Population Fit Threshold:

Fully Automated Signal Map generation for Baseline Correction in NMR-based Metabolomics

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Engineering

By

DANIEL CHARLES HOMER

B.S. Biology, Wright State University, 2003

2010

Wright State University
WRIGHT STATE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

November 16, 2009

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY Daniel Charles Homer ENTITLED Population Fit Threshold: Fully Automated SM generation for Baseline Correction in NMR based Metabolomics BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF Master of Science in Engineering

______________________________
Nicholas V. Reo, Ph.D.
Thesis Co-Advisor

______________________________
Julie A. Skipper, Ph.D.
Thesis Co-Advisor

______________________________
S. Narayanan, Ph.D.
Department Chair

Committee on Final Examination

______________________________
Nicholas V. Reo, Ph.D.

______________________________
Julie A. Skipper, Ph.D.

______________________________
Michael Raymer, Ph.D.

______________________________
Ping He, Ph.D.

______________________________
John A. Bantle, Ph.D.
Vice President for Research and Graduate Studies and Interim Dean of Graduate Studies
Abstract

Homer, Daniel Charles. M.S.Egr., Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 2010.

Population Fit Threshold: Fully Automated Signal Map generation for Baseline Correction in NMR-based Metabolomics

Baseline correction of NMR-based metabolomic spectra is a key step in data processing to elucidate biomarkers of diseased and toxic states. Automated baseline correction methods often use human-selected parameters. Presented is a method for automated baseline correction using parameters selected from and without manipulation of the spectrum. The focus is on generating an accurate signal map based on the differentiating characteristics of baseline noise and signal. The presented method of signal map generation and baseline correction was developed and tested on $^{13}$C and $^1$H NMR spectra. The spectra were as simple as chemical standards containing less than 25 signal peaks to very dense urinary metabolite profiles. This method has shown proper correction of even the most complicated NMR biofluid spectra and is acceptable for use in multivariate analyses. This technique may also be robust enough to utilize in other spectroscopic methods in which thermally generated baseline noise is present.
# Table Of Contents

I. Introduction .......................... 1

1) NMR Basics .......................... 1
2) Defining Metabonomics ................. 4
3) Assumptions About Noise in NMR Spectra ........................................ 6

II. Background .......................... 9

III. Data Collection ....................... 18

1) Sample Creation ....................... 18
2) Data Acquisition ...................... 19
3) Data Sets ............................ 19

IV. Computational Methods .................. 24

1) Determination of Optimal Packet Size ................. 27
2) Similarity to a Gaussian Population .................. 31
3) Error Model Generation .................. 36
4) Threshold Selection .................... 39
5) Signal Map Generation .................. 42
6) Baseline Estimate ...................... 44

V. Results ............................... 48

1) Baseline Correction/Baseline Recognition .............. 48
2) Metabolomic/Multivariate Analysis .................. 51
3) Common Noise Identification in Data Groups .......... 57
4) Comparison of Baseline Estimates: LOWESS vs. PFT .... 58

VI. Conclusions .......................... 60

VII. References .......................... 63
### Table Of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1:</td>
<td>Baseline Roll in $^1$H Spectroscopy</td>
<td>10</td>
</tr>
<tr>
<td>Figure 2:</td>
<td>CDCl$_3$ Peaks after Sum Normalization</td>
<td>11</td>
</tr>
<tr>
<td>Figure 3:</td>
<td>Binned Spectrum Without Baseline Correction</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4:</td>
<td>Binned Spectrum After Baseline Correction</td>
<td>13</td>
</tr>
<tr>
<td>Figure 5:</td>
<td>Segment of Noise</td>
<td>16</td>
</tr>
<tr>
<td>Figure 6:</td>
<td>Data Set 1; $^{13}$C Homogenized Tissue Lipid extract</td>
<td>20</td>
</tr>
<tr>
<td>Figure 7:</td>
<td>Data Set 2. $^1$H Urine Metabolite Profile</td>
<td>21</td>
</tr>
<tr>
<td>Figure 8:</td>
<td>Data Set 3. $^{13}$C Lipid Standard Profile</td>
<td>22</td>
</tr>
<tr>
<td>Figure 9:</td>
<td>Computational Method Flowchart</td>
<td>26</td>
</tr>
<tr>
<td>Figure 10:</td>
<td>$^1$H Metabolite Profile with Packet Containing Greatest Variance</td>
<td>29</td>
</tr>
<tr>
<td>Figure 11:</td>
<td>Packets Widths and $\sigma$</td>
<td>30</td>
</tr>
<tr>
<td>Figure 12:</td>
<td>Gaussian Model: LMS Error Generation</td>
<td>33</td>
</tr>
<tr>
<td>Figure 13:</td>
<td>$^{13}$C Standard Gaussian Measure Profile</td>
<td>36</td>
</tr>
<tr>
<td>Figure 14:</td>
<td>$^1$H Biofluid profile Error from Gaussian</td>
<td>40</td>
</tr>
<tr>
<td>Figure 15:</td>
<td>Signal Map Grooming</td>
<td>42</td>
</tr>
<tr>
<td>Figure 16:</td>
<td>Signal Map Expansion</td>
<td>44</td>
</tr>
<tr>
<td>Figure 17:</td>
<td>Pareto front from Fidelity/Smoothness Parameters</td>
<td>45</td>
</tr>
<tr>
<td>Figure 18:</td>
<td>$^{13}$C Lipid Profile SM and Corrected Profile</td>
<td>49</td>
</tr>
<tr>
<td>Figure 19:</td>
<td>$^{13}$C Lipid Profile SM and Corrected Profile</td>
<td>49</td>
</tr>
<tr>
<td>Figure 20:</td>
<td>$^1$H Urine metabolite Profile</td>
<td>50</td>
</tr>
<tr>
<td>Figure 21:</td>
<td>PCA of Expert hand corrected and Population fit threshold baseline corrected spectra</td>
<td>53</td>
</tr>
<tr>
<td>Figure 22:</td>
<td>PCA of 16:0 lipid standard: Baseline Corrected vs. Non-Baseline Corrected</td>
<td>54</td>
</tr>
<tr>
<td>Figure 23:</td>
<td>PCA of lipid standards: Sum Normalized</td>
<td>56</td>
</tr>
<tr>
<td>Figure 24:</td>
<td>PCA of lipid standards: Sum Normalized and Baseline Corrected</td>
<td>56</td>
</tr>
<tr>
<td>Figure 25:</td>
<td>Comparison of Baseline Estimates</td>
<td>59</td>
</tr>
</tbody>
</table>
I. Introduction

1) NMR Basics

Nuclear magnetic resonance (NMR) is a physical phenomenon that can be observed using a magnetic field and radio frequency (RF) pulses to manipulate and measure the atomic spin property of charge-unbalanced nuclei possessing the fundamental nuclear quantity of spin. The most prevalent nuclei measured in biologically relevant studies are $^1\text{H}$, $^{13}\text{C}$ and $^{31}\text{P}$. Each of these molecules has a spin quantity ($I$) of $\frac{1}{2}$. This means that there are two specific energy ($E$) states separated by an energetic quantity ($\Delta E$) that the nuclear spins can possess, referred to as spin up or spin down. The difference in population of these energy states is determined by the Boltzmann statistical model. The difference in spin populations determine the magnitude of a bulk magnetization vector created by the alignment of nuclear spins along the $B_0$ magnetic field. Application of an orthogonal energy pulse in the form of RF waves at the appropriate $\Delta E$ quantities will serve to flip the bulk
magnetization vector along transverse planes (\(u\) and \(v\) planes in a rotating frame of reference) for measurement.

These bulk vectors will precess (rotate) about the central \(Z\) plane, defined by the direction of the \(B_0\) magnetic field, at frequencies dependent on the strength of the applied magnetic field, the neighborhood and chemical connectivity of the nucleus under observation and that nuclei’s specific magneto-gyric ratio (\(\gamma\)). The magnitude of these bulk magnetization vectors represents the number of nuclei under the influence of the magnetic field. The frequency of precession is called the Larmor frequency (\(w_0\)) and is defined as \(w_0 = \gamma B_0\). The magnetization vectors will then relax back along the \(Z\) plane (T1 relaxation) or will undergo energetic phase dispersion within the \(u\) and \(v\) planes due to energy loss within the lattice (T2 relaxation). These relaxation processes and rotational frequencies are received by an RF coil and provide an exponentially decaying signal of complex sinusoids. The magnitude and frequencies of the relaxation signals can be extracted using a Fourier transformation.

The precessional rotation of the bulk magnetization vector along the \(u\) and \(v\) planes produces a measurable electro-motive force that is received by the excitation probe (RF coil) within the NMR machine. The measured signal
stemming from the atomic spins is called the Free Induction Decay (FID). It is a composite sinusoidal signal reflecting the resonance frequencies of the atoms decaying in an exponential fashion based on the decay equations from T1 and T2 relaxation.

Since all similar nuclei have the same magneto-gyric ratio, the differences in frequency from the prescribed resonance frequency of that molecule are caused by the local nuclear environment: to what it is covalently connected and, to a lesser extent, what is within the general area surrounding the nucleus. A specific example is the $^{13}$C profiles of fatty acids. In these spectra, each carbon atom within the long chain fatty acid (16 to 22 carbons in length) is usually individually distinguishable as a peak within the Fourier-transformed FID. If atoms have a similar environment, peaks may be indistinguishable aside from the increased area under that specific peak.

All carbon FIDs are observed while the $^1$H signal is decoupled. Decoupling is a method of saturating the spin states by constantly bombarding that specific molecule with its prescribed energy, altering the spin states so quickly that there is no coherent effect exhibited within the resonance spectrum. This provides single peaks instead of the 1:2:1 J-coupled triplets (for carbons bound to hydrogen
atoms) exhibited by molecules bound to spin ½ nuclei (such as $^1\text{H}$).

2) Defining Metabonomics

Metabonomics, metabolomics and metabolic profiling are all terms used to define a quantitative measurement of the altered metabolism of organisms under stress from various states of disease or toxicity. This science has shown that biological systems under the induced or natural stress of disease or toxic insult can be differentiated and mapped through multivariate analysis techniques, mainly principal component analysis (PCA), according the progression and severity of the acute state. The power in metabonomics lies in its ability to differentiate between states of pathophysiology present in biological samples.

Metabonomics, as a word, was specifically defined in 1999 by Nicholson, et al., 1999 as “the quantitative measurement of the time related multi-parametric metabolic response of living systems to patho-physiological stimuli or genetic modification.” Metabolomics has been defined separately by Fiehn, 2001 as the “comprehensive and quantitative analysis of all metabolites...” The distinction between the two terms appears subjective as presented in literature and is usually determined by the
preference of the author. There appears to be subtle differences in the ultimate goals of the two interconnected fields.

It is continually argued which field (metabolomics and metabonomics) belongs as a subset of the other (Reo), but the most simple differentiation of the groups may be identified through their final purposes. Metabonomics generally measures the timeline progression of metabolic products through physiologic states of stress. Metabolomics, specifically derived from metabolome, denotes a snapshot entirety of the biological chemical profile of a living organism. However, this is almost impossible to achieve. Each term still refers to quantification of organism and system level metabolic profiles in an attempt to garner information on the metabolic phenotypes of disease, toxic and acclimation states of an organism.

Metabolomics currently has many in use applications and theoretical directions. The data, either nuclear magnetic resonance (NMR) spectra from a myriad of spin-sensitive atomic molecules (mainly $^1$H, $^{13}$C ad $^{31}$P) at high tesla (11.7T to 18.7T), high resolution states, or various mass spectrometry (MS) experiments, are processed with various chemometric methodologies. The information garnered from these techniques expands from botanical and
environmental sciences into clinical applications. The identification of specific disease states and even the severity of the condition, toxicological and drug studies such as drug screening, biomarker identification in toxicity and underlying mechanisms of toxicity are all interests of metabolomics.

3) Assumptions About Noise in NMR Spectra

Halouska, et al., 2006 in their analysis of noise on PCA projection and group analysis, presented an interesting observation. It was shown that segments of noise in known chemical profiles showed the highest component loadings in PCA analysis, which artificially influenced the projection values and, ultimately, group clustering. Removal of these segments of noise brought the spectra back into a 95% confidence index for similar chemical profiles in the PCA space. The following method serves as a tool to distinguish noise from signal in NMR spectra. The basis of a population fit threshold (PFT) comes from certain assumptions about the nature of noise compared to signal in the NMR spectrum.

The first assumption is that noise present in NMR spectra is of a Gaussian nature. Grage, et al., 2003, in their paper involving a statistical analysis of NMR spectrometer noise, show that the assumption of
uncorrelated real and imaginary portions of the noise sequence (white noise) is valid for single receiving coil systems using quadrature detection (as are most NMR spectroscopy machines). The conclusions state that noise in complex NMR signal is generally uncorrelated and Gaussian in nature. This suggests that spectral noise is created thermally and is a product of the receiving coil amplifier.

The second assumption about the nature of NMR spectra is that segments of noise can be characterized as having a lower standard deviation (σ) than a sample population containing signal, even for very low amplitude signals. This assumption has been made previously in Koradi’s AUTOPSY program for automated peak picking in macromolecular NMR spectra. Here, Koradi, et al., 1998 assume that a baseline noise level can be determined by finding a portion of the spectrum representing 5% of the spectral width having the minimal σ of any linear set of the sampled population.

The initial intent in the undertaking of this project was to remove purported salient bins representing spectral segments containing only noise from the linear discriminate analysis (LDA) of normalized and data-reduced (binned) datasets. LDA was applied to extracted lipid spectra of 2,3,7,8-tetrachlorodibenzo-p-dioxin (TCDD) insulted mouse
liver in which the metabolic changes include a gross increase in fatty acid composition. This stems from an increased lipid metabolism and a decreased ability of the insulted liver to process the released fatty acids from TCDD poisoning. LDA was also applied to spectra from homogenized mouse brain segments under varying atmospheric conditions yielding spectral segments containing only baseline noise.

The end product of the investigation has become a general noise removal and baseline correction tool for NMR and, potentially, other spectroscopic methods. The method of noise identification is based on describing noise as having a specific statistical population distribution (Gaussian) and signal that can be classified as having a $\sigma$ higher than that of the sampled noise regions. This method is shown to overcome the limitations of linear threshold selection based on binned (reduced variability) spectra and the subjectivity of user defined noise regions. These two methods pose very serious risks in removing low level metabolites from the metabolic spectrum. Noise removal before the reduction of data variability is the most logical method for accurate noise removal. This method does not remove low lying metabolites or broad signals from large molecule residuals.
II. Background

Metabolomics, the science of mapping metabolite profile differences under states of physiological stress (disease, toxicity, environmental acclimation), is emerging as an important tool in clinical medicine. Multivariate analysis (MVA) techniques, such as principal component analysis (PCA) and other discriminate analyses, are the workhorses of these toolsets. However, all of these techniques rely on an accurate model of the spectroscopic data used to deduce information. Here is proposed a method for baseline correction, i.e., a process for flattening baseline for data interpretation, using parameters selected only from the data.

Nuclear magnetic resonance (NMR) spectroscopy is a powerful tool in small molecule metabolite profiling of body fluids and tissue extracts. These spectra are readily used in metabolite identification and quantification. However, to glean the useful information from sets of NMR spectroscopic metabolomic data, one must perform rigorous preprocessing techniques to prepare the data for further MVA techniques.
NMR spectra can be phase corrected, baseline corrected and normalized or scaled according to the types of information desired. Usually, data reduction techniques are applied (especially binning procedures) before application of MVA techniques. Each step has an impact on all subsequent procedures. Baseline roll is illustrated in Figure 1. Despite efforts to control baseline roll artifact in

![Figure 1](image.png)

**Figure 1**: Two $^1$H spectra showing baseline roll artifacts. The spectra have been expanded vertically to emphasize the baseline distortion. The baseline roll is greater than some known peaks.

NMR spectroscopy, sources still include analog filter applications for sweep width, acoustical ring within the coils of the magnet, any residual solid state (large molecule) material left in samples as well as FID artifacts from digitization.
Examples illustrating the importance of accurate baseline correction are shown in Figures 2 through 4. Figure 2 illustrates relative metabolite levels after sum normalization with and without baseline correction. The figure shows CDCl$_3$ peaks from the same lipid standard across multiple runs. Without proper baseline correction, same sample spectra will greatly vary in overall intensity.

**Figure 2:** CDCl$_3$ peaks from different runs of the same lipid mixture sample. Inset A shows ten phase corrected center CDCl$_3$ peaks that have been sum normalized without baseline correction (13.5 ± 0.55*). Peak heights and overall signal integral values are highly varied due to artifacts induced by sum normalization in combination with baseline roll of sparsely populated spectra. Inset B shows the same spectra after successful baseline correction and the same preprocessing techniques (9.38 ± 0.23*). * is mean ± σ of peak heights.
simply due to baseline distortion. In Figure 3 we see a

![Graph showing baseline distortion]

**Figure 3:** Data reduced (binned) $^{13}$C lipid standard profile over the original spectrum. Here one can visualize the increase in baseline distortion caused by binning a spectrum that has not been successfully baseline corrected. Binned (mean) data has been scaled from original value.

vertical expansion of binned baseline (orange) imposed over its original (unbinned) baseline (blue) from a $^{13}$C lipid profile (see Dataset 3). Figure 4 is the same spectrum after baseline correction.
In lieu of the existing body of literature dedicated to baseline correction of NMR spectroscopic profiles, we have decided to pursue another more simplistic approach to the identification and modeling of baseline distortions in NMR spectroscopic data. A standard method for correcting baseline in NMR-based spectroscopy involves the selection of signal and noise from the spectrum and subsequent creation of a binary signal map (SM). The SM is then used to construct a baseline model whereupon the model is subtracted from the original spectrum, yielding a flat baseline. The same general methodology is applied here and

Figure 4: The same spectra as in Figure 3 after successful baseline correction. Baseline correction removes the artifact roll induced by binning procedures.
is generally termed the Pearson method of baseline correction (G. A. Pearson). There exist other techniques for baseline correction which do not rely on a Pearson methodology. However, each of these methods relies on human based parameter selection in determination and correction of baseline in NMR spectroscopy. Smoothing parameters are tunable (Golotvin and Williams) and based on preferred visual outcome. Packet size (the choice of how large of sample to take) is also based on heuristic practice [ (Golotvin and Williams), (Chang, Banak and Shaw), (Xi and Rocke)]. Baseline model smoothness and fidelity parameter optimizations [ (Xi and Rocke), (Cobas, Bernstein and Marin-Pastor)] can be removed from analysis. Peak modeling of selected signal regions for model refinement (Chang, Banak and Shaw) and other human decided parameters can be removed from baseline correction methodology (Table 1).

The method described here uses two simple assumptions regarding signal and noise in NMR chemical profiles and metabolomic data sets. The first assumption is that noise in NMR and other spectroscopic methods is thermally generated. This means that baseline noise (sections of the spectroscopic profile not containing signal) will be of a white and Gaussian nature (Grage and Akke) (Figure 5).
Table 1: List of baseline correction methods along with noted strengths and weakness

<table>
<thead>
<tr>
<th>Method</th>
<th>Strength(s)</th>
<th>Weakness(es)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pearson (1977)</td>
<td>3-step methodology introduced</td>
<td>sparse signals only</td>
</tr>
<tr>
<td>CWT (2006)</td>
<td>very sensitive; better than Dietrich; good BL model</td>
<td>3 adjustable parameters; altered signal for SM; requires nearest neighbor or erosion filter</td>
</tr>
<tr>
<td>Chenomx (2007)</td>
<td>use of frequency filter</td>
<td>lorentzian fitting; arbitrary parameter selection</td>
</tr>
<tr>
<td>LOWESS (2008)</td>
<td>good signal model; well derived parameters</td>
<td>sigma based on LOWESS estimation; arbitrary sampling size</td>
</tr>
<tr>
<td>IFLAT (1995)</td>
<td>probability based estimation of BL vs. signal; weighted BL model construction</td>
<td>criticized as a post processing, solvent suppression technique; best implemented for 2D NMR</td>
</tr>
<tr>
<td>Golotvin (2000)</td>
<td>Intuitive; simple to implement</td>
<td>sparse signal only</td>
</tr>
<tr>
<td>AUTOPSY (1998)</td>
<td>intuitive</td>
<td>2D NMR; Peak selection, not baseline correction</td>
</tr>
</tbody>
</table>

White has two meanings here. The first, most intuitive definition says white noise fluctuates evenly about a mean baseline value (0 amplitude). In NMR spectroscopy using quadrature detection, white also means that the real and imaginary observation channels are uncorrelated. The former definition of white is the assumed in the algorithm development.
The second and most easily observable assumption is that a packet of noise has a much lower $\sigma$, defined in Eq. 1, than a similar representative sample of signal (Koradi, Billeter and Engeli). Specifically, white noise with amplitudes fluctuating about a mean value will have a lower $\sigma$ than rising and falling peak edges from NMR signal. It is hypothesized that use of these two differentiating factors

\[
\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2} / (N - 1)
\]  

(1)

Figure 5: Sample of baseline noise from a $^1$H NMR metabolite profile as in Dataset 2. The segment shown is representative of a large standard bin width used in data reduction for $^1$H based MVA. The spectrum has a 0.25 Hz per point resolution showing the 128 points associated with the bin.
will allow us to accurately determine baseline regions in any given NMR spectrum, thus allowing us to create an accurate SM.

The SM will then be used for baseline estimate construction in NMR metabolite profiles with the goal of comparing and quantifying differences. The goal is to perform this task without the use of heuristic and human decided parametric input.
III. Data Collection

1) Sample Creation

NMR metabolite profiles were obtained from rat and mouse biofluid samples while these animals were undergoing toxicity studies previously pursued by our laboratory. The profiles generated are $^1$H (densely populated spectra with approximately 32K points) and $^{13}$C (relatively sparsely populated spectra with approximately 130K points) spin experiments on urine and homogenized tissue extracts with a lipid phase extraction, respectively. Another sparsely populated set of spectra was generated from $^{13}$C lipid standards ranging from 16 to 22 carbons obtained from Sigma-Aldrich. All lipid standards used were >95% pure chromatography grade standards, and were prepared at a 50mM concentration in CDCl$_3$ solvent. Mixtures of lipids standards were created in a 1:1 mixture at a total concentration of 50mM. The internal reference standard, TSP (trimethylsilyl propionic (2, 2, 3, 3 d$_4$) acid), was added to lyophilized urine samples for reconstitution.
2) Data Acquisition

All spectra were acquired using a Varian Inova 600 NMR spectrometer (14.1T) operating at a $^1$H resonance frequency of 600 MHz. $^1$H spectra were acquired using standard NOESY pulse sequences with water suppression techniques. Generally, 400 transients were performed on $^1$H urine samples as long as signal to noise was acceptable. $^{13}$C spectra were acquired using 1407 transients to ensure excellent signal to noise ratios. A NOESY pulse sequence was used along with $^1$H decoupling. FIDs were processed using exponential multiplication with line broadening factors of 0.3 and 0.5 Hz for $^1$H and $^{13}$C spectra, respectively. Fourier transformations as well as visual phase correction were also performed using Varian 6.1C software before being written out as tab delimited frequency versus amplitude files.

3) Data Sets

The spectra used for this analysis are representative of NMR metabolite profiles and chemical standard profiles with a range of signal densities. The $^{13}$C lipid extract from homogenized mouse liver will be denoted as Dataset 1. The $^1$H urine metabolite profiles will be denoted Dataset 2 and the $^{13}$C lipid standard profiles will be denoted Dataset 3.
Visual illustrations of the spectra are presented in Figures 6 through 8.

Dataset 1 (Figure 6) is composed of moderately dense spectra. The spectra are most densely populated with single and double bonded carbons (aliphatic and olefinic carbons).

Figure 6: A representative spectrum from Dataset 1, $^{13}$C spectra of a homogenized mouse liver lipid extract at 14.1T (150.821 MHz). The resonance from the solvent, CDCl$_3$, is a triplet centered at 79.9 PPM. Methanol is also present as part of the solvent and resonates at about 50 PPM.

The spectra are most densely populated with single and double bonded carbons (aliphatic and olefinic carbons).
from fatty acids ranging from 20-40 and 120-140 PPM. Large expanses of noise are found between 80-120 PPM. Solvent peaks, methanol and deuterochloroform (CDCl₃), exist at 50 and 77 PPM, respectively. Dataset 2 (Figure 7) represents a very complicated ¹H urine metabolite spectrum. TSP is an internal standard referenced at 0 PPM. The residual signal

![Data Set 2, ¹H Urine Metabolite Profile](image)

**Figure 7:** A representative spectrum from Dataset 2, ¹H urine profiles from rats. The residual water signal is at 4.8 PPM (out of phase due to solvent suppression techniques) and the signal from the TSP internal standard is set to 0.0 PPM.
from water is located at 4.8 PPM, which has been experimentally suppressed in this spectrum. Dataset 3 (Figure 8) contains many of the same features as Dataset 1, though the spectra contain a mixture of known lipid compounds. These compounds include biologically relevant saturated and unsaturated fatty acids with 16 to 22 carbons, beginning with palmitic acid (16:0) and ending

Figure 8: $^{13}$C profile representative of Dataset 3, eicosatrienoic acid (20:3) at 50mM concentration with CDCl$_3$. 
with docosahexaenoic acid (22:6). CDCl$_3$ peaks, single and double bonded carbon signatures as well as similar regions of noise are present. 10 spectra for each lipid standard and mixture sample were run back to back over a 14 hour period.
IV. Computational Methods

A stepwise procedure for our methods of NMR SM determination and subsequent baseline correction (flattening) is given below and illustrated in Figure 9. Each of these procedures is described in detail and tested using NMR spectra \( (^1\text{H}, ^{13}\text{C}) \) obtained on a Varian Inova 600 spectrometer.

The computation procedure is as follows:

1. Determine packet size by finding packet size associated with maximum observed \( \sigma \) from the spectrum as outlined in Determination of Optimum Packet Size (section IV, 1).

2. Segment the spectrum into packets of equal width using the determined packet size.

3. Calculate the \( \sigma \) for each packet and sort packets in order of least (assumed noise) to greatest (assumed signal) \( \sigma \).

4. Add mean corrected amplitudes from \( \sigma \) sorted packets and calculate the kurtosis of the set (Section IV, 3).
5. Select a threshold based on the critical value of the kurtosis measure (Section IV, 4), delineating the packet on the boundary of noise and signal.

6. Groom the SM (Section IV, 5).

7. Check if the SM is the same as the last generated SM. If not, repeat steps 4 through 6 with packets selected as signal removed from further analysis. If yes, continue on.

8. Expand signal boundaries to include Lorentzian tails and groom again (Section IV, 5).

9. Estimate a baseline using the selected regions of noise with a smoothing function. (Section IV, 6) Areas of signal are interpolated with a linear segment.

10. Subtract the baseline estimate from the original spectrum to yield a flattened baseline.
Figure 9: Flowchart for baseline recognition and flattening algorithm
We will be using metabolite profiles represented in Datasets 1 through 3 to illustrate the method of packet size selection, SM generation and baseline correction.

1) *Determination of Optimal Packet Size*

Packet size is an important feature in several baseline correction algorithms [(Golotvin and Williams), (Chang, Banak and Shaw), (Xi and Rocke)], as it is in this one, an optimization routine was created for packet size determination. In determination of packet size, the packets need to be as small as possible to maintain the high-resolution detail contained in the original data (to minimize data reduction), but they must also be large enough to provide an accurate measure of the $\sigma$ reflecting the contents of the packet. This is important as $\sigma$ will be used to differentiate between packets containing signal and those containing noise.

The first procedure in several baseline correction algorithms involves dividing the spectrum into non-overlapping, equal size samples, or packets. Amplitudes contained in each packet are then subject to a measure of $\sigma$ followed by a cumulative packet evaluation of kurtosis to determine whether the packet is representative of signal or noise. In these procedures, the packets will be sorted by $\sigma$
and a specified value of packet $\sigma$ will become the packet inclusion threshold. Therefore, we vary packet size and analyzed the $\sigma$ values for a spectral data group in order to optimize the packet width selection.

An overlapping filter is used to select and calculate packet $\sigma$. The maximum $\sigma$ value, $SD_{\text{max}}$, is recorded for each packet size. The maximum recorded sigma for each spectrum corresponds to a specific packet width and is selected as the signal-noise differentiation threshold.

In Figure 10, an example of the packet with the largest $\sigma$ is highlighted against the rest of the spectrum. This packet is associated with rising or falling edges of the peak with the greatest change from baseline (greatest peakedness); this usually corresponds to the peak with largest amplitude. Nominally, the selected packet size is approximately twice the full width half maximum (FWHM) value of observed peaks. As illustrated in Figure 11, the maximum $\sigma$ value corresponds to the greatest difference between $SD_{\text{max}}$ and $SD_{\text{min}}$ of a particular packet width, offering the greatest differentiation of signal samples. Therefore, signal resolution plays the most important role in optimum packet size selection. Because the greatest $\sigma$ is associated with signals possessing greatest peakedness, it is important to remove solvent signals and remaining water
signal from the spectrum before determining packet size. Even though solvents do well to reflect the resolution of peaks, they do not always correspond to actual metabolite peaks.

The sliding filter was applied starting at a minimum packet width of approximately 1 Hz (five data points), or about the FWHM of a Lorentzian NMR signal in our profiles. \( \sigma \) was calculated at each position. This procedure was repeated as the packet width incrementally increased to a maximum of 7.5 Hz, or 30 points.

Figure 11 also shows a plot of the \( SD_{\text{max}}/SD_{\text{min}} \) ratio as well as \( SD_{\text{min}} \) for each spectrum in the dataset as a function
Figure 11: Panel A shows maximum $\sigma$ values corresponding to optimum packet sizes for a group of spectra as in Dataset 2. Selection of packet size is most strongly a function of spectral resolution and maximum peak height. The maximum SD value positions are marked by a black X. The maxima located at 18 and 19 points are reflective of spectra where the TSP peak is the dominating signal in the profile. Panel B shows the increasing $\text{SD}_{\text{min}}$ value against increasing packet size. Panel C shows the $\text{SD}_{\text{max}}/\text{SD}_{\text{min}}$ ratio and stabilization of $\sigma$ values over increasing packet width. The sampling resolution of the above $^1\text{H}$ spectra is 0.244 Hz/point.

of packet width. All spectra from similar cohorts show a similar profile, where the $\text{SD}_{\text{max}}/\text{SD}_{\text{min}}$ ratio is highest for small packet sizes and precipitously decreases against increasing packet width until it reaches a stable level at widths $\geq 20$ points (~5Hz). The $\text{SD}_{\text{max}}/\text{SD}_{\text{min}}$ ratio stabilizes over increasing packet size. For the $^{13}\text{C}$ spectral example
shown in Figure 3, the ratio stabilizes at 5 – 8 Hz (at least 20 data points). We repeated this procedure for spectra of other NMR nuclides (\(^{31}\)P and \(^{1}\)H) and obtained similar profiles. The ratio stabilizes at packet widths of 12 Hz (40 data points) for \(^{31}\)P spectra of homogenized brain tissue and at 4 – 15 Hz (at least 16 points) for the \(^{1}\)H urine metabolite profiles.

Also shown in Figure 11 is SD\(_{\text{min}}\) as it changes over increasing packet width. A continual increase over the selected packet widths suggests that there is no differentiating quality to the SD\(_{\text{min}}\) value. For this reason, we simply take the packet size corresponding to the maximum \(\sigma\) providing us with the greatest difference between SD\(_{\text{min}}\) and SD\(_{\text{max}}\).

Selection of the optimum packet size for each spectrum is easily determined by taking a sample of the spectrum around the peak representing greatest amplitude. This eliminates the need to analyze the entire spectrum and greatly reduces the computational time taken to determine optimum packet width.

2) Similarity to a Gaussian Population

Following segmentation based on packet size, comparison of the amplitudes within \(\sigma\)-sorted packets to a
Gaussian population is performed to determine a threshold for packets containing noise versus signal. The amplitudes within a given packet are corrected by the packet mean to prevent skew in the Gaussian model.

Packets, sorted by packet σ, are compared to a Gaussian model as they are added sequentially to the sample population. An intuitive histogram binning methodology and comparison to a predicted Gaussian population has been investigated and is presented in Figure 12; this approach is discussed more rigorously in Section IV, 3, Error Model Generation. To validate this model, an Anderson-Darling coefficient (Eqs. 2 and 3) can also be used to determine a threshold based on determined critical values for normal populations \(A^2 = 0.74\). In Eq. 2, \(A^2\) provides the critical value for null hypothesis determination of a system’s ability to not be rejected as normal. In Eq. 3, \(F\) denotes the cumulative distribution function. \(Y\) is the population

\[
A^2 = -n - S
\]  

\[
S = \sum_{k=1}^{n} \frac{2k-1}{n} \left[ \ln F(Y_k) + \ln(1 - F(Y_{n+1-k})) \right]
\]  

under investigation; in this case \(Y\) is the mean corrected amplitudes contained within the included packets. \(n\) denotes the number of samples in a population whereas \(k\) marks the progression through the samples. However, this
rejection of the null hypothesis – that the population under investigation cannot be rejected as normal – is a poor choice for selection of a Gaussian system.

Instead, to select the threshold, the kurtosis (Eq. 4)

Figure 12: Histogram of included packet-sorted, mean-corrected spectral amplitudes. 819 packets represent 25% of the lowest $\sigma$ packets included in the analysis with each subsequent inset representing another quartile inclusion of packets. The red line is the prescribed Gaussian population based on the maximum occurring bin and the $\sigma$ of the amplitudes included based on packet $\sigma$. It is evident that our sample populations begin to deviate from Gaussian distributions once between 50% and 75% packets have been included.
of the packet sorted amplitudes is calculated and a critical value (zero crossing) is selected. Kurtosis was chosen because it is the standard method of determination of a population’s similarity to a Gaussian distribution and because it offers computational efficiency. Here, \( n \) represents the total number of points in the analysis, and \( x_i \) is the value of the distribution at position \( i \). In this algorithm, \( \bar{x} \) will always be zero.

The histogram binning methods follow NIST recommended parameters for statistical population comparisons (NIST/SEMANTECH). One by one, the packets of spectral amplitudes are added to the histogram and the result is compared to a Gaussian system (Eq. 5) by means of a normalized sum of differences squared metric (Eq. 6). The predicted Gaussian model is generated using Eq. 5 where \( A \) is the maximum occurrence (usually at \( 0 \pm 0.15\sigma \) unless the sample size is small) in the histogram and \( \sigma \) is the \( \sigma \) of the spectral amplitudes currently in the histogram.

\[
K = \frac{\frac{1}{n} \sum_{n} (x_i - \bar{x})^4}{\left[ \frac{1}{n} \sum_{n} (x_i - \bar{x})^2 \right]^2} - 3 \tag{4}
\]

\[
G = A e^{-\left(\frac{(x)^2}{2\sigma^2}\right)} \tag{5}
\]
In Eq. 6, $E_i$ is the measured error from the predicted Gaussian population ($G$) at the $i^{th}$ packet inclusion. It is a sum of

$$E_i = \sum_{l-k}^{i} \frac{(n(x)-G(x))^2}{\sum n}$$

the squared differences of the histogram values and the corresponding Gaussian values at that histogram bin center ($x$), normalized to the total number of points included in the analysis, with $k$ and $l$ representing the minimum and maximum histogram bin centers, respectively.

Amplitudes contained within packets are always mean corrected to prevent skew of the Gaussian population. The maximum of the histogram usually occurs around the zero (center) bin. Histogram bin widths are set to 0.3σ and external histogram bin centers are set to ±5.85σ to encompass >99.99% of a standard Gaussian system. Figure 12 shows a histogram of packet amplitudes as they are inclusively fit and compared to a Gaussian model.

Packets with relatively low σ values yield an excellent fit to a Gaussian system, whereas addition of packets with greater σ cause the histogram to shift away from the predicted Gaussian model. This is consistent with our hypothesized model (NMR spectroscopic noise originating from thermally generated electronic noise sources like the
receiving coil amplifier producing a Gaussian distributed system of noise amplitudes).

3) Error Model Generation

Figure 13 shows three difference from Gaussian metrics against packet inclusion. The most striking feature of the plotted difference from Gaussian metrics is their similar placement of critical values for threshold selection. Also shown is that the LMS error from Gaussian metric is almost identical to the Anderson-Darling coefficient. The shape of

![Figure 13: Three measures of difference from a Gaussian population. These difference curves are characteristic of $^{13}$C profiles as described in Dataset 1. The included packet, error measure coordinate displayed shows the automated threshold selection based on the kurtosis of included signal amplitudes as included by packet $\sigma$.](image)
the LMS error and Anderson-Darling metric provide interesting insight into the nature of the statistical model presented. This error model is characteristic of fitting large systems of thermally generated noise (baseline from a sparsely populated spectrum). Other more densely populated spectra show similar shaped Anderson-Darling curves and very similar kurtosis measures, however, the LMS error from Gaussian begins to fluctuate creating several local minima before the critical inflection point. This fluctuation in the LMS error curve stems from fitting increasing packet amplitudes of Gaussian noise through the course of the spectrum and the subsequent filling of the histogram as more noise packets are included with similar distributions.

While the Anderson-Darling test statistic is not useful as a threshold selection tool due its nature of rejecting the possibility of a Gaussian (normal) distributed system, it does provide a statistical basis for understanding the complexity of the model. At the very low packet inclusions (low SD) we see erratic behavior of the error curve followed by a gently rising and falling section (1000 to 6500 packets). The error curve then increases abruptly at an included signal σ value of 0.49 (~6700 included packets) where included packet amplitudes begin to
deviate from the predicted Gaussian model. The erratic behavior at the beginning of the error plot comes from small sample size (low $N$). The very large rise at the right side of the curve shows where included packets deviate from a Gaussian nature disrupting the fit.

The gently rising and falling section before the last large increase in error suggests a model of competing variance and is a product of the packet sorting method. It is a balance between adding packets with increasing $\sigma$, and the law of large $N$. As packets are added to the model, the error from Gaussian initially increases. However, the error returns to a basal level as the number of included packets increase (law of large $N$). The minimum of the Anderson-Darling plot and the LMS error metric is indicative of the portion of signal providing the best fit to a Gaussian system and is based on the inclusion of the maximum number of packets representing pure noise.

The measured kurtosis of the sorted and inclusively fit packets provides the most straightforward and concrete understanding of the Gaussian nature of noise in NMR spectroscopy. The plot of kurtosis against increasing packet inclusion shows that the zero crossing is reached very close to the critical values provided by the LMS error and Anderson-Darling coefficients. There is generally only
one zero crossing from the measure of kurtosis. If there are more, they are either located with very low packet inclusion numbers (less than 5% of total spectral width) or within a few packets within the critical zero crossings near LMS and Anderson-Darling critical values.

4) Threshold Selection

The purpose of the difference from Gaussian metric is to differentiate packets as noise or signal based on their similarity to a Gaussian model. Using the LMS error method, there were several possible choices in the selection of a threshold position all based on commonly occurring phenomenon within the error plot (Figure 13 and 14). These initial choices included the local minimum of the error plot associated with a range representing the estimated percentage of baseline noise in the spectrum and, similarly, zero points within the first derivative of the error plot associated with estimated percentages of baseline noise. Both of the above methods require the additional parameter of estimated percent baseline, as in (Chang, Banak and Shaw), of the spectrum under investigation and not all of the error plots exhibited such behavior (i.e. more densely populated spectra). There is a point of inflection in the upward swing of the LMS
error curve that represents a very small reduction in error from Gaussian, but holds no statistical significance. It was for these reasons that the error metric was dismissed as a potential threshold selection method.

The Anderson-Darling coefficient also yields critical values. The least $A^2$ value representing the greatest inclusion of low $\sigma$ packets can be selected as the threshold position on the basis that it has the least likelihood of being rejected as a Gaussian system. However, the measure of Kurtosis allows the most straight-forward determination of critical value.

**Figure 14:** Three measures of difference from a Gaussian population. These difference curves are characteristic of $^1$H profiles as described in Dataset 2. The included packet, error measure coordinate displayed shows the automated threshold selection based on the kurtosis of included signal amplitudes as included by packet $\sigma$. 

![Graph showing error from Gaussian](image-url)
The zero crossing of the kurtosis value is the most suitable choice of threshold selection. Upon the first iteration of noise identification, a special precaution is made to remove only signal packets for further processing. Once the packet associated with the kurtosis zero crossing is determined, all packets having a lower σ value have the cumulative σ calculated. The initial threshold is six times the σ value (6σ) of the sample population up to the selected threshold. This is to ensure that only signal is selected and removed from further iterations. This provides a very conservative threshold necessary to the application of an iterative process. Upon further iterations, the critical kurtosis value then becomes the differentiating value for threshold selection. The SM usually converges within four or five iterations. The zero crossing for the kurtosis value is highlighted in Figures 13 and 14. The σ value associated with the packet incurring the critical value is denoted ζ.

In practice, it is more efficient to calculate the kurtosis values of the entire σ sorted population and remove packets as kurtosis is calculated. Since subsequent iterations contain less and less signal, the number of calculations to reach the zero crossing greatly decreases.
5) **Signal Map Generation**

A signal map is a binary mask applied to each spectrum under investigation. This is a simple way of denoting signal and noise within the spectrum. A Boolean value is assigned to regions found to be signal and those found to be noise. The packets with $\sigma \geq \zeta$ are labeled as signal (1 is selected signal), and those with $\sigma < \zeta$ are labeled as noise (0 is selected noise).

Upon each iteration, the SM is subjected to a simple grooming process. The checks made to the SM are two-fold. The first check looks for a minimum number of contiguous packets with $\sigma \geq \zeta$. The highlighted packets are those that have been removed and noise selected packets that do not represent at least four contiguous packet widths.

**Figure 15**: A SM grooming procedure performed on the SM of the first iteration of signal and noise selection. The green line represents the groomed SM while the red line shows the original selected SM from the selected Kurtosis measure. Highlighted are lone signal packets that have been removed and noise selected packets not representing at least four contiguous packet widths.
noise packets and the second looks for individual packets identified as signal. An example of the grooming process is given in Figure 15. Signal identified packets with no neighbors are reclassified as noise. Segments of noise less than three contiguous packets are reclassified as signal. This prevents identification of noise spikes and keeps densely populated shoulder regions from influencing the baseline model.

Tails of Lorentzian peaks are covered in thermally generated baseline noise (low signal to noise ratio) and, therefore, have a propensity to be included within the selected noise regions. When the tail is included as a noise region, the baseline model produces trapezoidal signal shapes instead of smooth Lorentzian line shapes as in Figure 16. To address this problem we increase the width of signal containing regions by 5 packet widths after convergence of the SM. Since packet widths generally run 1.5 to 2 times the FWHM, five packet widths beyond the border guarantee full inclusion of the Lorentzian peak. This preserves the desired Lorentzian line shape as illustrated in Figure 16.
Figur 16: A converged signal map (red) expanded by 5 packet widths in order to preserve desired Lorentzian line shape. The green line represents the expanded signal map while the red line is the original converged SM. This becomes computationally more efficient than modeling Lorentzian peaks and selecting the tail cutoff from the model.

6) Baseline Estimate

The iterative threshold selection provides a complete SM with noise and signal successfully defined. Much work has been done on the best way to generate a baseline estimate using polynomials, spline fitting and other more complicated mathematical models based on fidelity and smoothness parameter balancing (Cobas, Bernstein and Marin-Pastor). In reality, the struggle between fidelity and smoothness in a modeled baseline can be described by a
Pareto front with the only true distinction being between optimal and non-optimal solutions (Figure 17). In spectra with well differentiated baseline noise, it may be more beneficial to use an estimated baseline with greater fidelity in order to increase overall signal to noise ratios.

Fidelity of the baseline model \((M)\) to the original spectrum \((Y)\) is determined as in Eq. 7. The smoothness

![Figure 17: A Pareto front generated from repetitive smoothing of noise regions from a spectrum representative of Dataset 2. Fidelity and smoothness parameters are calculated as in equations 6 and 7 and plotted against each other to yield a Pareto front yielding several non-optimal solutions.](image-url)
of the original baseline model is calculated as in Eq. 8.

\[ S = \sum_{i} (M_i - Y_i)^2 \]  

The smoothness parameter (R) as defined in (Cobas, Bernstein and Marin-Pastor) is calculated as the sum of differences squared between each adjacent point of the spectrum (\(y_i\) and \(y_{i+1}\)). Fidelity (S) is measured as a sum of differences squared of the estimated baseline to the actual baseline. All fidelity and smoothness values were calculated using a Heaviside step function corresponding to the SM where only portions of identified baseline noise are considered. When plotted against one another, a very evident Pareto front is formed. Baseline estimate creation can then effectively be achieved by simple smoothing methodologies, such as a rectangular average filter (Golotvin and Williams) or spline fitting methods (polynomial spline interpolation) and will have the same fidelity/smoothness relationship. For ease of use, the Whittaker smoothing method outlined in (Cobas, Bernstein and Marin-Pastor) can also effectively be used with inclusion of a smoothness parameter.
In this method, a simple rectangular average filter is applied to segments of noise identified by the error from Gaussian plot. A greater smoothness is achieved by increasing the width of the filter (usually three to five packet widths) from its designated 2 packet widths as in (Golotvin and Williams) or repeated application of the rectangular smoothing filter. In (Golotvin and Williams), they select a packet size of 32 as a default effectively making the rectangular filter width 65 points (2M+1, with M being the selected packet width of 32 points). Greater fidelity of the baseline estimate to the noise in the spectra will be maintained by lessening the width of the rectangular filter. When subtracted from original spectra, a baseline estimate maintaining higher fidelity to the noise will have an overall higher signal to noise ratio for further MVA processing. However, a baseline estimate with lesser fidelity to the actual spectrum will ensure that any packets of signal misidentified as packets of noise will not be wiped out upon correction. In both the Whittaker smoother and rectangular filter methods, a line is interpolated across regions of selected signal based on the signal intensity average at the boundary packets surrounding the regions of selected signal.
V. Results

1) Baseline Correction/Baseline Recognition

The original dataset used to train this algorithm consisted of $^{13}$C profiles of lipid standards ranging from 16 to 22 carbons with varying degrees of unsaturation and mixtures thereof. These data provided a simple set (sparsely populated spectra) of test spectra for initial testing having, at most, 42 relatively discrete signals plus that of the solvent signal (CDCl$_3$). This provides a significant amount of baseline to identify, and little signal to separate. More complex spectra followed including $^1$H urine metabolite profiles, $^{31}$P metabolite profiles and homogenized tissue extracts of $^{13}$C, $^1$H and $^{31}$P nuclear spin profiles.

For metabolomics applications, much more complex spectra are generated from biofluids and homogenized tissue chemical extracts. Figures 18 and 19 show acceptable baseline correction for complex metabolite profiles, as in Dataset 1. Figure 20 shows SM selection and baseline correction for Dataset 2.
Figure 18: This expanded segment from the olefinic region of a $^{13}$C lipid spectrum shows selectivity of the population fit threshold algorithm. In grey is the original spectrum with baseline roll artifacts. The red line is the estimated baseline. The green line is the created SM based on iterative convergence. The blue spectrum has been corrected. Signal selection, though not completely accurate, is visually excellent meeting psycho-visual scrutiny.

Figure 19: The entire carbon lipid profile from the figure 18. Signal selection is excellent regardless of the inclusion of small segments of baseline representing sporadic, increased noise as noted in the areas defined as baseline noise in Dataset 1.

In Figure 19, the signal map (green) shows inclusion of several regions not necessarily belonging to expected
groupings of signals as described in Datasets 1 and 3. However, these inclusions are of such a small width that they do not affect the overall baseline correction procedure. These erroneous signal selections are most evident from -20 to 0 PPM and from 80 to 120 PPM.

Figure 20 is a $^1$H urine metabolite profile and represents one of the most complicated spectral types encountered in NMR-based metabolomics. The only potential issues with the baseline correction methodology lies in
selection of overlapping signal region at about 1 PPM, and incorrect identification of water signal ring artifacts. There are no selected regions of baseline noise surrounding the cropped water region and therefore the method does not flatten baseline around the residual water signal. Also, the selected cutoff may be too abrupt to accurately reflect the entire cumulative Lorentzian tails at about 1 PPM. The remaining ring of the water signal at ~5 PPM is not properly modeled but as this signal region is usually removed for further processing, it does not create a problem.

2) Metabolomic/Multivariate Analysis

In metabolomics analyses, differentiation and discrimination between metabolic profiles lies in statistical and pattern recognition techniques. PCA, discriminate analyses and other statistical pattern recognition algorithms are the cornerstones of metabolomic investigation and rely on accurate and reproducible models for reliable data interpretation. An un-flattened or poorly flattened baseline will affect (1) the normalization procedures, by contributing baseline amplitudes to the overall sum of the spectral intensities; (2) the PCA projection values, especially along PC2 (Halouska and
Powers); and (3) baseline differences between spectra may even be identified by MVA and discriminate analysis techniques as important differentiating markers.

An entire study group of twenty-two exponentially weighted, FT and phase corrected spectra, as in Dataset 3, were baseline corrected using (1) Varian software correction methodologies, and (2) using the baseline correction method described herein. Both data sets were subjected to PCA and visually displayed as a scores plot. Varian provided baseline correction methodologies simply involve the psycho-visual selection of spectral regions containing noise for use in fifth degree polynomial and spline correction. Figure 21 shows almost identical cluster patterns suggesting little difference in PCA applications of current expert performed baseline corrections using Varian 6.1C and the above described method.
Since signal portions of the $^1$H urine data overwhelm the PCA, lipid standards as in Dataset 3 were used as they possess little signal. Ten runs of the same 16:0 lipid standard at 50mM concentrations were used. The spectra were either sum normalized or baseline corrected and sum normalized. Binning was performed using dynamic adaptive binning procedures (Anderson, L. and V.) whereupon bin boundaries were hand refined. Bins refinements were made to include large expanses of noise, overlapping signal groups and individual signals. Bins were identical for each given spectrum.

Figure 21: Twenty-two $^1$H urine metabolic profiles, as in Dataset 2, were expertly hand corrected and corrected by the above algorithm. The first two principal components from each data set are plotted above. Each of the spectra overlaps almost perfectly suggesting that the algorithm is adequate for data intended to be used with PCA, other MVA and discriminate analysis methods.

Since signal portions of the $^1$H urine data overwhelm the PCA, lipid standards as in Dataset 3 were used as they possess little signal. Ten runs of the same 16:0 lipid standard at 50mM concentrations were used. The spectra were either sum normalized or baseline corrected and sum normalized. Binning was performed using dynamic adaptive binning procedures (Anderson, L. and V.) whereupon bin boundaries were hand refined. Bins refinements were made to include large expanses of noise, overlapping signal groups and individual signals. Bins were identical for each given spectrum.
Figure 22 shows the difference between properly baseline corrected spectra and poorly or uncorrected spectra. The within group scatter is greatly reduced. Large expanses of rolling baseline contributed to the overall differences observed in PCA as well as peak height variability induced by sum normalization of raw spectra.

![Figure 22](image)

**Figure 22**: PCA comparison of baseline corrected vs. non-baseline corrected spectra as in Dataset 3. The corrected (red x’s) and non-corrected (black circles) lipid standard profiles of 16:0 were sum normalized and dynamically binned before being subjected to PCA. The within class spread is greatly reduced due to baseline correction before sum normalization.

Figure 23 and 24 show a group of lipid standards that have or have not been baseline corrected. Sum normalization was applied before processing using PCA. The non-baseline corrected spectra show huge variation in PCA space stemming from baseline roll. Especially in sparsely populated
spectra, baseline roll contributes greatly to the overall intensity of the spectrum and outcomes of normalization routines will be affected if baseline roll is not corrected. It is evident from the outliers in Figure 23 (especially the outlier in group 22:3) that baseline correction before sum normalization greatly affects the end product of PCA. Three standards and 1:1 mixtures of those three standards were selected for analysis. Multiple profiles from 22:3, 16:0 and 18:1 and mixtures thereof were used for this example. The comparison, once again, is between baseline corrected and non-baseline corrected spectra. All spectra were sum normalized and binned with bin widths of 0.2PPM, a standard width for $^{13}$C spectra. The baseline roll in the outlier of the 22:3 lipid standard group was enough to induce a negative sum and create peak inversions in the spectrum upon sum normalization. Amplitude tails are easily observed for the rest of the spectral groups, especially mixtures of 16:0 and 18:1 as well as 16:0 and 22:3.

Once baseline has been corrected, not only do we observe an overall change in scale due to amplitude similarities, within group clustering becomes evident
Figure 23: Lipid standard profiles as in Dataset 3, sum normalized and binned in standard fashion before being subjected to PCA. Class separation suffers greatly from lack of baseline correction. Note the outlier from the 22:3 lipid standard group as well as the range of PC values.

Figure 24: The same cohorts as in Figure 23, only with baseline correction. Note the scale of PC values as well as the great improvement of within class clustering and between class scatter.
(Figures 22 and 24). The PCA plot even becomes suggestive as to composition of mixtures of lipid samples. There is virtually no group overlap and each cluster is uniquely defined in PCA space. The only exception is for mixed lipid standard 16:0 and 22:3.

3) Common Noise Identification in Data Groups

Since groups of data contain similar information patterns (similar metabolic profiles), it is expected that the patterns of metabolites do not change much between spectra within a group. A test was run using dense $^1$H urine spectra (Dataset 2) from a control group plotted in Figure 21. The previously aligned and phase corrected spectra were subjected to the above noise identification algorithm. The 22 control spectra had a common packet size of 24 points applied generating 1365 packets for sorting and analysis. Five-hundred and thirty-four packets were selected as baseline noise among all of the spectra. The most packets identified as baseline were 667. The commonly identified packets represent about a third of the total sweep width. This is expected as similar cohorts from similar studies should present overall similar metabolite profiles.
4) Comparison of Baseline Estimates: LOWESS vs. PFT

It was necessary to compare previous methods of baseline correction with the PFT method. For the comparison we decided to use the most recently published method, the LOWESS baseline estimation as described in (Xi and Rocke). A sample spectrum from Dataset 2 was selected and both methods of baseline estimation were performed.

LOWESS baseline estimation was performed using the parameters and variance estimation methods outlined in the original publication (Xi and Rocke). Inclusion of negative signal from biphasic or out of phase residual water signal during baseline estimate calculation creates a dip in the baseline estimate (Figure 25) massively distorting the spectrum upon correction. The size of this dip is a function of the negative amplitude and the smoothness penalty value used in baseline estimation. Reduction of the smoothness penalty reduces the width of the dip around the negative portion of the biphasic water peak but drives the baseline estimate closer to the original spectrum. Altering the derived negativity penalty simply shifts the position of the baseline estimate up or down within observed baseline regions.

PFT was performed using a packet width of 6 points selected using the optimization routine. Once again this
reflects an upslope of the doublet as shown in Figure 10. The baseline estimate is much more reasonable than the LOWESS method estimate.

Figure 25: A Spectrum from Dataset 2 (grey) showing baseline estimates from PFT (red) and LOWESS (green) baseline estimation methods. It is evident that inclusion of the biphasic water signal greatly distorts the LOWESS baseline estimate. LOWESS baseline estimates were performed using parameters suggested in (Xi and Rocke). The biphasic water signal does not reduce efficacy of baseline estimation using PFT methodologies. For PFT, a packet size of 6 points was selected for analysis.
VI. Conclusions

The baseline correction algorithm outlined above follows a typical three stage baseline identification algorithm involving SM creation, baseline estimation and subsequent subtraction of that baseline estimate. It has proven effective in signal identification and baseline correction for a range of samples across many nuclear spin profiles and signal densities. This method both accounts for the selected packet width used and requires no arbitrary parametric input. The selection of packet width is shown to reflect the use of \( \sigma \) as a threshold selection. Also, we have been able to successfully model the regions of baseline noise to a Gaussian system using a novel histogram fitting arrangement closely approximating the Anderson-Darling Coefficient. The three difference from Gaussian metrics provide information on noise observed in NMR metabolite and chemical standard profiles. LMS error and the Anderson-Darling coefficient show us that the total of the observed noise becomes Gaussian only when all of it is included in the analysis. Kurtosis, however, proves to
be the most straightforward method for threshold
determination and subsequent SM creation.

We have illustrated the nature of fidelity/smoothness
parameters as a Pareto front. It can be argued that the
best solution lies in a baseline model with greater
fidelity to the identified noise regions of the spectrum as
it increases the overall signal to noise ratio after
correction of the spectrum. An accurate signal map also
provides the opportunity to simply remove the selected
segments of noise from the corrected spectrum reducing data
spread in PCA, especially along PC2 (Halouska and Powers).

It was shown through PCA that expert manual noise
selection and the above described PFT selection and
baseline correction create identically separated profiles
in PCA space. This suggests that the population fit
selected threshold and subsequent baseline corrected
spectra are suitable for further use in MVA. It was also
shown how proper baseline correction facilitates pattern
recognition techniques such as PCA in accurate
classification of chemical profiles.

Comparison of the latest published baseline correction
method to PFT shows that inclusion of biphasic water
signals in analysis completely distorts the estimated
baseline. Removal of these negative signal regions provide
similar baseline estimates to PFT with the exception that LOWESS estimation method removes the overlapping tail regions associated with dense spectral regions.

In conclusion, the two initially stated assumptions - noise is white and Gaussian in nature, and that noise has a lower $\sigma$ than signal - are sufficient to generate accurate differentiation of signal from noise for SM generation in NMR spectroscopy. Application of very simple baseline modeling algorithm has shown to produce accurate correction of baseline roll, even in very complex spectra. Also, very similar results have been produced in PCA space when spectra with manually selected noise regions using 5th degree polynomial spline interpolations are compared with the above method baseline correction method.
VII. References


