight. It has also been found that the two tweaking parameters are not dependent on each other and that acceptable results can often be obtained with a range of parameters.

The application of the optimal selection of these parameters for different types of simulated baseline functions and different experimental signals will be presented and further discussed in Chapter 4.

Chapter 4
Applications of TFALS

4.1 Introduction

In chemical signal analysis, the point of interest for quantitation is generally the acquisition of accurate analyte peak height, peak area, or the signal amplitude information. To achieve acceptably accurate quantitative information, elimination of unwanted high frequency noise and high amplitude, highly variable, low frequency baseline components is necessary as discussed in Chapter 1. Since the baseline is relatively variable over the duration of the acquired signal, it can complicate both the absolute and relative (e.g. peak height ratios) quantitation of the analyte. To avoid this complication it is often necessary to choose an approach that can approximate the baseline as closely as possible over the entire signal duration and especially over the region of analyte peaks to provide accurate analyte signal for quantitation.

Many approaches published in the literature to estimate the low frequency baseline have been already discussed in Chapter 1 along with their strengths and weaknesses. Baseline estimation using asymmetric least square regression is a relatively recent approach that has been extensively studied and highly reported during the last decade. Since this approach has been studied by many groups and the proposed method in this thesis is also based on the same approach, this chapter will provide the comparison of proposed method with two existing and frequently cited asymmetric least squares based methods in the literature; asymmetric least squares smoothing (ALSS), [33], and adaptive iteratively re-weighted penalized least squares (airPLS), [37].

This chapter is divided into two parts. The first part provides a comprehensive simulation study for the qualitative and quantitative comparison of TFALS with the ALSS and airPLS approaches. Tables and figures are also included to validate the results of individual approaches for each data set. The second part of this chapter consists of the application of TFALS to experimental data sets. Results of different analytical signals are included to validate the baseline approximation appropriateness by TFALS, and for several data sets these results are qualitatively compared to published results from airPLS.

4.2 Experimental

4.2.1 Computational Aspects

All data processing was carried out using programs written by the author in MatLab® 2010b (MathWorks, Natick, MA) under Windows 7 Professional 2009© on a 2.10 GHz processor with 2.00 GB of memory. The two comparison algorithms were downloaded from online sources [122, 123].

4.2.2 Simulated Data Sets

Simulated data sets were intended to provide a qualitative and quantitative comparison of the TFALS with the two comparator methods. The nominal signal consisted of a vector of 2000 points with five Gaussian peaks (peak heights of 585, 243, 279, 522, 315; peak standard deviations of 28, 14, 16, 25 and 17; and peak locations of 200, 550,900, 1300 and 1750. Normally distributed random noise was added to the signal with a standard deviation of 6.

This signal was then superimposed on five different baseline functions: linear, sinusoidal, exponential, Gaussian and the sum of linear, Gaussian and exponential functions. The linear baseline function was generated with a slope of 0.174 and an intercept of 123.5 with x ranging from 1 to 2000. The exponential baseline was generated using a decay rate of 0.004 and an amplitude of 573. The sinusoidal baseline was generated using an angular frequency, 7.5×10^{-4} and an amplitude of 250. The Gaussian baseline was generated using a combination of two Gaussian functions (heights, $h=500,700$, standard deviations, $\sigma=750,250$) centered at indices 500 and 2200. A sum of previously generated linear, exponential and Gaussian functions is used as a combination baseline.

4.2.3 Experimental Data Sets

Four sets of experimental data were used to demonstrate the performance of proposed method for baseline elimination. The first data set included in this study consists of three consecutive Raman spectra of DNA on a gold surface. The gold surfaces used were Sphere Segment Void (SSV) nano-structured surfaces as described in the literature [124]. Following preparation, the surfaces were immersed in Tris buffer containing 1 M NaCl and 1 μM DNA, to allow adsorption of the DNA molecules to the gold surface [127]. The SERS (Surface Enhanced Raman Spectroscopy) data were collected using a Renishaw 2000 microscope with a 633 nm He-Ne laser, a 5 μm spot size and a 30 second accumulation time. These three spectra were considered as replicates and supposed to have similar but not identical baseline features. Therefore, this data set

was used to visualize the results and test the hypothesis that the same parameters can be used for replicate data.

The second data set consists of replicates of X-ray fluorescence spectra of tea [128]. This data set was also used to confirm that same TFALS parameter values provide the optimum results for replicates and also as an example of data taken from different analytical instrument source for baseline removal via TFALS. Since no standard data were available, only visual interpretation will be provided. The experiment was performed using a bench-top X-ray fluorescence instrument, the Shimadzu EDX 700 (Kyoto, Japan). For spectra acquisition, ≈ 200 mg of solid samples were put into a Teflon cell. This cell had an orifice diameter of 5 mm and samples were covered with a 3 μm thick Mylar film. In all cases spectra were recorded from 0 to 40 keV , with a resolution of 0.02 keV , resulting in 2047 points per spectrum.

The third data set, referred as the Raman minerals data, consists of Raman spectral data of six different minerals. These spectral data were taken from the *Handbook of Minerals Raman Spectra* [127]. The excitation wavelength was 514 nm for first five and 532 nm for the last spectrum, at an unstated power for this set of spectra. This spectral data set was used to demonstrate the application of proposed baseline elimination approach for different low frequency signals since these signals consists of visually different baseline artifacts. The first five minerals spectra have already been used by Rowlands for baseline removal [101].

The last set of data, taken from an online source [122], consist of NMR, chromatographic and Raman data, developed and used by Liang et al [37] for baseline

removal via the airPLS method. Three visually different Raman signals and one selected signal each for the NMR and chromatographic data are included to visually compare the results of two approaches and show the applicability of TFALS for experimental data.

4.3 Results and Discussion - Simulated Data Sets

Simulated data sets with different baseline functions were studied to test and validate the application of proposed baseline removal approach for different baselines. In addition, two other methods were compared in this study based on five different aspects; visual comparison of simulated data results, individual peak height estimation based on baseline approximation and total absolute error of peak height estimation, RMS error over the entire channel length of each signal, computation time for each estimation method, and number of adjustable parameters. Comparative results for each of these aspects will be presented and discussed in the following sections in detail.

4.3.1 Simulation Results - Qualitative

4.3.1.1 Visual Comparison

The baselines were estimated for all five simulated signals having different baselines using three methods: TFALS, airPLS and ALSS. The parameters for each signal were optimized using a grid search approach for all three baseline removal algorithms. The optimal parameters were chosen based on the minimum RMS error in baseline estimation (although this does not necessarily produce the smallest peak height errors). In this section, baseline estimation results of all five signals will be presented using three approaches by using their respective optimal parameters. The estimated

baselines, along with the raw signals and baseline corrected signals, will be presented for each signal using all three baseline approaches separately.

The first simulated signal is presented in Figure 4.1, having a linear baseline and uncorrelated white noise. The estimated baseline and baseline corrected signals using all three baseline estimation algorithms (TFALS, airPLS, ALSS) are shown in Figure 4.1. Figure 4.1(a) shows the result of TFALS, 4.1(b) shows the results of airPLS approach and Figure 4.1(c) is shows the baseline estimation and baseline corrected results for the linear baseline using asymmetric least squares smoothing approach. It is apparent from the estimated baseline and resulting corrected signals that all three approaches were able to provide very good results for the linear baseline. It is evident that the airPLS method forces the baseline to travel at the bottom the signal in contrast to the TFALS and ALSS approaches, where baseline seems to travel somewhere in the middle of the noise.

Figures 4.2(a), 4.2(b) and 4.2(c) represent the simulated signal with an exponential baseline and white noise, and the estimated and baseline corrected signals using TFALS, airPLS and ALSS algorithims respectively. All three methods produces a corner effect at the left hand side as a result of underestimating the baseline near channel 1. This is not surprising as the rapid decay of the exponential contains high frequency components not typical of a baseline. The greatest corner effect is exhibited by ALSS. TFALS also shows a small corner effect on the right hand side. All three methods tend to underestimate the baseline under the second peak by about the same amount. All three methods give similar good estimates of the baseline, but both airPLS and ALSS seem to slightly overestimate the baseline for peak 4 (where the baseline is nearly flat) and airPLS underestimates the baseline for peak 5.

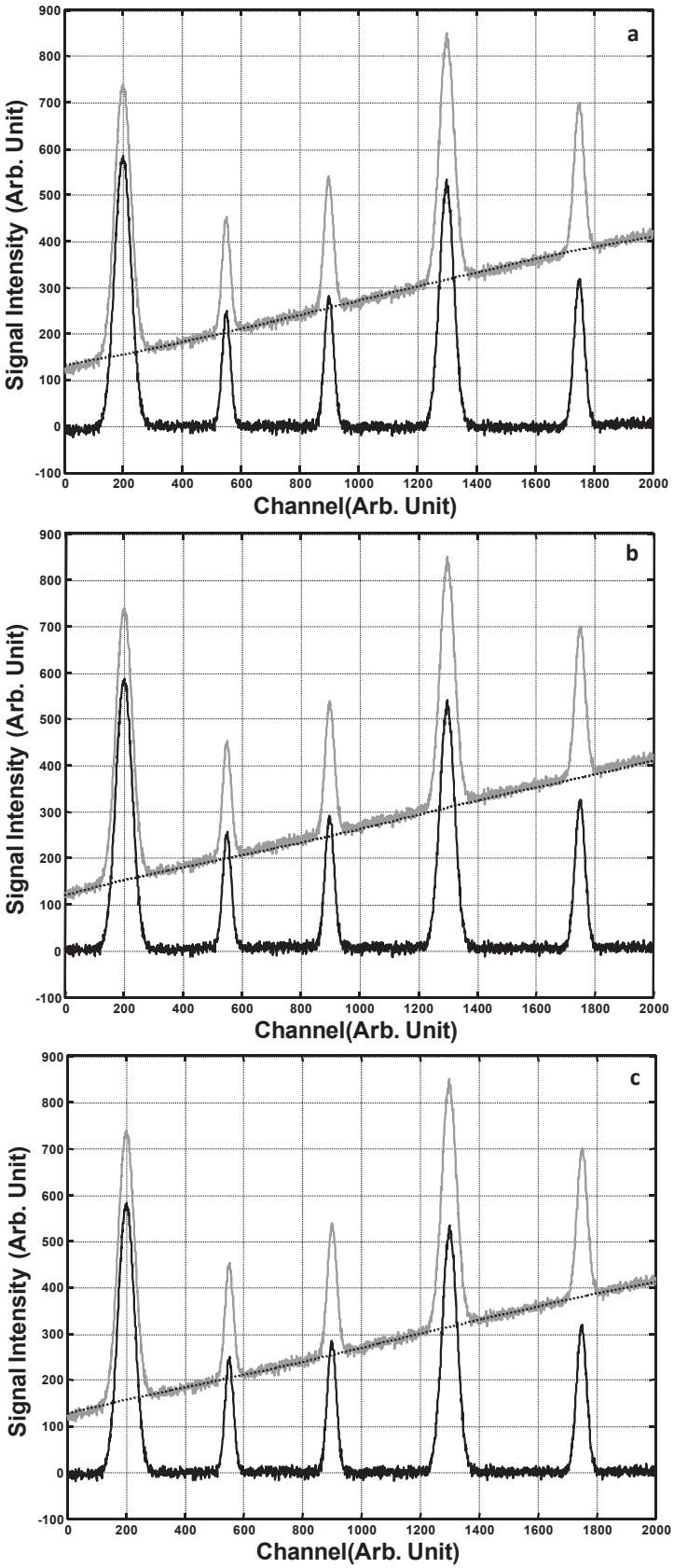

Figure 4.1 Raw signal with linear baseline, estimated baseline and baseline corrected signal using **(a)** TFALS **(b)** airPLS and **(c)** ALSS approaches.

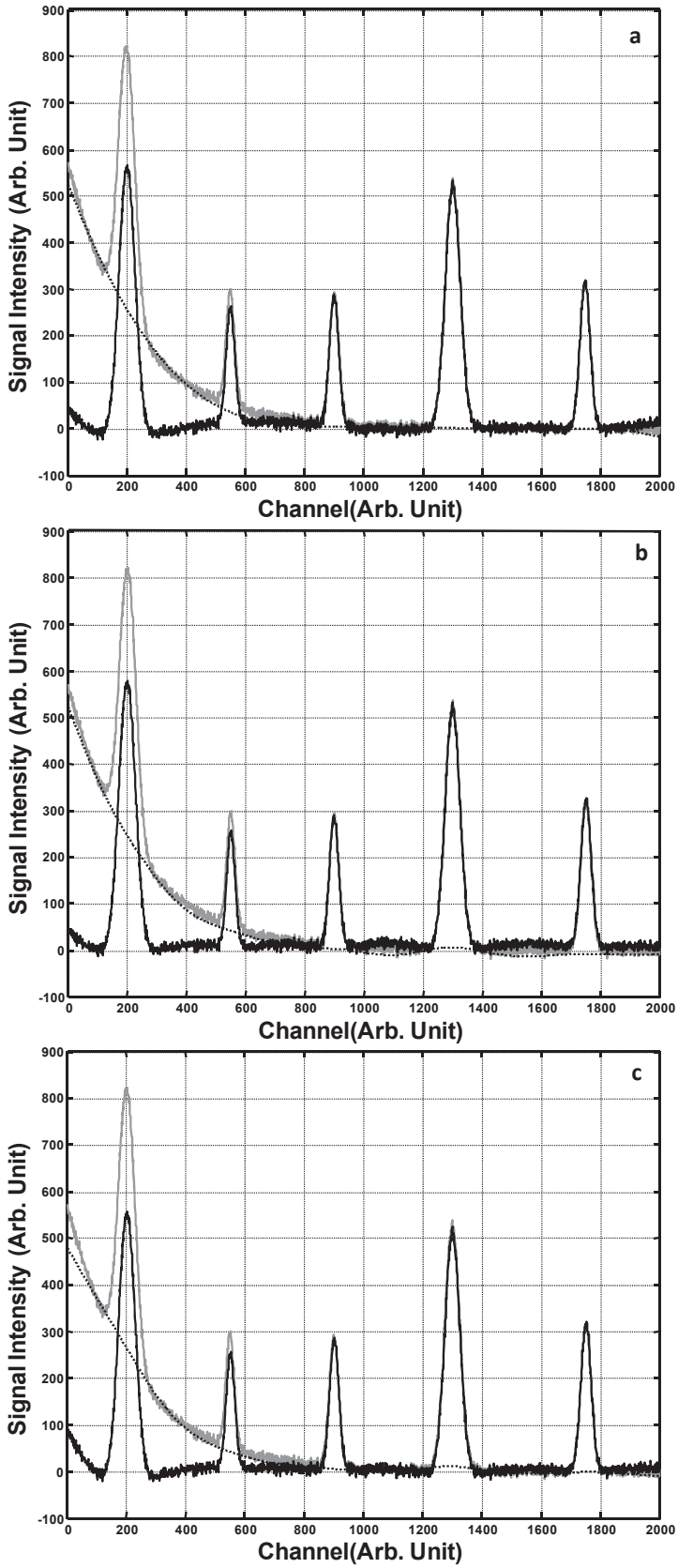

Figure 4.2 Raw signal with exponential baseline, estimated baseline and baseline corrected signal using **(a)** TFALS **(b)** airPLS and **(c)** ALSS approaches.

As before, airPLS tends to follow the bottom of the noise, especially between the peaks, and this has the effect of providing "humps" between the peaks. Overall, TFALS method seems to provide the most consistent modeling of the baseline between and under the peaks for this data set.

Figure 4.3 represents a simulated signal having sinusoidal baseline and white noise in gray. Dotted lines show the estimated baselines, and the baseline corrected signals are shown in black using TFALS, airPLS and ALSS approaches in Figures 4.3(a), (b) and (c) respectively.

It appears that all three approaches were able to estimate the sinusoidal baseline satisfactorily; however, airPLS again seems to force the baseline to stay at the bottom of noise whereas the estimated baselines for TFALS and ALSS travel closer to the middle of the noise, with ALSS perhaps the better of the two. Therefore, the estimated baseline via airPLS would probably affect the peak height estimation and give relatively higher peak height errors, as well as a higher RMS error in baseline estimation. These two errors in estimation will be analysed and presented in Sections 4.3.2 and 4.3.3.

Figure 4.4 presents a simulated signal having a combination of two Gaussians as a baseline with random noise. Dotted lines represent the estimated baselines, and the baseline corrected signals are shown in black using TFALS, airPLS and ALSS approaches in Figures 4.4(a), (b) and (c), respectively.

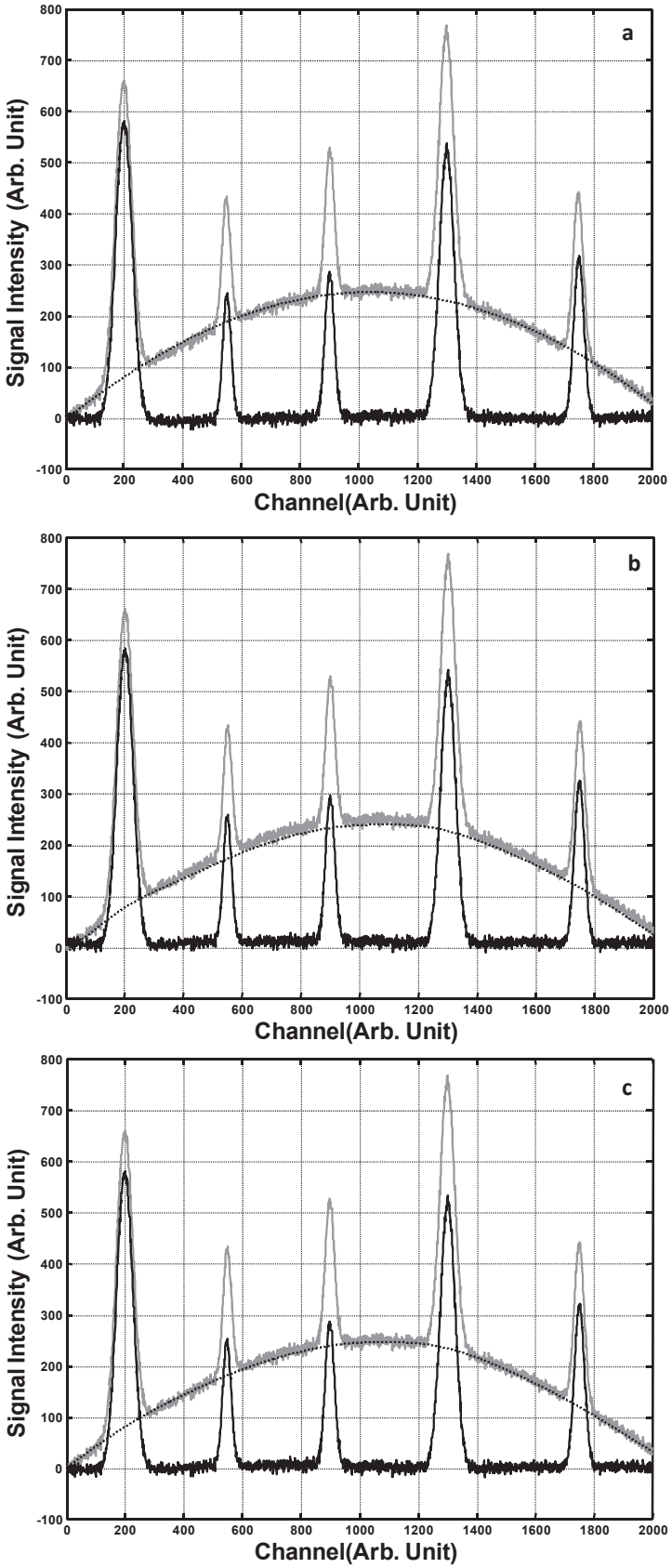

Figure 4.3 Raw signal with sinusoidal baseline, estimated baseline and baseline corrected signal using **(a)** TFALS. **(b)** airPLS and **(c)** ALSS approaches.

The baseline signal corrected using TFALS shows a minor corner effect on the left hand side and also underestimate the baseline slightly for the second and third peaks. Otherwise, the estimated baseline seems to pass through in the middle of the noise for the remainder of the signal.

The estimated baseline using airPLS, represented by the dotted line in Figure 4.4(b), once again seems to pass close to the bottom of the noise in the regions between the peaks and then rises slightly upward in the peak regions. This effect is more evident in wider peaks (first and fourth peaks from the left). It is the expected that this is the effect of second-order differential term in the objective function, since the trial baseline in the first step of airPLS algorithm uses equal weights in the penalized least squares. Due to this effect, the baseline corrected signal exhibits small amplitude arcs between the peaks and this might leads to higher errors in the peak height estimation. This aspect is discussed in Section 4.3.2.

The estimated baseline using ALSS method, represented by the dotted line in Figure 4.4(c), also seems to give this differential effect in first and fourth peaks from the left but the corrected signal shows smaller distortions between peaks than the airPLS corrected signal. This could be because the estimated baseline travels closer to the middle of the noise. The ALSS method also exhibits a corner effect, but in contrast to TFALS, it occurs on the right side.

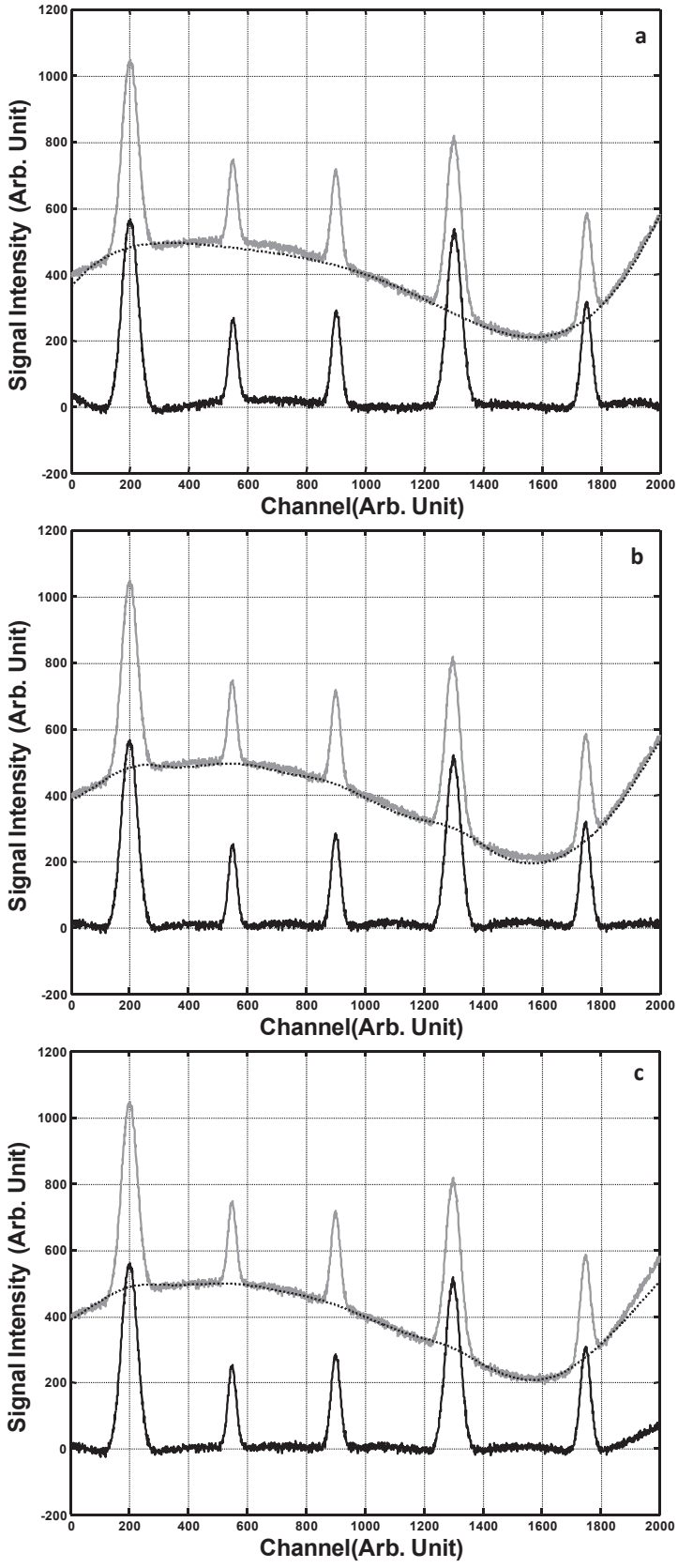

Figure 4.4 Raw signal with Gaussian baseline, estimated baseline and baseline corrected signal using **(a)** TFALS **(b)** airPLS and **(c)** ALSS approaches.

Figure 4.5 shows the raw signals with the combination baseline and random noise, along with the estimated and baseline corrected signals using TFALS, airPLS and ALSS algorithms. The baseline for this signal is a combination of linear, exponential and Gaussian functions. The dotted line in Figure 4.5(a) represents the estimated baseline using the TFALS method. It is the only one of the three methods that does not produce a noticeable corner effect on the left, but it does underestimate the baseline for peaks 2 and 3, as with the Gaussian baseline. Otherwise, the baseline seems to pass through the middle of the noise. The baseline subtracted signal, represented in black, confirms these observations. The estimated baseline using airPLS approach is represented by the dotted line in Figure 4.5(b). A relatively high corner effect is observed on the left and once again in this case the baseline follows the bottom of the noise between the peaks, leading to distorsions in the corrected signal. Overall, however, the baseline under the peaks is well-estimated. The estimated baseline using the ALSS approach is almost identical to that estimated by airPLS in this case, so the same comments apply. In this combined baseline case, the estimated baseline seems to stay at the lower edge of the signal because the smallest RMS error in baseline estimation was provided by using an asymmetry parameter value that was very small (0.001) and order of difference was 3. Th small asymmetry parameter tends to push the baseline down, as previously noted.

In overall visual comparison, all three baseline correction methods appeared to give generally satisfactory results. Although certain methods appeared to work better in certain cases, it is difficult to make definitive conclusions based on visual observations. In several cases, corner effects were observed, but no one method was superior in this regard and the effect is mainly aesthetic.

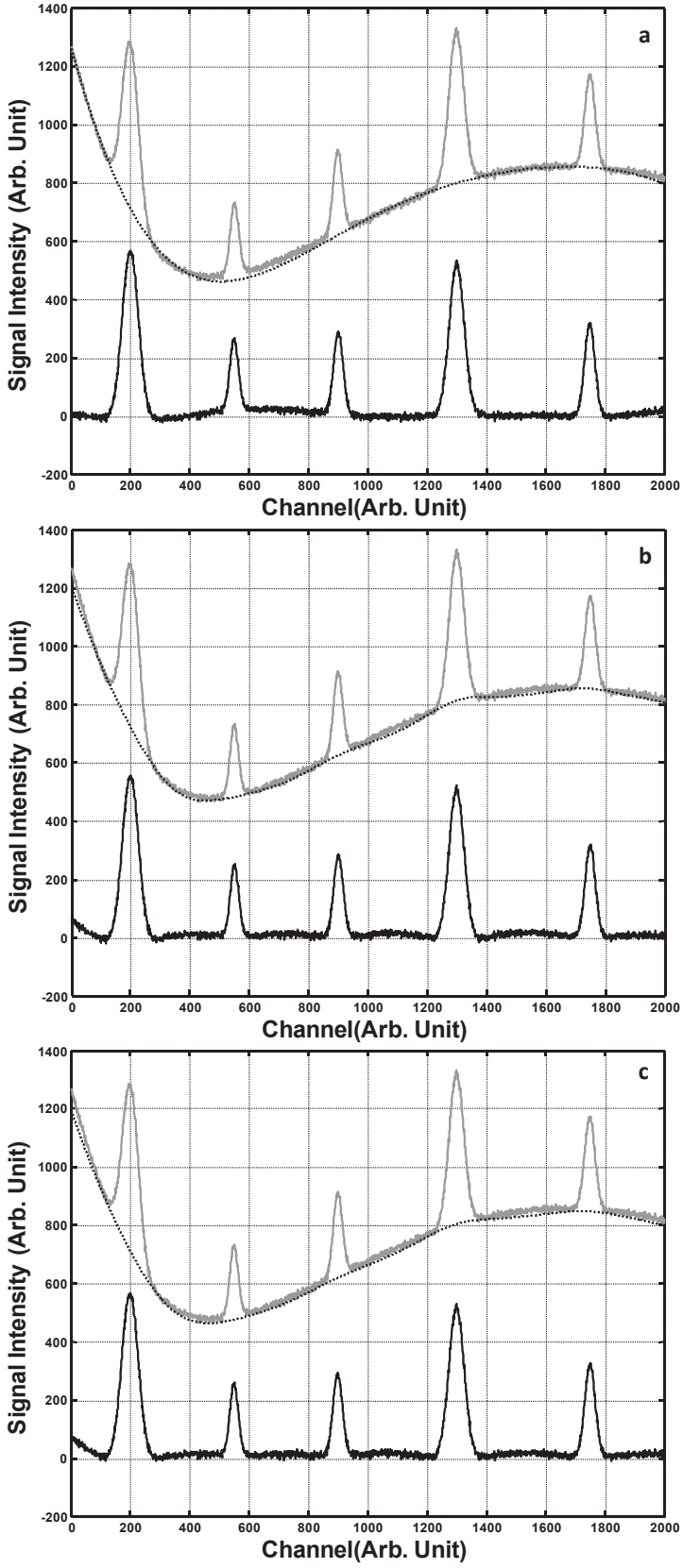

Figure 4.5 Raw signal with combination baseline, estimated baseline and baseline corrected signal using **(a)** TFALS **(b)** airPLS and **(c)** ALSS approaches.

For all five simulated signals, airPLS seems to force the baseline to stay at the bottom of the signal noise and as a consequence this produced some distorsion in the corrected signal between the peaks and was more evident for wider peaks. On the other hand, esimated baselines for TFALS and ALSS usually pass through in the middle of the signal noise. Error measurements in peak heights estimation will provide better comparative results in next section.

It is important to note that the comparisons made here were based on parameters optimized to fit the baseline, but such an optimaization is not possible for real signals. Therefore factors such as the number of adjustable parameters is an important consideration. Also, these simulations could have been optimized using other criteria, such as peak height estimation, but closeness to the true baseline seemed to be the more natural choice. More quantitative comparisons are made in the next section.

4.3.2 Simulation Results - Quantitative

4.3.2.1 Errors in Peak Height Estimation

To quantify the errors in peak height estimation, several matrices were used. In addition to the total error in peak height estimation for each peak, it was of interest to distinguish between the precision and bias (accuracy). In other words, it is useful to know whether methods were producing consistently low or high errors for each peak, or if the errors were more random in nature. Ten replicate data sets were employed ($N_{rep} = 10$) for each type of baseline evaluated. For each replicate, the signal and

baseline components remained the same, along with noise level ($\sigma = 6$), but a different realization of the noise sequence was used. For a given method and baseline, the error in peak j for replicate i, e_{ij}, was calculated as

$$e_{ij} = \left(h_j^o - \hat{b}_{ij}\right) - \left(h_j^o - b_j^o\right) = b_j^o - \hat{b}_{ij} \tag{4.1}$$

Here, h_j^o is the peak height for peak j with true baseline, b_j^o, but without the noise (i.e.($h_j^o - b_j^o$)) is the true peak heak height) and $\hat{b}_{ij}$ is the estimated baseline for peak j and replicate i. For a given peak, method and baseline, bias was assesed using the mean error, $\overline{e}$.

$$\text{mean error} = \overline{e} = \frac{\sum_{i=1}^{N_{rep}} e_i}{N_{rep}} \tag{4.2}$$

The standard deviation in this value, s_e, was also calculated.

$$s_e = \sqrt{\frac{\sum_{i=1}^{N_{rep}} \left(e_i - \overline{e}\right)^2}{N_{rep} - 1}} \tag{4.3}$$

This allowed as assesment of whether or not the bias was significantly different from zero. The total error can be evaluated by the root-mean-squared error (RMSE) around zero.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N_{rep}} e_i^2}{N_{rep}}} = \sqrt{\frac{\left(N_{rep} - 1\right)}{N_{rep}} s_e^2 + \overline{e}^2} \approx \sqrt{s_e^2 + \overline{e}^2} \tag{4.4}$$

As an overall assessment accross the five peaks for each method and baseline, the average RMSE is calculated from the individual RMSE for each peak, $RMSE_j$.

$$RMSE_{avg} = \frac{\sum_{j=1}^{5} RMSE}{5} \tag{4.5}$$

These results are presented in Table 4.1.

Table 4.1 Mean errors with standard deviation of mean in peak height estimation using three approaches for five different signals, and average RMS error in estimated peak heights using each of three approaches.

Baseline Function	Estimation Method	Peak Number/Mean Error ± SD					Average RMSE
		1	2	3	4	5	
	Actual PH	585	243	279	522	315	
Linear	TFALS	-3.78±0.41	0.41±0.22	-0.43±0.23	-2.46±0.26	0.98±0.3	1.65
	airPLS	2.82±1.59	7.30±1.39	8.01±1.22	6.00±1.50	7.87±1.26	6.57
	ALSS	-3.10±0.38	0.52±0.17	1.28±0.24	-0.50±0.24	2.15±0.32	1.54
Exponential	TFALS	-18.63±0.56	13.64±0.62	4.55±0.77	-0.17±0.48	0.57±0.57	7.63
	airPLS	-1.44±3.73	11.38±2.97	9.96±2.23	3.10±4.29	9.88±2.83	8.22
	ALSS	-28.58±0.49	8.66±0.59	3.29±0.52	9.29±0.37	1.19±0.47	10.23
Sinusoidal	TFALS	-5.66±0.41	-4.66±0.26	2.31±0.27	2.50±0.24	-2.01±0.33	3.44
	airPLS	-0.94±1.09	-11.19±0.78	10.31±0.81	6.73±0.64	7.75±0.80	7.50
	ALSS	-6.48±1.16	3.07±0.43	3.67±0.33	-2.89±0.90	2.87±0.51	3.86
Gaussian	TFALS	-18.40±0.63	17.48±0.78	5.43±0.79	3.13±0.45	-1.03±0.53	9.40
	airPLS	-20.52±2.39	3.21±1.32	1.66±0.58	-14.29±1.26	0.25±0.94	8.22
	ALSS	-26.22±0.91	2.78±0.81	-0.54±0.50	-17.45±0.61	-12.35±0.69	11.93
Combination	TFALS	-15.02±0.57	20.23±0.71	6.28±0.77	0.43±0.51	-0.50±0.63	8.60
	airPLS	-26.1±2.11	3.48±1.21	1.85±0.72	14.23±1.48	1.23±0.82	9.49
	ALSS	-18.64±3.03	10.15±5.15	7.12±3.75	-6.00±3.37	-5.62±5.02	10.34

It is noticeable that the mean errors are highly variable in sign and magnitude among the different peaks, methods and baselines, ranging from -28 to +18. For 44 out of 75 cases the mean errors are positive (low baseline), indicating that the baseline is more often under-estimated than over estimated. This is not surprising because of the nature of asymmetric least squares that tends to give smallest weight to the peak regions. For most of the cases, 69 out of 75, the largest component of the error arising in the baseline was the bias as opposed to the variance, giving significant differences in the peak height estimation. On observation of mean errors for individual peak, the direction of error

seems to be correlated to some extent with the position of the peak. For example, peak 1 exhibited negative errors (high baseline) for 14 out of 15 cases, while peak 2 and 5 exhibited positive errors (high baseline) for 13 out of 15 cases. This is consistent with the visual observations. In terms of overall average RMSE, TFALS gave the smallest values for three of the five baselines types and ranked second in the other cases. ALSS and airPLS gave smallest values for one case each. ALSS gave highest average RMSE error in three case and airPLS gave in the remaining two cases. Therefore, it can be concluded that the TFALS gives consistently good performance for peak height estimation.

4.3.2.2 RMS Error in Baseline Estimation

The RMS error in the baseline ($RMSE_b$) is the root-mean-squared (RMS) difference between the approximated and actual baseline at each sample point over the entire channel length.

$$RMSE_b = \sqrt{\frac{\sum_{i=1}^{N_{chan}} \left(b_i^o - \hat{b}_i \right)^2}{N_{chan}}} \tag{4.6}$$

Here $\hat{b}_i$ is the estimated baseline at channel i, b_i^o is the actual baseline at channel i, and N the number of channels The RMS error in the baseline measures the fidelity of the estimated baseline with the true baseline across the entire signal rather than just at the peaks. It is an indication of the "average" baseline error.

The results are tabulated in Table 4.2 for signals having different baseline functions along with each estimation approach. As previously noted, the adjustable parameters for each method were optimized to minimize the RMS error in the baseline, so

these should be optimum values for each method. The fit to the baseline will be affected by two factors. The first is the asymmetric or penalized least squares aspect, which will tend to push the fitted baseline below the true baseline, a common feature of all methods. The second is the ability of the underlying model to fit the true curve in the absence of noise. For TFALS, this is determined by the basis functions selected, while for the other methods it is imbedded in the smoothness criterion.

Table 4.2 RMS error in the baseline with each of three approaches for each baseline.

Baseline Function	Estimation Method	RMS Error
Linear	TFALS	2.711
	airPLS	6.287
	ALSS	1.823
Exponential	TFALS	9.140
	airPLS	10.581
	ALSS	13.775
Sinusoidal	TFALS	3.365
	airPLS	9.238
	ALSS	3.320
Gaussian	TFALS	10.589
	airPLS	10.264
	ALSS	14.442
Combination	TFALS	10.122
	airPLS	13.147
	ALSS	15.301

Of the methods and baselines examined, TFALS gives the lowest RMS error for 2 out of 5 cases (exponential and combination) and never gives the highest; instead, it is always very close to the best results (linear, sinusoidal and Gaussian). In contrast, airPLS gives the highest RMSE in the baseline for 2 out of 5 cases and gives the lowest errors only for the mixed Gaussian baseline. Its poor performance is likely due to its tendency

to model the bottom of the noise due to large weights imposed on negative residuals. Where it does perform better than other methods, this is likely due to the suppression of corner effects, but additional parameters need to be specified in the algorithm to achieve this. It can be concluded that, for the conditions examined in this study, TFALS seems to provide the most consistent estimation of the baseline by providing the best or close to the best overall fidelity with the true baseline.

4.3.2.3 Computation Time

The baseline estimation algorithm presented in this thesis is simple and converges relatively quickly. Computation times for each of three approaches were recorded with the simulated data set for different baseline functions and same channel lengths. Five different channel lengths of linear baseline were also used to estimate the relationship between the channel length and the computation time for individual approaches.

The computation time tests were performed under Windows 7 Professional 2009© on a 2.10 GHz processor with 2.00 GB of memory. The results of computation times for different channel lengths and different baseline functions with same channel length are tabulated in Table 4.3.

It can be noted from Table 4.3 that airPLS and TFALS computation times are comparable in all the cases; even for higher channel lengths, the TFALS computation time is very close to airPLS and their run times are also in the neighborhood of each other for different baseline functions at same channel length. On the other hand, ALSS can estimate the baselines only up to specific length. In this specific case of processor speed and memory, ALSS was unable to compute at channel lengths of 10,000 or higher.

Moreover, computation time for ALSS is at least 200 times more than TFALS for different baseline functions and observed channel lengths.

Table 4.3 Computation times for each of the three approaches at different channel lengths (left) and with different baseline functions (right).

Channel Length	Estimation Method	Run Time (sec)
1000	TFALS	0.059
	airPLS	0.074
	ALSS	2.484
5000	TFALS	0.104
	airPLS	0.074
	ALSS	178.082
10000	TFALS	0.116
	airPLS	0.119
	ALSS	Out of mem
50000	TFALS	0.320
	airPLS	0.321
	ALSS	Out of mem
100000	TFALS	0.365
	airPLS	0.260
	ALSS	Out of mem

Baseline Function	Estimation Method	Run Time (sec)
Linear	TFALS	0.0495
	airPLS	0.0413
	ALSS	13.111
Exponential	TFALS	0.0429
	airPLS	0.0395
	ALSS	8.7410
Sinusoidal	TFALS	0.0426
	airPLS	0.0387
	ALSS	9.3864
Gaussian	TFALS	0.0395
	airPLS	0.0377
	ALSS	10.476
Combination	TFALS	0.0402
	airPLS	0.0362
	ALSS	10.4756

4.3.2.4 Number of Adjustable Parameters

So far, the discussion has focused on methods where the parameters have been optimized by a grid search method to ensure that the comparisons are made under the best conditions for each method. However, this optimization is tedious and only possible under simulation conditions where the true baseline is known. In practice, the parameters must be optimized through a visual assessment of the estimated baseline and this becomes more difficult as the number of adjustable parameters increases and their interaction

becomes more complex. Since baseline removal should be a routine and simple task, increased time spent on parameters is undesirable.

Table 4.4 lists the adjustable parameters required for each of the method compared here. TFALS has the advantage of requiring only two adjustable parameters, the number of frequencies used and the asymmetry factor. Moreover, it was demonstrated in Chapter 3 that the two parameters have little interaction and the dependence on asymmetry parameter is small. This greatly simplifies the optimization of parameters for the proposed method.

Table 4.4 Adjustable parameters required for baseline correction methods.

Estimation Method	Number of Parameters	Parameters
TFALS	**2**	Asymmetric weight, number of basis functions
airPLS	**4**	Asymmetric weight, penalty factor, order of difference, weight exception proportion for for both ends
ALSS	**3**	Asymmetric weight, penalty factor, order of difference

The ALSS method requires three adjustable parameters to be optimized, an asymmetry parameter like the TFALS method, plus the penalty factor, λ, and the order of differential for the smoothness term. It could be argued that the order parameter is generally set to 2 and does not have a great range of values in any case, so the number of parameters is really only two. However, in this case, the optimization of the two parameters is closely linked since the negative residuals and the smoothness term balance each other. This interaction means that the parameters cannot be optimized independently, making the optimization more difficult.

Based on the original paper [37], the airPLS method should have only two parameters to be optimized, the penalty factor and the order associated with the smoothness term. The penalized least squares uses an exponential weight for negative residuals and a zero weight for positive residuals. However, the algorithm available in the public domain has two additional parameters. This is because a weighting parameter is applied to the positive residuals at the beginning and end of the signal, presumably to avoid corner effects. This requires the specification of the weighting factor and the fraction of the signal at the start and in to which it is to be applied. Thus, a total of four parameters need to be specified, all with some interaction. This makes the airPLS method the most demanding in terms of optimization.

4.4 Results and Discussion - Experimental Data

In Section 4.3, simulated data sets were used to demonstrate the performance of TFALS and compare it with two other approaches in the literature. While the TFALS method performs very well for simulated data, this does not guarantee that the proposed approach will also provide acceptable baseline estimation for experimental data, since experimental data often exhibit more complex artifacts. In this section, data sets from variety of analytical instruments containing visually different baseline artifacts are used to check the application and limitations of TFALS for experimental data. Since no standard baseline removed data were available, only a visual assessment of the results will be provided.

4.4.1 Replicated Data

It has been noted in Chapter 3 that the optimum number of frequencies used in TFALS is not highly dependent on peak height, width or location, whereas the optimum asymmetric weighting parameter is only mildly dependent on the signal-to-noise ratio. Here it is hypothesized that the same set of parameters would provide optimal results for replicate data obtained from same instrument under same analytical conditions. To validate this assumption, data sets from two different analytical approaches were used: Raman and X-ray fluorescence spectroscopy.

4.4.1.1 DNA Raman Spectra

Figure 4.6 shows the raw spectra (gray), estimated baselines (gray-black) and baseline corrected (black) Raman spectra of DNA. The two parameters for baseline estimation in TFALS were kept the same (number of frequencies, $n_{freq} = 3$ and weighting parameter, $p = 0.009$) for all three signals. It can be noticed that all three signals have somewhat different artifacts, which were successfully removed by TFALS using same parameters.

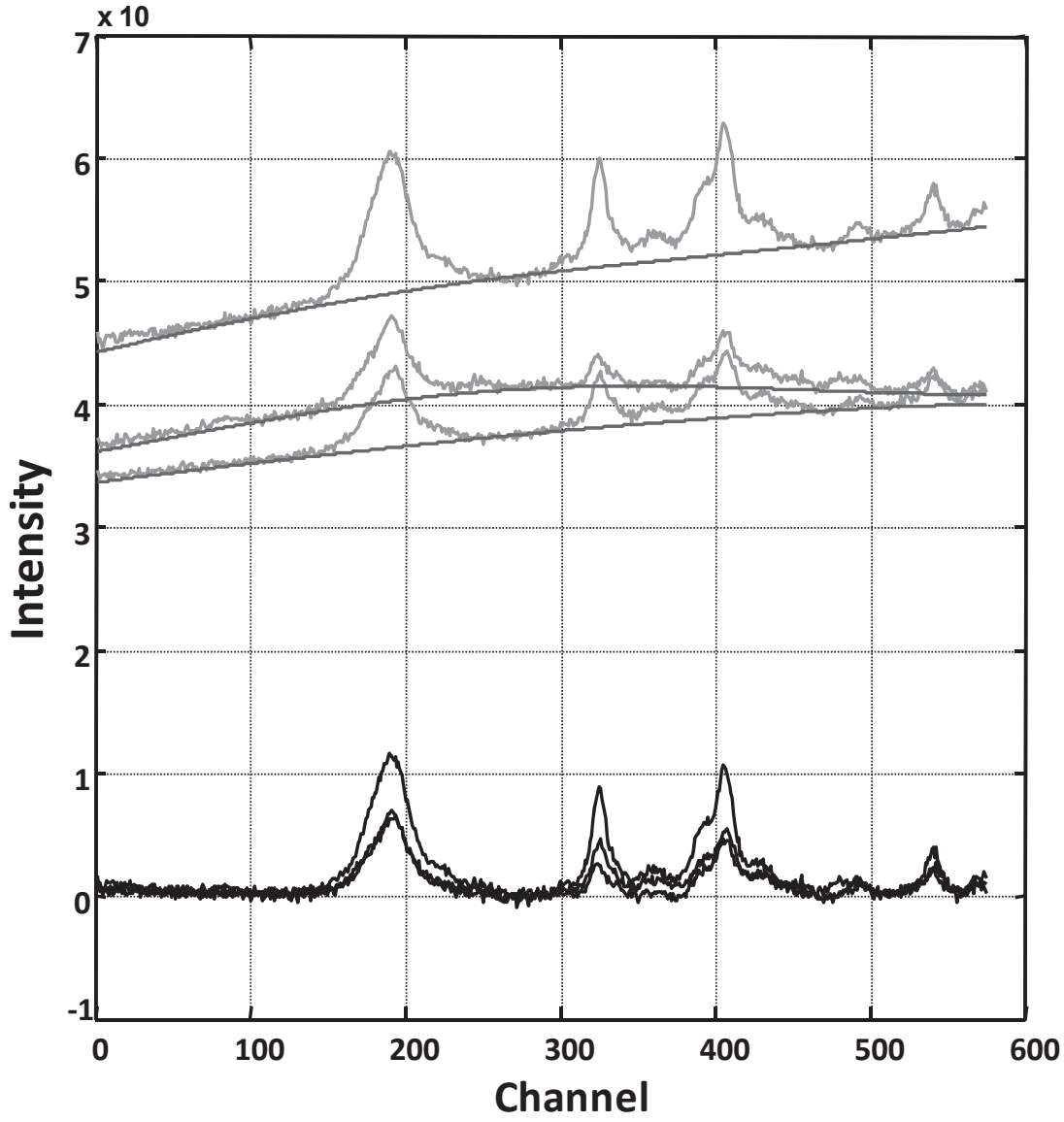

Figure 4.6 Original and baseline-corrected Raman spectra of DNA on a gold surface with the estimated baselines included.

4.4.1.2 X-ray Fluorescence Spectra

Figure 4.7 shows the results of replicate X-ray fluorescence replicate spectra with two different settings for the number of frequencies included in TFALS. In this data set, replicates contain minor differences in baseline features that are not apparent in the figure

and shown as a single condensed signal, but some interesting features of TFALS results

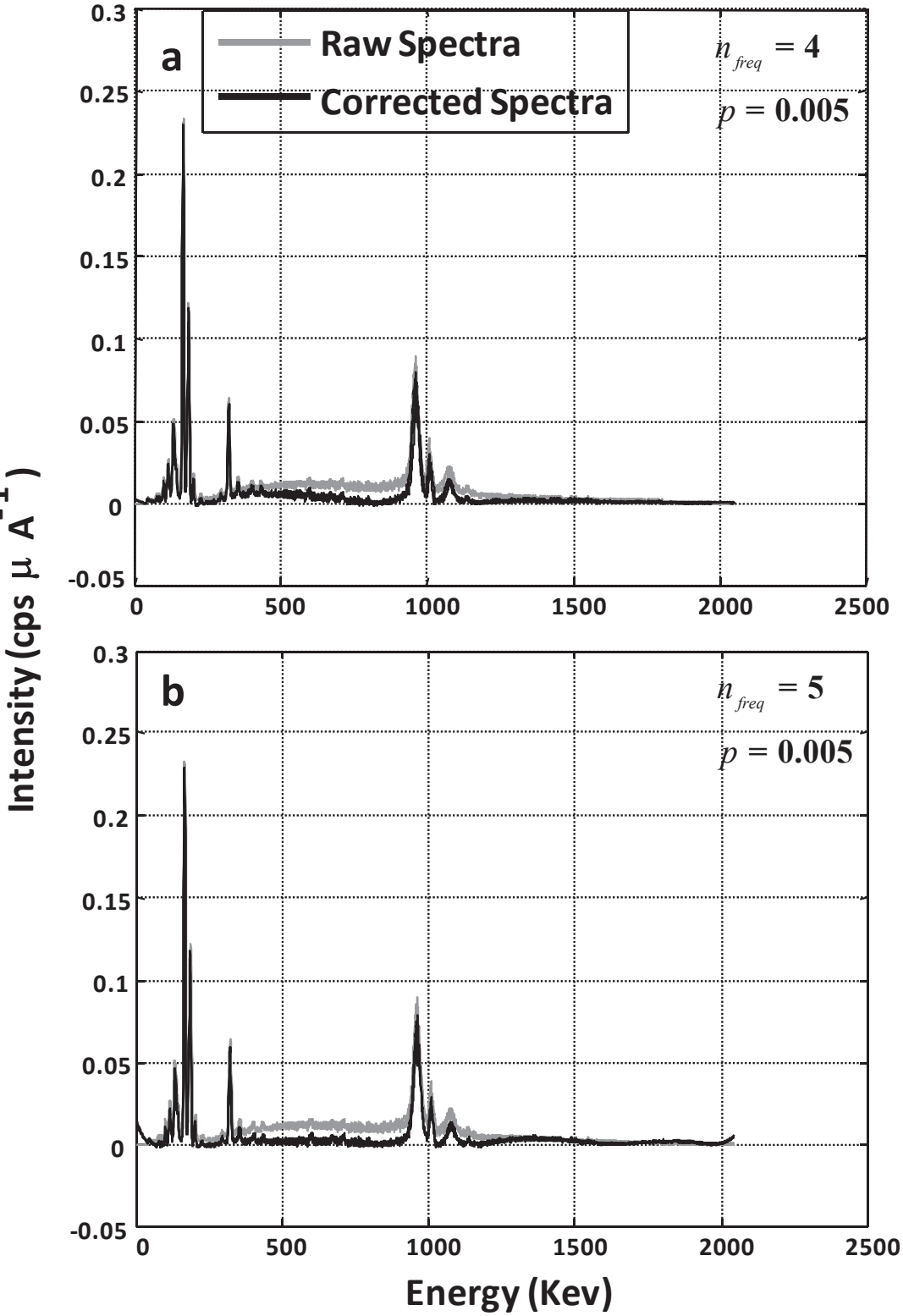

Figure 4.7 Results of three replicates of X-ray fluorescence spectra of tea with **(a)** 4 and **(b)** 5 baseline frequencies.

can be noted here using different values of number of frequencies parameter. It is apparent from the raw signals that only part of the signal contains baseline artifacts. Figure 4.7 (a) shows the results with 4 frequency values and asymmetry parameter of 0.005, whereas 4.7 (b) shows the results using 5 frequencies and the same asymmetry parameter. It can be noted from the two results that 4 frequencies were able to remove part of the baseline and left the non-peak area of spectra flattened, whereas 5 frequencies were able to completely remove the baseline artifacts from the peak area and resulted in a flat signal around peak area but left some artifacts in non-peak regions where no baseline were present in the raw signal. Hence, TFALS was able to remove baseline artifacts from the X-ray fluorescence replicate spectra effectively using same parameter values. However, it left over part of the baseline between 250 and 800 kev using $n_{freq} = 4$ or alternatively left some artifacts in non-peak areas using $n_{freq} = 5$ in the raw signals. This is one of the short comings of TFALS that will be further examined and discussed in the next section using signalscontaining a variety of baseline artifacts and signal-to-noise ratios.

4.4.2 Raman Minerals Data

A quantitative performance analysis of the proposed algorithm on experimental data is not possible in this case since no data were available in the absence of a baseline, but the performance of TFALS will be documented through a visual demonstration. The Raman signals of all six minerals and their baseline corrected signals are presented in Figure 4.8. The raw signals were offset slightly upward to clearly show the signal features and separate the baseline corrected signal. It appears that all six signals have somewhat different baseline artifacts.

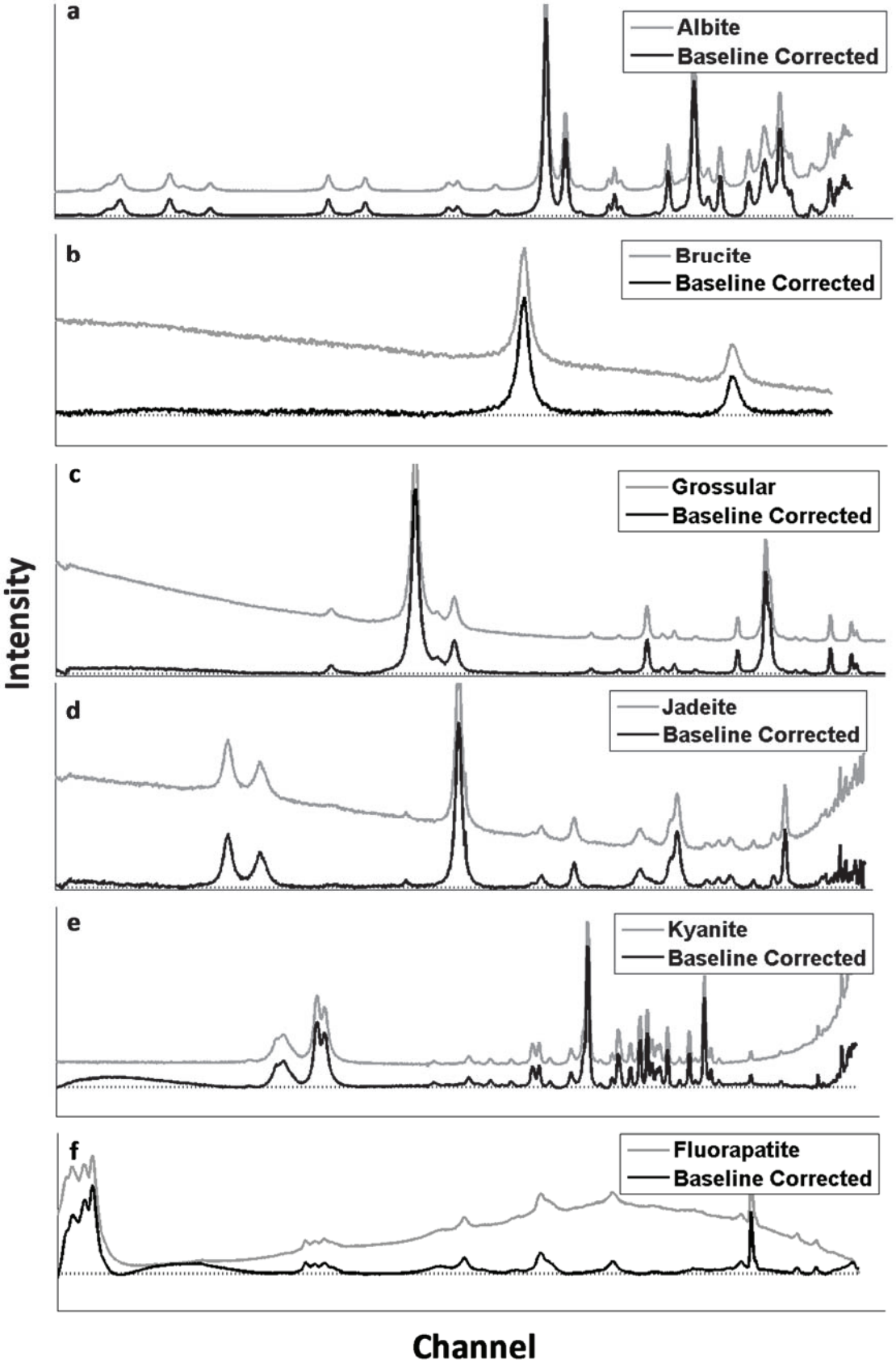

Figure 4.8 TFALS baseline correction results on Raman spectra of minerals.

It is evident that TFALS is able to remove the slowly varying baseline artifacts in the signals, although a few residual artifacts are apparent. For albite (4.8a) it could be argued that some residual baseline remains at the very right hand side of the signal, but without knowing the true signal, it is unclear whether there are small peaks on a rising baseline or components of a larger cluster of peaks. For signals 4.8 b-d, there appear to be no serious artifacts. For Kyanite in Figure 4.8e, the exponentially rising baseline at the right hand side is not completely removed and a slowly rising artifact (a "hump") is created in the flat region on the left hand side. Although distinguishable from the true peaks and in a region that contains no information, the appearance of this feature is still undesirable. Although its fundamental cause is unknown, it is speculated that it is a consequence of employing higher frequencies to model the rapidly decaying exponential on the right. This speaks to the difficulties of modeling extended baselines with mixed frequency components.

A similar small hump is observed on the left hand side for Fluorapatite in Figure 4.8 f. In this case, the source of the problem is likely to be the large peak cluster at the very left. As shown in the simulation in Chapter 3, peaks at the signal limit are problematic, since TFALS is more likely to attempt to model the peaks because there is no penalty region on the other side. As observed in simulations, this produces a negative-going oscillation on the near side of the peak, which can give rise to the observed artifact.

It was shown in the preceding chapter that the optimal asymmetry parameter in TFALS is dependent on the signal-to-noise ratio. Therefore, it is not surprising that Brucite, Figure 4.8 b, with a relatively lower signal-to-noise ratio, required a relatively

higher asymmetric weight ($p = 0.01$), whereas smaller weighting parameter values ($p = 0.001 \text{ to } 0.005$) usually provided acceptable results for the rest of the signals.

For this data set, three to five frequency components were enough to provide visually acceptable results. In the proposed augmented and truncated design, five frequency components corresponds to frequencies of 0, $0.25f_1$, $0.5f_1$, f_1 and $2f_1$. In the observed results, the artifacts seem to correspond to the higher frequency components of this sequence, as speculated in the discussion.

4.4.3 airPLS Data for Comparison

Raman, NMR and chromatographic data sets were taken from an online source used by the author [122, 37] in order to provide a comparative study of TFALS with an existing approach using experimental data. In this section, a visual comparison of TFALS with the airPLS method is provided using same parameter values provided by the author [96]. To provide a concise visual comparison, only a few signals were selected from all three experimental data sets as described in the experimental section.

One raw NMR signal with the baseline corrected signals using airPLS and TFALS is presented in Figures 4.9(a) and 4.9(b), respectively, along with the zoomed view of part of the signal with the estimated baselines. Both of the approaches seem to provide acceptable baseline removal; however on closer inspection of the estimated baseline for the two approaches, it is observed that airPLS follows the base of the peak and then cuts it off abruptly, resulting in a discontinuous baseline that includes a significant proportion of the peak. This type of behaviour was reported in the original paper as well. The amount of the peak that is removed depends on the peak height and width (since this approach

depends on the order of difference (see Section 1.3.7)). On the other hand, TFALS produces a much smoother estimate of the baseline that follows the signal more naturally. It is expected that the behavior of airPLS would be more likely to generate negative errors in the estimate peak height.

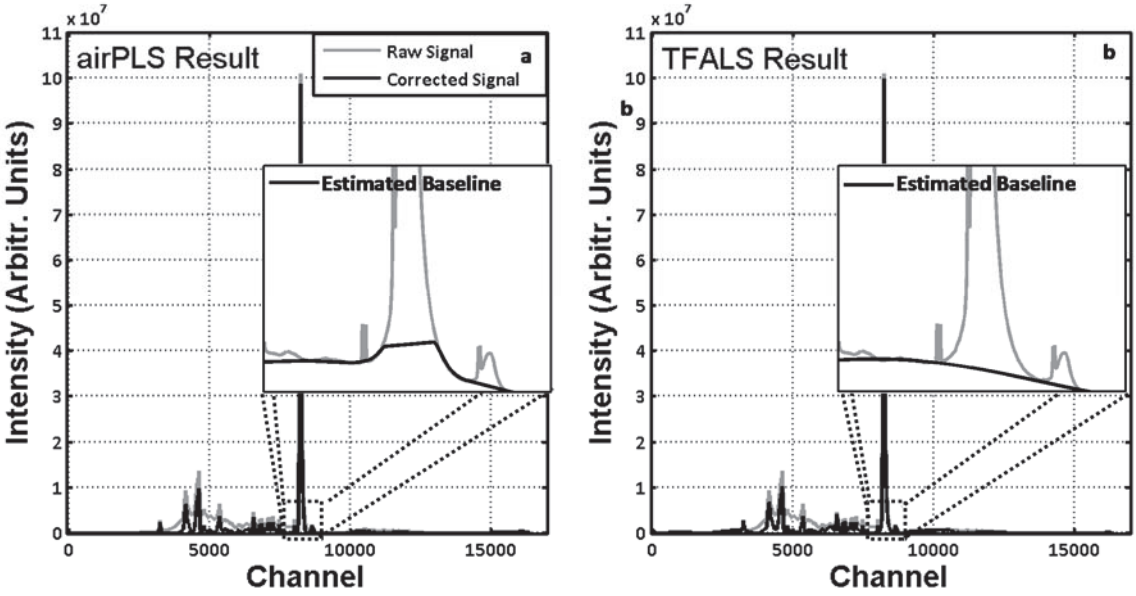

Figure 4.9 Raw and baseline corrected NMR signals using **(a)** airPLS and **(b)** TFALS algorithms along with a small zoomed view of the raw signal and estimated baseline.

Similar results were observed from the chromatographic data. Out of eight chromatographic signals, only one signal is presented here to closely compare the results of the two approaches. The raw and baseline-corrected chromatographic signals using airPLS and TFALS are presented in Figures 4.10(a) and 4.10(b) along with a zoomed view of the raw signal and estimated baselines with each approach. It can be clearly noted here that TFALS provided very smooth baseline whereas airPLS again removes a substantial portion of the base of the peak and as a result provides a relatively rough baseline with sharp features.

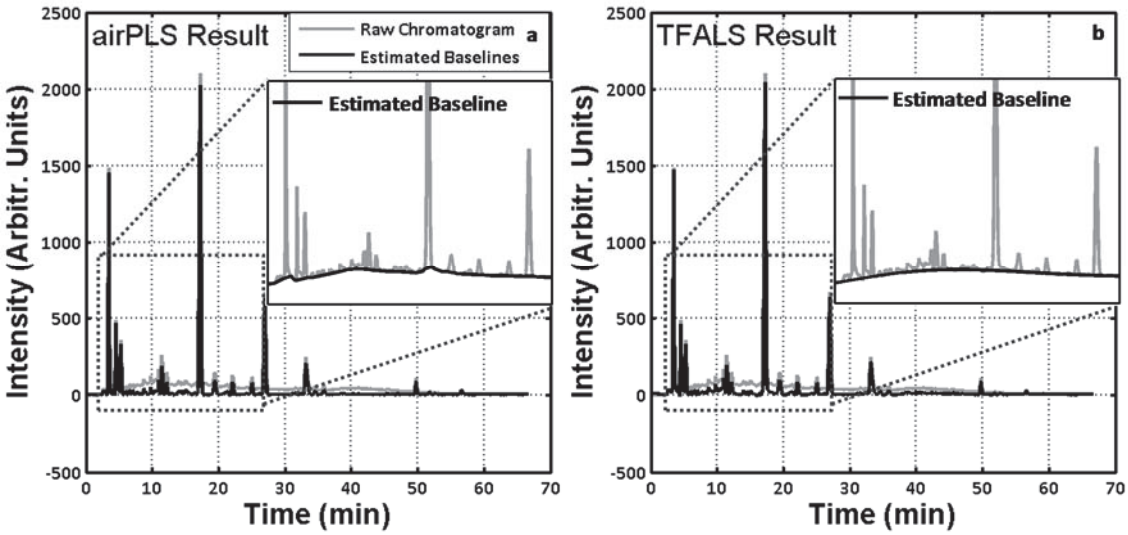

Figure 4.10 Raw and baseline corrected chromatographic signals using **(a)** airPLS and **(b)** TFALS algorithms along with a zoomed view of raw signal and estimated baseline.

The third data set consisted of three Raman spectra taken from the same data group published in the literature [96]. Estimated baselines using airPLS and TFALS are presented along with the raw signals in Figure 4.11(a) and 4.11(b) respectively. Here TFALS seems to exhibit minor corner effect at the left hand side, but overall the estimated baseline is very smooth and provides a very good approximation using identical

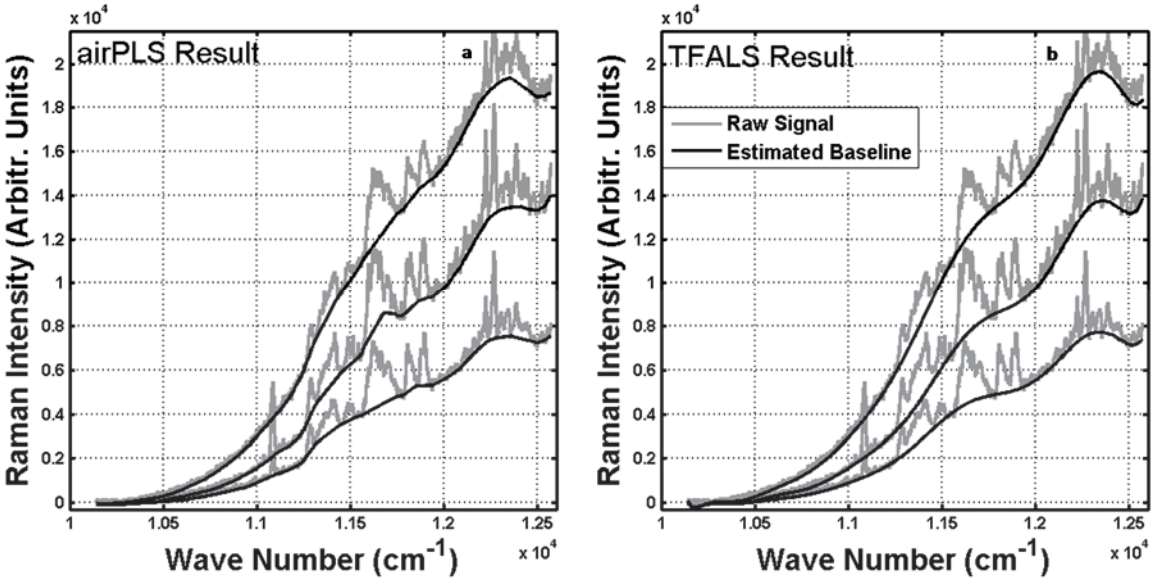

Figure 4.11 Raman spectra and estimated baselines using **(a)** airPLS and **(b)** TFALS algorithms.

parameter values for all three signals. On the other hand airPLS again seems to generate discontinuities in the baseline under peak regions.

4.5 Conclusions

In this chapter, the performance of TFALS was compared to two widely used methods using simulated and experimental data. Visually, all three methods provided acceptable results for five different simulated baselines. All methods displayed some "corner effects" (a deviation of the estimated baseline at the end of the signal) under certain conditions, but these were largely inconsequential. Quantitatively, all three methods were evaluated according to errors in peak height estimation and fidelity with the true baseline. In both aspects, all methods were roughly comparable, but TFALS most often produced the lowest RMS error in peak estimation and exhibited the most consistent baseline fidelity. Computation time and optimization efficiency were also considered. Asymmetric least squares smoothing (ALSS) was at a severe disadvantage in terms of its computational efficiency for large data sets, where TFALS and airPLS were both fast and comparable in their execution times. In terms of parameter optimization, TFALS has the smallest number of parameters to optimize and, because these are largely independent, it is likely to be the simplest method to use.

In the application to experimental data sets, TFALS appeared to perform well on a variety of data sets, although no quantitative comparisons could be made. For several sets of experimental data, TFALS was compared to airPLS. For high signal-to-noise data, the latter method seemed to incorporate part of the peak into the baseline and also produced a baseline estimate with discontinuities.

Chapter 5
Conclusions

5.1 Conclusions

The work presented in this thesis has described the development and optimization of a novel baseline estimation algorithm, called truncated Fourier asymmetric least squares (TFALS). It is often essential to estimate and remove the slowly varying, high magnitude baseline features from analytical signals prior to further data analysis. TFALS uses asymmetric least squares (ALS) along with the truncated and augmented Fourier basis functions to model the slowly varying baseline. Chapter 1 introduced the limitations and drawbacks of existing baseline estimation approaches. The challenge was presented as being able to estimate the baseline without extensive parameter adjustment or the requirement of representative baseline signals. Multiple parameters and the complex inter-dependency of these parameters can complicate the choice of optimal parameter values, and the acquisition of representative baseline signals is not always possible.

The development of an algorithm for baseline estimation using truncated Fourier functions along with some extra augmentations as basis functions for slowly varying baselines was described and it is shown that the use of augmented Fourier basis functions in conjunction with asymmetric least squares (ALS) provide an effective estimation of baselines. It was noted that TFALS uses only two adjustable parameters and that the optimum values of these parameters are not highly dependent on signal attributes. The optimum value of the first parameter, the number of frequencies, n_{freq} , is dependent on

the nature of the baseline and largely independent of the analytical signal. The optimum value of the second parameter, the asymmetry parameter, p, is also independent of the analytical signal, although it has a small dependence on the signal-to-noise ratio. It was found that the two parameters are not highly dependent on each other and can be optimized individually.

A qualitative and quantitative comparison of TFALS with the other two approaches showed that all three methods are roughly comparable, but TFALS most often produced the lowest RMS error in peak height estimation and also exhibited the most consistent fidelity with the baseline. Computation time to estimate the baseline via TFALS is also very small and comparable to the fastest of the other approaches. In terms of parameter optimization, TFALS seems to be the simplest approach, having only two largely independent parameters to optimize. Applications to a wide variety of experimental data sets were also presented and a qualitative comparison was included with another method in the literature for signals from three different analytical measurement systems. TFALS appeared qualitatively to perform well on a wide variety of data sets and perform better than the literature method in the experimental comparison study as well.

To estimate the slowly varying unwanted components of analytical signals, no additional baseline measurements or blank runs are required for the estimation of baseline using TFALS. Overall, it is a good general approach that requires minimal user input and parameter adjustments with fast execution.

5.2 Future Work

With the development and optimization of TFALS algorithm described in this work and its application to a wide variety of analytical data sets, there are a number of challenges that remain. It has noted that the augmentation of extra basis sets between DC and first Fourier frequency extensively improved the approximation of low frequencies (linear, sinusoid functions) in the absence of analyte peaks. However some artifacts remain in the presence of peaks in a few instances. These artifacts seem to appear when the gradient of baseline changes significantly or when the frequency difference between parts of the signal is high. These artifacts can be minimized by giving different asymmetric weights for those data points. Methods to provide a generalized criterion to choose the number of data points required to have different weights and asymmetry values for those points is one area that remains to be explored.

Although a quantitative simulated study is included in this thesis experimental calibration would give a better idea of TFALS accuracy in experimental conditions. In order to experimentally calibrate the proposed baseline estimation approach, a dependable experiment is required to perform relative quantitation. A pre-defined analyte concentration, peak area or peak height can be compared with the estimated results for this purpose. Relative percentage measurement error would provide the measure of accuracy in the baseline estimation by TFALS.

Finally, a persistent problem with all of the ALS methods studied here is the tendency to underestimate the baseline by following the lower limits of the baseline noise.

It would be useful to modify the proposed approach to allow it to track the centre of the baseline more reliably.

References

1. Robinson JW, Skelly EM, Frame GM; Undergraduate Instrumental Analysis. *Marcel Dekker Inc*., New York, Sixth edition.

2. Ewing GW; Analytical Instrumentation Handbook. *Marcel Dekker Inc.*, New York, Second edition.

3. Baumann F, Tao F; Digital integrators - effect of slope sensitivity filtering and baseline correction rate on accuracy. *Journal of Gas Chromatography.* **5:621-626** (1967).

4. Williams, DT, Hager RN; The derivative spectrometer. *Applied Optics.* **9:1597-1605** (1970).

5. Liland KH, Almoy T, Mevik BH; Optimal Choice of Baseline Correction for Multivariate Calibration of Spectra. *Applied Spectroscopy.* **64:1007-1016** (2010).

6. Schulze G, Jirasek A, Yu MML, Lim, A, Turner RFB, Blades MW; Investigation of selected baseline removal techniques as candidates for automated implementation. *Applied Spectroscopy.* **59:545-574** (2005).

7. Janssens, F, Francois JP; An optimized background correction algorithm in automated spectral-analysis based on convolution signals. *Applied Spectroscopy.* **46:283-292** (1992).

8. Marion, D, Bax A; Baseline distortion in real Fourier transform NMR-spectra. *Journal of Magnetic Resonance.* **79(2):352-356,** (1988).

9. Marion, D, Bax, A; Baseline correction of 2D FT NMR-spectra using a simple linear prediction extrapolation of the time-domain data. *Journal of Magnetic Resonance.* **83:205-211** (1989).

10. Friedrichs MS; A model-free algorithm for the removal of base-line artifacts. *Journal of Biomolecular NMR.* **5(2):147-153** (1995).

11. Keselbrener L, Keselbrener M, Akselrod S; Nonlinear high pass filter for R-wave detection in ECG signal. *Medical Engineering and Physics.* **19:481-484** (1997).

12. Lewis DM, Chatwin PC; The treatment of atmospheric dispersion data in the presence of noise and base-line drift. *Boundary-Layer Meteorology.* **72:53-85** (1995).

13. Kourkoumelis N, Polymeros A, Tzaphlidou M; Background estimation of biomedical Raman spectra using a geometric approach. *Spectroscopy-an International Journal.* **27:441-447** (2012).

14. Bao Q, Feng J, Chen F, Mao W, Liu Z, Liu K, Liu C; A new automatic baseline correction method based on iterative method. *Journal of Magnetic Resonance.* **218:35-43** (2012).

15. Coombes KR, Fritsche HA, Clarke C, Chen JN, Baggerly KA, Morris JS, Kuerer HM; 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** (2003).

16. Golotvin S, Williams A; Improved baseline recognition and modeling of FT NMR spectra. *Journal of Magnetic Resonance.* **146:122-125** (2000).

17. Xian F, Corilo YE, Hendrickson CL, Marshall AG; Baseline correction of absorption-mode Fourier transform ion cyclotron resonance mass spectra. *International Journal of Mass Spectrometry.* **325:67-72** (2012).

18. Chang D, Banack CD, Shah SL; Robust baseline correction algorithm for signal dense NMR spectra. *Journal of Magnetic Resonance.* **187:288-292** (2007).

19. Xi Y, Rocke DM; Baseline correction for NMR spectroscopic metabolomics data analysis. *BMC Bioinformatics.* **9:324-333** (2008).

20. Johnsen LG, Skov T, Houlberg U, Bro R; An automated method for baseline correction, peak finding and peak grouping in chromatographic data. *The Analyst.* **138:3502-3511** (2013).

21. Krishnan S, Vogels J, Coulier L, Bas RC, Hendriks MWB, Hankemeier T, Thissen U; Instrument and process independent binning and baseline correction methods for liquid chromatography-high resolution-mass spectrometry deconvolution. *Analytica Chimica Acta.* **740:12-19** (2012).

22. Phillips AJ, Hamilton PA; Improved detection limits in Fourier transform spectroscopy from a maximum entropy approach to baseline estimation. *Analytical Chemistry.* **68:4020-4025** (1996).

23. Jegla JD, Richardson RL, Griffiths PR; An automated baseline correction algorithm for high- and low-resolution open-path FT-IR measurements. *Proceedings of the SPIE-The International Society for Optical Engineering.* **2883:323-332** (1996).

24. Danielsson R, Bylund D, Markides KE; Matched filtering with background suppression for improved quality of base peak chromatograms and mass spectra in liquid chromatography-mass spectrometry. *Analytica Chimica Acta.* **454:167-184** (2002).

25. Li N, Li XY, Zou ZX, Lin LR, Li, YQ; A novel baseline-correction method for

standard addition based derivative spectra and its application to quantitative analysis of benzo(a)pyrene in vegetable oil samples. *Analyst.* **136:2802-2810** (2011).

26. Bialkowski SE; Finite impulse-response filters. *Analytical Chemistry.* **60:A355-A361** (1988).

27. Schulze HG, Foist RB, Okuda K, Ivanov A, Turner RFB; A Model-Free, Fully Automated Baseline-Removal Method for Raman Spectra. *Applied Spectroscopy.* **65:75-84** (2011).

28. Schulze HG, Foist RB, Okuda K, Ivanov A, Turner RFB; A Small-Window Moving Average-Based Fully Automated Baseline Estimation Method for Raman Spectra. *Applied Spectroscopy.* **66:757-764** (2012).

29. Weakley AT, Griffiths PR, Aston DE; Automatic Baseline Subtraction of Vibrational Spectra Using Minima Identification and Discrimination via Adaptive, Least-Squares Thresholding. *Applied Spectroscopy.* **66:519-129** (2012).

30. Solis A, Rex M, Campiglia AD, Sojo P; Accelerated multiple-pass moving average: A novel algorithm for baseline estimation in CE and its application to baseline correction on real-time bases. *Electrophoresis.* **28:1181-1188** (2007).

31. Prakash BD, Wei YC; A fully automated iterative moving averaging (AIMA) technique for baseline correction. *Analyst.* **136:3130-3135** (2011).

32. Boelens HFM, Dijkstra RJ, Eilers PHC, Fitzpatrick F, Westerhuis JA; New background correction method for liquid chromatography with diode array detection, infrared spectroscopic detection and Raman spectroscopic detection. *Journal of Chromatography A.* **1057:21-30** (2004).

33. Boelens HFM, Eilers PHC; Baseline correction with asymmetric least squares smoothing.http://www.science.uva.nl/~hboelens/publications/draftpub/Eilers_2005. Last access on April 1, 2012.

34. Dijkstra RJ, Boelens HFM, Westerhuis JA, Ariese F, Brinkman UAT, Gooijer C; Hyphenation of column liquid chromatography and Raman spectroscopy via a liquid-core waveguide: chemometrical elimination of spectral eluent background. *Analytica Chimica Acta.* **519:129-136** (2004).

35. Cobas CJ, Bernstein MA, Martin-Pastor M, Garcia TP; A new general-purpose fully automatic baseline-correction procedure for 1D and 2D NMR data. *Journal of Magnetic Resonance.* **183:145-151** (2006).

36. Peng JT, Peng SL, Jiang A, Wei JP, Li CW, Tan J; Asymmetric least squares for multiple spectra baseline correction. *Analytica Chimica Acta.* **683:63-68** (2010).

37. Zhang ZM, Chen S, Liang YZ; Baseline correction using adaptive iteratively reweighted penalized least squares. *Analyst.* **135:1138-1146** (2010).

38. Zhang ZM, Chen S, Liang YZ, Liu ZX, Zhang QM, Ding LX, Zhou H; An intelligent background-correction algorithm for highly fluorescent samples in Raman spectroscopy. *Journal of Raman Spectroscopy.* **41:659-669** (2010).

39. Devos O, Mouton N, Sliwa M, Ruckebusch C; Baseline correction methods to deal with artifacts in femtosecond transient absorption spectroscopy. *Analytica Chimica Acta.* **705:64-71** (2011).

40. Peng J, Peng S, Xie Q, Wei J; Baseline correction combined partial least squares algorithm and its application in on-line Fourier transform infrared quantitative analysis. *Analytica Chimica Acta.* **690:162-168** (2011).

41. de Rooi JJ, Eilers PHC; Mixture models for baseline estimation. *Chemometrics and Intelligent Laboratory Systems.* **117: 56-60**, (2012).

42. de Rooi JJ, Devos O, Sliwa M, Ruckebusch C, Eilers PHC; Mixture models for two-dimensional baseline correction, applied to artifact elimination in time-resolved spectroscopy. *Analytica chimica acta.* **771:7-13** (2013).

43. Mazet V, Carteret C, Brie D, Idier J, Humbert B; Background removal from spectra by designing and minimising a non-quadratic cost function. *Chemometrics and Intelligent Laboratory Systems.* **76:121-133** (2005).

44. Xu Z, Sun X, Harrington PdB; Baseline Correction Method Using an Orthogonal Basis for Gas Chromatography/Mass Spectrometry Data. *Analytical Chemistry.* **83:7464-7471** (2011).

45. Spaink HA, Lub TT, Otjes RP, Smit HC; Baseline correction method for second-harmonic detection with tunable diode lasers. *Analytica Chimica Acta.* **183:141-15,** (1986).

46. Dietrich W, Ruedel CH, Neumann M; Fast and precise automatic baseline correction of one- and two-dimensional NMR spectra. *Journal of Magnetic Resonance.* **91:1-11** (1991).

47. Lunga GD, Pogni R, Basosi R; A simple method for baseline correction in EPR spectroscopy. *Journal of Magnetic Resonance, Series A.* **108:65-70** (1994).

48. Brown DE; Fully automated baseline correction of 1D and 2D NMR spectra using Bernstein polynomials. *Journal of Magnetic Resonance, Series A.* **114:268-270** (1995).

49. Kolibal J, Howard D; MALDI-TOF baseline drift removal using stochastic Bernstein approximation. *Eurasip Journal on Applied Signal Processing*.

2006:1-9 (2006).

50. Cao A, Pandya AK, Serhatkulu GK, Weber RE, Dai H, Thakur JS, Freeman DC; A robust method for automated background subtraction of tissue fluorescence. *Journal of Raman Spectroscopy.* **38:1199-1205** (2007).

51. Kuligowski J, Quintas G, Garrigues S, de la Guardia M. New background correction approach based on polynomial regressions for on-line liquid chromatography Fourier transform infrared spectrometry. *Journal of Chromatography A.* **1216: 3122-3130** (2009).

52. Gan F, Ruan GH, Mo JY; Baseline correction by improved iterative polynomial fitting with automatic threshold. *Chemometrics and Intelligent Laboratory Systems.* **82:59-65** (2006).

53. Mosierboss PA, Lieberman SH, Newbery R; Fluorescence rejection in Raman-spectroscopy bt shifted spectra, edge-detection, and fft filtering techniques. *Applied Spectroscopy.* **49:630-638** (1995)

54. Liu BF, Sera Y, Matsubara N, Otsuka K, Terabe S; Signal denoising and baseline correction by discrete wavelet transform for microchip capillary electrophoresis. *Electrophoresis.* **24:3260-3265** (2003).

55. Chen D, Shao XG, Hu B, Su QD; A Background and noise elimination method for quantitative calibration of near infrared spectra. *Analytica Chimica Acta.* **511: 37-45** (2004).

56. Hu Y, Jiang T, Shen A, Li W, Wang X, Hu J; A background elimination method based on wavelet transform for Raman spectra. *Chemometrics and Intelligent Laboratory Systems.* **85:94-101** (2007).

57. Galloway CM, Le Ru EC, Etchegoin PG; An Iterative Algorithm for Background Removal in Spectroscopy by Wavelet Transforms. *Applied Spectroscopy.* **63: 1370-1376** (2009).

58. Wan BY, Small GW; Wavelet analysis used for spectral background removal in the determination of glucose from near-infrared single-beam spectra. *Analytica Chimica Acta.* **681:63-70** (2010).

59. Chen D, Chen ZW, Grant E; Adaptive wavelet transform suppresses background and noise for quantitative analysis by Raman spectrometry. *Analytical and Bioanalytical Chemistry.* **400:625-634** (2011).

60. Ma CX, Shao XG; Continuous wavelet transform applied to removing the fluctuating background in near-infrared spectra. *Journal of Chemical Information and Computer Sciences.* **44:907-911** (2004).

61. Paredes JL, Sosa E; A baseline correction algorithm for capillary electrophoresis data using local optimization of the legend algorithm in the wavelet domain. *Interciencia.* **34:556-562** (2009).

62. Shin H, Sampat MP, Koomen JM, Markey MK; Wavelet-Based Adaptive Denoising and Baseline Correction for MALDI TOF MS. *Omics-a Journal of Integrative Biology.* **14:283-295** (2010).

63. Gorski L, Ciepiela F, Jakubowska M, Kubiak WW; Baseline Correction in Standard Addition Voltammetry by Discrete Wavelet Transform and Splines. *Electroanalysis.* **23:2658-2667** (2011).

64. Pearson GA; General baseline-recognition and baseline-flattening algorithm. *Journal of Magnetic Resonance.* **27:265-272** (1977).

65. Ingle JD, Crouch SR; Spectrochemical Analysis. *Englewood Cliffs. Prentice-Hall*, New Jersey, (1988).

66. McKinnon GC, Burger C, Boesiger P; Spectral baseline correction using CLEAN. *Magnetic Resonance in Medicine*. **13:145-149** (1990).

67. Otting G, Widmer H, Wagner G, Wuthrich K; Origin of T1 and T2 ridges in 2D NMR-spectra and procedures for suppression. *Journal of Magnetic Resonance.* **66:187-193** (1986).

68. Moskau D; Application of real time digital filters in NMR spectroscopy. *Concepts in Magnetic Resonance.* **15:164-176** (2002).

69. Lewis IR, Edwards HGM; Handbook of Raman spectroscopy: From the research laboratory to the process line. *Marcel Dekker Inc.,* New York, (2001).

70. Cappiello A, Famiglini G, Palma P, Pierini E, Termopoli V, Trufelli H; Overcoming Matrix Effects in Liquid Chromatography-Mass Spectroscopy. *Analytical Chemistry.* **80:9343-348** (2008).

71. Liang YZ, Kvalheim OM, Rahmani A, Brereton RG; A 2-way procedure for background correction of chromatographic spectroscopic data by congruence analysis and least-squares fit of the zero-component regions - comparison with double-centering. *Chemometrics and Intelligent Laboratory Systems.* **18:265-279** (1993).

72. Vogt F, Steiner H, Booksh K, Mizaikoff B; Chemometric correction of drift effects in optical spectra. *Applied Spectroscopy.* **58:683-692** (2004).

73. Burgess DD; A comparison of methods for baseline estimation in gamma-ray spectrometry. *Nuclear Instruments & Methods in Physics Research Section A-Accelerators Spectrometers Detectors and Associated Equipment.* **221:593-599**

(1984).

74. Malitesta C, Rotunno T; Quantitative resolution of x-ray photoelectron spectra of mixtures of chromium compounds by the Kalman filter after cubic spline background removal. *Surface and Interface Analysis.* **17:251-258** (1991).

75. Krappe HJ, Rossner HH; Bayesian approach to background subtraction for data from the extended x-ray-absorption fine structure. *Physical Review B, Condensed Matter and Materials Physics.* **70:1041021-104102,** (2004).

76. Virtanen J, Noponen T, Kotilahti K, Virtanen J, Ilmoniemi RJ; Accelerometer-based method for correcting signal baseline changes caused by motion artifacts in medical near-infrared spectroscopy. *Journal of Biomedical Optics.* **16:0870051-087005,** (2011).

77. Zhao J, Lui H, McLean DI, Zeng H; Automated autofluorescence background subtraction algorithm for biomedical Raman spectroscopy. *Applied Spectroscopy.* **61:1225-1232** (2007).

78. Wilson JD, McInnes CA J; Elimination of errors due to baseline drift in measurement of peak areas in gas chromatography. *Journal of Chromatography.* **19:486-494** (1965).

79. Deighton MO; Minimum-noise filters with good low-frequency rejection. *IEEE Transactions on Nuclear Science.* **NS16:68-75** (1969).

80. Statham PJ; Deconvolution and background subtraction by least-squares fitting with pre-filtering of spectra. *Analytical Chemistry.* **49:2149-2154** (1977).

81. Wang J, Dewald HD; Background-current subtraction in voltammetric detection for flow-injection analysis. *Talanta.* **31:387-390** (1984).

82. Inczedy J, Molnar M; Baseline correction of chromatograms using numerical filtering. *Magyar Kemai Folyoirat*. **91:429-431** (1985).

83. Profio AE, Balchum OJ, Carstens F; Digital background subtraction for fluorescence imaging. *Medical Physics.* ***13*:717-721** (1986).

84. Brown, SD, Rutan SC; Adaptive Kalman filtering. *Journal of Research of the National Bureau of Standards.* **90:403-407** (1985).

85. Rutan SC, Brown SD; Simplex optimization of the adaptive Kalman filter. *Analytica Chimica Acta.* **167:39-50** (1985).

86. Rutan SC; Adaptive Kalman filtering. *Analytical Chemistry.* **63:1103-1109** (1991).

87. Rutan SC, Bouveresse E, Andrew KN, Worsfold PJ, Massart DL; Correction for drift in multivariate systems using the Kalman filter. *Chemometrics and Intelligent Laboratory Systems*. **35:199-211** (1996).

88. Gerow DD, Rutan SC; Background subtraction for fluorescence detection in thin-layer chromatography with derivative spectrometry and the adaptive Kalman filter. *Analytica Chimica Acta.* **184:53-64** (1986).

89. Gerow, DD, Rutan SC; Background correction for fluorescence detection in thin-layer chromatography using factor analysis and the adaptive Kalman filter. *Analytical Chemistry.* **60:847-852** (1988).

90. Kalman RE; A new approach to linear filtering and prediction problems. *Journal of Basic Engineering.* **82:35-45** *(*1960).

91. Savitzky A, Golay MJE; Smoothing + differentiation of data by simplified least squares procedures. *Analytical Chemistry.* **36:1627-1639** (1964).

92. Bauchspiess KR; EXAFS background subtraction using splines. *Physica B (Amsterdam),* **208:183-184** (1995).

93. Saffrich R, Beneicke W, Neidig KP, Kalbitzer HR; Baseline correction in n-dimensional NMR spectra by sectionally linear interpolation. *Journal of Magnetic Resonance, Series B.* **101:304-308** (1993).

94. Lunga GD, Basosi R; A simple method for baseline correction in EPR spectroscopy. 2. The use of cubic spline functions. *Journal of Magnetic Resonance, Series A.* **112:102-105** (1995).

95. Malloni WM, De Sanctis S, Tome AM, Lang EW, Munte CE, Neidig KP, Kalbitzer HR; Automated solvent artifact removal and base plane correction of multidimensional NMR protein spectra by AUREMOL-SSA. *Journal of Biomolecular NMR.* **47:101-111** (2010).

96. De Sanctis S, Malloni WM, Kremer W, Tome AM, Lang EW, Neidig KP, Kalbitzer HR; Singular spectrum analysis for an automated solvent artifact removal and baseline correction of 1D NMR spectra. *Journal of Magnetic Resonance.* **210:177-183** (2011).

97. Bernabe-Zafon V, Torres-Lapasio JR, Ortega-Gadea S, Simo-Alfonso EF, Ramis-Ramos G; Capillary electrophoresis enhanced by automatic two-way background correction using cubic smoothing splines and multivariate data analysis applied to the characterisation of mixtures of surfactants. *Journal of Chromatography A.* **1065:301-313** (2005).

98. Kuligowski J, Carrion D, Quintas G, Garrigues S, de la Guardia M; Cubic smoothing splines background correction in on-line liquid chromatography-Fourier transform infrared spectrometry. *Journal of Chromatography A*. **1217:6733-6741** (2010).

99. Palacky J, Mojzes P, Bok J; SVD-based method for intensity normalization, background correction and solvent subtraction in Raman spectroscopy exploiting the properties of water stretching vibrations. *Journal of Raman Spectroscopy*. **42:1528-1539** (2011).

100. Gorski L, Ciepiela F, Jakubowska M, Kubiak WW; Baseline Correction in Standard Addition Voltammetry by Discrete Wavelet Transform and Splines. *Electroanalysis*. **23:2658-2667** (2011).

101. Rowlands C, Elliott S; Automated algorithm for baseline subtraction in spectra. *Journal of Raman Spectroscopy*. **42:363-369** (2011).

102. Wentzell PD, Brown CD; Signal processing in Analytical chemistry. *Encyclopedia of Analytical Chemistry*. **11:9764-9800** (2000)

103. Newey WK, Powell JL; Asymmetric least-squares estimation and testing. *Econometrica*. **55:819-847** (1987).

104. Whittaker ET; On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*. **41:63-75** 1922.

105. Linag YZ, Leung AKM, Chau FT. A roughness penalty approach and its application to noisy hyphenated chromatographic two-way data. *Journal of Chemometrics*. **13:511-524** (1999)

106. Eilers PHC; Parametric time warping. *Analytical Chemistry*. **76:404-411** (2004).

107. Kauppinen J, Partanen J; Fourier transforms in spectroscopy. *Wiley-VCH*, New York, (2001).

108. Morrison N; Introduction to Fourier analysis. *Wiley*, New York, (1994).

109. Douglas AL; Discrete Fourier transform, part 4: spectral leakage. *Journal of object technology*. **8:23-34** (2009).

110. Brigham EO; The fast Fourier transform. *Englewood Cliffs*, Prentice-Hall, New Jersey, (1974).

111. Smith SW; Digital signal processing: A practical guide for engineers and scientists. Elsevier publishers, Amsterdam, Newnes, (2003).

112. Poole D; Linear algebra: A modern introduction. *Thomson Brooks/Cole,* Belmont, CA, (2005).

113. Beebe KR, Kowalski BR; Nonlinear calibration using projection pursuit regression - application to an array of ion-selective electrodes. *Analytical Chemistry*. **60:2273-2278** (1988).

114. Gnanadesikan R; Methods for Statistical Data Analysis of Multivariate Observations. *Wiley*, New York, (1977).

115. Cauchy AL; On the equation with which the secular inequalities of planetary motion is determined. Exercises in mathematics. *Oeuvres*. **2: 172-175**, (1829).

116. Pearson K; On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*. **2:559-572** (1901).

117. Hotelling HJ; Analysis of a complex of statistical variables into principal components. Educational Psychology. **24: 417-441,** (1933).

118. Malinowski ER; Factor analysis in chemistry. *Wiley*, New York, (2002).

119. Adcock RJ; A problem in least squares. *The Analyst*. **5:53-54** (1878).

120. Newey WK, Powell JL; Asymmetric least squares estimation and testing. *Econometrica*. **55:819-847** (1987).

121. Koenker R, Bassett GW; Regression quantiles. *Econometrica*. **46:33-50** (1978).

122. https://code.google.com/p/airpls/downloads/list. Last access on February 2013.

123. http://pubs.acs.org.ezproxy.library.dal.ca/doi/suppl/10.1021/ac034800e. Last access on May 2013.

124. Pereira FMV, Pereira-Filho ER, Bueno MIMS; Development of a methodology for Calcium, Iron, Potassium, Magnesium and Zinc quantification in teas using X-ray spectroscopy and multivariate calibration. *Journal of Agricultural and Food Chemistry*. **54:5723-5730** (2006).

125. Abdelsalam ME, Bartlett PN, Baumberg JJ, Cintra S, Kelf TA, Russell AE; Electrochemical SERS at a structured gold surface. *Electrochemical Communication.* **7:740-744** (2005).

126. Johnson RP, Gale N, Richardson JA, Brown T, Bartlett PN; Denaturation of sdDNA immobilised at a negatively charged gold electrode is not caused by electrostatic repulsion. *Chemical Science.* **4:1625-1632** (2013).

127. Handbook of Minerals Raman Spectra, Laboratories de sciences de la Terre CNRS, ENS Lyon. http://www.ens-lyon.fr/LST/Raman. Last access on May 2012.

Appendix A

TFALS.m MatLab® Code

```
%  TFALS  Truncated Fourier Asymmetric Least Squares method of baseline estimation
%  TFALS estimates the baseline of a signal vector xab.
%  Returns the estmated baseline xb and basline corrected vector xa
%
%  Input:
%  xab=Raw signal vector [variable,sample];Column vector
%  nf=Total number of frequency components needs to accomodate baseline
%  p=Asymmetry parameter (0.001>=p<=0.1)
%
% Choice of nf depends on the shape/frequency of baseline; linear baseline usually
% requires nf=2, more variable baseline requires higher value of nf.
%
%  Value of p depends on the noise level of raw signal vector; Reletivly lower p
%  value require for signal with high noise or higher value for vice versa
%
%  Output:
%  xb=(Estimated baseline)
%  xa=(Baseline corrected data)
%
function [xa,xb]=TFALS(xab,nf,p)
if nargin==2
   p=0.001;
end


%% Normalized Fourier basis sets
N=size(xab,1);                    % Calculate number of data points in xab
t=0:N-1;
nb=nf*2+1;
z=zeros(nb,N);
z(1,:)=ones(1,N);
fs=0.25;
for i=1:nf
   z(2*i,:)=cos((2*pi*fs*t)/N);       % calculate basis functions
   z(2*i+1,:)=sin((2*pi*fs*t)/N);     % calculate basis functions
   if fs<0.5
      fs=0.25+fs;
   elseif fs==0.5
      fs=0.5+fs;
   else
      fs=1+fs;
   end
```

```
end
for m=1:nb;
   z(m,:)=z(m,:)/norm(z(m,:));        % Normalize basis functions
end

%% Orthonormal Basis sets
[~ , ~, P]=svd((z),'econ');           % Orthogonalize basis functions
%% Asymmetric least squares
w=ones(N,1);
target=1;
while target
   W=spdiags(w,0,N,N);                % Generate sparse matrix of asymmetric weights
   bw=(P'*W);
   q=(bw*P)\(bw*xab);                 % perform least squares fit
   xb=P*q;                            % Estimate baseline
   w0=w;
   w(xab>(xb))=p;
   w(xab<=(xb))=(1-p);
   target = sum(abs(w - w0)) > 0;
end
xa=xab-xb;                            % Calculate baseline corrected signal vector
```