ALGORITHMIC TECHNIQUES EMPLOYED IN THE QUANTIFICATION AND CHARACTERIZATION OF NUCLEAR MAGNETIC RESONANCE SPECTROSCOPIC DATA

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

By

PAUL EDWARD ANDERSON
M.S.C.S., Wright State University, 2006
B.S.C.E., Wright State University, 2004

2010
Wright State University

______________________________
Michael L. Raymer, Ph.D.
Dissertation Director

______________________________
Arthur A. Goshtasby, Ph.D.
Director, CSE Graduate Program

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

Committee on Final Examination

______________________________
Michael L. Raymer, Ph.D.

______________________________
Travis E. Doom, Ph.D.

______________________________
Mateen M. Rizki, Ph.D.

______________________________
Nicholas V. Reo, Ph.D.

______________________________
Nicholas J. DelRaso, Ph.D.
ABSTRACT


Nuclear magnetic resonance (NMR) based metabolomics is a developing research field with broad applicability, including the identification of biomarkers associated with pathophysiologic changes, sample classification based on the mechanism of toxicity, and clinical diagnosis. Intrinsic to these applications is the need for statistical and computational techniques to facilitate the associated data analysis. Further, a typical $^1$H NMR spectrum of pure proteins, biofluids, or tissue may contain thousands of resonances (i.e., peaks), thus, a pure visual inspection is insufficient to fully utilize the spectral information.

Common practice within the NMR-based metabolomics community is to evaluate and validate novel algorithms on empirical and simplified simulated data. Empirical data captures the complex characteristics of experimental data; however, evaluations on empirical data often rely on indirect performance metrics because the optimal or correct output is difficult to obtain a priori. To overcome the drawback of relying on indirect performance metrics, researchers often evaluate their algorithms on simplified simulated data. The conclusions derived from this type of data can be difficult to generalize to true experimental data. This dissertation combines the advantages of both empirical and simplified simulated data by generating exacting synthetic data sets that emulate the salient features of experimental data.

The analysis of NMR metabolic spectroscopic data can be divided into four steps: (1) standard post-instrumental processing of spectroscopic data; (2) quantification of spectral features; (3) normalization and scaling; and (4) multivariate statistical modeling of data. Quantification of
spectral features, step (2), is a key step in the development of classification algorithms and biomarker identification (i.e., pattern recognition). Algorithms for spectral quantification are designed to enhance the efficacy of pattern recognition and multivariate statistical techniques for metabolomics. This is accomplished by reducing the dimensionality of the spectra, while retaining salient information and mitigating peak misalignment.

This dissertation develops two novel spectral quantification techniques: Gaussian binning and dynamic adaptive binning. Gaussian binning utilizes a kernel-based binning algorithm to decrease the sensitivity to peak misalignment. Dynamic adaptive binning optimizes the bin boundaries through an objective function using a dynamic programming strategy. Both Gaussian binning and dynamic adaptive binning are compared to common spectral binning techniques by analyzing their ability to reduce the probability of peaks spanning bin boundaries and increase the interpretability of the results. Finally, a case study is presented to show the ability of dynamic adaptive binning and Gaussian binning to enhance the analysis of a $^1$H NMR-based experiment to monitor rat urinary metabolites following exposure to the toxin α-naphthylisothiocyanate.
Table of Contents

1 Introduction .................................................................................................................................................. 1

1.1 Problem description.................................................................................................................................. 1

2 Background .................................................................................................................................................. 3

2.1 Metabolomics vs. metabonomics ........................................................................................................... 3

2.2 Metabolomics .......................................................................................................................................... 3

2.2.1 Nuclear magnetic resonance ........................................................................................................... 4

2.3 Bionomics: Genomics, Proteomics, and Metabolomics ......................................................................... 7

2.4 Principal component analysis .................................................................................................................. 8

2.5 Wavelet transform .................................................................................................................................. 9

3 Literature review ....................................................................................................................................... 11

3.1 Metabolomics ........................................................................................................................................ 11

4 Dissertation Overview ............................................................................................................................... 21

5 Characterization of $^1$H NMR spectroscopic data and the generation of synthetic validation sets .. 25

5.1 Introduction ............................................................................................................................................ 25

5.2 Methods .................................................................................................................................................. 26

5.2.1 $^1$H Spectroscopic data .................................................................................................................... 26
5.2.2 Spectra characterization ................................................................. 28
5.2.3 Generating a synthetic spectral data set .............................................. 39
5.2.4 Case study: Comparing alignment algorithms .................................... 45
5.3 Results and discussion .......................................................................... 47
5.3.1 Parameters ....................................................................................... 47
5.3.2 Case study: Comparing alignment algorithms .................................... 49
5.4 Discussion ............................................................................................ 51
6 Gaussian binning .................................................................................... 53
6.1 Introduction .......................................................................................... 53
6.2 Materials and Methods ......................................................................... 54
6.2.1 Empirical data ................................................................................... 54
6.2.2 Binning ............................................................................................. 54
6.2.3 Evaluating binning methods ............................................................... 57
6.3 Results ................................................................................................ 62
6.3.1 Loss of information .......................................................................... 62
6.3.2 Sensitivity to peak shift ................................................................. 68
6.4 Discussion ............................................................................................ 70
6.5 Concluding remarks ............................................................................ 74
List of Figures

Figure 1. Sample urine $^1$H NMR spectrum 24 hours following exposure to the toxin $\alpha$-naphthylisothiocyanate (ANIT) [30]. ................................................................. 6

Figure 2. Example selection of the first principal component ........................................... 8

Figure 3. Graphical representation of the construction of a Gaussian-Lorentzian peak ........... 29

Figure 4. Sample region of a spectrum decomposed into 7 peaks modeled as Gaussian-Lorentzian functions ........................................................................................................ 31

Figure 5. Illustration of a region showing the residual (i.e., signal minus model) .................... 35

Figure 6. Illustration of a set of Gaussian-Lorentzian peaks divided into three groups: foreground, background, and baseline ................................................................. 36

Figure 7. Generation of a new peak by sampling the distributions for the height, width at half height, fraction Lorentzian, and distance between adjacent peaks ................................. 40

Figure 8. Process of generating piecewise baseline ($\Delta = 0.5$ ppm) ................................. 42

Figure 9. Two simplified spectra with their associated base spectrum ................................. 43

Figure 10. Sample region of one of the synthetic data sets used in the evaluation of the alignment algorithms ......................................................................................... 50

Figure 11. The process of uniform binning over a region of the $^1$H spectrum of ethylbenzene containing the quartet from the ethyl group at 2.56 ppm ........................................ 56

Figure 12. The process of Gaussian binning over a region of the $^1$H spectrum of ethylbenzene containing the quartet from the ethyl group at 2.56 ppm ........................................ 58

Figure 13. Region of the ANIT data set for comparing uniform and Gaussian binning using the cluster error metric .................................................................................. 62
Figure 14. Percent variance in the first two principal components, $PC_1$ and $PC_2$, for (a) Gaussian binning as a function of the step size (ppm) and the standard deviation as a fraction of the step size (fss); and (b) uniform binning as a function of the step size................................. 64

Figure 15. The cluster equality criterion, $J_2$, for (a) Gaussian binning as a function of the step size (ppm) and the standard deviation as a fraction of the step size (fss); and (b) uniform binning as a function of the step size................................................................. 65

Figure 16. The cluster criterion, $J_2$, plotted as a function of the Gaussian standard deviation ($\sigma$) for the ANIT $^1$H spectral data set using a bin step size of 0.04 ppm............................................. 66

Figure 17. The PCA scores plot for (a) Gaussian binning with a step size of 0.04 ppm and a standard deviation of $0.29 \times$ step size, and (b) uniform binning with a step size of 0.04 ppm. ..... 67

Figure 18. The signal variance, $SV$, for (a) Gaussian binning as a function of the step size (ppm) and the standard deviation as a fraction of the step size (fss); and (b) uniform binning as a function of the step size................................................................. 68

Figure 19. Sensitivity to peak shift................................................................................................................. 69

Figure 20. Cluster error for Gaussian ($\sigma = 0.4 \times$ step size) and uniform binning graphed as a function of the step size................................................................................................................. 70

Figure 21. Process of binning 6 observed peaks from two spectra............................................................... 85

Figure 22. The subproblem of binning the first 4 observed peaks (a) is used to compute the binning of the first 5 observed peaks (b) and the first 6 observed peaks (c). ................................. 86

Figure 23. Sample regions of $^1$H spectroscopic data demonstrating the advantages of dynamic adaptive binning (A) over uniform binning (B). ................................................................................................................. 98

Figure 24. Principal component scores (means and standard error) after dynamic adaptive binning (A), uniform binning (B), adaptive binning (C), and adaptive uniform binning (D) for several
non-lethal ANIT doses (control: black; 20 mg/kg ANIT: blue; 50 mg/kg ANIT: red; and 100 mg/kg ANIT: magenta)
List of Tables

Table 1. Summary of the different components used to create a synthetic spectrum and their associated parameter distributions................................................................. 48

Table 2. Summary of the relationships between the parameters for each of the components: (a) Signal peaks, (b) baseline peaks, (c) piecewise baseline, and (d) noise.................................................. 49

Table 3. Average and standard deviation of the relative increase ($RI$) of the ASSE for 30 synthetic data sets........................................................................................................ 50

Table 4. Typical values of standard deviation................................................................................................................................. 72

Table 5. Recurrence relationship that recursively defines the optimal binning of $n$ maxima ...... 83

Table 6. Top 10 peak identification wavelet parameter configurations, ranked according to the average (AVG) of percentage of peaks missed ($PM$) and percentage of extra peaks ($PE$)............ 91

Table 7. The performance of the binning algorithms and their parameters as measured by the mean CPU seconds/data set relative to the fastest algorithm................................................. 93

Table 8. The performance of the binning algorithms and their parameters as measured by (A) normalized number of observed peaks per bin, and (B) median distance from observed peak to nearest boundary.......................................................................................................................... 94
1 Introduction

1.1 Problem description

Nuclear magnetic resonance (NMR) based metabolomics is a developing research field with broad applicability. Existing applications include the identification of biomarkers associated with responses to toxin and pathophysiologic changes [1-3], sample classification based on the type of toxic exposure [4], large scale human studies [5], clinical diagnosis [6, 7], the study of genetic disorders [7], and genetic classification [8, 9]. Intrinsic to these applications is the need for statistical and computational techniques to facilitate the associated data analysis. With the development of novel alternative techniques, the ability to reliably compare these techniques is of critical importance to researchers.

The ability of a researcher to select the most appropriate NMR processing or analysis technique is limited by the number of exacting comparisons available in the literature. The types of comparisons that are commonly available are based on empirical and simplified simulated data. Empirical data captures the complex characteristics of experimental data; however, evaluations on empirical data often rely on indirect performance metrics because the optimal or correct output is difficult to obtain \textit{a priori}. To overcome the drawback of relying on indirect performance metrics, researchers often evaluate their algorithms on simplified simulated data. The conclusions obtained on this type of data can be difficult to generalize to true experimental data.

A typical \textsuperscript{1}H NMR spectrum of pure proteins, biofluids, or tissue may contain thousands of resonances (i.e., peaks), thus, a pure visual inspection is insufficient to fully utilize the spectral information [2]. Further, many metabolomic experiments have a small sample pool, where a
A typical data set consists of a large number of features and a small number of samples. This is primarily due to the expensive and time-consuming nature of conducting large sample size experiments. This type of data set presents problems to traditional pattern recognition techniques and is known as the “curse of dimensionality” [10, 11].

The quantification of spectral features is a key step in the development of classification algorithms and biomarker identification (i.e., pattern recognition). The goal of spectral quantification is to transform the spectroscopic data into feature vectors that enhance pattern recognition and multivariate statistical techniques. These techniques are designed to overcome the innate difficulties of NMR spectral data, including the misalignment of peaks resulting from experimental conditions (e.g., pH, ionic strength, and composition) [12, 13], noise, and congestion (i.e., overlapping peaks).
2 Background

2.1 Metabolomics vs. metabonomics

There have been two terms used to describe the quantitative analysis of metabolites: metabonomics and metabolomics. The definition of metabonomics is a quantitative measurement of the multiparametric metabolic response of biological systems to pathology or genetic modification [14]. It was originally defined by Jeremy Nicholson, Elaine Holmes, and John Lindon of Imperial College (London) and is based on the Greek roots “meta” – change and “nomos” – rules or laws [15]. The term metabolomics is defined as a comprehensive analysis in which metabolites of a biological system are identified and quantified [16]. The origin of this term is suspected to stem from the concept of metabolome, which is an attempt to identify and define the total small molecule component of the cell [17, 18]. Both terms have been described as a subset of each other [19, 20]. For this dissertation, the terms will be considered interchangeable, but metabolomics is chosen to avoid confusion.

2.2 Metabolomics

Any technique that can quantify metabolites can be used for metabolomics, but there are two primary techniques seen in the literature: nuclear magnetic resonance (NMR) [2] and mass spectrometry with a prior on-line separation step such as high performance liquid chromatography (HPLC) [21] or gas chromatography (GC) [22]. While neither technique is strictly superior, each technique has its own advantages and disadvantages [19, 20]. Some of the main benefits of mass spectrometry include sensitivity, selective identification of specific metabolites, and more mature data analysis algorithms and software. Further, some of the benefits of NMR include reproducibility within and across labs, versatility, sample preparation, non-
destructive, sample analysis automation, and the lack of sample bias. Specifically, NMR allows a wide range of metabolites to be measured with no sample preparation and without pre-selection [23]. For a more thorough treatment of the differences between mass spectrometry and NMR-based metabolite analysis see Robertson [19], and see Idborg et al. for more information on processing mass spectrometry data for metabolomics [24, 25]. The work in this dissertation will focus on NMR-based metabolomics applied to toxicology applications.

2.2.1 Nuclear magnetic resonance

Nuclear magnetic resonance (NMR) spectroscopy is a technique that exploits the properties of an atom’s nucleus. It can be used to obtain information about the concentration and structure of molecules. The phenomenon of NMR was discovered simultaneously by Bloch et al. [26] and Purcell et al. [27]. NMR studies magnetic nuclei by applying a static magnetic field followed by applying a second oscillating magnetic field [28]. Only nuclei with an odd number of protons or neutrons can be measured using NMR; however, the two most common atoms studied are $^1\text{H}$ and $^{13}\text{C}$.

Neutrons and protons have an intrinsic quantum mechanical property called spin, and the overall spin of a nucleus is known as the spin quantum number, $I$. A non-zero spin is associated with a non-zero magnetic moment by the following equation:

$$\mu = \gamma I,$$

where $\gamma$ is a constant known as the magnetogyric ratio. The values of $I$ are quantized and are related to the magnetic quantum number, $m$ by the following equation:

$$I_z = m \hbar,$$
where $\hbar$ is Dirac’s constant, and $I_z$ is the z component of the angular momentum associated with the spin. The magnetic quantum number, $m$, can take the integral values from $+I$ to $-I$, thus, for a given nucleus, there are a total of $2I + 1$ angular momentum states. If the magnetic field is oriented along the z axis, the energy of a state is given by the following equation:

$$E = -\mathbf{B}_0 \mu = -\mu_z B_0 = -\gamma I_z B_0,$$

where $\mu_z$ is the z component of the magnetic moment, which is equal to $\gamma I_z$. The energy between two states is given by:

$$\Delta E = \hbar \gamma B_0,$$

When electromagnetic radiation of the correct frequency is applied to match this energy difference, resonant absorption will occur, causing the distributions of atoms in the states to change. Because the energy of a photon is $E = h\nu$, the absorption will occur when the frequency, $\nu$, is equals:

$$\nu = \frac{\Delta E}{\hbar},$$

where $\hbar$ is Planck’s constant. These equations appear to state that the resonance frequency for all nuclei of the same nuclide is the same; however, this is not the case because the effective magnetic field felt by each nucleus is different. This is a result of shielding, which is caused by the surrounding electrons.

Fourier NMR spectroscopy consists of three steps [29]. The first step is the application of a static magnetic field, $B_0$. In the presence of this field, the atoms will distribute themselves between the different energy states according to the Boltzmann distribution. The second step is
the resonance step. An exciting radio frequency (RF) field is applied as a pulse. The energy excites all of the NMR transitions between energy states, perturbing the previously mentioned balanced states. This results in tilting the magnetization vector out of alignment with the static magnetic field. The third step is relaxation, where the system returns to the original state. During which, the magnetization vector precesses about the static magnetic field at the frequency of the spins. This induces a current in a nearby pickup coil, which results in an oscillating electrical signal known as the free induction decay (FID). The signal consists of a sum of the NMR responses from all of the excited spins. This time domain signal is Fourier transformed, resulting in a spectrum (sample shown in Figure 1).

![Sample urine $^1$H NMR spectrum 24 hours following exposure to the toxin $\alpha$-naphthylisothiocyanate (ANIT) [30].](image-url)
2.3 Bionomics: Genomics, Proteomics, and Metabolomics

Genomics is a semi-quantitative approach to the measurement of gene expression that identifies differentially expressed genes as a result of genetic modification, disease, or xenobiotic toxicity (i.e., a foreign compound) [2, 30]. This promising and advancing field of research is viable because of the development of a device known as a gene chip, which measures the mRNA levels of a cell [31]. A single gene chip simultaneously measures thousands of genes, but the cost of each individual gene chip keeps the number of samples and replicates small. This leads to data sets that contain a large number of features and small sample size, similar to NMR-based metabolomic experiments. Thus this type of data also suffers from the “curse of dimensionality” and presents a problem to traditional pattern recognition techniques [10, 11].

Another bionomics technology is proteomics, which measures the production of cellular proteins in response to drug exposure and other pathophysiological processes [32, 33]. Proteomics employs a variety of techniques but generally centers around a protein separation method (typically 2D gel-electrophoresis) and then analysis by mass spectrometry. While cheaper than gene chips, it is more labor intensive and difficult to tie to toxicological mechanisms due to the difficulty to obtain a detailed time-course of the proteomic response to disease or drug exposure [2].

Unlike proteomics and genomics that assess intermediate products, metabolomics assesses the end product of cellular function, metabolites. Changes occurring at the level of genes and proteins (assessed by genomics and proteomics) may or may not influence a variety of cellular functions. But metabolomics, by contrast, assesses the end products of cellular metabolic function, such that the measured metabolite profile reflects the cellular metabolic status. For instance, a disease or
foreign compound may interfere at the genomic or proteomic level, while it will always manifest itself at the metabolomic level.

### 2.4 Principal component analysis

Principal component analysis (PCA) is an unsupervised method commonly used to visualize NMR-derived metabolomic data [34, 35]. PCA transforms the data into a new coordinate system, where each new dimension is computed as a linear combination of the original values. These new dimensions are called principal components, and are orthogonal vectors selected to best explain the overall variance in the data. This is illustrated in Figure 2, where the data set shown has two dimensions. Each sample, shown as a black dot in the figure, is projected onto the first principal component, where the principal component points along the direction of greatest variance.

---

**Figure 2.** Example selection of the first principal component.
2.5 Wavelet transform

The potential of wavelets to remove noise from NMR spectra has been explored by Cancion-De-Greiff et al. [36], and wavelets have been combined with pattern recognition techniques such as artificial neural networks [37]. Further, wavelets have been used for de-noising and variable selection of chemometric analysis [38-41].

In signal processing, the Fourier transformation is commonly used to analyze the frequency components of a signal; however, it does not provide any time (location) information (i.e., at what time does a certain frequency occur). Currently, one of the most popular solutions designed to overcome this problem is the wavelet transform. The wavelet transform has been employed by numerous disciplines with a wide variety of applications, including image processing, signal noise analysis, data compression, and speech recognition[42]. While Fourier analysis represents a signal as a sum of a series of sinusoids, wavelet analysis represents a signal as a sum of a series of scaled and translated wavelets, which are functions of finite length that are fast decaying and have an average value of zero. This allows wavelet analysis to successfully reconstruct finite, non-periodic and/or non-stationary signals, while Fourier analysis has trouble representing signals with discontinuities and sharp peaks. A stationary signal is one that infinitely repeats with the same periodicity.

There are two types of wavelet transforms: continuous and discrete. The continuous wavelet transform (CWT) operates over every combination of translation and scale, while the discrete wavelet transform (DWT) operates on a subset of translations and scales. If the energy of the signal is finite, then the DWT is sufficient to completely encode the signal when the scales and positions are chosen as powers of two. When applying a wavelet transformation, the signal is broken up into shifted and scaled wavelets, where scaling is defined as stretching or compressing
the wavelet and is analogous to changing the frequency. The signal is decomposed into a number of levels, which can be broken up into coefficients that represent the approximate signal and coefficients that represent the details of the signal. The approximations are high-scale low-frequency components. The high-scale coefficients are processed further until the optimal number of levels is reached. The optimal number of levels can be chosen from a suitable criterion, such as entropy. For a more in-depth discussion of wavelets see the following books [43-45].
3 Literature review

3.1 Metabolomics

The first use of multivariate statistical techniques to high resolution NMR spectra was the classification of spectra from rat urine according to type of organ toxin which had been administered [46, 47]. NMR-based metabolomics has also been applied to reveal patterns of metabolic markers associated with various toxins and pathophysiologic changes [1-3, 6]. It has also been applied to large-scale human studies [5]. Further, chemometric methods have been developed to classify metabolomic data based on the type of toxin exposed [4, 48]. NMR-based metabolomics has also been used in clinical diagnosis [6, 49], to study genetic disorders [7], and genetic classification [8, 9]. All of the above applications rely on algorithmic spectral processing techniques to be successful.

There have been multiple initiatives to standardize metabolic experiments and the associated data processing [50-53]. Typically, NMR metabolic spectroscopic data is analyzed as follows: (1) standard post-instrumental processing of spectroscopic data, such as Fourier transform, phase adjustment, and baseline correction [54]; (2) quantification of spectral features; (3) normalization; (4) scaling; and (5) multivariate statistical modeling of data [52].

Quantification of spectral features, step (2), is a key step in the development of classification algorithms and biomarker identification. A common method of quantification employed by the NMR community is known as uniform or fixed width binning, which divides an NMR spectrum into several hundred non-overlapping regions or bins of equal size [4, 6, 46-48, 55-61]. This technique is performed to (1) minimize effects from variations in peak positions caused by sample pH, ionic strength, and composition [12, 13]; and (2) reduce the data size for multivariate
statistical analyses. To help mask these variations in peak positions the NMR peaks can be artificially broadened before Fourier analysis is applied. This technique is known as apodisation [62, 63]. The result of binning is a data set with fewer features, thereby, increasing the tractability of pattern recognition techniques.

One example of uniform bins employed in a recent NMR experiment is Ebbels et al. who utilize a probabilistic classifier based on the probabilistic neural network to predict the most likely toxicology effect for new compounds [48]. Their method is named Classification Of Unknowns by Density Superposition (CLOUDS). It is implemented as a non-neural network implementation of a probabilistic neural network [64]. The authors use a data set of ca.100,000 spectra of urine and blood serum from animals treated with ca 150 model toxins assembled by the consortium for metabonomic toxicology (COMET) [4]. This paper confirms the good classification rates obtained by a previous probabilistic neural network, and shows that probabilistic techniques have important advantages over traditional classifiers, such as the addition of non-absolute classification (probabilities). These probabilities allow a much more in-depth interpretation of the results. This technique is also able to detect outliers as those samples with consistently low class probabilities.

Webb-Robertson et al. investigate the classification accuracy of two sample NMR spectral data sets as a function of the uniform bin width and normalization approaches [65]. The authors find that the accuracy of a PCA k-nearest-neighbors classifier is relatively independent of the width of the uniform bin. They also note that there is a slight increase in the percent variance explained by the first principal component as the bin width is increased. The analysis did not reveal an optimal bin width for classification of their data set.
Spraul et al. compare two spectral quantification techniques: uniform binning and peak height measurement [13]. Uniform binning or fixed width binning is a common method of quantification employed by the NMR community, which divides a NMR spectrum into several hundred non-overlapping regions or bins of equal size [49, 55-59]. The authors also study the advantages of allowing overlap between the bins, and find that it does not affect the classification and PCA results. They observe that using the peak heights instead of binning could miss peaks that are on the shoulder of larger peaks (i.e. peaks that are partially or fully hidden). It is noted that uniform binning has been shown to be robust and fast in many applications; however, binning is susceptible to baseline distortions, so baseline correction is important. Further, manual data reduction is time consuming and subject to bias, thus, it may only contain a subset of the important information.

Trbovic et al. compare the clustering of groups after PCA for uniform binning with and without noise removal via a wavelet transform [66]. They also compare these results to the clustering results after applying PCA to the coefficients from the wavelet transform with noise removal and multiresolution analysis (MRA) [67]. With MRA only a subset of the resolution levels are used to restore the signal, which can be used to suppress low or high frequency components. The wavelet de-noising algorithm uses soft-thresholding of the coefficients, where the coefficients are shrunk toward zero by subtracting a constant threshold $\lambda$. The threshold is determined using a technique known as ‘universal thresholding’ [68] with a soft-thresholding value of $\lambda = \sqrt{2 \log N}$, where $N$ is the total number of data points. This is opposed to hard-thresholding that removes those wavelet coefficients with a value lower than a threshold $\lambda$. The paper shows that incorporating wavelet de-noising before uniform binning improves the clustering after applying PCA. Further improvement in the clustering was observed when applying PCA to the wavelet coefficients with noise removal and MRA.
The potential of wavelets to remove noise from NMR spectra has also been explored by Cancion-De-Greiff et al. [36], and wavelets have been combined with pattern recognition techniques such as artificial neural networks [37]. Further, wavelets have been used for de-noising and variable selection of chemometric analysis [38-41], and as a result tutorials are available for removing noise with wavelets for chemometric analysis [69]. For an implementation of de-noising NMR spectra, see the MATLAB toolbox implemented by Gunther et al. [70].

Davis et. al. introduce an adaptive binning method using the undecimated wavelet transform [71]. The algorithm utilizes the undecimated wavelet transform to smooth and remove noise from a composite spectrum, so that peak detection is more accurate and robust. The bins are dynamically chosen based on a reference spectrum that is calculated by taking the maximum value for each chemical shift. A maximum composite spectrum is constructed to avoid missing peaks that are only visible in one spectrum. The composite spectrum is then smoothed with the undecimated wavelet transform, and the minimums are used to define the bins.

The authors employ hard-thresholding to remove noise regions with a wavelet transform as opposed to soft-thresholding [66, 72]. The threshold is determined using a technique known as ‘universal thresholding’ [68] with a value of $\lambda = \sigma \sqrt{2 \log N}$, where $\sigma$ is the standard deviation of highest decomposition level and $N$ is the total number of data points. This technique shows a significant increase in the ratio of inter-class to intra-class variation, which leads to improved classification [71]. The authors employ synthesized data to illustrate the problem of peaks spanning the boundaries of uniform bins. The very simple basis functions of Haar [73] and Daubechies’ family [74] achieve the best results. The authors report similar classification results with increased interpretability, along with the avoidance of boundary problems.
The authors of adaptive binning use the undecimated wavelet because it is shift-invariant also known as time-invariant, or maximal overlap [75]. A fast wavelet transform is shift variant, meaning that if the starting point is changed, the transform will change (i.e., it is not just shifted). Finally, the undecimated wavelet transform is also used because it retains more information, thus, it is more suitable for noise removal.

Another dynamic binning algorithm is adaptive, intelligent binning, which recursively identifies bin edges in existing bins [76]. The algorithm begins by binning the entire spectra into a single bin. Every data point in the bin is then examined as a potential boundary (split) for two new bins. A decision rule based on the relative height of the maximum intensity in a bin is employed to determine whether a bin should be split. The authors demonstrate improved performance of adaptive, intelligent binning over uniform binning in a hypertension study performed on a subset of subjects in a longitudinal population study. The algorithms benefits include the avoidance of a composite spectrum and the avoidance of arbitrary parameters.

In addition to the more computationally intensive techniques described above, variable width bins are available in the commercial package from ACD/Laboratories (www.acdlabs.com) [77]. The algorithm allows for smaller or larger bins within a predefined range. The technique, known as intelligent bucketing, defines the bin edges based on local minima within the predefined range.

There are several alternatives to spectral binning that still provide data dimension reduction. Examples of these include PARS [78, 79], curve-fitting method for direct quantification [80], peak alignment tools in HiRes [81], and targeted profiling [82]. These techniques identify peaks or specific peak patterns in the spectra that are conserved across spectra. After the patterns have been identified, they are quantified by determining the peak area or amplitude. The accuracy of these algorithms is dependent on the quality of the peak alignment and the pattern databases.
While binning has been shown to be useful in the development of classification models and to guide the biomarker identification process, a full resolution spectrum should be examined before the results are interpreted from a biological perspective. Recently, there have been several promising full resolution techniques developed [83, 84] and applied [85, 86]. In most cases, these techniques require the spectra to be preprocessed by an alignment algorithm, providing a “cleaner” data set [87-92]. Further, several examples of artifacts from unaligned NMR signals have been reported, thus, proper alignment is a critical problem when applying full resolution techniques [88, 90, 93-95], and is necessary to correct for differences that reflect the individual variation of the individual’s metabolism [14]. These can complicate multivariate statistical techniques, such as PCA and nonlinear techniques, such as artificial neural networks [96]. In addition, PCA has also been commonly used as a method for evaluating alignment algorithms [88, 94, 97-102].

Cloarec et al. develop a full resolution technique that leverages the mutlicolinearity of the intensity variables in a set of spectra that displays the correlation among intensities of various peaks across the entire sample, termed statistical total correlation spectroscopy (STOCSY) [83]. This technique is combined with a supervised pattern recognition technique called orthogonal projection on latent structure (O-PLS) [103, 104]. They demonstrate the technique’s ability to find resonances from the same metabolite and to highlight molecules involved in related pathways.

Another full resolution technique is developed by Stoyanova et al. where a Bayesian spectral decomposition is used to classify and cluster metabolomic NMR data [84, 105]. This technique allows the decomposition of spectral data sets into underlying spectral shapes. These features were used to define two patterns, aberrant and normal, that exhibit different temporal behavior in
treated and control animals. The aberrant pattern was characterized by relatively short-lived changes and rapid recover, while the normal pattern’s recovery is significantly delayed, thus, demonstrating the power of their technique to utilize the full resolution of NMR spectra to perform clustering. The results of this analysis were significantly improved after spectral alignment [88].

Forshed et al. developed an alignment algorithm that utilizes a genetic algorithm (GA) for automatic phasing and shift alignment [89, 90, 92]. The algorithm divides the spectra into several segments, and each segment is shifted sideways, and stretched or shrunk. The fit is evaluated by a correlation to a reference spectrum. The GA finds the optimal shrinking and stretching parameters. The authors summarize several alignment algorithms that have had success in other domains, including dynamic programming based techniques [106, 107], that have been applied to GC and HPLC chromatogram data [108, 109]; however, the authors assert that these techniques are too resource intensive for NMR spectroscopic data sets. The authors also discuss other alignment algorithms, including Kassidas et al. that aligns profiles from industrial batch processes of polymerization based on a time-warping algorithm first put forth by Wan and Isenhour [110, 111]. The signals are warped nonlinearly to fit a reference signals on a point by point basis until the peaks are aligned. This alignment is evaluated by the cumulative distance between peaks, which means that it is sensitive to peak height differences [90].

Other alignment algorithms have also been developed for NMR metabolomics data sets, including partial linear fit described by Vogels et al. [87, 91], which automatically picks windows of size D depending on the peak frequency and shifts them left or right. Every possible and relevant combination of window size and shift is tried until the sum of squared differences in the spectrum compared to the reference is minimized.
Stoyanova et al. introduce a technique that automatically aligns peaks in an NMR spectroscopic data set by detecting the peak-regions where shifts occur and then aligning the peaks within those regions [88]. The regions are detected by sliding first derivatives of a variety of peak shapes along the second principal component and calculating the resulting correlation. High correlation values (>0.8) are indicators of peak shifts. This algorithm assumes that peaks are well separated and that baseline variation is minimal. For example, shifts for strongly overlapping peaks do not result in derivative shapes in the second principal components. This alignment algorithm does not align to a reference spectrum, and thus, their algorithm is not dependent on the degree of correctness of this spectrum.

After spectral quantification, which may include an alignment or binning algorithm, normalization is performed on a per-spectrum basis to make the samples directly comparable to each other [112]. This is designed to remove artificial differences, such as variable dilution of the samples, which is a common problem in NMR spectroscopic studies of urine where many toxins can cause large increases or decreases in urinary volume. However, normalization is less important for experiments where samples are highly regulated, such as plasma, which is highly regulated by homeostasis (i.e. maintenance of physiological conditions required to maintain life). Further, when a tissue of known weight can be analyzed normalization is less important.

The high variability of urine concentration arises because urine excretion serves to eliminate waste and regulate blood volume. Experimental procedures, such as freeze drying and reconstituting the sample, can help normalize the variability between samples. Another technique to normalize a sample is to normalize to a “housekeeping” metabolite [113]. In most cases, Creatintine is considered a constant, and is therefore, ideal for normalization. However, some studies have shown it to be related to muscle mass, particularly in children [113] and the elderly.
If the concentration of a specific metabolite (e.g. glucose) can be determined independently, then this provides a reference value. The most common normalization technique is to represent each data point as a fraction of the total spectral intensity, which is referred to as normalization to a constant sum. This technique performs great when a group of spectra have similar internal peak ratios but differing total intensities; however, it will break down when different samples have large differences in peak ratios. For example, when applied to a spectrum with a large peak relative to other spectra, the smaller peaks will appear to have decreased due to the height of the large peak. This can present a major difficulty when interpreting pattern recognition results.

Scaling is typically applied after normalization and is designed to control the weighting of features before a multivariate statistical or pattern recognition technique is applied. The idea is that small changes in feature with a large average value can suppress relatively large changes in features with a small average value. Scaling techniques are applied to the entire data set on an individual feature basis. A number of scaling techniques are commonly used, including mean-centering, auto-scaling, pareto scaling, and logarithmic scaling. With mean-centering, the mean of each feature is subtracted from each feature value. Auto-scaling is mean-centering with an additional step where each feature is divided by its standard deviation. Pareto scaling is the same as auto-scaling, but instead of dividing by the standard deviation, each feature value is divided by the square root of the standard deviation. Logarithmic scaling is taking the logarithm of each feature. An additional goal of some scaling techniques, such as mean-centering, is to help produce a parsimonious model. Both normalization and scaling are highly context dependent, and therefore, no single approach is optimal for all types of experiments.
The specific challenges of analyzing bionomics data, such as high dimensionality with small sample sizes, are exacerbated by poor feature quantification. To enhance the effectiveness of NMR-based metabolomics, several approaches for improving feature quantification are actively researched by the community. Some of these techniques are based on combining peak alignment and data reduction methods. The results of which are dependent on accuracy of the alignment and the completeness of the pattern databases. Other techniques attempt to mitigate the misalignment of peaks and avoid incomplete pattern databases by binning the spectra. The applicability of these two methodologies depends on the objectives of the specific metabolomic experiment. This dissertation will focus on algorithms to improve spectral quantification via binning. These techniques are designed to improve the accuracy and power of multivariate statistical and pattern recognition techniques for biomarker identification.
4 Dissertation Overview

Analyzing NMR spectroscopic data requires a variety of algorithms that range from signal processing to pattern recognition techniques. An integral part of the data analysis workflow is the quantification of spectral features. This dissertation develops a methodology for the characterization of NMR spectroscopic data sets and the creation of synthetic validation sets. In addition, this dissertation develops novel techniques for improved spectral quantification.

The infrequent use of exacting synthetic data is often due to the difficult nature of capturing the pertinent characteristics of the experimental data. Identifying the relevant characteristics is one of the first steps of generating synthetic data. For example, the absolute scale of the peaks in a spectral synthetic data set is unimportant for validating the accuracy of an alignment algorithm, but the relative magnitude of the peaks with respect to each other and the baseline noise is important for validating the accuracy of an algorithm that separate signal from noise.

This dissertation develops a technique for creating synthetic validation sets based on NMR spectroscopic data of rat urine samples from a metabolomics experiment. The validation sets were developed by characterizing the salient distributions of empirical spectroscopic data set by modeling each spectrum as a combination of Gaussian-Lorentzian peaks and a piecewise cubic interpolated baseline. The model for each spectrum is accepted when the residual can be decomposed into regions that follow a normal distribution ($\mu = 0$). Using this technique, several validation sets are constructed with a variety of characteristics and are available for download [119]. A case study is then presented to show how the validation sets can be used to compare the relative accuracy of several alignment algorithms [120, 121].
Modeling each spectrum using the previously described methodology is a time consuming process that requires manual adjustments; therefore, alternative high-throughput spectral quantification techniques are used in practice. A common application of spectral quantification is the identification of biomarkers for a specific toxin. Here a biomarker is defined as a set of NMR signals that change after exposure. Such an experiment will have at least two groups (e.g., pre-dose and post-dose) for which spectroscopic data is compiled. Spectral quantification transforms every sample, represented as an NMR spectrum, into a feature vector. Biomarker identification can then be defined as finding a set of features that explain a differentiation among groups, thus, the success of biomarker identification is directly related to the quality of the feature vectors.

The techniques employed for spectral quantification are designed to overcome the innate difficulties of NMR spectral data, including misalignment resulting from experimental conditions (e.g., pH), noise, and congestion (i.e., overlapping peaks). Traditionally, these difficulties have been mitigated by dividing each spectrum into uniformly spaced equal sized integral segments or bins. This technique is known as uniform or fixed width binning, and while it mitigates the effects from variations in peak positions, shifts occurring near the boundaries can result in dramatic quantitative changes in the adjacent bins due to the non-overlapping boundaries.

To overcome the drawbacks of uniform binning, several dynamic binning algorithms have been developed: adaptive binning [71]; adaptive, intelligent binning [76]; and intelligent bucketing [77]. Further, there are several alternatives to spectral binning that still provide data dimension reduction. Examples of these include PARS [78, 79], curve-fitting method for direct quantification [80], peak alignment tools in HiRes [81], and targeted profiling [82]. These techniques identify peaks or specific peak patterns in the spectra that are conserved across spectra. After the patterns have been identified, they are quantified by determining the peak area.
or amplitude. The accuracy of these algorithms is dependent on the quality of the peak alignment and the breadth of pattern databases. In addition, these techniques often require time consuming manual interaction that reduces their usefulness as high-throughput spectral quantification techniques.

The decision to employ spectral binning or an alternative technique is dependent on the research objective and experimental design. For example, targeted profiling, which attempts to quantify specific metabolites, might be the best choice if the metabolites of interest are known \textit{a priori} and available in a pattern database. For this case, it is also important for the signals representing these metabolites to be distinguishable among the congestion and misalignment typical of NMR spectra. Spectral binning does not require the metabolites of interest to be known \textit{a priori}. It can help direct the researcher to areas of the spectra that are identified as potential biomarkers. The metabolites contributing to these regions can then be identified. Both of these methodologies have been successfully employed in metabolomics experiments [47, 49, 55, 57-59, 76, 78-80, 122-124].

The second contribution of this dissertation is the development of a new Gaussian binning method for processing NMR spectroscopic data for multivariate analysis. This technique is compared to the traditional uniform binning technique by analyzing sensitivity to peak shift and loss of information content. Data for this comparison are derived from a $^1$H NMR-based experiment to monitor rat urinary metabolites following exposure to the toxin α-naphthylisothiocyanate (ANIT) [125]. The Gaussian binning method showed a decrease in the sensitivity to peak shift and performed comparable to uniform binning in retaining spectral information. Finally, some guidelines for using this methodology in NMR-based metabolomics data sets are presented.
In addition to developing a kernel-based binning algorithm, this dissertation also develops a novel dynamic binning method: dynamic adaptive binning. With this technique bin boundaries are dynamically determined through the optimization of a heuristic based objective function. This technique is compared to the traditional uniform binning technique, adaptive binning, and adaptive intelligent binning by analyzing their ability to find responding peaks in a synthetic spectral data set. The methods are further compared on their likelihood to place the same peak in different bins. This detailed comparison is capable due to the aforementioned synthetic data set developed in this dissertation. Finally, a case study is presented to show the ability of heuristic adaptive binning to enhance the analysis of a $^1$H NMR-based experiment to monitor rat urinary metabolites following exposure to the toxin α-naphthylisothiocyanate (ANIT) [125].

This dissertation will include three chapters dedicated to describing the three major research contributions. The first of these chapters will describe the techniques used to characterize and create synthetic data for comparing spectral quantification algorithms, which has been published in the journal *Bioinformatics*. In addition, this chapter will describe the empirical data sets. The second chapter will cover Gaussian binning, which has been published in the journal *Metabolomics*. The third chapter will cover the development of a novel dynamic binning algorithm (dynamic adaptive binning). This chapter will also contain a detailed comparison of dynamic adaptive binning with uniform binning, adaptive binning, and adaptive, intelligent binning using the synthetic validation sets. A journal article detailing the contributions of this final chapter has been submitted to the journal *Metabolomics*.
5 Characterization of $^1$H NMR spectroscopic data and the generation of synthetic validation sets

5.1 Introduction

Inherent to the data-driven applications of metabolomics (e.g., biomarker identification, sample classification, and clinical diagnosis) is the need for statistical and computational techniques to facilitate the associated data analysis. As such, metabolomics is particularly subject to the proliferation of data preparation and analysis methods. The selection of the most appropriate data analysis techniques is a common problem for researchers working in the “omics” fields (e.g., metabolomics, proteomics, and genomics). The interpretation of results requires in-depth knowledge of both the biological aspects and the analytical methods. As with other modern assays, there are a wide variety of potential data-transformation methods at each of the many data analysis steps. In current practice, selection methods are based upon the type of experiment, the specific hypothesis, expediency, and investigators’ background, experience, and preference. The multivariate nature of these data can yield varied results dependent upon the choice of analytical method, and are highly subject to misinterpretation.

In order for comparisons of technique performance on simulated data to be of value, the data must emulate the salient features of experimental data. Identifying the pertinent characteristics is the most critical step in generating realistic synthetic data. Ideally, synthetic data should be indistinguishable from empirical data, yet retain a “known” best analysis.

Herein, we propose a technique for creating realistic synthetic metabolomics validation sets based on NMR spectroscopic data. The validation sets are developed by characterizing the salient distributions in sets of empirical spectroscopic data. Each spectrum is modeled as a combination
of Gaussian-Lorentzian peaks and a piecewise cubic interpolated baseline. Using this technique, several validation sets are constructed with a variety of characteristics present in “real” data [119]. A case study is presented to compare the relative accuracy of several alignment algorithms using the increased precision afforded by these synthetic data sets [120, 121].

5.2 Methods

5.2.1 1H Spectroscopic data

The following animal work was conducted by the Air Force Research Laboratory, Biosciences and Protection Division, Wright-Patterson AFB, and the subsequent NMR acquisition was conducted by Nicholas V. Reo’s laboratory at Wright State University, Cox Institute.

5.2.1.1 Animals

All protocols for handling laboratory animals were approved by the Wright-Patterson Institutional Animal Care and Use Committee (IACUC) and meet appropriate Federal guidelines. Fisher 344 rats (approx. 250 g) were obtained from Charles Rivers Laboratory equipped with jugular vein catheters, and allowed to acclimate for seven days prior to the start of the study. Animals were then housed individually in metabolism cages and given ad libitum access to food (Purina Certified Rat Chow # 5002) and water. The housing environment was maintained on a 12 hour light-dark cycle at 25 °C, and all animals were examined by Vivarium personnel twice daily to ensure their health and well-being. Urine was collected into containers chilled on dry ice and containing 1 ml of 1% sodium azide. All urine samples were stored at -40 °C prior to analysis by NMR spectroscopy.
5.2.1.2 Urine sample preparation

Urine samples for NMR analysis were prepared as described by Robertson et al. [122] and modified as follows [125]. Samples were thawed at 4 °C overnight then allowed to equilibrate to room temperature just prior to NMR sample preparation. A 600 µL aliquot of urine was mixed with 300 µL of a phosphate buffer (0.2 M monosodium phosphate and 0.2 M disodium phosphate, pH 7.4) and allowed to equilibrate for ten minutes. Samples were then centrifuged at 5000 rpm (2300 rcf) for ten minutes to remove any particulate matter and 550 µL of supernatant was transferred to a 5mm NMR tube. An internal standard consisting of 150 µL of trimethylsilylproponionic (2, 2, 3, 3 d₄) acid (TSP) dissolved in deuterium oxide was added at a final concentration of 2 mM.

5.2.1.3 ¹H NMR spectroscopy

Proton NMR spectra were acquired at 25 °C on a Varian INOVA operating at 600 MHz. Water suppression was achieved using the first increment of a NOESY pulse sequence, which incorporated saturating irradiation (on-resonance for water) during the relaxation delay (7.0 s total; 2 s with water presaturation) and the mixing time (50 ms total; 42 ms with water irradiation). Data were signal averaged over 64 transients using a 4.0 s acquisition time and interpulse delay of 11.05 s.

5.2.1.4 Data set

NMR spectral data were processed using Varian software and employing exponential multiplication (0.3 Hz line-broadening), Fourier transformation, and baseline flattening (fifth-order polynomial and spline fitting routines). The TSP signal was used as an internal chemical shift reference, and the regions surrounding the residual water signal (≈4.8 ppm) and the urea signal (≈5.8 ppm) were excluded from the analyses. The vertical shift of the entire spectrum was
adjusted such that the mean of the intensities between 11.6 and 10 ppm was zero. Then the peak intensities of each spectrum were normalized to a constant sum. The final data set consists of 22 $^1$H spectra from individual normal healthy rats and is analogous to a set of control samples for a typical urinary metabolomics study using a rate animal model.

### 5.2.2 Spectra characterization

#### 5.2.2.1 Modeling the spectra

Each spectrum is characterized by decomposing it into its constituent components: peaks, noise, and baseline. The peaks are modeled by Gaussian-Lorentzian functions that are defined by the standard deviation of the Gaussian ($\sigma$), the center ($x_c$), the width at half height of the Lorentzian ($\Gamma$), and the magnitude ($M$):

$$S([M, \sigma, P, x_c], x) = P \times L([M, \Gamma, x_c], x) + (1 - P) \times G([M, \sigma, x_c], x),$$  \hspace{1cm} (6)

$$L([M, \Gamma, x_c], x) = \frac{M \times \Gamma^2}{4(x - x_c)^2 + \Gamma^2},$$  \hspace{1cm} (7)

$$G([M, \Gamma, x_c], x) = M \exp\left\{ -\left(\frac{(x - x_c)^2}{2 \times \sigma^2}\right) \right\},$$  \hspace{1cm} (8)

$$\Gamma = 2\sqrt{2\ln2\sigma},$$  \hspace{1cm} (9)

where $P$ is a real value between 0.0 and 1.0 that weights the contribution of the Lorentzian ($L(\ldots)$) and Gaussian ($G(\ldots)$) functions. The mixture of the Gaussian and Lorentzian peaks is selected to provide a flexible peak shape. The relationship between the width at half height of the Lorentzian peak and the standard deviation of the Gaussian peak is fixed by assuming that both the height and the width at half height are the same for both peaks. This simplifies the model by avoiding a separate parameter for both the standard deviation and width at half height. A graphical representation of the Gaussian-Lorentzian peak is shown in Figure 3.
The first step in decomposing a spectrum is to divide it into regions separated by signal that has been removed (e.g., the water signal). For the spectra considered in this paper, the signals that have been removed are the water, urea, and TSP signals. This results in two independent regions divided by the water and urea signals. These regions are then divided into non-overlapping segments ranging from 0.05 to 0.15 ppm in size. The width of a segment is varied to avoid placing a boundary in the middle of a peak. In congested areas of the spectrum, each segment encompasses several peaks. The segments remain small to allow the initial fitting routines to be performed interactively. The algorithm for determining the segments is as follows:
1. Divide the spectrum into uniform segments of size 0.01 ppm
2. \( i \leftarrow 2 \)
3. \( \text{left} \leftarrow \text{x-location (ppm) corresponding to the minimum intensity in segment } i-1 \)
4. \( \text{right} \leftarrow \text{x-location (ppm) corresponding to the minimum intensity in segment } i+1 \)
5. Store the adjusted segment \([\text{left}, \text{right}]\)
6. \( i \leftarrow i + 2 \)
7. Repeat steps 3 – 7 if \( i \) < number of segments

Following the creation of the segments, the initial locations of the peaks are interactively selected. The final locations of the peaks and their parameters (e.g., width, height) are determined algorithmically by solving the corresponding nonlinear curve-fitting problem. The parameters of the nonlinear curve-fitting problem are estimated by a subspace trust-region method based on the interior-reflective Newton method [126, 127]. The parameters are adjusted to minimize the function:

\[
\frac{1}{2} \sum_{i} (F(\beta, x_i) - y_i)^2, \tag{10}
\]

where \( x_i \) and \( y_i \) are the chemical shift and intensity of the \( i \)-th point in the segment, \( m \) is the number of data points in the segment, \( \beta \) is a vector of the parameters, and \( F \) is the model that will be fit by the algorithm, which is composed of Gaussian-Lorentzian peaks and a baseline offset:

\[
F(\beta, x_i) = \sum_{j=1}^{N} S([M_j, \sigma_j, P_j, x_{c_j}], x_i) + O, \tag{11}
\]

where \([M_j, \sigma_j, P_j, x_{c_j}]\) and the baseline offset \( O \) refer to parameters in the vector \( \beta \). The parameters \( M_j, \sigma_j, P_j, \) and \( x_{c_j} \) refer to the height, standard deviation, fraction of Lorentzian, and center of the \( j \)-th peak, respectively. An illustration of this model is shown in Figure 4, where a region of a spectrum is modeled as a combination of 6 peaks.
Figure 4. Sample region of a spectrum decomposed into 7 peaks modeled as Gaussian-Lorentzian functions. Figure 4(a) shows the original region, and Figure 4(b) shows the individual peaks and the result of combining them to fit the intensities.

The nonlinear curve-fitting algorithm estimates the optimal model parameters using their initial values and bounds. The initial location, $x_{c,j}$, of each peak is manually selected. The initial height, $M_j$, of each peak is defined as the difference between the maximum and minimum intensities in the region surrounding the peak. The initial value of the width at half height, $\Gamma_j$, is defined as double the distance (ppm) between the maximum intensity in the region and the location of the peak’s half height (i.e., initial height divided by 2). The initial standard deviation, $\sigma_j$, can then be computed from Equation (9). The initial fraction Lorentzian, $P_j$, of each peak is defined as 0.5. The initial offset, $O_j$, is defined as the minimum intensity in the segment. The lower and upper bounds for parameters are defined as:
\begin{align}
0 &< M_j \leq MAX_j, \\
0 &< \sigma_j \leq |s_L - s_R|, \\
0 &\leq P_j \leq 1.0, \\
\alpha_j &< x_{c,j} < \omega_j, \\
0 &\leq O \leq MAX_j,
\end{align}

where $MAX_j$ is the maximum height in the $j$-th segment, and $s_L$ and $s_R$ are the left and right boundaries of the segment. The boundaries for location of each peak, $[\alpha_j, \omega_j]$, are defined as the locations corresponding to the minimum intensities between the current peak and the adjacent peaks. In the special cases of the first and last peaks of each segment, the segment boundary is used to define the region.

After defining the initial values and bounds for the parameters, the nonlinear curve-fitting algorithm optimizes the parameters to minimize the difference between the model and the original data measured by Equation (10). The resulting parameters are then used as inputs to a second iteration of the nonlinear curve-fitting algorithm. Additionally, the newly optimized peak locations are used to update the lower and upper bounds of $x_{c,j}$. This second iteration enhances the nonlinear curve-fitting algorithm’s ability to find the global optimum. Following this second iteration, the results are visually inspected. Each segment is then adjusted by adding, removing, and modifying the locations of the peaks. This procedure is repeated until the model passes a visual inspection. At this point in the characterization, the goal is an approximate model for each segment. These segments will be combined to form a global model, which will be adjusted until the residual can be decomposed into independent normally distributed regions, each with a mean of zero.

After the segments are modeled individually, all of the segments are combined to obtain a global model, which is defined as follows:
where $G(\beta, x_i)$ is the global model with the model parameters, $\beta$. Further, $N$ is the number of peaks in the entire spectrum, thus, $M_j$, $\sigma_j$, $P_j$, and $x_{c_j}$ refer to the height, standard deviation, fraction of Lorentzian, and center of the j-th peak. The baseline model, $baseline(\beta, x_i)$, is the piecewise cubic interpolation of baseline intensities (i.e., height of the baseline) spaced 0.05 ppm apart \[128\]. The baseline intensities are parameters of the model ($\beta$), and thus, they are determined by the nonlinear curve-fitting algorithm.

Due to the large number of peaks (i.e., parameters) in the rat urine spectra described above, the global model is fit in an iterative fashion. First, the peaks determined from independently fitting the segments are held constant as the baseline model is fit. The initial values of the baseline intensities are the offsets of the independent segments. These baseline intensities are uniformly spaced at an interval of 0.05 ppm. These intensities are interpolated to create a smooth baseline. Second, with the baseline held constant, the peaks are fit using a sliding window of width 0.04 ppm. The window is used to select those peaks that will be fit during the current iteration. Those peaks outside of the window are held constant. A step size of 0.01 ppm is used to provide overlap between adjacent windows. Finally, after the sliding window has covered the entire spectrum, the baseline is updated again with the peaks held constant. This procedure results in the first global model.

The global model is then interactively modified until the residual can be decomposed into independent normally distributed regions ($\mu = 0$). The Anderson-Darling test is used to determine if each region follows a normal distribution ($\alpha = 0.01$) \[129-132\], and the t-test is used to determine if a normally distributed region has a mean of zero ($\alpha = 0.01$). The minimum width of
each region is 0.025 ppm (60 data points). Each region is extended until it no longer follows a normal distribution with a mean of zero. To provide flexibility, a number of small (< 0.01 ppm) non-normal segments are allowed between the normally distributed regions. The number of non-normal segments is determined by the following formula:

\[
\alpha \times \frac{(x_{\text{max}} - x_{\text{min}})}{0.01 \times \alpha},
\]

where \( \alpha \) is the significance level, and \( x_{\text{max}} \) and \( x_{\text{min}} \) are the maximum and minimum chemical shift values of the spectrum, respectively. An example region is shown in Figure 5.

In addition to defining a stopping condition for the interactive procedure described above, analyzing the residual can also be used to refine the model for each spectrum. Where two models satisfy the requirement that the residual can be decomposed into independent normally distribution regions equally well, the more parsimonious model is preferred. To achieve this objective, each peak (smallest to largest) is tested for removal from the model until the residual no longer satisfies the stopping condition. Further, a second condition is added to check the local region around selected peak, specifying that a region of 0.15 ppm centered on the peak can be decomposed into independent normally distributed regions with a mean of zero with no exceptions (i.e., flexibility).

This process is repeated until no additional peaks can be removed. Once this is finished, a single peak is considered as a replacement for every pair of adjacent peaks. Two potential peaks are fit independently as a single Gaussian-Lorentzian peak. The two adjacent peaks are then replaced by the single peak if the two stopping conditions are met and the \( R^2 \) value is above 0.98. This is repeated until no two peaks can be combined.
Once all of the spectra are modeled by a set of Gaussian-Lorentzian peaks and a piecewise cubic interpolation baseline model, the peaks are separated into three groups: baseline, background, and foreground. The distinction between background and foreground facilitates the characterization of within-peak variation. A heuristic is also used to identify baseline peaks whose width at half height is greater than six times their height. The background and foreground peaks are distinguished by the minimum distance between a maximum and its corresponding minima, where maxima and peaks are matched based on proximity. The minimum distance from maximum to minimum is calculated from the model consisting of Gaussian-Lorentzian peaks and piecewise cubic interpolated baseline. If this distance is above four times the standard deviation of the entire residual, then it is considered a foreground peak (i.e., observable). A sample illustration of a set of Gaussian-Lorentzian peaks divided into groups is shown in Figure 6. Such
real spectral features arise since the $^1$H spectra of biofluids are very congested with multiple overlapping peaks, or can sometimes contain naturally broad signals from proteins or lipids (more prevalent in blood samples). In urinary spectra, these broad signal regions are mostly due to numerous overlapping metabolite signals that are at or near the limits of NMR detection (sometimes referred to as chemical noise). In practice, measurement of these signals is not possible because they are too weak and poorly resolved, but their presence tends to distort the baseline; therefore, our peak-fitting algorithm must address these spectral features.

![Figure 6](image)

**Figure 6.** Illustration of a set of Gaussian-Lorentzian peaks divided into three groups: foreground, background, and baseline.

5.2.2.2 **Characterizing the variability of the spectra**

Each spectrum is now decomposed into a set of peaks such that the residual can be broken into independent normally distributed regions ($\mu = 0$). All of the peaks are further divided into foreground, background, and baseline. The baseline peaks are peaks whose width at half height is
greater than eight times their height. These peaks are assumed to vary between spectra. The remaining peaks are further divided based on whether they are clearly observable. The peaks that satisfy this condition (foreground) are assumed to be observable throughout the entire spectral data set, which is composed of urinary spectra from 22 control animals. These peaks are used to approximate the within peak variability.

The peak parameters \( (M_j, \sigma_j, P_j) \) for the signal peaks (combination of foreground and background peaks) and the baseline peaks are tested using the Anderson-Darling statistical test \( (\alpha = 0.05) \) to determine if they follow one of the following parametric distributions: Weibull, exponential (specific case of the Weibull distribution), normal, lognormal, and Gumbel (also known as the extreme value type 1 distribution) [133].

The peak parameters are common to both the signal and baseline peaks; however, the signal and baseline peaks are analyzed independently. Further, there are parameters that are specific to each group. This is a result of the process that will be used to create a synthetic spectrum, where the signal peaks are placed first, followed by the piecewise interpolated baseline, and baseline peaks. The distance between adjacent peaks, the distance from the start of the spectrum to the first peak, and the distance from the end of the spectrum to the last peak are calculated to characterize the signal peaks.

The baseline intensities (i.e., height of the baseline) for the piecewise cubic interpolated baseline are calculated in relationship to the number of signal peaks per ppm, and the previous baseline intensity. The baseline peaks are then determined in relation to the number of signal peaks per ppm and the baseline intensity. These values are calculated for each baseline segment of size 0.05 ppm. In addition, the distance to the first baseline peak and the distance to the last
baseline peak is measured. Finally, the normalized sum of squared error is calculated to capture the within-baseline variability using the following formula:

\[ NSSE = \sum_i \left( \frac{\mu_i - s_i}{\mu_i} \right)^2, \]  

(15)

The residual is characterized by employing a sliding window of size 0.1 ppm with a step size of 0.05 ppm to calculate the standard deviation of the residual along the spectrum. The number of signal peaks, and the number of baseline peaks per ppm are calculated for each window.

For all of the components (peaks and baseline), the relationships between the parameters must be determined to create an accurate synthetic spectrum. This relationship is evaluated using the Spearman rank correlation (\( \alpha = 0.05 \)) [134], if the parameters do not follow a parametric distribution; otherwise, the correlation is evaluated using the Pearson correlation coefficient.

The distributions described above detail the components of a single spectrum. The baseline peaks and residual are independently generated for each spectrum; however, the variation of the signal peaks and the piecewise cubic interpolated baseline between spectra must be estimated. The degree of this variability can be modified when creating a validation set. The variability within each signal peak can be approximated from the foreground peaks, which can be matched between spectra. After the peaks are matched the task of modeling the within-peak variation is straightforward; however, the results of a peak matching algorithm cannot be verified on the experimental data set. This type of evaluation will be available after the creation of a synthetic data set. The goal of characterizing the within-peak variation is to provide an approximation that will be used as a basis for the synthetic data sets. The resulting within-spectrum distributions can be varied to create several synthetic data sets to achieve a more robust validation.
The peak matching algorithm begins by arbitrarily selecting one of the spectra to serve as a reference spectrum. The rest of the spectra are then matched to this spectrum by matching its foreground peaks to the nearest peak in the reference spectrum. If two or more peaks from the same spectrum are matched to the same reference peak, these matches are ignored in future calculations. This algorithm will result in a set of peaks that have been matched between spectra that characterize within-peak variation. The within-peak distributions include distance from center, difference from average height, difference from average width, and the difference from average fraction Lorentzian. The Anderson-Darling statistical test ($\alpha = 0.01$) is repeated for each peak and each of the aforementioned distributions. If less than 1% of the tests are significant (i.e., does not follow the distribution), then the parameter is assumed to follow the test distribution.

5.2.3 Generating a synthetic spectral data set

Any number of synthetic data sets can be generated from the characteristics of the experimental $^1$H NMR spectroscopic data set. A synthetic data set is based on a single base spectrum. The base spectrum is constructed in two stages: (1) generation of the signal peaks and (2) generation of the piecewise cubic interpolated baseline. The data set is then constructed by modifying the base spectrum to introduce between spectra variability to emulate the $^1$H spectral data set. Specifically, the height, width, fraction Lorentzian, and location of the peaks are altered from the base spectrum to simulate real experiments. In addition, the piecewise cubic interpolated baseline is varied between spectra. Finally, the baseline peaks and Gaussian noise are independently generated for each spectrum.

5.2.3.1 Signal peaks

During the first stage, the signal peaks are generated by sampling the corresponding characteristic parameter distributions for the height, width, fraction Lorentzian, and location. For example, the
The positions of the peaks are determined by sampling the distance between adjacent peaks distribution, and the heights of each peak are selected by sampling the peak height distribution. The generation of a new peak is illustrated in Figure 7. The location of the first and last signal peaks are selected by sampling the corresponding distributions.

Figure 7. Generation of a new peak by sampling the distributions for the height, width at half height, fraction Lorentzian, and distance between adjacent peaks.

5.2.3.2 Baseline

The second component, the baseline, is composed of a piecewise cubic interpolated baseline of uniform segments of size 0.05 ppm and baseline peaks. The baseline is divided into three regions to accurately model segments of the baseline with different characteristics. The first and second regions contain the baseline intensities from the beginning of the spectrum to the first peak and the baseline intensities from the end of the spectrum to the last peak, respectively. The third region consists of the intervening baseline intensities. The first and second regions remain relatively flat, while the third region contains the majority of the baseline distortion. The process of generating a synthetic baseline is shown in Figure 8, where the first step is to generate a basic
baseline that will serve as a base for the baselines of the individual spectra. The basic baseline is generated by smoothing an intermediate baseline that is created by sampling the baseline intensities distributions. The basic baseline intensities, \( s_i \), are computed as the weighted average of the adjacent intermediate baseline intensities, \( u_j \), within a minimum distance, \( \Delta \):

\[
 s_i = \sum_j w_{ij} u_j ,
\]

(16)

\[
 w_{ij} = \frac{1 - |x_i - x_j|}{\sum_j w_{ij}} ,
\]

(17)

where \( |x_i - x_j| \) is the distance between the baseline intensities. Further, the degree of variation of a baseline can be controlled by modifying the minimum distance, \( \Delta \) (i.e., for a gradual baseline use a large minimum distance).

The individual spectrum baselines are generated to conform to the overall shape of the basic baseline. The amount of variation from the basic spectrum is determined by generating a target \( \text{NSSE} \). A specific baseline is then generated from the basic baseline by individually adjusting its intensities using their corresponding standard deviations (Figure 8(b)). The standard deviations control the regions of the spectrum that have higher variability (i.e., the third region). These intermediate intensities are then smoothed according to Equation (17). The smoothed intensities are then iteratively adjusted until they reach the target \( \text{NSSE} \).
Figure 8. Process of generating piecewise baseline ($\Delta = 0.5$ ppm). (a) Generation of the basic baseline by applying a weighted mean to the intermediate baseline. (b) Generation of a specific baseline for a spectrum from the basic baseline.

The baseline peaks are introduced to each spectrum by selecting the number of baseline peaks per segment with relation to the number of signal peaks and the baseline intensity. The baseline peaks are generated by sampling the characteristic parameter distributions for their height, width, and fraction Lorentzian. The locations of the baseline peaks are randomly selected within each segment. The location of the first and last baseline peaks are selected by sampling the corresponding distributions.

5.2.3.3 Noise

The standard deviation of the noise is not constant throughout the spectrum. This may be the result of a mixture of chemical noise in some regions, and true white thermal noise in other regions. This is modeled by estimating the standard deviation of the noise every 0.05 ppm with respect to the number of signal peaks in the neighborhood (0.1 ppm). These estimates are then interpolated to determine the standard deviation of the noise along the entire spectrum.
5.2.3.4 Within-spectra variability

Each spectrum in the synthetic data set is constructed by adding the spectrum independent components (baseline peaks and noise) and by modifying the base spectrum. The within-peak variability is introduced to the signal peaks, and finally, the piecewise baseline of the base spectrum is modified for each spectrum. A simplified base spectrum and two synthetic spectra with peak variability are shown in Figure 9.

![Figure 9. Two simplified spectra with their associated base spectrum.](image)

After adding the spectrum independent components, the within-peak variability is introduced by adjusting the peak parameters based on the matched foreground peak distributions. The parameter values of the matched peaks are normalized as fractional differences from their means. Then for each signal peak, a matched peak is randomly selected as a model for its within-peak variability.
The last step to creating a synthetic spectrum is to introduce variability to the piecewise baseline. The variability of the baseline is modeled by the sum of squared error from the mean baseline of the empirical data. For each baseline in the synthetic data set, a target sum of squared error is estimated. The intensities are modified according to their standard deviation until the sum of squared error from the baseline of the base spectrum reaches the target.

5.2.3.5 Generating parameters

Due to the large number of peaks (approx. 1,500) in each of the 22 spectra, sampling directly from the parameter values approximates the actual distribution. The method for selecting a parameter (e.g., peak height and location) for the synthetic spectrum depends on whether that parameter is correlated with one or more parameters, and whether the values of any of these parameters are preexisting. For example, if the height and width of a peak are correlated, they must be selected from an appropriate multivariate distribution. An example of the second case is if the height and width of a signal peak are correlated with the fraction Lorentzian, but that the distance between adjacent peaks is correlated with the height and width but not the fraction Lorentzian. To solve this problem, the height, width, and fraction Lorentzian are selected from a multivariate distribution. Then the height and width are used as preexisting values to select the distance between adjacent peaks.

The correlated parameters are drawn from a multivariate distribution represented as a table of values. If these parameters are not correlated to any preexisting parameters, then they can be selected from a table that captures the underlying multivariate distribution. The final value is determined by sorting the values for each parameter independently and then generating a uniform random number between previous and next parameter. This resolves the problem of fixing the
exact values of the parameters. When there are preexisting parameters, they constrain the range of values that can be selected from the table.

5.2.3.6 Available data sets

The procedure to generate spectral data sets can be modified to produce validation sets with different characteristics. Some of these modifications include selecting a fraction of the peaks to create a sparser spectrum, selecting a subset of the peaks to be consistent across spectra, and modifying the distributions via transformations (e.g., multiplication, addition, logarithm, exponential). In addition to generating control data sets, treatment data sets are also created with varying degrees of response. These data sets are available for download and have been organized according to their characteristics [119].

5.2.4 Case study: Comparing alignment algorithms

Three alignment algorithms were chosen to illustrate the advantages of using synthetic validation sets that accurately capture the characteristics of empirical data [120, 121]. These three alignment algorithms are available in the spectral processing software suite: SpecAlign [120]. These algorithms were developed specifically for the alignment of SELDI and MALDI type clinical proteomics data. Thus, this case study will also provide an evaluation of their applicability to NMR spectral data. The three algorithms include alignment algorithms based on peak matching or fast Fourier transform cross-correlation.

The first algorithm aligns peaks that have been automatically selected in each spectrum. Potential peaks are selected by sliding a window across the spectra to determine if there is a change in the gradient from positive to negative. These peaks are selected if they are above the baseline cutoff and also above the average intensity across the local region of the spectrum. The baseline cutoff is defined as the fraction of the baseline under the baseline intensity at which the
algorithm should ignore picking peaks. The baseline is automatically determined via a restrained moving average, where only values less than the local average are added to the global moving average. The local average is defined as \(1/100\)th the size of the entire spectrum. For the peak picking algorithm, the default parameters were used (window: 21, baseline cutoff: 0.5, height ratio: 1.5). After the peaks have been identified, each spectrum is aligned to an arbitrarily chosen target spectrum. For each spectrum, the peaks are individually aligned to the closest peak in the target spectrum. The alignment is performed by adjusting the minima adjacent to the selected peaks, where points that are inserted are estimated by least-squares fitting about the neighboring points.

The next two alignment algorithms are based on the fast Fourier transform cross-correlation. These alignment algorithms are the peak alignment by fast Fourier transform (PAFFT) and the recursive alignment by fast Fourier transform (RAFFT). These two algorithms do not depend on peak picking and are therefore more suitable to highly congested spectra [121]. These algorithms divide the spectra into segments before the evaluation of the best shift via the fast Fourier transform cross-correlation. The recursive alignment by fast Fourier transform (RAFFT) extends PAFFT by recursively searching for the optimal minimum size to divide the spectra (i.e., segment size). Both algorithms require the maximum shift of a segment to be specified. This comparison used a maximum shift of 20 points (approximately 0.01 ppm).

The use of a synthetic data set facilitates the development of metrics that can directly measure the relative performance of the algorithms. For the alignment algorithms, the optimal alignment is known \textit{a priori}. To compute the optimal alignment the peak shift is removed from each spectrum to align the peaks with the target spectrum. This alignment can then be directly compared to the
alignment results from the aforementioned algorithms. This is quantified by the average sum of squares error that is defined as follows:

\[
ASSE = \frac{1}{M} \sum_{j=1}^{M} \sum_{i=1}^{N} (y_{j,i} - a_{j,i})^2 / M,
\]

where \(y_{j,i}\) is the perfectly aligned value of the i-th data point in the j-th spectrum, and \(M\) and \(N\) are the number of spectra to align to the target and the number of data points in each spectrum, respectively. The \(ASSE\) of the unaligned spectra is compared to the \(ASSE\) after alignment. The relative increase in alignment is measured as follows:

\[
RI = \frac{ASSE_u - ASSE_a}{ASSE_u},
\]

where \(ASSE_u\) is the average sum of squares error for unaligned spectra and \(ASSE_a\) is the average sum of squares error for aligned spectra.

### 5.3 Results and discussion

#### 5.3.1 Parameters

The creation of synthetic spectral data set begins by characterizing the underlying parameter distributions. These distributions are extracted using the procedure described in Section 5.2.3. The components of a synthetic spectrum are the signal peaks, baseline peaks, baseline intensities that define the cubic interpolated baseline, and the noise. These components and their associated parameter distributions are shown in Table 1. Most of the parameters do not follow one of the parametric distributions listed in Section 5.2.2; therefore, they are treated as nonparametric. The exceptions include the baseline intensities and sum of squared error from the mean baseline; these parameters follow a normal distribution (\(\alpha = 0.05\)).
Table 1. Summary of the different components used to create a synthetic spectrum and their associated parameter distributions.

<table>
<thead>
<tr>
<th>Parameters</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Signal peaks</td>
<td>$M, \sigma, P$, distance between adjacent peaks, distance to first signal peak, distance to last signal peak</td>
</tr>
<tr>
<td>Baseline peaks</td>
<td>$M, \sigma, P$, # signal peaks per ppm, # baseline peaks per ppm, distance to the first baseline peak, and distance to the last baseline peak</td>
</tr>
<tr>
<td>Piecewise baseline</td>
<td># signal peaks per ppm, baseline intensity, sum of squared error from mean baseline</td>
</tr>
<tr>
<td>Noise</td>
<td>Standard deviation of the residual, # signal peaks per ppm, # baseline peaks per ppm</td>
</tr>
</tbody>
</table>

After analyzing each parameter individually, the relationship between parameters for each component was tested using the Spearman rank correlation. These relationships will determine the details of how a synthetic spectrum is constructed. For example, if the peak height and width are not correlated, then they can be selected independently. This procedure is described in detail in Section 5.2.3. The significant correlations ($\alpha = 0.05$) for each component are shown in Table 2.
Table 2. Summary of the relationships between the parameters for each of the components: (a) Signal peaks, (b) baseline peaks, (c) piecewise baseline, and (d) noise.

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Parameter 1</th>
<th>Parameter 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height, $M$ ↔ Width, $\sigma$</td>
<td></td>
<td>Height, $M$ ↔ Width, $\sigma$</td>
</tr>
<tr>
<td>Height, $M$ ↔ Fraction Lorentzian, $P$</td>
<td></td>
<td>Width, $\sigma$ ↔ Fraction Lorentzian, $P$</td>
</tr>
<tr>
<td>Height, $M$ ↔ Distance between adjacent signal peak</td>
<td></td>
<td># baseline peaks per ppm ↔ # signal peaks per ppm</td>
</tr>
<tr>
<td>Width, $\sigma$ ↔ Fraction Lorentzian, $P$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Width, $\sigma$ ↔ Distance between adjacent signal peaks</td>
<td></td>
<td># baseline peaks per ppm ↔ baseline intensities</td>
</tr>
<tr>
<td>(a) Signal peaks</td>
<td></td>
<td>(b) Piecewise baseline</td>
</tr>
<tr>
<td>Baseline intensity ↔ # signal peaks per ppm</td>
<td></td>
<td>Standard deviation ↔ # signal peaks per ppm</td>
</tr>
<tr>
<td>Baseline intensity ↔ Previous baseline intensities</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(c) Baseline peaks</td>
<td></td>
<td>(d) Noise</td>
</tr>
</tbody>
</table>

5.3.2 Case study: Comparing alignment algorithms

The case study illustrates the advantages of using the synthetic validation sets to directly compare algorithms. Three spectral alignment algorithms were selected to test their applicability to NMR spectral data. The algorithms were compared on 30 synthetic validation sets each containing 5 spectra. A sample region of one of these data sets is shown in Figure 10. Each alignment algorithm was applied to the data sets using the first spectrum in the data set as the reference. The algorithms are tested to determine if a statistically significant positive change in $RI$ is observed. These results are shown in Table 3.
The PAFFT and RAFFT correlation alignment algorithms show a significant positive change after alignment ($\alpha = 0.05$). The peak matching alignment algorithm fails to improve the alignment as measured by ASSE. This is most likely a result of the congestion typical of $^1$H spectra according to the authors [121]. A second comparison between RAFFT and PAFFT using the two-sample t-test showed that RAFFT was significantly better than PAFFT ($\alpha = 0.05$).

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Average</th>
<th>Standard deviation</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Peak matching method</td>
<td>3.59</td>
<td>12.33</td>
<td>0.0935</td>
</tr>
<tr>
<td>PAFFT correlation method</td>
<td>2.78</td>
<td>3.09</td>
<td>0.0002</td>
</tr>
<tr>
<td>RAFFT correlation method</td>
<td>5.16</td>
<td>7.30</td>
<td>0.0016</td>
</tr>
</tbody>
</table>
5.4 Discussion

Researchers in the “omics” fields (e.g., metabolomics, proteomics, and genomics) must select the most appropriate algorithms to facilitate data analysis. The selection of these techniques is complicated by the diverse array of algorithms available. These algorithms are designed for the complex characteristics of specific experimental data. To select from these algorithms, researchers must rely on domain expertise and published comparisons.

These data-driven algorithms are commonly compared and evaluated on empirical and simulated data. While empirical data captures the complex characteristics of experimental data, comparisons are often formed on indirect performance metrics because the optimal or correct output is difficult to obtain a priori. Simulated data is commonly employed to overcome the drawback of relying on indirect performance metrics; however, the conclusions obtained on this type of data depend on the ability of the simulated data to capture the salient features of empirical data.

In this manuscript, we develop a technique for creating synthetic validation sets that characterize the salient features based on NMR spectroscopic data of rat urine samples from a metabolomics experiment. The validation sets were developed by modeling each spectrum as a combination of Gaussian-Lorentzian peaks and a piecewise cubic interpolated baseline. Each spectrum was constructed such that the residual could be decomposed into regions that follow a normal distribution, each with a mean of zero. Several validation sets were constructed with a variety of characteristics and are publicly available [119]. Three alignment algorithms are compared to illustrate the applicability of the validation sets to compare algorithms.

Two of the alignment algorithms based on the cross-correlation (PAFFT and RAFFT) showed a significant positive change after alignment ($\alpha = 0.05$). The peak matching alignment algorithm
fails to improve the alignment as measured by ASSE. According to the authors, this is a result of the congestion typical of $^1$H spectra [121]. Between the PAFFT and RAFFT alignment algorithms, the RAFFT algorithm was significantly better than PAFFT using the two-sample t-test ($\alpha = 0.05$). The RAFFT or recursive alignment fast Fourier transform alignment algorithm optimizes the minimum segment size employed during the alignment.

The case study illustrates the advantages of the synthetic data sets to compare and validate algorithms. This data will facilitate the development of novel algorithms in addition to improving the quality of algorithm comparisons. The availability of these data sets significantly improves the ability of researchers to select the most appropriate algorithms for their data analysis needs.
6 Gaussian binning

6.1 Introduction

Analyzing NMR spectroscopic data requires a variety of signal processing and pattern recognition techniques. Quantification of spectral features is a key step in the development of classification algorithms and biomarker identification (i.e., pattern recognition). A common method of quantification employed by the NMR community is known as uniform or fixed width binning, which divides a NMR spectrum into several hundred non-overlapping regions or bins of equal size [47, 49, 55, 57-59, 122-124]. This technique is performed to (1) minimize effects from variations in peak positions caused by sample pH, ionic strength, and composition [13]; and (2) reduce the dimensionality for multivariate statistical analyses. The result is a data set with fewer features, thereby, increasing the tractability of pattern recognition techniques, such as principal component analysis (PCA) [34, 35] and partial least squares discriminant analysis (PLS-DA) [135, 136].

While uniform binning mitigates the effects from variations in peak positions, shifts occurring near the boundaries can result in dramatic quantitative changes in the adjacent bins due to the non-overlapping boundaries. This problem can be countered by incorporating a kernel-based binning method that weights the contribution of peaks by their distance from the center of the bin. A typical kernel shape is the Gaussian probability density function, where the amount of overlap between each bin is controlled by its standard deviation.

Herein we propose a new Gaussian binning method for processing NMR spectroscopic data for multivariate analysis. This technique is compared to the traditional uniform binning technique by analyzing sensitivity to peak shift and loss of information content. Data for this comparison
are derived from a $^1$H NMR-based experiment to monitor rat urinary metabolites following exposure to the toxin $\alpha$-naphthylisothiocyanate (ANIT) [125]. The Gaussian binning method showed a decrease in the sensitivity to peak shift and performed comparable to uniform binning in retaining spectral information. Finally, some guidelines for using this methodology in NMR-based metabolomics data sets are presented.

6.2 Materials and Methods

6.2.1 Empirical data

Alpha-naphthylisothiocyanate (ANIT) was purchased from Sigma Chemical Company (95% purity). All other chemicals were purchased from standard sources and were of reagent grade. Data acquisition and preparation is discussed in detail in Section 5.2.1. Treated animals were administered a single 50 mg/kg dose of $\alpha$-naphthylisothiocyanate (ANIT) in corn oil vehicle via gavage, while control animals were given an identical volume of corn oil vehicle only. Urine was collected into containers chilled on dry ice and containing 1 ml of 1% sodium azide. Urine samples were obtained pre-dose and at 24, 48, 72, and 96 hours post-dose. All urine samples were stored at -40 °C prior to analysis by NMR spectroscopy.

6.2.2 Binning

There are two main alternatives for analyzing NMR spectroscopic data: peak extraction/identification and spectral binning. Peak extraction attempts to identify individual signals in the spectrum, typically employing an alignment algorithm. Spectral binning divides a spectrum into regions (bins) and replaces each bin by its integral value. This increases the tractability of pattern recognition techniques by reducing the number of variables and mitigates problems associated with peak shift. Two binning techniques, uniform and Gaussian binning, are described below using the $^1$H spectrum from a standard sample of ethylbenzene (NMR
calibration/test standard; 0.1% ethylbenzene in CDCl$_3$). For the purposes of these binning descriptions, only the ethyl quartet (ca. 2.56 ppm) is shown for simplicity.

### 6.2.2.1 Uniform binning

Before applying pattern recognition and data mining techniques, each spectrum is commonly binned to reduce the number of variables (i.e., increase tractability). The first step to uniform binning is to divide the spectrum into non-overlapping regions/bins of fixed size (Figure 11(a)). Then, the collective data points in each region/bin are replaced by the integral value for the bin. The new integral value is placed at the middle of the region (Figure 11(b)). This process is referred to as uniform binning or fixed width binning, and the value of each bin can be calculated as follows:

\[
u_i = \sum_{j=1}^{\lvert R_i \rvert} I_j ,
\]

where \( u_i \) is the value of the \( i \)-th bin and \( I_j \) is the \( j \)-th intensity of \( R_i \), which is the set of individual preprocessed intensities in the \( i \)-th region.

### 6.2.2.2 Gaussian binning

Uniform binning methods typically divide the spectrum into non-overlapping regions. Overlapping bins, however, can be successfully employed using a kernel-based method. To apply kernel based binning, a function (or kernel) is used to weight the data points in each bin. For each data point, the kernel (any continuous function) is multiplied by the signal and the resulting value replaces the original data point. Generally a kernel is selected that assigns high weights to data points near the center of the bin, and decreasing weights to data points closer to the bin boundaries. A commonly used kernel for pattern recognition applications is the Gaussian probability density function. The mean of the distribution is placed at the center of the bin to be
evaluated. The Gaussian kernel has two tunable parameters: the center-to-center separation between adjacent kernels (step size), and the width represented by the standard deviation of the Gaussian distribution. Together these describe the degree of overlap and determine the resulting number of bins for a given spectral width. In comparison, the step size (bin width) is the only tunable parameter of uniform binning.

![Figure 11](image1.png)

**Figure 11.** The process of uniform binning over a region of the $^1$H spectrum of ethylbenzene containing the quartet from the ethyl group at 2.56 ppm. (a) Shows the region of the spectrum divided into equal width bins, and (b) shows this region after uniform binning.

To determine the binned value using a Gaussian kernel, the signal is first transformed by multiplying each data point by the univariate Gaussian probability density function with $\mu$ equal to the center of the bin, and a specific user defined value for $\sigma$ (Figure 12(a) and Figure 12(b)). The transformed curve is then integrated over the interval around the bin center (Figure 12(c)). The integration interval can be written $\mu \pm c\sigma$, where $c$ is chosen such that all values of the transformed signal beyond the integration boundaries are very close to zero. For the work presented here, a value of $c = 3$ was used. Thus, each bin value is calculated as follows:
where $g_i$ is the value of the $i$-th bin, $I_j$ is the value of the $j$-th intensity in $E_i$, which is the set of all preprocessed intensities in the interval $\mu_i \pm c \sigma$, $s_j$ is the chemical shift of $j$-th intensity in the $i$-th bin, $\mu_i$ is the center of the bin (i.e., $s_j - \mu_i$ is the distance from the center of the $i$-th bin), $\sigma$ is the standard deviation of the Gaussian weighting function, and $weight(\mu_i, \sigma, s_j)$ is the Gaussian weighting function.

6.2.3 Evaluating binning methods

In this study, Gaussian binning and uniform binning were evaluated and compared on two criteria: loss of information and sensitivity to peak shift. When a spectrum is binned there is an implicit loss of information that occurs due to the reduction in the number of data points. Binning also reduces the number of variables, which increases the tractability of pattern recognition techniques. The other criterion, sensitivity to peak shift, is a measurement of the robustness of a binning method to handle peak shift due to various factors, typically sample pH, ionic strength, and composition.
Figure 12. The process of Gaussian binning over a region of the $^1$H spectrum of ethylbenzene containing the quartet from the ethyl group at 2.56 ppm. (a) Shows the original region of the spectrum divided into Gaussian bins; (b) shows the second bin of the quartet before and after applying the Gaussian weighting function; and (c) shows the corresponding uniform and Gaussian binned spectra.

6.2.3.1 Loss of information

All binning techniques inherently result in a loss of information due to the reduction in spectral resolution (number of points/ppm). This loss of information should be weighed against other criteria, such as sensitivity to peak shift and increased tractability. To assess the loss of information, Gaussian and uniform binning methods were evaluated by examining PCA results and by measuring the degree of smoothing. PCA is an unsupervised method commonly used to
visualize NMR-derived metabolomic data [34, 35]. PCA transforms the data into a new coordinate system, where each new dimension is computed as a linear combination of the original values. These new dimensions are called principal components, and are orthogonal vectors that best explain the overall variance in the data. Various literature reports describe the use of PCA to illustrate the clustering of NMR spectral intensities using the top principal components (i.e., the principal components containing the largest parts of the total variance of the data set). [137-140] The underlying idea is that the metabolite profiles observed in the spectra can distinguish among experimental groups based on their ability to differentiate as a possible biomarker.

The results of PCA were analyzed in two ways: first, the percent data variance explained in the first two principal components is graphed as a function of the parameters, standard deviation and step size; second, the quality of the clusters in the first two principal components was measured using the $J_2$ criterion. A high value of $J_2$ indicates well separated and tightly clustered groups of samples. $J_2$ is defined as follows:

$$J_2 = \frac{|S_w + S_b|}{|S_w|}, \quad (23)$$

Here $S_w$ is the within-class scatter and $S_b$ is the between-class scatter. $S_w$ is a measure of the diffuseness of the data within each cluster, and is calculated as follows:

$$S_w = \sum_{i=1}^{M} P_i S_i, \quad (24)$$

where $P_i$ is the proportion of the data in the $i$-th cluster, $M$ is the number of clusters, and $S_i$ is the covariance matrix of the $i$-th cluster. $P_i$ and $S_i$ are calculated as follows:
\[ P_i = n_i / N, \]  \hspace{1cm} (25)

\[ S_i = E[(\bar{x} - \bar{\mu}_i)(\bar{x} - \bar{\mu}_i)^T], \]  \hspace{1cm} (26)

where \( N \) is the number of samples, \( n_i \) is the number of samples in the \( i \)-th cluster, \( \mu_i \) is the mean of the \( i \)-th cluster, \( ^T \) is the transpose operator, and \( E \) is the expectation function. The between-class scatter, \( S_b \), is defined as follows:

\[ S_b = \sum_{i=1}^{M} P_i (\bar{\mu}_i - \bar{\mu}_0)(\bar{\mu}_i - \bar{\mu}_0)^T, \]  \hspace{1cm} (27)

\[ \bar{\mu}_0 = \sum_{i=1}^{M} P_i \bar{\mu}_i \]  \hspace{1cm} (28)

To further measure the information loss, the amount of variability in the spectrum was calculated. As the degree of smoothing increases for a spectrum, the normalized variance from the mean will decrease. This is referred to as signal variance (SV) and is calculated as follows:

\[ SV = \frac{\sum_{v \in A} \left( \frac{v - m}{m} \right)^2}{|A| - 1}, \]  \hspace{1cm} (29)

\[ m = \frac{\sum_{v \in A} v}{|A|}, \]  \hspace{1cm} (30)

where \( A \) is the set of intensities in a spectrum, \( v \) is the intensity at a given point in a spectrum, and \( m \) is the mean intensity across the entire spectrum.
6.2.3.2 Sensitivity to peak shift

Sensitivity to peak shift is a measurement of the robustness of a binning method. To assess this, $^1$H spectra of ethylbenzene were acquired at various offset frequencies by varying the centerband frequency in steps of 0.5 Hz over a range of 24 Hz (this range covers one full bin width in this example). The error for each bin was then measured as a difference from the 0.0 Hz offset spectrum (reference) as follows:

$$E_i(p) = \left( \frac{v_{i,p} - v_{i,0,0}}{v_{i,0,0}} \right)^2$$  \hspace{1cm} (31)

Here $v_{i,p}$ is the value of the $i$-th bin in the spectrum with offset $p$ and $v_{i,0,0}$ is the value of the $i$-th bin in the spectrum with a 0.0 Hz offset.

The sensitivity to peak shift of a binning algorithm was also analyzed using a region of the ANIT data set (6.62 – 6.58 ppm), which is shown in Figure 13(a). The signal at ca. 6.6 ppm is tentatively assigned to overlapping peaks from trans-aconitate (partially resolved triplet) and fumarate (singlet), using Chenomx NMR Suite software (Edmonton, Alberta, Canada). These resonances are pH sensitive and the composite signal shown in Figure 13(a). Researchers are typically interested in clustering based on differences in peak heights, as opposed to differences resulting from peak shift. To measure sensitivity to peak shift on the clustering results, the inter-peak distances of the corrected (i.e., aligned) data was compared to the inter-peak distances of the uncorrected (i.e., misaligned) data. The corrected spectral region is shown in Figure 13(b). The comparison is quantified by computing the cluster error ($CE$), defined as follows:
\[ CE = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left| \frac{c_{i,j} - u_{i,j}}{c_{i,j}} \right| \]  

(32)

where \( N \) is the number of bins, and \( c_{i,j} \) and \( u_{i,j} \) are the distances between \( i \)-th and \( j \)-th data points in the PCA scores plot of the corrected and uncorrected data, respectively.

---

**Figure 13.** Region of the ANIT data set for comparing uniform and Gaussian binning using the cluster error metric. (a) Shows the uncorrected peaks; and (b) shows the corrected (i.e., aligned) peaks.

---

### 6.3 Results

#### 6.3.1 Loss of information

Loss of information was assessed by analyzing the PCA results for urinary metabolite data obtained before and after exposure to ANIT (50 mg/kg dose). Our laboratory and others [122, 125, 141, 142] have shown that ANIT induces a maximal effect on the urinary metabolite profile in rats at 48 hours post-exposure. This data set is described in detail in Section 5.2.1. For the purposes of this paper, a subset of the data was analyzed, specifically the spectra obtained at pre-dose and 48 hours post-dose. This data set comprises 24 spectra from 12 animals.
The information captured by a binning method should be robust to parameter changes (i.e., increasing the step size). Further, a binning procedure should maintain any natural clusters resulting from differences in peak heights between groups. This information can be lost when more than one peak is contained within a bin. The resulting loss of information can increase the percent variance explained by the first two principal components. As an example, consider two adjacent uncorrelated peaks that contribute to the first and second principal components. If these peaks are binned individually, one will contribute primarily to the first principal component while the other will contribute primarily to the second principal component. If the two peaks are binned together, then some of their orthogonality (i.e., information) will be lost. This will manifest itself in an increase in either the variation explained by the first or second principal component.

For Gaussian binning, the percent variance explained in the data by the first two principal components, $PC_1$ and $PC_2$, was graphed as a function of the kernel standard deviation ($\sigma$) and step size (Figure 14(a)). In general, it is desirable to encompass as much of the variance in these first two principal components to allow for easy visualization and pattern analysis. For the ANIT data set analyzed in this paper, greater than 80% of the variation in the data was explained by the first two principal components. In addition to maintaining the natural clusters of the data, a binning algorithm should be robust to parameter changes. This can be observed by examining the percent variance as a function of these parameters. These results show that the variance in $PC_1$ increases as the standard deviation of the kernel increases, which can be attributed to a smoothing of each spectrum (i.e., a loss of information). Here the standard deviation is expressed as a fraction of the step size. Initially, the percent variance in $PC_1$ increases as the value of $\sigma$ increases, but it then levels off and remains fairly stable. Further, as the step size becomes large relative to the kernel standard deviation, the percent variance increases. This is due to the kernels being spaced too far apart, which leads to intervening segments of the spectrum being omitted in
the binning process. Figure 14(b) illustrates the sensitivity of uniform binning to the step size. As the step size increases, the uniform binning procedure decreases the number of bins in a spectrum, which leads to a smoothing effect that produces an increase in the percent variance in $PC_1$. Much of the work in the field of NMR-metabolomics has typically used a step size of 0.04 ppm for $^1$H data, which appears to be a nearly optimal value (for this data set) with regard to the percent variance in $PC_1$ as shown in Figure 14(b).

![Figure 14. Percent variance in the first two principal components, $PC_1$ and $PC_2$, for (a) Gaussian binning as a function of the step size (ppm) and the standard deviation as a fraction of the step size (fss); and (b) uniform binning as a function of the step size. Results obtained from the $^1$H NMR ANIT data set.](image)

The next metric used to assess the loss of information was the $J_2$ criterion, which measures quality of the clustering. The clusters associated with experimental procedures (i.e., treatments) are the basis of biomarker identification, and therefore, preservation of cluster integrity is an important metric for the quality of binned data. Specifically, this metric measures the within-cluster scatter and between-cluster separation (see Equation (23)). An increase in the value of $J_2$ indicates more between-class separation and/or less within-class scatter. Thus, higher values indicate well-separated compact clusters. The effects of varying the Gaussian binning parameters
(standard deviation and step size) on $J_2$ are shown in Figure 15. As the standard deviation increases, the value of $J_2$ for Gaussian binning reaches a maximum and then gradually decreases (Figure 15(a)). Figure 15(b) shows how the value $J_2$ for uniform binning is affected by the choice of step size.

Figure 15. The cluster equality criterion, $J_2$, for (a) Gaussian binning as a function of the step size (ppm) and the standard deviation as a fraction of the step size (fss); and (b) uniform binning as a function of the step size. Results obtained from the $^1$H NMR ANIT data set.

For $^1$H spectral data sets, a step size of 0.04 ppm is common for a resonance frequency of 500-600 MHz (typical for NMR-based metabolomics experiments). The other parameter needed for Gaussian binning, standard deviation, can be chosen by examining the cluster criterion, $J_2$. In Figure 16, the $J_2$ metric for the ANIT $^1$H spectral data set is plotted as a function of the Gaussian standard deviation, expressed as a fraction of the binning step size (0.04 ppm). From this graph, the optimal value for the standard deviation is $0.29 \times \text{step size}$. This value of standard deviation, $\sigma$, yields approximately 8% overlap with the adjacent bins. The dashed line marks the uniform binning value of $J_2$ for the same data set with a step size of 0.04 ppm. The PCA scores plots for (a) Gaussian binning with a step size of 0.04 ppm and a standard deviation of $0.29 \times \text{step size}$ and
(b) uniform binning with a step size of 0.04 ppm are shown in Figure 17. Further, the associated PCA loadings plots for (c) Gaussian binning and (d) uniform binning are shown in Figure 17.

![Figure 16](image)

**Figure 16.** The cluster criterion, $J_2$, plotted as a function of the Gaussian standard deviation ($\sigma$) for the ANIT $^1$H spectral data set using a bin step size of 0.04 ppm. Here $\sigma$ is expressed as a fraction of the bin step size (fss). The dashed line is the value of $J_2$ for uniform binning with a step size of 0.04 ppm.

Loss of information was also measured by computing the signal variance, which calculates the variability in a spectrum. Figure 18(a) shows the average signal variance ($SV$) as a function of kernel standard deviation and step size. The signal variance decreases as the kernel standard deviation increases, as a result of the smoothing effect of Gaussian binning. A similar effect is shown in Figure 18(b) for uniform binning.
Figure 17. The PCA scores plot for (a) Gaussian binning with a step size of 0.04 ppm and a standard deviation of $0.29 \times \text{step size}$, and (b) uniform binning with a step size of 0.04 ppm. The associated loadings plots for (c) Gaussian binning and (d) uniform binning. Data are derived from the urinary spectral analyses for rats before (pre-dose/control) versus 48 hours after exposure to 50 mg/kg ANIT.
Figure 18. The signal variance, $SV$, for (a) Gaussian binning as a function of the step size (ppm) and the standard deviation as a fraction of the step size (fss); and (b) uniform binning as a function of the step size. Results obtained from the $^1$H NMR ANIT data set.

6.3.2 Sensitivity to peak shift

Sensitivity to peak shift is an important feature in a binning method. The sensitivity was assessed by comparing binning results for a peak-shifted data set. The peak shifted data set was obtained by offsetting the reference frequency from 0-24 Hz for $^1$H NMR spectra of ethylbenzene. A segment of the spectrum (the ethyl group quartet), shown in Figure 19(a), was then selected for assessing sensitivity. In this example, a reference spectrum with an offset of 0.0 Hz is shown together with two additional spectra in which the offset frequency was shifted by 8.0 and 12.0 Hz, relative to the reference. The bin width for this demonstration is set at 24 Hz (0.04 ppm). The sensitivity is measured as the error of a bin resulting from a given peak shift (see Section 6.2.3.2 for more information). The results, shown in Figure 19(b), illustrate that Gaussian binning is less sensitive to peak shift relative to uniform binning. Further, increasing the standard deviation increases the robustness of the algorithm to peak shift. The uniform binning procedure shows sharp jumps as peaks shift outside of the fixed bin width range.
Figure 19. Sensitivity to peak shift. (a) Shows the ethyl quartet from the $^1$H spectrum of ethylbenzene acquired at three different offset frequencies (0, 8, and 12 Hz) to generate a peak-shifted data set; (b) shows the error for the first bin with respect to peak shift.

The sensitivity to peak shift was also assessed by the clustering of data in a PCA scores plot for a region of the ANIT data set (shown in Figure 13(a)). To measure sensitivity to peak shift on the clustering results, the inter-peak distances of the corrected (i.e., aligned) data was compared to the inter-peak distances of the uncorrected (i.e., misaligned) data. This is measured by the cluster error metric, and the error for uniform and Gaussian binning is plotted as a function of the step size in Figure 20. The range of the step size was selected as 0.001 to 0.009, so that the selected region is divided into multiple smaller bins. The standard deviation used for Gaussian binning was $0.4 \times \text{step size}$. 
6.4 Discussion

Spectral binning is a common preprocessing technique applied to NMR spectral data sets to reduce the dimensionality of the data. Additionally, binning is useful to minimize problems associated with peak alignment between spectra due to slight shifts in signal resonance frequencies. Such shifts are typically observed for peaks that are sensitive to pH, ionic strength, and sample composition. But binning procedures are still susceptible to shifting peaks, especially uniform binning, which divides the spectrum into regularly spaced non-overlapping bins.

The hard boundaries of the bins present a problem because peaks can shift between adjacent bins. The bin boundaries are regular and independent of the observed spectrum, and therefore, it is possible and even likely that individual peaks will span bin boundaries. This can be problematic, because a signal from a specific metabolite may be divided between several neighboring bins. Further, this divided and thus weakened signal may become indistinguishable from noise. This effect is illustrated in Figure 19, where a slight peak shift results in a large change in bin values. For example, the error for uniform binning increases by 150% between a reference frequency of 8 and 12 Hz (i.e., a peak shift of 8 to 12 Hz with a bin width of 24 Hz),
while the error for Gaussian binning with a kernel standard deviation of 8 Hz increases by only 50% (Figure 19(b)). To avoid signal loss due to bin boundaries, overlapping bins may be used in combination with an appropriate kernel to weight the data. Gaussian binning avoids hard boundaries by incorporating a Gaussian weighting function, thus, resulting in bins that are resilient to peak shifts. The sensitivity to peak shift is also measured by the cluster error metric. Figure 20(b) shows that uniform binning is more sensitive to changes in the step size and that Gaussian binning performs better or equivalently to uniform binning for varying step sizes on the region shown in Figure 20(a).

The Gaussian binning procedure introduces overlap between adjacent bins, which has both a positive and a negative impact on the resulting data set. Overlap between bins improves the binning process by decreasing sensitivity to peak shift. But it can also hinder subsequent pattern recognition by duplicating signals when peaks lie in the overlapping region between bins. This loss of information is controlled by the binning parameters: standard deviation and step size. In the case of uniform binning, as the step size increases, the width of each bin increases, thus increasing the likelihood of bins containing more than one peak, which results in a higher probability of losing information from individual peaks. With Gaussian binning, increasing both the standard deviation and the step size also results in a higher probability of losing information from individual peaks. This loss of information as a function of the binning parameters was measured by the metrics $J_2$ and signal variance ($SV$) (Figure 15 and 6), and by the percent variance explained in the first two principal components (Figure 14). These figures show that increases in these parameters for either Gaussian binning or uniform binning increases the loss of information, but this is accompanied by decreasing the sensitivity to peak shift when employing Gaussian bins.
Another consequence of employing Gaussian binning is that the contribution of individual data points to the bins is dependent on its distance from the center of the bin. This means that some of the data points contribute more to a bin than others. A similar problem is evident with uniform binning because peaks that are near the boundaries can be split between bins, thus, the contribution of this peak to the bins is diminished. This shortcoming is the result of naively placing the bins at regular intervals. The solution, which is a direction for future work, is to dynamically place the bins, so that the center of the bin is located on the top of a peak.

As more information is retained by using a smaller standard deviation and/or a smaller step size, the sensitivity to peak shift increases. This tradeoff must be taken into consideration when choosing the appropriate parameters. Further, some pairs of the parameters, standard deviation and step size, can lead to some data points being ignored all together. A set of recommended values for the standard deviation and step size is shown in Table 4. This table shows values of standard deviation corresponding to varying degrees of overlap for standard values of the step size (i.e., uniform bin width). The overlap of a bin is calculated as the amount of area under that Gaussian kernel that overlaps adjacent bins.

**Table 4. Typical values of standard deviation.**

<table>
<thead>
<tr>
<th>Overlap (percentage)</th>
<th>0.04</th>
<th>0.16</th>
<th>0.20</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>0.0102</td>
<td>0.0408</td>
<td>0.0510</td>
</tr>
<tr>
<td>10</td>
<td>0.0122</td>
<td>0.0488</td>
<td>0.0610</td>
</tr>
<tr>
<td>20</td>
<td>0.0156</td>
<td>0.0624</td>
<td>0.0780</td>
</tr>
<tr>
<td>30</td>
<td>0.0193</td>
<td>0.0772</td>
<td>0.0965</td>
</tr>
<tr>
<td>40</td>
<td>0.0238</td>
<td>0.0952</td>
<td>0.119</td>
</tr>
</tbody>
</table>

The $J_2$ metric plotted as a function of the standard deviation (Figure 16) can be used to find an optimal range of values for the standard deviation. Any value of standard deviation with a value of $J_2$ at or above the value $J_2$ for uniform binning would produce an equivalent or better clustering.
in a PCA scores plot. The corresponding values of the standard deviation range from $0.16 \times step\ size$ to $0.45 \times step\ size$. Equivalently, this can be specified as a range of bin overlap from $0.1\%$ to $26\%$. Due to the fact that larger values of standard deviation result in bins resilient to shift, a standard deviation greater than or equal to the optimal value of $0.29 \times step\ size$ but with a corresponding value of $J_2$ above that of uniform binning is recommended. Thus, the recommended range of the standard deviation is $0.29 \times step\ size$ to $0.45 \times step\ size$ or $9\%$ to $26\%$ overlap. Using the optimal value of the standard deviation ($0.29 \times step\ size$), PCA was performed on the ANIT data set (Figure 17). These figures illustrate that the two binning techniques produce approximately the same PCA scores and loadings plots. Further, it is desirable for a PCA scores to show a large separation between tightly clustered groups. This is the case for both the Gaussian and uniform binning results. The cluster integrity and separation of the two methods are very similar; however, the key advantage of Gaussian binning is its resilience to peak shift, as illustrated in Figure 19(b).

Spectral binning is a single step in the processing of NMR-based metabolomics data. The interaction between the processing algorithms, such as binning, noise removal, and scaling, should be considered when analyzing spectral data. While most processing procedures that include noise removal and scaling in conjunction with binning will not affect the results of pattern recognition techniques, some techniques, such as auto-scaling, have a potential to produce spurious results. Auto-scaling is designed to remove the tendency of bins with large peaks to dominate pattern recognition results [112]. Thus changes in signal amplitudes for a peak across the samples are made relative to each other, and are independent of the absolute scale of the change. This is accomplished by scaling each bin to have unit variance and a mean of zero. In some cases, depending on the noise removal technique, not all of the noise is successfully removed. This may result in bins containing spurious 'noise spikes' that survive a noise removal
procedure, such as limit of detection. These bins could become significant after auto-scaling. This is a problem for both uniform and Gaussian binning; however, due to the overlapping bins of Gaussian binning, multiple bins can be affected by a single noise spike, as opposed to one for uniform binning. When using either Gaussian or uniform binning, these problems can be avoided by using a scaling technique that does not divide by the standard deviation, by using a noise removal technique that does not leave spikes, or by removing noise after binning.

The ability of these pattern recognition techniques to classify a sample or help identify putative biomarkers is influenced by the quality of the spectral features, typically represented as bins. While Gaussian binning helps mitigate the problem of peak shifts occurring near the boundaries of uniform binning, further improvement could be gained by dynamically placing the bins. With Gaussian binning this would require changing the step size of the bin and the standard deviation of the kernel to match spectral features. Dynamic bin selection would increase the amount of information retained from the high resolution NMR spectra, but would also increase the difficulty of comparing bins across multiple spectra. However, various techniques are currently being developed to overcome the difficulties in comparing binned results. The Gaussian binning method could also be incorporated into proprietary binning (bucketing) procedures, such as the IntelliBucket™ (http://www.bio-rad.com). The IntelliBucket™ allows the widths of the bins, which could be Gaussian or uniform, to be modified based on the Overlap Density Heatmap.

6.5 Concluding remarks

The Gaussian binning method described in this paper is shown to be robust to changes in peak positions and to retain the information needed by classification and multivariate statistical techniques for NMR spectroscopic data sets. The spectral quantification provided by the Gaussian binning technique reduces the number variables for pattern recognition methods and mitigates the
effects from variations in peak positions, including shifts that occur near the boundaries. The sensitivity to peak shift can be lessened by increasing the Gaussian kernel standard deviation, but an increase in the kernel standard deviation can also result in an increased loss of information. A guideline for picking this parameter is developed, and shown to be effective for a $^1$H NMR-based data set from an experiment to monitor rat urinary metabolites following exposure to the toxin $\alpha$-naphthylisothiocyanate (ANIT). In regards to both sensitivity to peak shift and loss of information, the Gaussian binning method is shown to be superior to the traditional technique of uniform binning.
7 Dynamic adaptive binning

7.1 Introduction

Despite the development of targeted quantification techniques, binning remains a commonly used simple and effective high throughput quantification technique for the NMR community [47, 49, 55, 57-59, 122-124]. The previous chapter demonstrated that dramatic quantitative errors resulting from shifts occurring near boundaries can be countered by incorporating a kernel-based binning method that weights the contribution of peaks by their distance from the center of the bin [143]. While this is an improvement over traditional fixed-width binning, dynamically determining the size and location of each bin mitigates the non-uniform spacing of NMR peaks.

One such dynamic binning algorithm is adaptive intelligent binning, which recursively identifies bin edges in existing bins [76]. Another dynamic binning method is adaptive binning, which uses the undecimated wavelet transform to smooth a composite spectrum. The observed peaks and minima of the smoothed composite spectrum are then used to dynamically bin the spectra [144]. The composite spectrum is smoothed to remove multiple observed peaks that arise from misaligned peaks. When using a composite spectrum, adjacent peaks in the same spectrum can be misinterpreted as a single peak due to misalignment. To overcome this drawback, the bin boundaries can be dynamically determined by optimizing a heuristic based objective function that utilizes individual spectra smoothed via a wavelet transform.

Herein we propose a novel dynamic binning method, dynamic adaptive binning, for processing NMR spectroscopic data for multivariate analysis. With this technique bin boundaries are dynamically determined via dynamic programming by optimizing an objective function that measures the quality of the bin configuration. This technique is shown to be superior to the
traditional uniform binning technique, adaptive binning, and adaptive intelligent binning based on their ability to create bins containing a single peak and maximize the distance from peak to bin boundary. This comparison is facilitated by synthetic data sets that capture the salient characteristics of $^1$H NMR spectroscopic data from a urinary profile [119]. Finally, a case study demonstrates the capabilities of dynamic adaptive binning in comparison to uniform binning on a $^1$H NMR-based experiment to monitor rat urinary metabolites.

7.2 Methods

The technique of spectral binning is a general signal processing technique that reduces the dimensionality of spectroscopic data while attempting to retain the pertinent information and mitigate quantitative effects of peak misalignment. Spectral quantification transforms every sample, represented as an NMR spectrum, into a feature vector. Biomarker identification can then be defined as finding a set of features that describe a pattern between groups, thus, the success of biomarker identification is directly related to the quality of the feature vectors. Here a biomarker is defined as a set of NMR signals that change relative to some reference (i.e., before and after exposure to a toxin). Such an experiment will have at least two groups (e.g., pre-dose and post-dose) for which spectroscopic data is compiled.

7.2.1 Spectroscopic Data

Both empirical and synthetic spectroscopic data are employed to show the application of dynamic adaptive binning. The synthetic spectroscopic data sets are based on urine $^1$H spectra and were developed by characterizing the salient distributions in empirical spectroscopic data [119]. Each spectrum is modeled as a combination of Gaussian-Lorentzian peaks and a piecewise cubic interpolated baseline. These synthetic data sets enable the use of exacting performance metrics because the true location and size of each peak is known a priori. By using the synthetic data sets,
metrics are developed that directly measure the ability of a spectral binning algorithm to create bins containing a single observed peak, while minimizing the probability of splitting peaks between bins. In addition to comparing spectral binning algorithms on synthetic data sets, this manuscript demonstrates the application of dynamic adaptive binning on empirical data from a $^1$H NMR-based experiment to monitor rat urinary metabolites after exposure to α-naphthylisothiocyanate (ANIT).

The following animal work was conducted by the Air Force Research Laboratory, Biosciences and Protection Division, Wright-Patterson AFB, and the subsequent NMR acquisition was conducted by Nicholas V. Reo’s laboratory at Wright State University, Cox Institute.

Animals were given a single administration, via oral gavage at 10 mL/kg, of ANIT in corn oil vehicle at one of the following doses: 20, 50, and 100 mg/kg. Control animals received corn oil only at 10 mL/kg. Sample size was 5-9 per group. All protocols for handling laboratory animals were approved by the Wright-Patterson Institutional Animal Care and Use Committee (IACUC) and meet appropriate Federal guidelines. Fisher 344 rats (approx. 250 g) were obtained from Charles Rivers Laboratory equipped with jugular vein catheters, and allowed to acclimate for seven days prior to the start of the study. Animals were then housed individually in metabolism cages and given \textit{ad libitum} access to food (Purina Certified Rat Chow # 5002) and water. The housing environment was maintained on a 12 hour light-dark cycle at 25 °C, and all animals were examined by Vivarium personnel twice daily to ensure their health and well-being. Urine was collected into containers chilled on dry ice and containing 1 ml of 1% sodium azide. All urine samples were stored at -40 °C prior to analysis by NMR spectroscopy.

Urine samples for NMR analysis were prepared as described by Robertson \textit{et al.} [122] and modified as follows [125]. Samples were thawed at 4 °C overnight then allowed to equilibrate to
room temperature just prior to NMR sample preparation. A 600 µL aliquot of urine was mixed with 300 µL of a phosphate buffer (0.2 M monosodium phosphate and 0.2 M disodium phosphate, pH 7.4) and allowed to equilibrate for ten minutes. Samples were then centrifuged at 5000 rpm (2300 rcf) for ten minutes to remove any particulate matter and 550 µL of supernatant was transferred to a 5mm NMR tube. An internal standard consisting of 150 µL of trimethysilylpropionic (2, 2, 3, 3 d₄) acid (TSP) dissolved in deuterium oxide was added at a final concentration of 2 mM.

Proton NMR spectra were acquired at 25 °C on a Varian INOVA operating at 600 MHz. Water suppression was achieved using the first increment of a NOESY pulse sequence, which incorporated saturating irradiation (on-resonance for water) during the relaxation delay (7.0 s total; 2 s with water presaturation) and the mixing time (50 ms total; 42 ms with water irradiation). Data were signal averaged over 64 transients using a 4.0 s acquisition time and interpulse delay of 11.05 s.

NMR spectral data were processed using Varian software and employing exponential multiplication (0.3 Hz line-broadening), Fourier transformation, and baseline flattening (fifth-order polynomial and spline fitting routines). The TSP signal was used as an internal chemical shift reference (set at 0.0 ppm), and the regions surrounding the residual water signal (≈4.8 ppm) and the urea signal (≈5.8 ppm) were excluded from the analyses. The vertical shift of the entire spectrum was adjusted such that the mean of the intensities between 11.6 and 10 ppm (a region containing spectral noise) was zero. Then the peak intensities of each spectrum were normalized to a constant sum.
7.2.2 Algorithm

Dynamic adaptive binning determines the optimal bin configuration of $n$ observed peaks as measured by an objective function. This process is divided into two steps: (1) determining the location of the observed peaks in each spectra and (2) finding the optimal bin boundaries with respect to the objective function. The identification of the observed peaks in each spectrum is accomplished by identifying local maxima after smoothing via a wavelet transform [36, 38-41]. After the observed peaks of each spectrum have been determined, the algorithm determines the optimal bin configuration using a dynamic programming strategy to efficiently find the best solution. These bin boundaries can then be used to quantify additional spectra.

7.2.2.1 Optimizing bin boundaries using dynamic programming

The complexity of identifying significantly responding metabolites (i.e., biomarkers) is increased when multiple peaks fall in the same bin. Ideally, each bin should contain a single peak from each spectrum representing the same metabolite. In $^1$H NMR spectra, a peak representative of a single type of proton in a molecule (i.e., methine, methyl, etc.) can sometimes be split into a multiplet (i.e., doublet, triplet, etc) due to J-coupling. Our approach does not attempt to address this issue, but rather, bin boundaries are selected to ideally contain only a single peak. The user, however, can interactively modify bin boundaries in an effort to combine signals that are identified as a J-coupled multiplet. The degree to which a bin approaches this ideal is approximated by counting the number of observed peaks within its boundaries for each spectrum. This can be quantified by a bin heuristic objection function ($BHOF$) that is calculated as follows:

$$BHOF(\alpha, \omega) = \sum_s |l - N_s|,$$

(33)
where \( N_s \) is the number of observed peaks in spectrum \( s \) for the region defined by the bin boundaries, \([\alpha, \omega]\). A \( BHOF \) value of 0 indicates that for the bin \([\alpha, \omega]\) each spectrum has one observed peak.

The bin heuristic objective function measures the fitness of an individual bin. For a set of bins, \( \beta \), a global heuristic objective function (\( GHOF \)) is calculated as follows:

\[
GHOF = \sum_{[\alpha, \omega] \in \beta} BHOF(\alpha, \omega),
\]

(34)

where \( \beta \) is the set of all bins and \([\alpha, \omega]\) are the boundaries of a bin. Thus, two or more sets of bin boundaries, \{\( \beta_1\), ..., \( \beta_n\)\}, can be ranked according to their \( GHOF \) scores. The \( GHOF \) score represents the cumulative score of the individual bins. The fitness of an individual bin is measured as the degree to which it conforms to the ideal that a bin contains one observed peak from each spectrum.

The \( GHOF \) score is a discrete function, where different sets of bin boundaries can yield the same \( GHOF \) score. To decide between these configurations, various tiebreaking heuristic objective functions (\( THOF \)) can be developed. One way to distinguish between these configurations is by the number of bins, where \( THOF = \|\beta\| \). The choice to maximize or minimize \( THOF \) will depend on the preference of the researcher, as increasing the number of bins increases the probability of peaks spanning bin boundaries; however, decreasing the number of bins increases the probability of two or more peaks residing in a single bin. For this manuscript, the \( THOF \) metric is maximized. If two bin configurations have an equal number of bins, then the configuration that maximizes the margins between adjacent bins is selected, where the margin between two adjacent bins is the minimum distance between their observed peaks. The average margin (\( AVGM \)) is defined as follows:
\[
AVGM = \frac{1}{\|\beta\|} \sum_{i=2}^{k} \text{margin}(i-1,i),
\]

where \( \text{margin}(i-1,i) \) is the margin between the previous and the \( i \)-th bin. Thus, the best binning solution is found by minimizing \( GHOF \) using \( THOF \) as a metric to distinguish between equivalent configurations.

In addition to these two objective functions (\( GHOF \) and \( THOF \)), unrealistic bin boundaries are avoided by including two additional parameters: the maximum distance between observed peaks in the same bin (\( W \)) and the minimum distance between an observed peak and a boundary (\( D \)). The first parameter is designed to avoid undesirably large bins, such as combining the entire spectrum in a single bin. The second parameter controls the minimum desired distance between observed peaks in adjacent bins, thus, preventing the algorithm from splitting peaks deemed too close by the user.

The optimal binning of \( n \) observed peaks is found via a dynamic programming strategy that minimizes \( GHOF \) with \( THOF \) breaking ties. Specifically, the optimal solution for binning \( n \) observed peaks is obtained by incorporating the \( n \)-th observed peak into the optimal solution for binning previous observed peaks (illustrated in Figure 21). Proofs verifying optimal substructure and overlapping subproblems can below. Formally, the recurrence relationship for incorporating the \( n \)-th observed peak is given in Table 5.

The bin boundaries are defined by the first and last observed peaks in the bin, which are known as the base observed peaks. The left-boundary is defined by finding the location of the minimum intensity of a maximum composite spectrum between the first observed peak in the bin and the previous observed peak. This procedure is repeated with the right-boundary using the last observed peak in the bin and the next observed peak. If this location results in an observed peak
to boundary distance below the user-defined threshold, then the boundary is centered between the adjacent observed peaks. The exceptions include the left-boundary of the first bin and the right-boundary of the last bin, which are set to a distance of half the maximum distance between observed peaks. Note that adjacent observed peaks closer than two times the threshold between observed peaks can be ignored because a valid boundary splitting the observed peaks does not exist.

**Table 5.** Recurrence relationship that recursively defines the optimal binning of \( n \) maxima.

<table>
<thead>
<tr>
<th>Description</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. ( n )-th maximum is a base for a new bin</td>
<td>( GHOF_n = GHOF_{n-1} + BHOF(\alpha_n, \omega_n) )</td>
</tr>
<tr>
<td>2. ( n )-th and ( n-i ) maxima are the bases for a new bin if the distance between the two maxima is less than the maximum bin width, where ( i ) is an integer greater than 0.</td>
<td>( GHOF_n = GHOF_{n-i-1} + BHOF(\alpha_{n-i}, \omega_n) )</td>
</tr>
</tbody>
</table>

**7.2.2.2 Dynamic programming strategy**

For a problem to lend itself to a dynamic programming strategy it must exhibit optimal substructure, contain overlapping subproblems, and the dynamic programming solution must outperform a greedy solution. Finding the optimal binning of \( n \) observed peaks reuses the optimal solution of binning \( n-i \) observed peaks (detailed in Table 5); therefore, the problem contains overlapping subproblems (example shown in Figure 23). The optimal solution is often reached by skipping a potential boundary, thus, a greedy approach is not guaranteed to find an optimal solution. An example of an optimal solution that is found by skipping a potential boundary is shown in Figure 22, where the last boundary in Figure 22(b) is skipped in the optimal solution (Figure 22(c)). For a problem to exhibit optimal substructure, an optimal solution to the problem comprises optimal solutions to subproblems. Thus, an optimal solution to binning \( n \) observed
peaks must comprise the optimal solution for binning subproblems. This is proved by contradiction:

**Optimal substructure proof:**

Let $GHOF_6$ be the optimal binning of the 6 observed peaks shown in Figure 22(c).

From Table 1, $GHOF_6 = GHOF_4 + BHOF(\alpha_5, \omega_6)$.

Assume $GHOF_4$ (Figure 22(a)) is not the optimal binning of the first 4 observed peaks, then there exists a $GHOF_4'$ that is less than $GHOF_4$.

Using $GHOF_4'$, $GHOF_6' = GHOF_4' + BHOF(\alpha_5, \omega_6)$, which is less than $GHOF_6$.

This is a contradiction because $GHOF_6$ is optimal, thus, $GHOF_4$ is the optimal binning of the first 4 observed peaks.

**General proof:** Let $GHOF_n$ be the optimal binning of the $n$ observed peaks.

From Table 5, $GHOF_n = GHOF_{n-i-1} + BHOF(\alpha_{n-i}, \omega_n), i \geq 0$.

Assume $GHOF_{n-i-1}$ is not the optimal binning of the first $n - i - 1$ observed peaks, then there exists a $GHOF_{n-i-1}'$ that is less than $GHOF_{n-i-1}$.

Using $GHOF_{n-i-1}'$, $GHOF_n' = GHOF_{n-i-1}' + BHOF(\alpha_{n-i}, \omega_n)$, which is less than $GHOF_n$.

This is a contradiction because $GHOF_n$ is optimal, thus, $GHOF_{n-i-1}$ is the optimal binning of the first $n-i-1$ observed peaks.
Figure 21. Process of binning 6 observed peaks from two spectra. (a) The optimal binning of the 1st observed peak. The configuration shown in (b) is unrealistic because the 2nd and 3rd observed peaks are too close to be in separate bins. Figures (c) and (d) both have identical GHOF scores for binning the first three observed peaks, but (c) would be preferred due to its higher THOF score. Figures (e) and (f) both show equivalent configurations for binning the first three observed peaks, but (f) is preferred due to its higher THOF score. Figure (g) is invalid due to the distance between the first and last observed peaks. The optimal binning of the first 5 observed peaks is shown in (h), and finally, the optimal binning of the first 6 observed peaks is shown in (j).
Figure 22. The subproblem of binning the first 4 observed peaks (a) is used to compute the binning of the first 5 observed peaks (b) and the first 6 observed peaks (c).

7.2.2.3 Selecting the parameters for identifying observed peaks

The procedure for determining the location of the observed peaks begins by smoothing each spectrum using a decimated wavelet transformation [36, 38-41]. A smooth spectrum is created by deconstructing each spectrum using a specific wavelet and then zeroing the wavelet coefficients that are designated as noise. There are several options to consider when smoothing via a wavelet transform, including the selection of the wavelet, the threshold selection rule, soft or hard threshold, and whether or not to include multiplicative threshold scaling.

The wavelets selected for evaluation in this study include the commonly used Haar, Daubechies, Symlets, and Coiflets wavelets [145]. Threshold selection rules based on either Stein’s unbiased risk (rigsure), a heuristic variant of Stein’s unbiased risk (heursure), a universal threshold (sqtwolog), or minimax thresholding (minimaxi) are evaluated. The value of using hard
or soft thresholding is also evaluated, along with the benefit of using multiplicative threshold rescaling. The threshold rescaling techniques evaluated include no rescaling (one), rescaling using a single estimation of level noise based on first level coefficients (sln), and rescaling done using level dependent estimation of noise (mln). Further, baseline variations and incorrectly smoothed regions can result in spurious observed peaks; therefore, only those observed peaks significantly above the noise of the spectrum are retained. The threshold is calculated as $n$ times the standard deviation of a region of noise. All permutations of the aforementioned wavelet parameters and $1 \leq n \leq 6$ are evaluated to determine the optimal wavelet configuration.

The performance of a wavelet smoothing technique is evaluated by comparing the observed peaks to the correct locations of those peaks. The correct locations are determined from the synthetic data set using spectra without noise. The optimal alignment between the computed and correct observed peaks is then calculated to provide the average distance ($AD$) from the correct observed peaks. The score for matching two observed peaks is equal to the absolute value of the distance between observed peaks, and the penalty for skipping an observed peak is defined as 0.01 ppm. After an optimal alignment is determined, the average distance between matched observed peaks is calculated as follows:

$$AD = \frac{1}{N_{matched}} \sum_{i} |cx_i - sx_i|,$$

where $N_{matched}$ is the number of observed peaks matched between the correct and smoothed spectra, and $cx_i$ and $sx_i$ are the locations of the $i$-th correct and smoothed observed peaks, respectively. In addition, the alignment provides the percentage of missed peaks ($PM$), and extra ($PE$) peaks are also computed:
\[ PM = \frac{N_{\text{missed}}}{N_{\text{correct}}} , \]  

(37)

\[ PE = \frac{N_{\text{extra}}}{N_{\text{smooth}}} , \]  

(38)

where \( N_{\text{correct}} \) is the number of correct observed peaks, \( N_{\text{missed}} \) is the number of correct observed peaks that are not matched to a smooth observed peaks, \( N_{\text{extra}} \) is the number of smooth observed peaks not assigned to a correct observed peaks, and \( N_{\text{smooth}} \) is the number of smooth observed peaks.

### 7.2.3 Evaluating and comparing binning algorithms

#### 7.2.3.1 Metrics applied to synthetic data sets

Synthetic data sets are employed to provide a statistical basis for comparing binning algorithms. For any statistical comparison on synthetic data to be useful, the synthetic data must accurately characterize the salient features of real data. The synthetic data used in this manuscript are based on urine \(^1\text{H}\) nuclear magnetic resonance data [119]. In total, 40 data sets each with 20 control and 20 treatment spectra were used to compare three binning algorithms: dynamic adaptive binning (DAB), uniform binning, adaptive binning (AB), and adaptive intelligent binning.

For the application considered here – identification of biomarkers of toxicity – the objective of a binning technique is to increase the effectiveness of biomarker identification. The result of such analysis is a set of bins that have been labeled as significantly responding (i.e., responsive). These responsive bins are then examined to determine which metabolites are reflected by each bin. The complexity of this analysis is increased when multiple observed peaks from a single spectrum reside in a single bin. Further, the closer an observed peak is to a boundary the more its effects
are distributed across adjacent bins, and the higher the probability that individual peaks will span bin boundaries.

As spectra contribute more than one peak to a bin, the more difficult it becomes to interpret the results. The ability of a binning technique to achieve this ideal is measured by penalizing each extra or missing observed peak in a bin. This metric is called the normalized number of observed peaks per bin ($NNP_s$). While similar in calculation to the $BHOF$ score described in the methods, the $NNP_s$ metric is calculated using clean synthetic spectra from which noise has been removed. Thus, the exact locations of the observed peaks are known when calculating $NNP_s$. These are not known to the dynamic adaptive binning technique. Further, to fairly compare algorithms, the set of bins included in the $NNP_s$ metric is limited to those bins containing at least one observed peak (i.e., empty bins are excluded). The normalized number of observed peaks per bin is defined as follows:

$$NNP_s(\alpha, \omega) = \left| \frac{1}{s} \sum_{i=1}^{s} N_i \right|,$$

(39)

where $N_i$ is the number of observed peaks in the clean spectrum $s$ for the region defined by bin $[\alpha, \omega]$. A $NNP_s$ value of 0 indicates that for the bin $[\alpha, \omega]$ in spectrum $s$ contains one observed peak. When calculating the number of observed peaks, the bins are restricted to those containing at least one observed peak.

In addition to measuring the number of observed peaks per bin, the probability of peaks spanning boundaries must be considered when evaluating binning algorithms. The probability of peaks spanning boundaries is approximated by calculating the distance from each observed peak to the nearest boundary ($DPB$):
\[ DPB_i = |\omega_i - p_i|, \]  

where \( p_i \) is the location of the \( i \)-th observed peak, and \( \omega_i \) is the location of the nearest boundary. Finally, the time complexity of an algorithm is of practical importance. To measure this, the CPU seconds the algorithm spent in user mode is studied for all 40 data sets.

### 7.2.3.2 Comparing binning algorithms on empirical \( ^1H \) data set

The dynamic adaptive binning method is compared to uniform, adaptive, and adaptive intelligent binning on its ability to analyze a \( ^1H \) toxicology data set. The motivation of an adaptive binning technique is demonstrated on two sample regions of spectroscopic data. Further, the results after principal component analysis (PCA) – a common unsupervised latent vector visualization technique – are analyzed for each of the aforementioned binning algorithms. The ability of a binning technique to enhance the results of PCA by improving between group separation and within group scatter is illustrated by the PCA scores plots. The parameters for each algorithm are selected from the results of the normalized number of peaks per bin.

### 7.3 Results

#### 7.3.1 Peak identification via wavelet smoothing

All combinations of the wavelet parameters and techniques previously described were evaluated and ranked according to the average of the percentage of peaks missed and the percentage of extra peaks (AVG). The top 10 peak identification configurations are shown in Table 6. The most accurate configuration of wavelet parameters, as measured by the average of \( PE \) and \( PM \), is wavelet: sym7, thresholding: soft, rescaling: rigrsure, level: 1, rescaling: one, and number of noise standard deviations: 5. While not significantly different from the other top wavelet
configurations, this configuration is assumed for all future analyses. In practice, any of the top configurations would produce similar results.

Table 6. Top 10 peak identification wavelet parameter configurations, ranked according to the average (AVG) of percentage of peaks missed (PM) and percentage of extra peaks (PE).

<table>
<thead>
<tr>
<th>Wavelet</th>
<th>Thres.</th>
<th>Selection Technique</th>
<th>L</th>
<th>Rescaling</th>
<th>#</th>
<th>PM (µ, σ)</th>
<th>PE (µ, σ)</th>
<th>AVG</th>
</tr>
</thead>
<tbody>
<tr>
<td>sym7</td>
<td>s</td>
<td>rigsure</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym7</td>
<td>h</td>
<td>rigsure</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym7</td>
<td>s</td>
<td>heursure</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym7</td>
<td>h</td>
<td>heursure</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym7</td>
<td>s</td>
<td>sqtwolog</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym7</td>
<td>h</td>
<td>sqtwolog</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym7</td>
<td>s</td>
<td>minimaxi</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym7</td>
<td>h</td>
<td>minimaxi</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.485</td>
<td>1.611</td>
<td>8.224 1.048 6.354</td>
</tr>
<tr>
<td>sym8</td>
<td>s</td>
<td>rigsure</td>
<td>1</td>
<td>one</td>
<td>5</td>
<td>4.749</td>
<td>1.637</td>
<td>8.125 0.921 6.437</td>
</tr>
</tbody>
</table>

7.3.2 Evaluating and comparing binning techniques

Each spectral binning algorithm is analyzed as a function of their tunable parameters. The process of uniform binning is measured as a function of the bin width. For $^1$H NMR spectra, a standard bin width is 0.04 ppm. For this analysis, the bin width is varied from 0.01 to 0.06 ppm by 0.01 ppm. The performance of adaptive binning (AB) was measured as a function of the level of the wavelet transform, which is varied between 1 and 6 for adaptive binning. For adaptive intelligent binning (AIB), the parameter R is set to 0.15, 0.5, and 0.85. For dynamic adaptive binning (DAB), the maximum bin width is set to 0.04 ppm and the minimum distance from boundary to observed peak is varied from 0 to 0.004 by increments of 0.001. The wavelet parameters for DAB were selected as the best results from Table 6. For a statistical comparison, the algorithms were ranked according to their performance on the four metrics recorded for the synthetic data sets described in 7.2.2.3.
A detailed comparison of the four binning algorithms based on the metrics previously described was carried out using several standard statistical tests. First each of the metrics was tested for normality using the Anderson-Darling test, which rejected normality for the number of observed peaks per bin and the distance from observed peak to nearest boundary \((\alpha = 0.05)\). The test failed to reject normality for the other two metrics, and thus, they are assumed to come from a normal distribution. Further, the Levene test showed that the variance of the CPU seconds exhibited heterogeneity \((\alpha = 0.05)\).

Using Welch’s variance-weighted one-way ANOVA, the mean the CPU seconds per data set was significant using an alpha of 0.05. The Games-Howell multiple comparison test was used to determine significant differences between algorithms. Algorithm and parameter pairs were sorted according to their mean rank using the Kruskal-Wallis test with multiple comparisons. These results are summarized in Table 7, where significantly different results are shown with different adjacent shading. To simplify the table, the mean rank was normalized by dividing each rank by the minimum mean rank.

Using the Kruskal-Wallis test (nonparametric one-way ANOVA), the mean ranks of both the distance from observed peak to nearest boundary and the normalized number of observed peaks per bin were significant using an alpha of 0.05. A multiple comparison test on the mean ranks (analogous to the Tukey-Kramer method) was used to determine any significant differences between algorithms. Algorithm and parameter pairs were ranked according to these tests and are summarized in Table 8, where algorithms with differing ranks are significantly different.
Table 7. The performance of the binning algorithms and their parameters as measured by the mean CPU seconds/data set relative to the fastest algorithm. Algorithm and parameter pairs were sorted according to the multiple comparison tests on the mean ranks. Significantly different results are shown with different adjacent shading.

<table>
<thead>
<tr>
<th>Method &amp; Parameters</th>
<th>Mean CPU seconds relative to fastest algorithm</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>AB, 5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>AB, 3</td>
<td>1.05</td>
<td>1.25</td>
</tr>
<tr>
<td>AB, 4</td>
<td>1.05</td>
<td>1.39</td>
</tr>
<tr>
<td>AB, 6</td>
<td>1.05</td>
<td>1.47</td>
</tr>
<tr>
<td>AB, 2</td>
<td>1.12</td>
<td>2.60</td>
</tr>
<tr>
<td>AB, 1</td>
<td>1.29</td>
<td>3.39</td>
</tr>
<tr>
<td>Uniform, 0.06</td>
<td>1.82</td>
<td>3.98</td>
</tr>
<tr>
<td>Uniform, 0.05</td>
<td>2.18</td>
<td>4.61</td>
</tr>
<tr>
<td>Uniform, 0.04</td>
<td>2.71</td>
<td>5.22</td>
</tr>
<tr>
<td>Uniform, 0.03</td>
<td>3.65</td>
<td>5.84</td>
</tr>
<tr>
<td>Uniform, 0.02</td>
<td>5.41</td>
<td>6.45</td>
</tr>
<tr>
<td>Uniform, 0.01</td>
<td>10.88</td>
<td>7.06</td>
</tr>
<tr>
<td>DAB, 0.004</td>
<td>473.12</td>
<td>8.33</td>
</tr>
<tr>
<td>DAB, 0.003</td>
<td>473.43</td>
<td>8.55</td>
</tr>
<tr>
<td>DAB, 0.002</td>
<td>474.71</td>
<td>9.39</td>
</tr>
<tr>
<td>DAB, 0.001</td>
<td>512.94</td>
<td>10.13</td>
</tr>
<tr>
<td>DAB, 0</td>
<td>9406.47</td>
<td>10.75</td>
</tr>
<tr>
<td>AlB, 0.15</td>
<td>42841.35</td>
<td>12.04</td>
</tr>
<tr>
<td>AlB, 0.5</td>
<td>47335.35</td>
<td>12.07</td>
</tr>
<tr>
<td>AlB, 0.85</td>
<td>50460.65</td>
<td>12.09</td>
</tr>
</tbody>
</table>
Table 8. The performance of the binning algorithms and their parameters as measured by (A) normalized number of observed peaks per bin, and (B) median distance from observed peak to nearest boundary. Algorithm and parameter pairs were sorted according to the multiple comparison tests on the mean ranks. Significantly different results are shown with different adjacent shading.

<table>
<thead>
<tr>
<th>Method &amp; Parameters</th>
<th>Median normalized # of observed peaks/bin</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB, 0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Uniform, 0.01</td>
<td>1</td>
<td>1.96</td>
</tr>
<tr>
<td>Uniform, 0.02</td>
<td>1</td>
<td>2.09</td>
</tr>
<tr>
<td>DAB, 0.001</td>
<td>1</td>
<td>2.09</td>
</tr>
<tr>
<td>AB, 3</td>
<td>1</td>
<td>2.21</td>
</tr>
<tr>
<td>AB, 4</td>
<td>1</td>
<td>2.21</td>
</tr>
<tr>
<td>AB, 5</td>
<td>1</td>
<td>2.21</td>
</tr>
<tr>
<td>AB, 6</td>
<td>1</td>
<td>2.24</td>
</tr>
<tr>
<td>AIB, 0.15</td>
<td>0</td>
<td>2.24</td>
</tr>
<tr>
<td>AIB, 0.5</td>
<td>1</td>
<td>2.26</td>
</tr>
<tr>
<td>AIB, 0.85</td>
<td>1</td>
<td>2.26</td>
</tr>
<tr>
<td>AB, 2</td>
<td>1</td>
<td>2.26</td>
</tr>
<tr>
<td>AB, 1</td>
<td>1</td>
<td>2.26</td>
</tr>
<tr>
<td>DAB, 0.002</td>
<td>1</td>
<td>2.72</td>
</tr>
<tr>
<td>Uniform, 0.03</td>
<td>3</td>
<td>2.91</td>
</tr>
<tr>
<td>DAB, 0.003</td>
<td>2</td>
<td>2.98</td>
</tr>
<tr>
<td>DAB, 0.004</td>
<td>2</td>
<td>3.01</td>
</tr>
<tr>
<td>Uniform, 0.04</td>
<td>4</td>
<td>3.39</td>
</tr>
<tr>
<td>Uniform, 0.05</td>
<td>5</td>
<td>3.86</td>
</tr>
<tr>
<td>Uniform, 0.06</td>
<td>7</td>
<td>4.07</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method &amp; Parameters</th>
<th>Median dist. from observed peak to nearest boundary</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAB, 0.004</td>
<td>0.48916</td>
<td>4.25</td>
</tr>
<tr>
<td>DAB, 0.003</td>
<td>0.16632</td>
<td>3.91</td>
</tr>
<tr>
<td>DAB, 0.002</td>
<td>0.05086</td>
<td>3.73</td>
</tr>
<tr>
<td>Uniform, 0.06</td>
<td>0.01507</td>
<td>3.19</td>
</tr>
<tr>
<td>Uniform, 0.05</td>
<td>0.01249</td>
<td>3.06</td>
</tr>
<tr>
<td>Uniform, 0.04</td>
<td>0.01001</td>
<td>2.89</td>
</tr>
<tr>
<td>DAB, 0.001</td>
<td>0.00584</td>
<td>2.67</td>
</tr>
<tr>
<td>Uniform, 0.03</td>
<td>0.00738</td>
<td>2.66</td>
</tr>
<tr>
<td>AIB, 0.85</td>
<td>0.00459</td>
<td>2.39</td>
</tr>
<tr>
<td>Uniform, 0.02</td>
<td>0.00499</td>
<td>2.31</td>
</tr>
<tr>
<td>AIB, 0.5</td>
<td>0.00250</td>
<td>1.85</td>
</tr>
<tr>
<td>Uniform, 0.01</td>
<td>0.00252</td>
<td>1.70</td>
</tr>
<tr>
<td>DAB, 0</td>
<td>0.00208</td>
<td>1.58</td>
</tr>
<tr>
<td>AIB, 0.15</td>
<td>0.00167</td>
<td>1.52</td>
</tr>
<tr>
<td>AB, 6</td>
<td>0.00083</td>
<td>1.02</td>
</tr>
<tr>
<td>AB, 5</td>
<td>0.00083</td>
<td>1.02</td>
</tr>
<tr>
<td>AB, 4</td>
<td>0.00083</td>
<td>1.02</td>
</tr>
<tr>
<td>AB, 2</td>
<td>0.00083</td>
<td>1.02</td>
</tr>
<tr>
<td>AB, 3</td>
<td>0.00083</td>
<td>1.01</td>
</tr>
<tr>
<td>AB, 1</td>
<td>0.00083</td>
<td>1</td>
</tr>
</tbody>
</table>

A. | B.
7.3.3 Empirical $^1$H NMR data set

The dynamic adaptive binning method and the traditional uniform binning method were applied to a $^1$H NMR-based experiment to monitor rat urinary metabolites. The parameters were selected based on their ability to minimize the normalized number of observed peaks per bin (Table 8b). Figure 23 illustrates the motivation of dynamic binning techniques on two sample regions using a uniform bin width of 0.01 ppm and dynamic adaptive binning parameters $D = 0$ and $W = 0.04$ ppm. The results of principal component analysis on samples from several non-lethal doses of ANIT (20 mg/kg, 50 mg/kg, and 100 mg/kg) are shown in Figure 24.

7.4 Discussion

Spectroscopic binning algorithms attempt to enhance the effectiveness of pattern recognition techniques by reducing problem dimensionality with minimal loss of information. One application of a binning algorithm is the determination of biomarkers associated with toxic exposure. The complexity of this analysis increases when multiple peaks fall in the same bin and span bin boundaries, both of which result in a loss of information. Thus, a binning algorithm attempts to minimize the number of dimensions, while maximizing pertinent information and mitigating peak misalignment. The results of this analysis are then analyzed post-hoc to determine the specific metabolites contributing to an individual bin.

The performance of binning algorithms was measured by the minimum distance from observed peak to boundary ($DPB$) and the normalized number of observed peaks per bin ($NNP$) (Table 8). In summary, the dynamic adaptive binning algorithm ($D = 0$) has a significantly better mean rank of normalized number of observed peaks per bin than all other algorithm and parameter pairs. It should also be noted that while the median normalized number of observed peaks per bin for AIB ($R = 0.15$) is 0, its mean rank is used in the multiple comparison test. When
increasing the user-defined parameter, minimum distance from observed peak to boundary \((D = 0, 0.001, 0.002, 0.003, \text{ and } 0.004)\), the probability of peaks spanning bin boundaries decreases, but the normalized number of observed peaks per bin also increases. This increases the complexity of determining the metabolites reflected by each bin. Using a minimum distance from boundary to observed peak of 0.001 ppm balances these two goals.

Secondary to the performance of the algorithms as measured by \(DPB\) and \(NNP\), the computational complexity as measured by the CPU seconds spent in user mode is also important. In summary, all methods required significantly less CPU seconds than adaptive intelligent binning \((R = 0.15, 0.5, \text{ and } 0.85)\). In addition, the CPU seconds required by adaptive intelligent binning \((D = 0)\), were significantly greater than uniform binning, adaptive binning, and dynamic adaptive binning \((D = 0.001, 0.002, 0.003, 0.004)\).

While the advantages of dynamic adaptive binning are quantified using the synthetic spectral data sets, the performance on an experimental \(^1\text{H}\) data set is illustrated in Figure 23. Uniform binning successfully mitigates misalignment when peaks fall in the center of the bin; however, it creates boundaries at fixed intervals, regardless of the environment. This can lead to peaks spanning adjacent bins, as shown in Figure 23. The probability of a peak spanning bin boundaries decreases as the bin width increases; however, this also increases the probability of multiple peaks residing in a single bin. In general, uniform binning lacks the flexibility to deal with the complexities of a \(^1\text{H}\) NMR spectrum.

The ability of each binning technique to enhance subsequent pattern recognition techniques by improving within and between group scatter is demonstrated by analyzing the PCA results on the \(^1\text{H}\) toxicology data set (Figure 24). This scores plot shows that dynamic adaptive binning provides equivalent or better separation measured by the Euclidian distance between means for
each ANIT dose (20 mg/kg, 50 mg/kg, and 100 mg/kg) throughout the time course (day-1, day-2,
day-3, and day-4). A specific example of this improvement is the increase in separation between
100 mg/kg ANIT samples at day-3 (d3) and 50 mg/kg ANIT samples at day-2 (d2). The
separation between these two groups is 28, 4, 4, and 33 for dynamic adaptive binning, uniform
binning, adaptive binning, and adaptive intelligent binning, respectively. Examining this
separation shows the advantages of dynamic adaptive binning and adaptive intelligent binning
versus uniform and adaptive binning. Further, by examining the separation between the 50 mg/kg
ANIT samples at day-2 and day-3 for the binning algorithms shows that dynamic adaptive
binning, uniform binning, and adaptive binning provide a 2-fold increase in group separation
versus adaptive intelligent binning. The exact distances are 42, 39, 42, and 17, respectively.

In comparison to dynamic adaptive binning, adaptive binning and adaptive intelligent binning
have fewer user defined parameters. In addition, these algorithms avoid the problem of
determining the location of observed peaks; however, finding the locations of the observed peaks
has several advantages, including the ability for the user to filter the observed peaks of interest
(i.e., based on height). Using the observed peaks also provides the user with domain specific
parameters, such as minimum distance from observed peak to the nearest boundary. Finally, the
inclusion of observed peaks will facilitate the development of more sophisticated objective
functions that can improve quantification by identifying multiplets and assisting in further
deconvolution. Specifically, the identified peaks may be supplied as input to a targeted approach
that removes metabolites identified with high confidence. The updated spectra could then be
processed by a binning approach. The software was written in MATLAB and is available for
Figure 23. Sample regions of $^1$H spectroscopic data demonstrating the advantages of dynamic adaptive binning (A) over uniform binning (B).
Figure 24. Principal component scores (means and standard error) after dynamic adaptive binning (A), uniform binning (B), adaptive binning (C), and adaptive uniform binning (D) for several non-lethal ANIT doses (control: black; 20 mg/kg ANIT: blue; 50 mg/kg ANIT: red; and 100 mg/kg ANIT: magenta). Each dose is measured as a function of time (e.g., d2 is 2 days post-dose). See text for details.
8 Conclusion

8.1 Contributions

The three main chapters of this dissertation have resulted in three journal publications (1 currently in submission), two refereed conference papers published at the IEEE Bioinformatics and Bioengineering conference (BIBE05 and BIBE07) [146, 147], and several conference poster and platform presentations [146-148]. Specifically, a novel kernel-based quantification technique, Gaussian binning, has been published in the journal *Metabolomics* [143]. This manuscript demonstrated the ability of Gaussian binning to counter dramatic quantitative errors in uniform binning. In addition, the characterization and generation of NMR spectroscopic validation sets has been published in the journal *Bioinformatics*. This research provides the first attempt to create a standard set of synthetic validation sets for the NMR community. Finally, an improved dynamic binning technique, dynamic adaptive binning, has been submitted to the journal *Metabolomics*. The research in this dissertation demonstrates its ability to improve spectral quantification over alternative dynamic binning techniques, including adaptive binning and adaptive intelligent binning. Additionally, several additional journal manuscripts are in preparation as part of the ongoing collaboration with the Dr. Reo’s lab. These manuscripts utilize the novel computational techniques described herein through a research portal developed throughout the process of this dissertation ([http://birg.cs.wright.edu/omics_analysis](http://birg.cs.wright.edu/omics_analysis)). In conclusion, the techniques described in this dissertation have a direct impact on the NMR metabolomics community.

8.2 Future work

The characterization and generation of synthetic validation sets can be extended to additional $^1$H $^{13}$C spectroscopic data sets. In addition, the salient distributions can be combined with freely
available metabolite databases (e.g., Human Metabolome Project and NMR Shift DB) to provide synthetic validation sets for targeted quantification approaches.

The dynamic adaptive binning approach to quantification can be incorporated as the first step to a local deconvolution algorithm. The dynamically sized and placed boundaries can serve as a guide for localizing the deconvolution process. Specifically, the affects from adjacent bins can be removed by fitting these regions to Lorentzian peaks using the techniques described in the characterization of NMR spectra. Removing the influences of adjacent peaks will greatly enhance the specificity and sensitivity of biomarker identification. Further, by dividing the deconvolution process in this manner, the problem can be solved in a parallel manner using a MapReduce paradigm.

While localized deconvolution can remove influences from adjacent peaks, the researcher will ultimately want to identify the specific metabolites of interest. Removing the adjacent effects will enhance a targeted approach by facilitating peak identification in congested regions of the spectrum. Currently, the identification process requires significant human interaction and time consuming algorithms. The parallel nature of peak identification can be exploited by a MapReduce or alternative paradigm. Further, identifying specific metabolites can be done in an iterative fashion, where identified metabolites are removed from the spectrum, thus, removing congestion. In practice, there will remain regions of the spectrum that are not automatically identifiable. These regions still contain useful information, and therefore, a future research area is the integration of a targeted and global approach to quantification that will enable robust high-throughput spectral quantification.
9 References


[104]


[77] ACD/Laboratories, "Intelligent bucketing."


