Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Anna Jolly Blackstock Date

Techniques for Pattern Recognition in High-Throughput Metabolomic Data

By

Anna Jolly Blackstock

Doctor of Philosophy

Biostatistics

Amita K. Manatunga, Ph.D.
Advisor

Tianwei Yu, Ph.D.
Advisor

F. DuBois Bowman, Ph.D.
Committee Member

Dean P. Jones, Ph.D.
Committee Member

Accepted:

Lisa A. Tedesco, Ph.D.
Dean of the James T. Laney School of Graduate Studies

Date

Techniques for Pattern Recognition in High-Throughput Metabolomic Data

By

Anna Jolly Blackstock

B.S., Vanderbilt University, 2002
M.S., Emory University, 2010

Advisors:
Amita K. Manatunga, Ph.D.
Tianwei Yu, Ph.D.

An Abstract of
A dissertation submitted to the Faculty of the Graduate School of Emory University
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy
in Biostatistics
2011

Abstract

Techniques for Pattern Recognition in High-Throughput Metabolomic Data

By Anna Jolly Blackstock

Nuclear magnetic resonance spectroscopy (NMR) and mass spectrometry (MS) have been introduced to studies of metabolite composition of biological fluids and tissues, providing information relating to disease states and other conditions. The collection of data from many types of biological samples is both simple and inexpensive, and most of these substances require very little pre-processing. However, the best methods for processing and analyzing complex metabolomic data are still being sought. We present NMR data pre-processing methods appropriate for metabolomic studies and techniques for recognizing patterns of metabolite levels in time course and cross-sectional data.

First, methods for overcoming issues that complicate metabolomic data are discussed, with a focus on NMR data. While NMR spectroscopy and MS provide a wealth of information about the collection of metabolites found in biological substances, the resulting data are extremely complex. Small changes in conditions can lead to significant shifts in NMR spectra, making it critical that metabolomic data be appropriately processed before analysis. NMR data are used to demonstrate existing pre-processing methods and to introduce a set of techniques appropriate for studies aiming to characterize the behavior of individual features.

Next, the possible periodic behavior of NMR features is assessed using time course data. Metabolite levels change throughout the day, and the identification of features with sinusoidal periodic behavior is of interest. Periodic regression is used to obtain estimates of the parameter corresponding to period for individual NMR features. A mixture model is then used to develop clusters of peaks, taking into account the variability of the regression parameter estimates. Methods are applied to NMR data collected from human blood plasma over a 24-hour period, and simulation results are presented.

Lastly, we present a method for investigating age-related changes in metabolite levels using MS data. Metabolite levels associated with some biological processes are thought to experience shifts at different times in life. An algorithm is used to identify metabolite level changes in blood plasma collected from marmosets of different ages. Clusters of breakpoints with similar locations of metabolite level shifts are determined, and metabolites corresponding to MS features with breakpoints are identified.

Techniques for Pattern Recognition in High-Throughput Metabolomic Data

By

Anna Jolly Blackstock

B.S., Vanderbilt University, 2002

M.S., Emory University, 2010

Advisors:

Amita K. Manatunga, Ph.D.

Tianwei Yu, Ph.D.

A dissertation submitted to the Faculty of the Graduate School of Emory University

in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in Biostatistics

2011

Acknowledgments

I owe thanks to many people for their support in the completion of this dissertation. First, I must thank my advisors in the Department of Biostatistics and Bioinformatics, Amita Manatunga, Ph.D., and Tianwei Yu, Ph.D. Without their ideas and feedback throughout this process, this research would not have been possible. Dr. Manatunga guided me since I began working on this project long ago, and the addition of Dr. Yu as a second advisor was her idea and proved to be a very wise one. He helped me tremendously and was available for the many questions I had along the way, both large and small. I would also like to thank my committee members DuBois Bowman, Ph.D., of the Department of Biostatistics and Bioinformatics and Dean Jones, Ph.D., of the Department of Medicine. The expertise of Dr. Jones was very helpful in formulating plans and interpreting results, and Dr. Bowman provided a different and useful perspective on the work.

In addition to my committee, several people should be recognized for sharing helpful information and data that made this project possible. Youngja Park, Ph.D., and Quinlyn Soltow, Ph.D., of the Department of Medicine exhibited great patience in explaining NMR and MS technologies and the interpretation of the data these analytical techniques produce. I would also like to thank Keith Mansfield, D.V.M., and Lynn Wachtman, D.V.M., of the New England Primate Research Center, Harvard Medical School, for the contribution of the Marmoset Healthy Aging LC-FTMS Data and Brian Schmotzer, formerly of the Department of Biostatistics and Bioinformatics, for his assistance with R and for sharing helpful R code.

Especially in the early years of my career at Emory, I received a great deal of support from my colleagues. There were many extensive study sessions, particularly in the

second year as we tackled intense theory classes, and the support of my classmates was very important to me. Years from now, I know that we will remain professional colleagues as well as friends. I would also like to thank the professors who shared their time and knowledge with me over the years, as well as the Department of Biostatistics and Bioinformatics for providing the type of environment that made me feel like I was part of a team.

Lastly, I would like to thank my friends and family. My parents provided me with support from day one, literally, and allowed me to pursue my academic interests without additional stressors that I know others must endure. I have received support from all of my friends and family, and although many of them still do not have any idea what I do, they always encouraged me and made me feel like I could and would be successful. My biggest thanks must go to my husband, Lindsey Blackstock. I didn't know when we met in fourth grade that he would have such a huge impact on my life, but here we are! Since I began at Emory, Lindsey has moved to Atlanta (with my help, the week before my second-year qualifying exams), and we have gotten married. Now we are very excited to be starting a family together. I am so fortunate to have a true partner in life, one who takes care of me after back injuries, cooks and cleans when I am busy with my dissertation, and lightens the mood when I get a little too uptight. I am where I am and who I am today thanks to him.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

In the last decade, the relationship between molecular science and medicine has been strengthened by the Human Genome Project and other advancements (Fernald et al., 2011), encouraging hope in the field of personalized medicine. However, using full genomic sequences is not common in medicine years later, and personalized medicine, while used in some cases, has not progressed as people had hoped. One reason that genomic information is not used more in practice is that genes alone do not determine phenotype, as the environment plays a large role as well (Soltow et al., 2010). Instead of focusing solely on the genome, there is increasing interest in studying the metabolome.

Metabolites are small cellular molecules that participate in or are produced by metabolic processes and can include amino acids, vitamins, environmental toxins, and pharmaceutical components, to name a few. When taken together, these small molecules form the metabolome (Nicholson et al., 1999; Beckonert et al., 2007). Metabolites can be found in tissue extracts and biofluids such as blood plasma, urine, and digestive

fluids, many of which are very easily and inexpensively collected. Studies of the metabolome often employ a systems approach and can be used to compare disease classes, to measure toxicity, to analyze digestive properties, and to investigate many biological processes. It is hoped that information collected from genomic, proteomic, and metabolomic studies can together produce a more complete picture of human biological processes. Since the metabolome is furthest downstream from genetic information, metabolomic studies could produce results that are affected by both genetic information and the environment (Soltow et al., 2010).

Different studies concerned with metabolites may have different goals and thus may be considered metabolic profiling, metabolomic analysis, or metabonomic analysis. In metabolic profiling, individual metabolites are pre-specified, and the goal is to identify as well as quantify these substances from a biological material (Fiehn, 2002). Metabolomics and metabonomics, on the other hand, address the entire metabolome. Although referring to the same type of data and often used interchangeably, metabolomic and metabonomic studies were originally characterized differently. Metabonomics has been described as "the application of data-rich analytical chemistry technology to study the multi-component metabolic composition of biofluids"(Bollard et al., 2005b) and was concerned with changes in the metabolome over time or comparisons of data collected under different conditions. Metabolomic studies, however, concentrated on describing the metabolome (Bollard et al., 2005b). Since these terms are now considered equivalent, we will use the more prevalent "metabolomics." Data used in these types of studies commonly result from mass spectrometry (MS) and nuclear magnetic resonance (NMR) spectroscopy.

In this chapter, I will give an introduction to NMR spectroscopy and MS technologies. A description of the numerous challenges presented by metabolomic data will be provided, along with applications and statistical techniques commonly used for this type of study. Two datasets will be introduced, one with information collected from

human blood plasma of individuals over the course of a day, and the other containing information from blood plasma from marmosets of different ages. These are the motivating datasets for the next three chapters. In Chapter 2, methods for the extensive processing necessary before any metabolomic study can be carried out will be detailed. Also, specific steps used to pre-process the human blood plasma dataset before it is analyzed will be developed. This dataset is analyzed in Chapter 3, where the periodic behavior of individual NMR peaks is of interest. A method for detecting clusters of peaks with similar periodic behavior will be presented. In Chapter 4, a study aiming to detect ages at which system-level shifts in metabolite levels occur will be detailed.

1.2 Nuclear Magnetic Resonance (NMR) Spectroscopy and Mass Spectrometry (MS)

Nuclear Magnetic Resonance (NMR) Spectroscopy and Mass Spectrometry (MS) can be used to analyze the same substances and have complementary results (Beckonert et al., 2007). MS measures mass-to-charge ratios (m/z) of ions, or charged particles, while NMR measures the effective field strength of nuclei (McMurry, 1996). Mass spectrometers measure m/z in various ways, depending on the type. In time of flight (TOF) spectrometers, particles are accelerated using an electric field. The velocity of an ion and thus the time it takes it to travel through a flight tube corresponds to its mass-to-charge ratio. Fourier Transform (FT)-based machines use the frequencies of the orbits of molecules to determine their mass-to-charge ratios (Eckel-Passow et al., 2009). Since the number of charges on each ion is usually one, m/z is generally equal to mass. The intensity values for MS data represent the number of ions with a particular m/z value.

If MS is used in metabolomic studies, metabolites are first separated using gas chromatography (GC) or liquid chromatography (LC). This step involves dissolving the substance of interest in a solvent (mobile phase) and passing this solution over an adsorbent material (stationary phase). Different components of the substance adsorb to the stationary phase to varying degrees and thus travel through this phase at different rates, emerging at "retention times" inversely proportional to these rates (McMurry, 1996). Non-polar molecules generally elute faster than polar molecules. This technique succeeds in separating metabolites from other substances found in biological samples (Dunn et al., 2011). While results of MS are commonly presented in a graph with mass-to-charge values on the horizontal axis and intensity on the vertical axis, the retention time is also available when LC-MS is used. These retention times provide additional information and can be used to produce more accurate data in terms of feature detection, quantification, and alignment (Yu et al., 2009).

The other analytical technique we focus on, ^{1}H NMR spectroscopy, is related to the magnetic field strength of hydrogen nuclei in organic compounds. These nuclei act as if spinning on an axis, and in the absence of an external magnetic field, these spins are oriented randomly. In NMR spectroscopy, substances are placed between poles of a strong magnet, causing nuclei to either align their magnetic fields with the magnet's field (parallel orientation) or away from it (antiparallel orientation). These oriented nuclei are disturbed by an electro-magnetic pulse, usually a radio frequency pulse. This is done until the nuclei with parallel orientation, which is the lower energy state, flip to the higher energy state, or antiparallel orientation (McMurry, 1996). When this happens, they are said to be in "resonance" with the applied energy. This will occur at different frequencies depending on the strength of the magnetic field as well as on the properties of the nuclei themselves, allowing one to distinguish between different types of hydrogens present in a compound and leading to identification of a substance. NMR results are often presented in a plot with intensity on the vertical

axis and chemical shift on the horizontal axis, where chemical shift is a measure of the effective field strength of a nucleus. The chemical shift of a peak is given in parts per million (ppm), corresponding to one millionth of the spectrometer operating frequency (McMurry, 1996). The area under each peak corresponds to the number of ^{1}H protons with the same chemical shift and thus is an indication of the concentration of a substance in a sample.

1.3 Challenges of Metabolomic Data

The analysis of metabolomic data is complicated by many factors. Data resulting from MS and NMR spectroscopy are very complex, with a single spectrum consisting of thousands, or even millions in the case of high resolution MS, of data points. Needless to say, the number of data points will almost surely be greater than the number of subjects in a study, making traditional statistical analysis of complete spectra impossible. Another factor complicating NMR data is that a single substance will often be represented by multiple groups of peaks with different chemical shifts in NMR spectra. This does not make the identification of substances impossible, but it does make the process more complex and could affect the manner in which NMR data should be analyzed.

Unwanted variation introduced by a myriad of sources for both mass and NMR spectra further complicates metabolomic data. Differences in temperature, salt content, and pH level in samples can lead to chemical shift differences in spectra, even though the same molecules are represented (Alm et al., 2009). The electric field used in some FT-based mass spectrometers causes frequency shifts and thus affects the m/z values for molecules (Eckel-Passow et al., 2009). Although technicians try to keep variation due to controllable conditions to a minimum by regulating sample conditions

and machine settings, some variation is inevitable. This makes it impossible to compare chemical shift or mass-to-charge ratio across spectra without first making adjustments. Many processing techniques are being used, but there is no standard methodology as methods used are dependent upon the nature of a study.

1.4 Analysis of Metabolomic Data

Metabolomic studies have been conducted with many different goals in mind. Data collected in such studies have been analyzed in various manners as well. There are, however, some common themes in terms of both applications of such data and techniques for analysis.

1.4.1 Applications of Metabolomic Methodology

Metabolomic studies can be used in many different types of investigations. Studies can focus on topics such as the characterization of metabolic processes, the toxic effects of drugs or drug interactions, the effects of injuries or interventions such as surgeries and diets, or the detection of markers for various diseases, just to name a few.

Some studies aim at using metabolomic data to investigate toxicity in response to the administration of certain drugs. Studies of toxicity *in vivo* are generally very expensive and can be lengthy (Ebbels et al., 2007), so an alternative approach to these types of studies would be welcomed. The effects of hydrazine toxicity in rats and mice are studied in Bollard et al. (2005a), and it is shown that the toxic response varies by species. Ebbels et al. (2007) investigate the use of metabolomic methods for toxic screening using 80 different treatments in rats, concluding that such screening

programs could be used to detect and classify toxic responses. The metabolic effects of drug doses, both toxic and non-toxic, in rats are studied by Beckonert et al. (2003), who believe based on their results that drug toxicity can be predicted. Studies such as these may help scientists better understand toxic effects of drugs and could one day lead to non-invasive early detection of toxicity.

Metabolic changes can occur not only due to drug disturbances but also to injury, dietary change, or other alteration in circumstance. In Viant et al. (2005), changes in metabolic behavior in rats following traumatic brain injury (TBI) are examined. Differences in brain metabolites due to injury, both global and specific, are found. Bertram et al. (2006) investigate the biochemical effects of a whole grain diet in pigs, attempting to identify metabolic evidence that might explain the known advantages of a whole grain diet. Hopefully, some of this type of research can be extended to humans and can help identify those with certain types of injuries or dietary risk factors so that interventions or treatments could be more suitable to patients.

Differences in metabolic profiles may also be attributed to disease status. If certain metabolic patterns or individual metabolite markers were found to be indicative of particular diseases, metabolomic methods could provide a non-invasive disease detection method. Metabolomic studies could also provide information on disease processes, aiding in the treatment of such diseases. Odunsi et al. (2005) show that metabolomic methods are promising in the early detection of epithelial ovarian cancer. In Denkert et al. (2006), metabolite differences in invasive ovarian carcinomas and ovarian borderline tumors are demonstrated, and differences between colon carcinoma and normal colon tissue are found in Denkert et al. (2008).

As mentioned above, many of metabolomic studies involve comparing normal samples to disturbed or diseased samples. These samples may be collected from subjects that differ in diet, disease status, drug dose, or some other factor. In order to accurately

compare these samples, the underlying normal behavior of metabolites must be studied. Bollard et al. (2005b) demonstrate that while metabolic profiles are different among species, the variation of metabolites within the human species is considerable. They go on to characterize differences based on age, gender, stress levels, diurnal effects, and other factors. In Park et al. (2009), differences in macronutrient regulation among subjects throughout the course of a day are examined. Although regulation varies from individual to individual, overall time-of-day patterns are identified.

1.4.2 Statistical Tools for Analysis

Metabolomic studies use statistical methods dependent upon the specific focus of the research. Most studies aim to use spectra to differentiate between different groups, to identify individual features that may distinguish one group from another, or both. It is possible for spectra to be separated on the basis of many qualities such as injury status (Viant et al., 2005), drug treatment groups (Keun et al., 2002; Forshed et al., 2003; Bollard et al., 2005b), disease groups (Odunsi et al., 2005), species (Bollard et al., 2005b), and country of origin (Dumas et al., 2006). Common methods used in order to separate these groups include K-means clustering, hierarchical clustering, principal components analysis (PCA), and partial least squares regression discriminant analysis (PLS-DA).

K-means clustering uses the squared Euclidean distance between points to determine the best cluster profile for a given number of clusters. Hierarchical clustering, on the other hand, does not require a pre-specified number of clusters. In the most common form of this clustering method, observations that are most similar in terms of some dissimilarity matrix are grouped together, then those groups that are similar are grouped together, and so on until one cluster containing all data remains. The number of clusters chosen is determined by how similar the user would like elements

within the same cluster to be relative to other clusters (Hastie et al., 2001). In PCA, principal components (PCs) are calculated as linear combinations of the data so that the first linear component explains the as much as possible of the variation present in the data, and the subsequent PCs explain less and less variation. All PCs are orthogonal to one another, and plots of various PCs can reveal clustering patterns (Hastie et al., 2001). PLS-DA also involves calculating a linear combination of data but uses features as well as class membership in this process (Barker and Rayens, 2003). Similar to PCA, scores resulting from this method can be plotted and can reveal separation of groups. Along with providing a means for separation, PCA and PLS-DA also allow for the identification of important features that differ among groups. In PCA the coefficients for the linear combination of variables, or "loadings," convey how much of the variation explained by a PC is due to each variable. Loadings resulting from PLS can be used similarly.

While K-means and hierarchical clustering methods provide a means for separation, PCA and PLS-DA also allow for the identification of important features that differ among groups. In many studies, combinations of all of these methods are used to determine the number of clusters, cluster membership, and important features contributing to cluster differences. In Park et al. (2009), metabolite patterns over the course of a day are investigated using PCA to determine the number of clusters, followed by K-means clustering to determine group membership.

1.5 Proposed Methods

We intend to develop methods to characterize healthy metabolite patterns, as understanding healthy behavior will aid in distinguishing healthy from abnormal behavior in the future. In this section, I will first describe the datasets that motivate this

research. Brief summaries of the material presented in the following chapters will then be provided.

1.5.1 Motivating Datasets

NMR Human Blood Plasma Data

Blood samples were drawn from eight subjects every hour, beginning at 8:30 in the morning and ending at 8:30 the next morning, for a total of 25 samples per person. From the blood collected, high resolution ^{1}H NMR spectra were obtained for analysis. Conditions were held constant for the subjects, and meals were administered at 9:30 am, 1:30 pm, 5:30 pm, and 9:30 pm. Half of the subjects were male and half were female, and they were of various ages. For more details, see Park et al. (2009).

LC-FTMS Marmoset Blood Plasma Data

Blood plasma was collected from 76 marmosets at the New England Primate Research Center. These marmosets ranged in age from 2 to 13 years and and had similar living conditions. Mass spectra were obtained using anion exchange (AE) liquid chromatography coupled with Fourier transform mass spectrometry (LC-FTMS). For each sample, spectra were produced in order to increase the number of metabolites detected for each marmoset. See Soltow et al. (2011) for more details.

1.5.2 Peak Identification and Pre-processing Methods for NMR Data

Methods for overcoming the complex and sensitive nature of NMR data are discussed. These methods include phase and baseline correction, normalization and scaling,

binning, using a calibration peak, spectral alignment, and the elimination of chemical shift ranges. The NMR human blood plasma dataset is used to demonstrate some of these existing methods, and a set of techniques appropriate for the identification of metabolomic NMR peaks is established.

1.5.3 Clustering on Periodicity Using Metabolomic Time Course Data

Metabolic information gathered by NMR spectroscopy over time can be used to study metabolite patterns of the human body. Metabolite levels change with time, and the identification of peaks with sinusoidal periodic behavior is of interest. We consider NMR human blood plasma data, obtained from eight subjects every hour over a 24-hour period. We use periodic regression and obtain estimates of the parameter corresponding to period for the various peaks. A mixture model is then used to develop clusters of peaks, taking into account the variability of the regression parameter estimates. Simulation studies under various conditions are presented.

1.5.4 Using Mass Spectra to Determine Possible Ages of Metabolic System-Level Shifts

Metabolite levels are thought to change with age, and metabolomic data will be used to investigate the locations of possible metabolic system-level shifts. The LC-FTMS marmoset blood plasma dataset will be analyzed, eliminating those marmosets considered unhealthy at the time of the blood draws. We use an algorithm to determine age locations at which metabolite levels change for individual MS features. Clusters of breakpoints are then determined, and features and corresponding metabolites with similar shift locations are identified.

Chapter 2

Peak Identification and Pre-processing Methods for NMR Data

2.1 Introduction

Unwanted variation is inevitably present in metabolomic data resulting from nuclear magnetic resonance (NMR) spectroscopy. This variation can be caused by sample differences such as pH level and temperature as well as by differences in machines or experimental conditions. These factors affect the resulting spectra from these methods, making direct comparisons of raw spectra unwise. Analysis of metabolomic data is further complicated by its size since spectra are often comprised of thousands of data points. For these reasons, pre-processing of this type of data is mandatory, and many context-dependent pre-processing schemes have been proposed.

In this chapter, existing methods for carrying out pre-processing steps are described. The pre-processing of an NMR blood plasma dataset is then outlined in detail, using

both existing and developed methods. Although all metabolomic data must be pre-processed to overcome the complicating issues listed above, we focus on NMR data. The pre-processing of NMR spectra can include a multitude of steps but usually includes phase and baseline correction, alignment of spectra, elimination of uninformative peaks, and normalization to an internal standard. Often, a data reduction step is used as well.

2.2 Existing Adjustment and Processing Techniques

2.2.1 Shimming, Phasing, Baseline Correction, and Assigning a Calibration Peak

Shimming, phasing, baseline correcting, and using calibration peaks are all adjustments used to make NMR spectra more accurate and comparable by controlling for different factors. Differences in magnetic field over a sample are called "gradients" and can affect the accuracy of NMR spectra if not addressed. Adjusting magnets called "shims," or shimming, can correct this problem if done properly. Shimming must be performed every time a sample is run due to slight changes in sample temperatures, NMR tubes, sample levels, and other seemingly small differences so that all spectra are of good quality and are comparable (Pearson, 1991).

NMR spectra can be distorted in several ways and must be adjusted appropriately. Zero-order and first-order phase errors can be present, zero-order arising from the fact that the x-axis is not perfectly in line with the reference, and first-order errors being introduced since machines cannot begin to collect data at the exact moment of an electro-magnetic pulse (Sanders, 1993). These errors cause peaks to change shape, a problem that can be corrected by "phasing" the spectra. For the zero-order

correction, the largest peak is usually selected, and the phase is adjusted until it appears to have the correct shape. This process is repeated with another peak for the first-order correction.

Another problem is that NMR spectra have "rolling baselines," meaning that the baseline rises and falls throughout the spectrum, making it difficult to recognize true signals. This type of baseline distortion can be caused by the recovery time necessary after a radio frequency pulse or by a phase change in the filters used to exclude noise (Sanders, 1993). Baseline correction can be used to eliminate this problem so that signals can be identified and spectra can be compared. One method of baseline correction uses a polynomial fit to estimate the baseline and then subtracts this from the spectrum.

Calibration peaks are used to define the zero chemical shift location of spectra. A small amount of a substance not naturally present in a sample is added to produce this peak. An ideal substance would produce only one peak, which would have a lower ppm value than other naturally-occurring peaks. Also, peaks of naturally-occurring components would not be affected by the addition of this substance. For these reasons, tetramethylsilane (TMS) is a good substance to use to produce calibration peaks in ^{1}H NMR spectra. The substance 3-trimethylsilylpropionic acid (TSP) is also used for this purpose.

2.2.2 Correcting Chemical Shift Differences

Small differences in magnetic field strengths, sample conditions such as temperature, concentration, and pH levels, or other factors can cause a peak to present at different chemical shifts across spectra (Forshed et al., 2003; Craig et al., 2006). Since it is impossible to obtain spectra under exactly the same conditions, this issue must be addressed in most if not every study. Some propose aligning spectra using various

algorithms, and others think chemical shift differences can be minimized by integrating spectra over a number of spectral regions, a process often called "binning." The choice of method is guided by the goals of the study and the type of analysis.

Binning

"Binning" or "bucketing" involves summing peaks over spectral segments and is fast and easy to implement. Commonly, bins of 0.04 ppm are used for urine samples, which seems to correct for small shifts that may be present for peaks (Craig et al., 2006; Odunsi et al., 2005). Bins of different sizes can be used, however, and some studies use bins as small as 0.005 ppm for tissue and plasma samples (Viant et al., 2005). Summing over groups of peaks simplifies the data, but information is lost in the process. Also, inaccuracies can arise since unrelated peaks can be placed in the same bin and peaks can be split between different bins (Alm et al., 2009). Some try to get around this problem by combining bins known to contain the same or related substances (Keun et al., 2002; Viant et al., 2005). This method is used in classification studies, but since peak locations are combined, biomarker identification is not possible using such a method (Craig et al., 2006). With modern computing power, other more sensitive methods such as spectral alignment are available.

Spectral Alignment

Alignment algorithms generally shift and/or stretch and shrink specified spectral segments until a certain criterion or stopping point is met, and some sort of distance measure is used to measure the success of alignment. Forshed et al. (2003) suggest using a genetic algorithm for alignment of predetermined regions of spectra to the corresponding regions of a reference spectrum. The alignment and thus the correlation coefficient improves with each iteration, or "generation," and the algorithm is stopped

after a certain number of generations. The correlation coefficients for the reference spectrum with both the complete unaligned spectra and the aligned spectra are used to measure the success of alignment. This algorithm is relatively efficient and is appropriate for complex NMR data, unlike other alignment methods. Lee and Woodruff (2004) propose a new and faster method based on the Forshed et al. (2003) method but replacing the genetic algorithm with a beam-search algorithm. Recent alignment attempts use algorithms originally used for image analysis (Alm et al., 2009), and new alignment methods will probably continue to be introduced.

2.2.3 Elimination of Uninformative Peaks

Often, some peaks should be removed from NMR spectra in the processing stage. Biofluids such as urine and blood plasma have high water content relative to the other molecules present. If the water signal were not suppressed, NMR peaks corresponding to water protons would be so large that other substances would not be visible. The effects of this water suppression are not constant, however, causing resulting water-related peaks to be meaningless. These peaks, generally occurring between 4.5 and 6 ppm, are often removed from NMR spectra for this reason (Antti et al., 2002; Keun et al., 2002; Odunsi et al., 2005; Bertram et al., 2006; Dumas et al., 2006). Regions corresponding to other substances such as contaminants (Viant et al., 2005) or substances that are not of interest (Keun et al., 2002; Park et al., 2009; Nicholson et al., 1995) are often removed as well.

2.2.4 Normalization and Scaling

Normalization generally refers to a subject operation, and scaling refers to a feature or peak operation. In normalization, the spectra are divided by their total sums or

some other spectrum-dependent value. The result is that relative concentrations of substances are found, and thus samples are comparable. This can be particularly important in urine samples, where the volume and thus the concentration of substances can depend heavily on a drug being taken or on a toxicity level (Craig et al., 2006). If the spectra intensity totals are used in normalization, an increase in one peak will make it appear that other peaks have decreased even if they have not, causing problems. Instead, some prefer normalizing to an internal standard.

When normalizing to an internal standard, reference compounds that do not naturally occur in a substance such as a biofluid are added to samples before obtaining NMR spectra. A known amount is added to each sample, so the size of the peak corresponding to this substance can be used to normalize spectra in a study. The height or the area of the internal standard peak can be used to define its size, although area may be more appropriate since it better corresponds to the concentration of a substance.

Several different substances can be used as the internal standard for NMR spectra. In aqueous solutions, the sodium salt of 3-trimethylsilylpropionic acid (TSP) is often used. However, this compound should not be used as a reference standard in samples with a high protein content such as blood plasma samples. Since it can bind to proteins, it will affect the results of the NMR, but formate can be used instead (Beckonert et al., 2007). Other substances that can be used include Tetramethylsilane (TMS) (Park et al., 2009) and 3-trimethylsilylpropyl (TMSP) (Viant et al., 2005).

Scaling, a peak-specific operation in our case, can also be used. Often, spectral positions are mean-centered and scaled to unit variance. They can also be Pareto-scaled by dividing by the square root of the standard deviation of the position values or logarithmically scaled (Craig et al., 2006). There is much variation in how normalization and scaling procedures are carried out. Normalization and scaling

can both be used on a dataset, but the types of samples and studies should dictate which methods are used.

2.3 Application of Processing Techniques: Blood Plasma NMR Data

Processing of the blood plasma data is described below, from initial basic processing steps to spectral denoising and peak identification. The dataset is processed for a peak-specific study, the goal of which is to characterize the periodic patterns of increase and decrease for metabolite peaks. Different methods would be more appropriate for other types of studies with classification or other goals.

2.3.1 Initial Processing

Blood samples were drawn from eight subjects every hour, beginning at 8:30 in the morning and ending at 8:30 the following morning, for a total of 25 samples per person (see Park et al., 2009, for more details). From the blood collected, high resolution ^{1}H NMR spectra were obtained for analysis, and these original spectra consisted of 16384 data points each. The water signal (4.8 ppm) was suppressed and parameters were standardized so that spectra could be compared. Also, the location of a chemical shift of zero was defined using the internal standard tetramethylsilane (TMS). The NMR spectra were phase and baseline corrected using NUTS NMR data processing software (Acorn NMR, Inc). Suitable areas were chosen to represent the baseline signal, and baseline correction was implemented using least squares polynomial regression. The spectra were then aligned using a beamsearch algorithm maximizing the correlation between a reference spectrum and the rest of the spectra (Lee and Woodruff, 2004).

Also, the water region was eliminated (4.8 - 5 ppm), and the spectra were normalized using the size of the reference tetramethylsilane (TMS) peak. The spectra were limited to the range of 0 to 5.5 ppm since it was determined that no usable signal existed for chemical shifts above this range (Nicholson et al., 1995; Park et al., 2009). Limiting to this range decreased the number of data points per spectrum to 8192. Figure 2.1 includes a plot of one of the raw spectra as well as a plot of the same spectrum after these processing steps. You can see that the water peaks between 4.8 and 5 ppm dominate this spectrum.

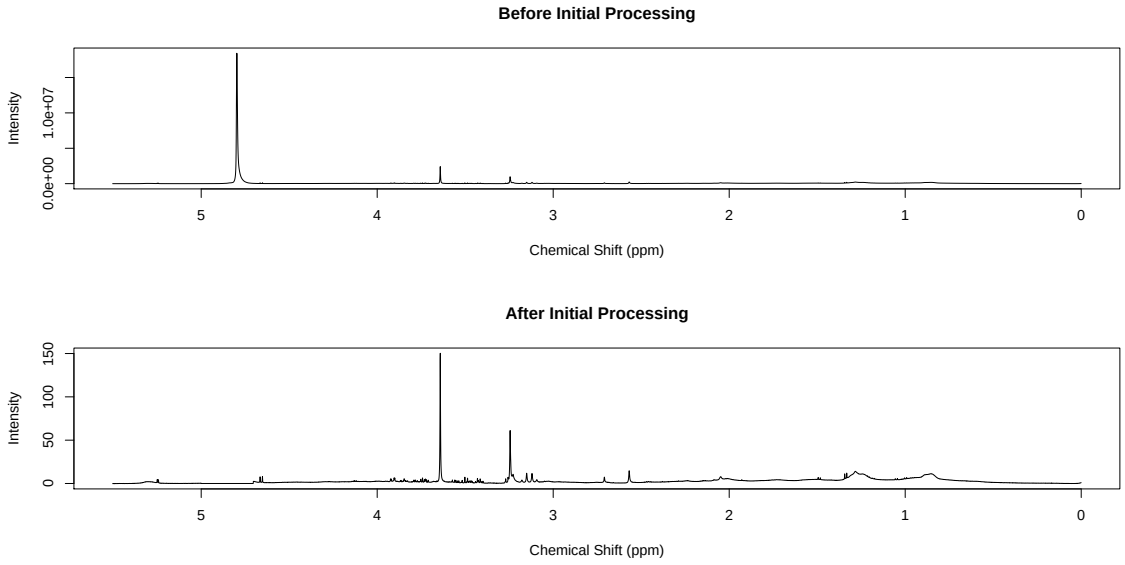

Figure 2.1: The top plot displays a raw spectrum, and the bottom plot shows the same spectrum after initial processing steps. Details are more visible after the elimination of the dominant water peak between 4.8 and 5 ppm.

2.3.2 Wavelet Smoothing

Initial pre-processing takes care of many problems, but NMR spectra still display considerable noise (Figure 2.2). In an effort to try to isolate the true signal in our data, the spectra were wavelet denoised. This process involves using a wavelet filter to

first transform and decompose a signal into different levels of detail. Assuming that all coefficients are not needed to accurately represent the true signal, wavelet coefficients that are not larger in magnitude than a threshold value are set to zero to denoise data (Strang, 1996). If those coefficients that are larger than the threshold value are retained, this is called "hard thresholding." In "soft thresholding," coefficients larger in magnitude than the threshold value are shrunk toward zero by the threshold value while other values are set to zero. Hard thresholding was used to denoise our data, as shrinking of the signals was not necessary. The threshold value was defined as the standard deviation of the level of finest detail.

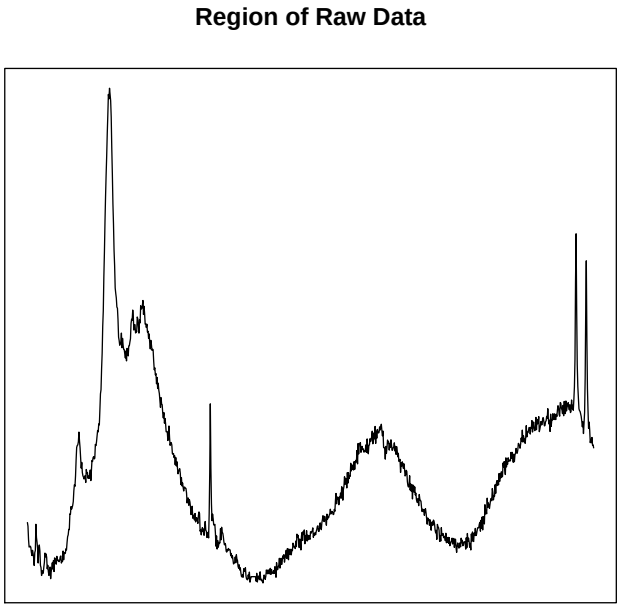

Figure 2.2: NMR spectra exhibit considerable noise.

There are many different wavelets that can be used to transform signals, allowing an appropriate wavelet for a particular type of signal to be chosen (Burrus, 1998). The Haar wavelet is the simplest and can be used for "blocky" data since it is discontinuous, but it is often not practical (Hastie et al., 2001). The Daubechies

family of wavelets is more complicated, with the members having different orders indicating their complexity. The Daubechies wavelet of order two is the same as the Haar wavelet, and the maximum order for Daubechies wavelets is ten. The Symmlet family of wavelets includes modifications to the Daubechies family, and these two groups have very similar properties. The Symmlet wavelet of order four is widely used and is often appropriate for denoising images. The largest order for Symmlet wavelets is eight.

Figure 2.2, as mentioned above, shows a small range of an NMR spectrum before denoising. Figure 2.3 shows the same region along the top row, and the results of wavelet smoothing with different wavelet filters are given below. The Haar wavelet result shows that this wavelet is not a good choice for our data. Other results vary in the level of smoothness. The goal was to choose a wavelet filter that preserved the visible peaks, meaning that the image was not oversmoothed but did not keep too much detail. This is very subjective, but with the NMR data, the decision was fairly straightforward. The Symmlet 4 wavelet was chosen as the best to transform the NMR spectra for denoising using these criteria. Figure 2.3.2 shows another portion of the NMR spectrum along with its denoised counterpart. It looks as if major peaks are preserved while not including too much detail, so the denoising wavelet choice seems appropriate.

2.3.3 Identification of Peaks

After wavelet smoothing of the spectra, peaks were identified for each spectrum. Observation of the spectra indicated that peaks would be appropriately represented by relative maxima. All relative maxima and relative minima for each spectrum were identified, and a sample of these results can be seen in Figure 2.5. We defined peak

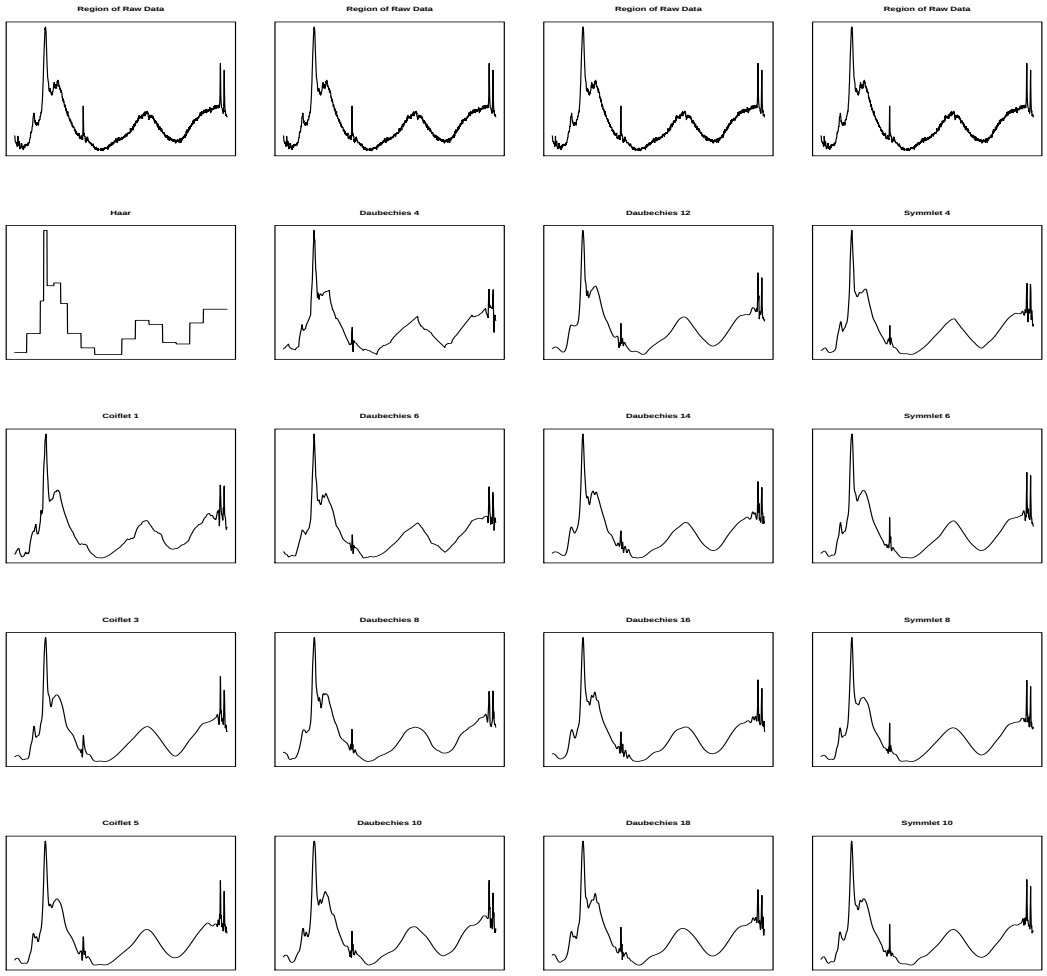

Figure 2.3: Region of NMR spectrum before denoising is depicted on the top row. Other rows show results of wavelet denoising using various wavelet filters.

magnitudes to be the differences between the relative maxima and the corresponding relative minima. Of course, all but the smallest and largest relative maxima were sandwiched between two relative minima. The smaller of the relative minima around these maxima were used to calculate peak magnitudes. The total number of peaks after using this algorithm was 44183 amongst all spectra, or about 221 per spectrum. Only peaks greater than 5 times the standard deviation of the noise were kept, where the standard deviation of the noise was defined as the standard deviation of the region of the spectrum with no discernible signal. This threshold was chosen to eliminate many of the noisy or very small peaks without removing peaks with a strong signal.

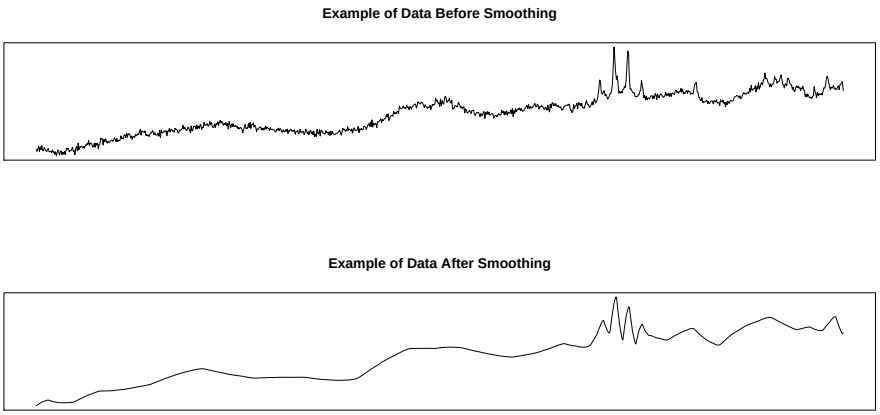

Figure 2.4: The top plot shows a portion of one of the spectra before wavelet denoising. The same portion after denoising is shown in the bottom plot.

Also, only unique peaks were kept, meaning that peaks occurring at the same exact position as peaks on other spectra were eliminated. This brought the total peak count down even further to 1242.

2.3.4 Determination of Peak Locations

Although 1242 unique peak locations were found amongst all spectra, some of these peak locations were very similar. The spectra were aligned, as mentioned above, but no alignment algorithm is perfect. Some peaks representing analogous protons could have surfaced close to one another but not at exactly the same chemical shift. Examining the data revealed that several spectra showed peaks that were determined to be the same peak but were about 6 units (about 0.004 ppm) apart. Therefore, if peaks on multiple spectra were similar in chemical shift, they were classified as the same peak. Peaks that were the closest to one another were combined first, and the range of peaks identified as representing the same peak did not exceed 0.005 ppm (8 units). This window is slightly larger than the 0.004 ppm observed between peaks so is

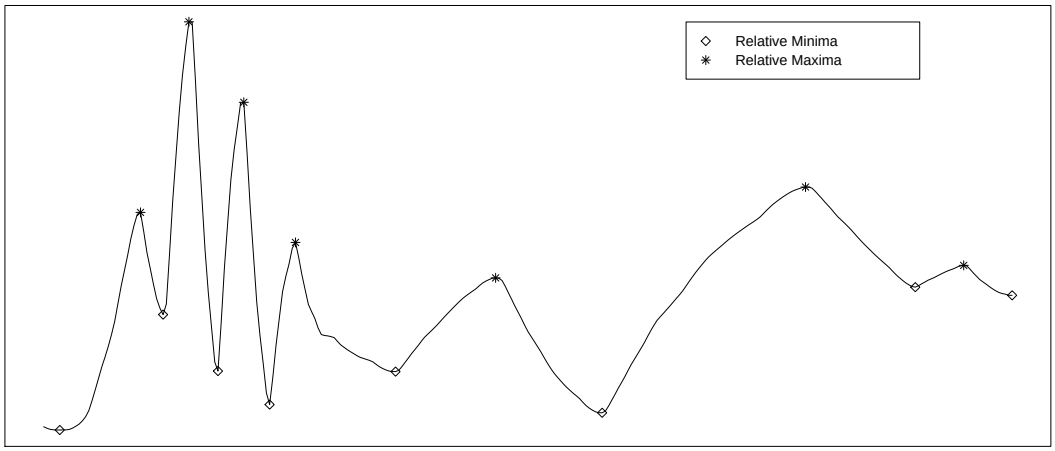

Figure 2.5: Relative maxima and relative minima were identified for each spectrum and used to define peak magnitudes.

a little more flexible. A final list of peaks from all spectra was produced. Spectra peaks were standardized by subtracting the mean and dividing by the standard deviation. For every location on the peak location list, the maximum value within 4 units (about 0.0025 ppm) in either direction was found in each spectrum. This value could be a peak for some spectra and a non-peak for others, but the values kept for each spectrum were in the same locations. The peaks were then normalized. After all processing, each spectrum, which originally contained 16384 data points, was whittled down to only 259 peaks. Figure 2.6 shows a histogram of the identified peak locations and how it corresponds to one of the processed NMR spectra. You can see that regions where peaks are visible in the spectrum did in fact have peaks identified in their areas, so the method appears to have performed well.

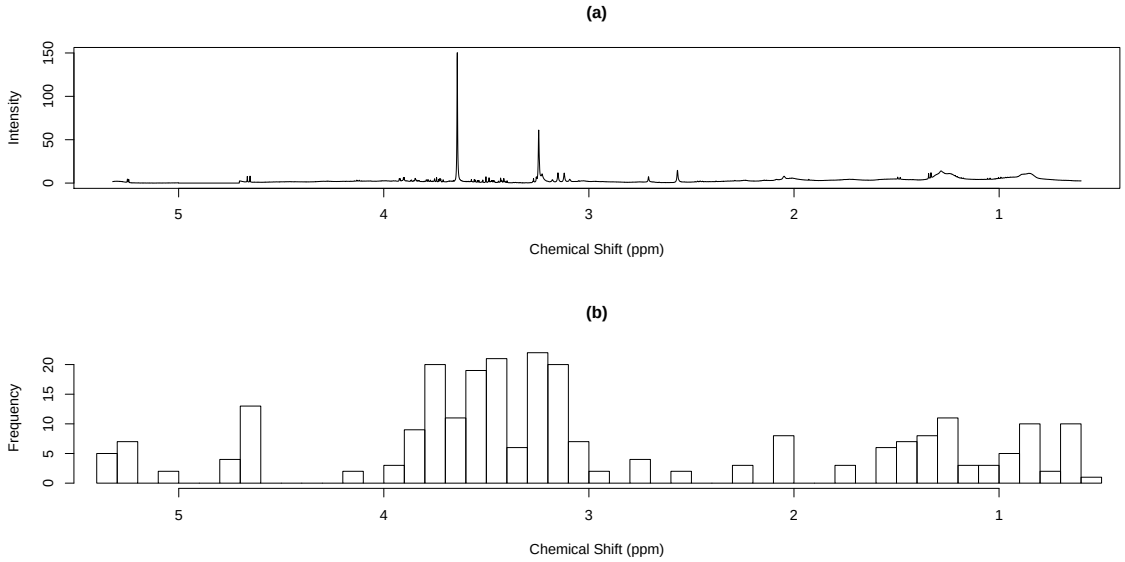

Figure 2.6: One of the processed NMR spectra is shown in the top panel, and the bottom panel is a histogram of the final peak list found.

2.4 Discussion

Variation present in all metabolomic data, along with the sheer number of data points, can make analysis of such data very difficult. Therefore, pre-processing this type of data is absolutely necessary. Pre-processing techniques used for metabolomic data often include baseline correction, alignment of spectra, elimination of uninformative peaks, and normalization to an internal standard. These methods among others are detailed above, and then multiple pre-processing steps of a particular NMR dataset are described.

Although some pre-processing techniques are thought to be more appropriate than others, no automatic complete pre-processing techniques are recommended or are currently available, and this issue is being addressed (Trainesel et al., 2007). Techniques chosen depend on the resources available to a researcher as well as on the particular features of study. The processing of the human plasma dataset detailed above focuses on achieving a final dataset comprised of information for individual

peaks rather than complete spectra. This focus influenced the choice of techniques, but others with similar goals may not choose to use the same methods for their data. For instance, other smoothing methods or normalization techniques could be used. Although different, both sets of tools could be equally valid. Thus, at least for now, it is up to the researcher to choose appropriate methods for his or her data and analyze the resulting data accordingly.

Chapter 3

Clustering on Periodicity Using Metabolomic Time Course Data

3.1 Introduction

Originally used for structure determination in organic chemistry, nuclear magnetic resonance (NMR) spectroscopy is also a useful tool in the biosciences as it can be used to investigate the metabolic composition of biofluids (Beckonert et al., 2007). NMR spectroscopy provides a tool for characterizing the nuclei found in a substance in terms of relative field strength and concentration (McMurry, 1996). Results are often presented in a plot with intensity on the vertical axis and chemical shift on the horizontal axis, where chemical shift is a measure of the effective field strength of a nucleus. In ^{1}H NMR spectroscopy, the area under each peak corresponds to the number of ^{1}H protons with the same chemical shift and thus is an indication of the concentration of a substance in a sample. NMR spectra taken from biological samples could provide information relating to diseases and disorders. Collecting blood plasma and urine samples is relatively simple and inexpensive, and most of these substances

require very little pre-processing. For these reasons, NMR data is an attractive source of information.

Information collected from NMR data has already been used for several purposes. NMR spectra have been used to investigate metabolic and other types of disorders (Odunsi et al., 2005; Denkert et al., 2006, 2008), to determine whether or not individuals were in a normal metabolic state following surgery, dietary intervention, or injury (Beckonert et al., 2003; Bertram et al., 2006; Viant et al., 2005), and to detect toxic drug levels in individuals (Ebbels et al., 2007; Bollard et al., 2005a; Beckonert et al., 2003). Distinguishing between different groups using average NMR spectra is often of interest. Methods commonly used include K-means clustering, hierarchical clustering, principal components analysis (PCA), and partial least squares regression discriminant analysis (PLS-DA). While K-means and hierarchical clustering methods provide a means for separation, PCA and PLS-DA also allow for the identification of important features that differ among groups. In many studies, combinations of all of these methods are used to determine the number of clusters, cluster membership, and important features contributing to cluster differences.

One area of current research using NMR data focuses on the characterization of metabolite time-of-day patterns. Some biochemical, physiological, and behavioral processes in the human body display circadian rhythm, having cycles repeating about once every 24 hours (Leikin, 2003; Blanco et al., 2007). Melatonin, which affects many physiological actions, is under circadian regulation and is produced mainly at night (Simko and Pechanova, 2009). Disruptions in circadian rhythm and thus melatonin levels are thought to lead to increased risk of cardiovascular disease (Simko and Pechanova, 2009), metabolic syndrome (De Bacquer et al., 2009), sleep disorders (Blask, 2009), and even some types of cancer (Blask, 2009; Stevens, 2009). The study of the daily patterns of physiological systems has resulted in treatment regimens that take into account the circadian variation in phenomena such as blood pressure (Simko

and Pechanova, 2009).

The influence of circadian forces on metabolite levels and the possible presence of other periodic patterns of metabolites are investigated in Park et al. (2009). In order to observe metabolite patterns over time, Park et al. (2009) collected blood plasma samples from subjects every hour, beginning at 8:30 am, over a 24-hour period. Meals were administered to study participants at 9:30 am, 1:30 pm, and 5:30 pm, and a snack was given at 9:30 pm. Each sample collected was used to produce a ^{1}H NMR spectrum. After averaging spectra for each time point, Park et al. (2009) find that average spectra can be classified into groups corresponding to morning, afternoon, and night using PCA and K-means clustering. Summed metabolic signal intensities in certain NMR regions differ when comparing spectra collected during morning, afternoon, and night hours, and graphical evidence suggests the existence of periodic time-of-day patterns of metabolites.

Identifying periodic metabolic peaks could aid in the understanding of the behavior of the human body, providing information useful in diagnosing and treating many conditions. In order to supply investigators with detailed time-of-day information from metabolic signals, we propose a method that will provide a means for extraction of NMR peaks exhibiting periodic behavior and that will characterize this behavior. Techniques mentioned above succeed at finding groups with similar spectra and can even identify features that most contribute to differences in groups. However, our goal is to estimate the period of individual NMR peaks and to find groups of peaks with similar behavior. A new method is thus developed.

Periodic regression is used to model the behavior of identified peaks in the NMR spectra, and the regression coefficients are then used to identify groups with different periodic behavior. While some methods use the joint distribution of the parameter estimates in the clustering procedure (Qin and Self, 2006), we propose using only the

marginal distribution of the parameter associated with periodicity. It is of scientific interest to find clusters of peaks with similar periodic patterns, regardless of amplitude and phase shift. Even if clusters existed such that peaks within a cluster shared the same amplitude, phase shift, and period, these clusters would not provide the desired result: clusters of peaks with homogeneous periodic patterns. The parameters corresponding to phase and amplitude can therefore be ignored during the clustering process. This will allow peaks with different phase shifts but similar periods to be classified together. We want to allow for the possibility that patterns in one cluster could be opposite of one another. This is because, similar to other biological processes (Spellman et al., 1998; Yu, 2010), functionally related metabolites may have similar periodicity yet differ in phase. Metabolic regulation involves inhibition of the production of a metabolite once that substance reaches high levels. Thus, as the substance increases, intermediate metabolites involved in the production of the substance may decrease.

Additionally, since values for the period-related parameter are estimated in the periodic regression step, we propose accommodating the variability associated with the parameter estimates of interest in the clustering process. The method is applied to our sample data, extracted from ^{1}H NMR spectra collected from eight subjects every hour for 25 hours (spectra also used in Park et al. (2009)), and the periodic behavior of the identified peaks is assessed. Simulation studies are used to show that when there is an extra variance component due to estimation of parameter estimates, it should be accounted for in clustering.

3.2 Data Pre-Processing and Peak Identification

The 25 ^{1}H NMR spectra collected from each subject were phase and baseline corrected using NUTS NMR data processing software (Acorn NMR, Inc). The spectra were then aligned using a beamsearch algorithm maximizing the correlation between a reference spectrum and the rest of the spectra (Lee and Woodruff, 2004). Also, the water region was eliminated, and the spectra were normalized using the size of the reference tetramethylsilane (TMS) peak (Kim and Park, 2005). The spectra were limited to the range of 0 to 5.5 ppm since metabolites are known to present in this region (Nicholson et al., 1995) and since the signal-to-noise ratio is high here. The data were denoised using a Symmlet wavelet transformation of order four with hard thresholding. The standard deviation of the level of finest detail was used as the threshold level, meaning that only wavelet coefficients larger than this value were retained. The wavelet denoising process resulted in much smoother data.

The relative maxima of the smoothed spectra were considered to be peaks. The differences between the relative maxima and the smaller of the relative minima around these maxima were taken to be the magnitudes of the peaks. Only peaks with intensities greater than 5 times the standard deviation of the noise were used in the analysis, where the region of the spectrum with no discernible signal was used to define the noise. After imposing this threshold, redundant peak locations were removed. If peaks on multiple spectra were similar in chemical shift, they were classified as the same peak, and a total peak location list was found. The range of peaks identified as being the same did not exceed 0.005 ppm. Spectra were standardized, and for every location on the peak location list, the maximum value within about 0.0025 ppm was found in each spectrum. At a particular location, the maximum value within the 0.0025 ppm range could be classified as a peak for some of the spectra and a non-peak for other spectra, but values for each peak location were found in each spectrum.

The peaks in the final peak list were then normalized. For more details, see Chapter 2.

3.3 Methods

3.3.1 The Model

Our goal is to detect clusters of peaks so that peaks within a cluster have similar periodicity but those in different clusters can be dissimilar in this regard. All peaks are not expected to behave periodically, and preliminary plots of the data show that a sine curve is appropriate for a subset of peaks (Figure 3.1). It is not of interest to find groups of peaks in which each group has the same amplitude, phase shift, and period. Although some peaks may exhibit the same periodic pattern, they may have different amplitudes and phase shifts, some perhaps beginning at lower levels and reaching higher levels later in the day, and some having the opposite pattern. These peaks could be affected by the same mechanism and should be classified in the same group. Thus, we focus our attention on estimating the parameter associated with periodicity, β_p. Instead of fitting a model in which the joint distribution of the parameters associated with the amplitude, periodicity, and phase shift is explicitly defined, we fit a simplified model for each peak in which the marginal distribution of β_p is specified. After estimating the β_p values, these estimates along with the (known) variability associated with estimating them are used to find clusters of peaks.

Suppose an NMR spectrum consists of P peaks and that we have spectra collected from N subjects at multiple times. Thus, we have a measurement for peak p ($p = 1, 2, ..., P$) for each subject i ($i = 1, 2, ..., N$) at each time j ($j = 1, 2, ..., J$). Let Y_{ijp} be the observed peak magnitude for subject i at time j for peak p, and let t_{ij} be the time of the measurement for subject i at time j. Let X_i be an indicator variable for person

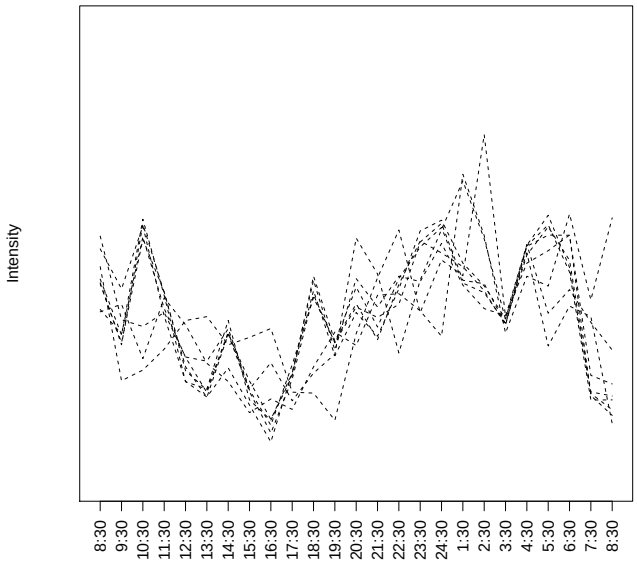

Figure 3.1: Plot of intensity vs. time for subset of peaks. It appears that a sine curve would be a good fit for these peaks.

i, where $X_i = 1$ for person i and 0 otherwise and using $i = N$ as the reference subject. Consistent with the notation in Qin and Self (2006), assume there are K classes of peaks with distinctive periodic behavior, and let $u_p = k$ $(k = 1, 2, ..., K)$ denote class membership for peak p. Alternatively, one can use $\mathbf{u}_p = (u_{p1}, u_{p2}, ..., u_{pK})^T$ to indicate class membership, where $u_{pk} = 1$ if peak p belongs to class k and 0 otherwise. Let $\boldsymbol{\pi} = (\pi_1, \pi_2, ...\pi_K)^T$ be the vector of probabilities associated with each class membership vector $\mathbf{u}_p$, where $\pi_k \geq 0$ and $\sum_{k=1}^{K} \pi_k = 1$.

The model can be written as, given $u_p = k$:

$$\begin{aligned} Y_{ijp} &= \alpha_p sin(\beta_p t_{ij} + \gamma_p) + \delta_{ip} X_i + \epsilon_{ijp}, \\ \epsilon_{ijp} &\sim N(0, \sigma^2), \\ \beta_p &= \mu_{\beta_k} + e_p, \\ e_p &\sim N(0, \sigma_k^2), \\ \mathbf{u}_p &\sim Multinomial(\boldsymbol{\pi}). \end{aligned} \tag{3.1}$$

Parameters α_p, β_p, and γ_p are the regression parameters associated with the sine function for peak p. The amplitude of the sine curve for peak p is given by $|\alpha_p|$, the period is $2\pi/|\beta_p|$, and the phase shift for the curve is given by $-\gamma_p/\beta_p$ (Swokowski, 2009). We assume that the subject effect δ_{ip} is fixed, and the error values are given by ϵ_{ijp} are assumed to be independent and identically normally distributed with mean zero and constant variance σ^2. β_p has mean μ_{β_k} and residual e_p, which is normally distributed with mean zero and variance σ_k^2. The class membership vector, $\mathbf{u}_p$, has a multinomial distribution with parameter $\boldsymbol{\pi}$.

With this model, the periodic behavior of peaks within a cluster is similar, but periodicity can be different across clusters. Heterogeneity is allowed in the parameters associated with amplitude and phase shift since these curve attributes are not of interest. The joint distribution of α_p, β_p, and γ_p is not specified in Model 3.1, so the complete maximum likelihood procedure cannot be carried out for this simplified model since the full likelihood is unknown. Note that a full model incorporating the joint distribution of the sine curve parameters would imply the simpler Model 3.1, but the reverse is not necessarily true.

3.3.2 Estimation

The estimate $\hat{\beta}$ for each peak has mean μ_{β_k}, the mean value for the peaks in class k, with a traditional residual component e_p as well as a random effect signifying the extra variation due to the estimation of β_p denoted by ψ_p. We rewrite the model as $\hat{\beta}_p = \mu_{\beta_k} + e_p + \psi_p$, where $\psi_p \sim N(0, s_p^2)$, and where s_p^2 is the variability associated with the estimation of β and is assumed to be known. It follows from these relationships and from Model 3.1 that, given $u_p = k$, $\hat{\beta}_p$ has the distribution:

$$\hat{\beta}_p \sim N(\mu_{\beta_k}, \sigma_k^2 + s_p^2). \tag{3.2}$$

Given class k, $\hat{\beta}_p$ is normally distributed with mean μ_{β_k} and with a known peak-specific variance component as well as an unknown class-specific variance component.

The β_p values are first estimated using nonlinear least squares regression with the Gauss-Newton algorithm with Levenberg-Marquardt modifications for global convergence (Matlab, The MathWorks, Inc.). Starting values and iteration limits are changed as necessary to successfully model as many peaks as possible. The β_p values and their variances are of primary interest in terms of the periodic behavior of the peaks.

After estimating the β_p values, the expectation-maximization (EM) algorithm (Dempster et al., 1977) is used to find groups of peaks with similar periodic behavior while considering the variability in the parameter estimates of interest. This algorithm's first step is the expectation step (*E step*), which involves computing the expected value of the log-likelihood of the complete data using current parameter estimates. The next step, or maximization step (*M step*), finds maximum likelihood parameter estimates using the expected likelihood calculated in the E step.

Let $\Delta_{pk} = 1$ if $\hat{\beta}_p$ is in class k and 0 otherwise. Let π_k be the probability that $\hat{\beta}_p$ is in

class k ($\pi_1 + \pi_2 + ... + \pi_K = 1$), so that $Pr(\Delta_{pk} = 1) = \pi_k$ for each $\hat{\beta}_p$. Let η_{pk} be the expected value of Δ_{pk} given parameter estimates θ and observed data $\mathbf{Z}$, as detailed in (and with similar notation as) Hastie et al. (2001), such that:

$$\eta_{pk} = E(\Delta_{pk}|\theta, \mathbf{Z}) = Pr(\Delta_{pk} = 1|\theta, \mathbf{Z}) = \frac{Pr(\mathbf{Z}|\Delta_{pk} = 1, \theta)Pr(\Delta_{pk} = 1)}{Pr(\mathbf{Z}|\theta)}. \tag{3.3}$$

The vector θ includes parameter estimates for μ_k and σ_k^2 values, and $\mathbf{Z}$ consists of the $\hat{\beta}_p$ values. Substituting for these quantities for our particular situation, we get

$$\eta_{pk} = \frac{\pi_k \phi_{\theta_k}(\hat{\beta}_p)}{\sum_{k=1}^{K} \pi_k \phi_{\theta_k}(\hat{\beta}_p)} \tag{3.4}$$

where $\phi_{\theta_k}(\hat{\beta}_p)$ is the normal density for $\hat{\beta}_p$ with mean μ_{β_k} and variance $\sigma_k^2 + s_p^2$. More specifically,

$$\phi_{\theta_k}(\hat{\beta}_p) = \frac{1}{\sqrt{2\pi(\sigma_k^2 + s_p^2)}} e^{-((\hat{\beta}_p - \mu_{\beta_k})^2)/(2(\sigma_k^2 + s_p^2))} \tag{3.5}$$

Using this information, we can get the expected log-likelihood at each iteration j+1 (*E-Step*):

$$E(\ell(\theta; \mathbf{T})|\mathbf{Z}, \hat{\theta}^{(j)}) = \sum_{p=1}^{P} \sum_{k=1}^{K} [(\hat{\eta}_{pk} log \phi_{\theta_k}(\hat{\beta}_p)) + (\hat{\eta}_{pk} log(\pi_k))], \tag{3.6}$$

where $\mathbf{T}$ is the complete data and includes the unknown classes, $\mathbf{Z}$ is the observed data, and $\hat{\theta}^{(j)}$ represents the parameter estimates at iteration j. The estimate $\hat{\eta}_{pk}$ is found by replacing π_k and θ_k in (3.4) with their current estimates $\hat{\pi}_k$ and $\hat{\theta}_k$. This allows us to get maximum likelihood estimates for $\hat{\mu}_{\beta_k}$, $\hat{\sigma}_k^2$, and $\hat{\pi}_k$ at each iteration using the following equations (*M-Step*):

$$\hat{\mu}_{\beta_k} = \frac{\sum_{p=1}^{P} \hat{\eta}_{pk} \hat{\beta}_p / (\sigma_k^2 + s_p^2)}{\sum_{p=1}^{P} \hat{\eta}_{pk} / (\sigma_k^2 + s_p^2)} \tag{3.7}$$

$$\sum_{p=1}^{P} \frac{\hat{\eta}_{pk}}{\hat{\sigma}_k^2 + s_p^2} - \sum_{p=1}^{P} \frac{\hat{\eta}_{pk}(\hat{\beta}_p - \mu_{\beta_k})^2}{\hat{\sigma}_k^2 + s_p^2} = 0 \tag{3.8}$$

$$\hat{\pi}_k = \frac{\sum_{p=1}^{P} \hat{\eta}_{pk}}{P} \tag{3.9}$$

At each step, equations 3.7 and 3.8 are solved iteratively until the sum over all k of the squared changes in estimates $\hat{\mu}_{\beta_k}$ and $\hat{\sigma}_k^2$ are less than a certain tolerance level. The overall stop criterion is that the sum of the squared changes over k for all three estimates ($\hat{\mu}_{\beta_k}$, $\hat{\sigma}_k^2$, and $\hat{\pi}_k$) is less than the tolerance level for consecutive iterations. Note that when s_p^2 is set to zero, the equations for mean and variance estimates (equations (3.7) and (3.8)) reduce to those of a simple Gaussian mixture model. For more derivation details, please see Appendix A.2.

3.4 Application to Blood Plasma NMR Data

The blood plasma data consists of high resolution ^{1}H NMR spectra from eight people obtained from blood samples collected every hour over a period of 24 hours. After smoothing, peak identification, and peak consolidation, it was determined that 259 peaks were represented in at least one of the spectra. Before modeling individual peaks as described above, hierarchical clustering was used to find preliminary groups of peaks. A one-sine model, ignoring the person effect, was then fit for each of the preliminary clusters to obtain starting values for the periodic regression models for the individual peaks. Out of the 259 peaks, 258 were fit with a one-sine model (3.1), for a total of 258 values for $\hat{\beta}$ and variance estimates for each. The one peak for which sine curve parameters were not identifiable was ignored in the analysis.

Peaks with amplitude or frequency close to zero do not have levels that significantly

change with time and thus do not effectively behave periodically. These peaks were ignored. Out of the 258 peaks, 159 (61.6%) were found to have amplitude and frequency significantly different than zero (t-test, $\alpha = 0.05$) and to be in the reasonable range of $-\pi$ to π (Quinn and Thomson, 1991). Since $sin(\beta t+\gamma) = sin(-\beta t-(\gamma+\pi))$, meaning that models with an apparent negative β can be transformed to an equivalent model with a positive β, the absolute value of $\hat{\beta}$ was used for all peaks in the clustering procedure. Also, clustering was performed using log-transformed values so that the normality assumption was better satisfied. Variances for the log-transformed values were found using the delta method (Casella and Berger, 2002). The EM algorithm was implemented to find clusters of peaks, with the number of total clusters ranging from one to five. Results were given in terms of period τ, where $\tau = 2\pi/|\beta|$, so that the results could be more easily interpreted and compared.

For each number of clusters, the log-likelihood was calculated using the resulting parameter estimates. Since the log-likelihood increases as the number of clusters increases, it cannot be used to determine which model is best. Instead, the Bayesian Information Criterion (BIC), which penalizes models based on the number of parameters they include, was used (Schwarz, 1978). Although all necessary conditions for the BIC are not met by mixture models, the use of the BIC in this context is supported (Fraley and Raftery, 1998). The BIC was calculated as $2\ell(\theta) - m\log(n)$, where $\ell(\theta)$ is the log-likelihood, m is the number of parameters, and n is the number of observations, or in our case, peaks. The number of clusters corresponding to the first relative maximum of the BIC values is recognized to be the best (Fraley and Raftery, 1998). The three- and four-cluster results were found to have BIC values that were almost identical (-324.5 and -324.9, respectively), so the simpler three-cluster model was determined to be the best (Table 3.1).

Details of the results are given in Table 3.1. In the table, the means of the clustered values are labeled as "$\hat{\mu}$" and are given for each cluster. The transformed mean values,

Table 3.1: Analysis of blood plasma data, including extra variation[a]

Number of Clusters	Cluster Number	N (% of 258)	$\hat{\mu}$ (Period)	$\hat{\sigma}^2$	$\hat{\pi}$	BIC (ℓ)
1	1	159 (61.6)	-1.420 (36.3)	0.666	1	-403.2 (-196.5)
2	1	47 (18.2)	-2.505 (77.1)	0.011	0.285	-364.3 (-169.5)
	2	112 (43.4)	-0.991 (19.6)	0.288	0.715	
3	1	51 (19.8)	-2.470 (74.9)	0.019	0.319	-324.5 (-142.0)
	2	64 (24.8)	-1.227 (21.4)	0.0004	0.358	
	3	44 (17.1)	-0.589 (12.8)	0.244	0.324	
4	1	51 (19.8)	-2.468 (74.8)	0.019	0.321	-324.9 (-134.6)
	2	67 (26.0)	-1.222 (21.3)	0.002	0.414	
	3	25 (9.7)	-0.668 (12.3)	<0.0001	0.134	
	4	16 (6.2)	-0.213 (8.8)	0.237	0.132	
5	1	36 (14.0)	-2.568 (81.9)	<0.0001	0.220	-333.0 (-131.0)
	2	14 (5.4)	-2.229 (58.4)	<0.0001	0.102	
	3	68 (26.4)	-1.222 (21.3)	0.002	0.413	
	4	25 (9.7)	-0.668 (12.3)	<0.0001	0.134	
	5	16 (6.2)	-0.212 (8.7)	0.236	0.132	

[a] Of 258 possible peaks, 159 (61.6%) met all criteria and were used in the analysis.

corresponding to the periods for the clusters, are given in parentheses. Also given are the values for the variance, $\hat{\sigma}^2$, and the mixing probability, $\hat{\pi}$, for each cluster. The mixing probability represents the percentage of 159 peaks that fall into each cluster, and the percent of the total 258 peaks is listed as well.

As Figure 3.2 shows, there is a clear separation of clusters when the extra variation is taken into account. With mean periods of 21.4 and 12.8, the peaks in two of the clusters have patterns that repeat roughly once or twice per day, respectively. The other cluster has mean period of 74.9 hours, or about three days. Since the data spans only 24 hours, the periodic pattern seen in this cluster is outside the scope of the data. Thus, a total of 41.9% of the peaks express periodic behavior which can be explained using 24-hour data: about 24.8% of the 258 peaks have the once-per-day pattern, and 17.1% have a pattern that repeats about twice per day.

Heatmaps provide visual representations of data, assigning a range of colors to values so that the more similar values are, the more similar their colors are as well (Eisen et al., 1998). This type of graphic could give us an idea of how well our clustering

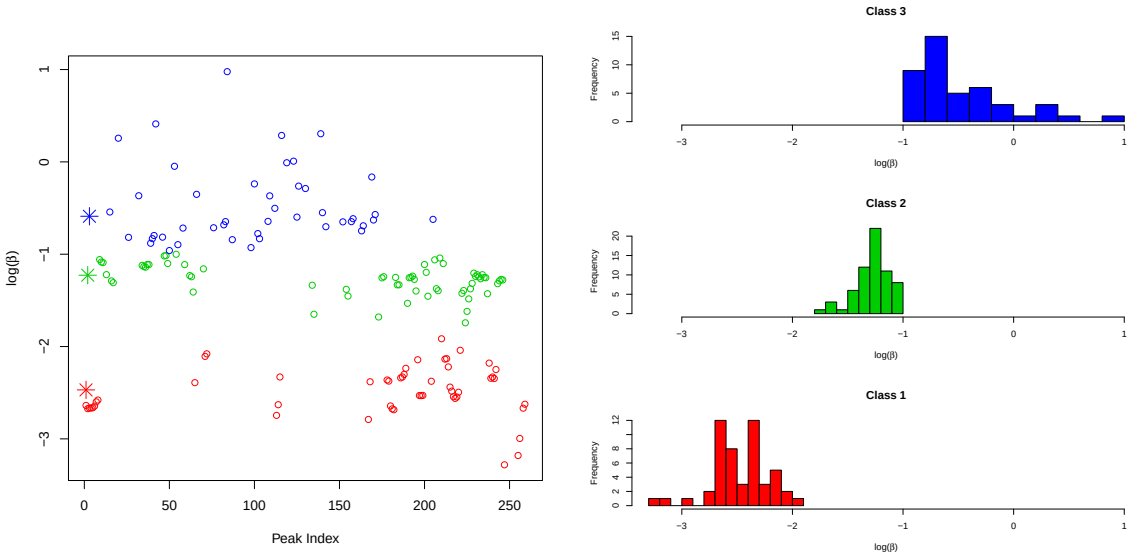

Figure 3.2: Scatter plot and histogram of $log(\beta)$ (periodicity parameter) for three clusters. The '*' for each cluster denotes the location of the mean.

algorithm grouped similar peaks by allowing us to view the behavior of the peaks within each cluster over time. Heatmaps of the processed data for the 24 hours of our study, averaged over the eight subjects, were produced for each cluster (Figure 3.3). Each heatmap has time in order across the horizontal axis. The individual peaks make up the rows and have been rearranged based on hierarchical clustering results to aid in visual interpretation. The heatmap for Class 2, which has a period of 21.4, clearly shows a strong pattern across the 24 hours (Figure 3.3). The patterns within Class 3 are not as clear, perhaps due to different phase shifts in the data. It does, however, look as if a twice-repeating pattern is visible in Class 3, which was found to have a period of 12.8 hours. Even though the presence of a 74.9-hour period for Class 1 cannot be evaluated with data spanning only 24 hours, a strong pattern appears to exist in this cluster. These peaks have levels that increase throughout the course of the day and return to lower levels the following morning.

Figure 3.4 displays a pre-processed spectrum along with the peaks identified as being in Classes 1, 2, and 3. It appears that there are overlapping regions for the three

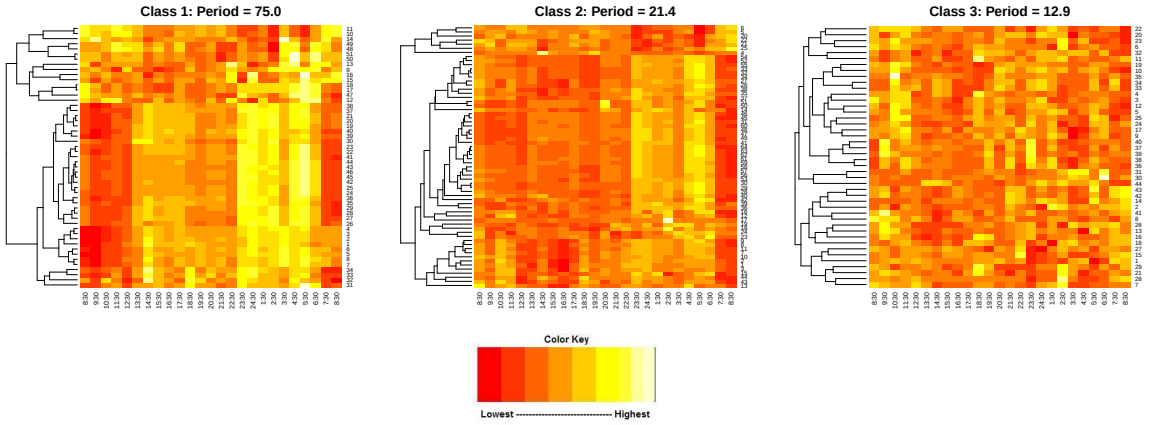

Figure 3.3: Heatmaps of observed data for three classes, averaged over eight subjects. Times (in order) are on the horizontal axis, and peaks are on the vertical axis.

classes, but there are regions that seem closely related to the classes as well. In their 1995 article, Nicholson et al. identify resonances from ^{1}H-NMR spectroscopy which can be assigned to particular metabolites (Nicholson et al., 1995). One can compare Figure 3.4 and the figure in which Nicholson et al. label identified metabolites in an NMR spectrum (Figure 1 from Nicholson et al. (1995)).

This comparison reveals that peaks in the region of glucose-β1 are mainly found in Classes 2 and 3, and most of the peaks in the glucose-β4, glucose-β2, and choline region are found in Class 3. The region of albumin lysyl looks most closely matched by Class 2, and it appears that Classes 1 and 2 contain the peaks in the lipid, acetyl signals from α_1-acid glycoprotein (NAC1 and NAC2), VLDL, LDL, HDL, valine, and alanine regions. The peaks for lactate appear to be in Class 3.

3.5 Simulation Results

As detailed above, the clustering method accounted for the extra variability associated with the estimation of periodicity. Simulation studies were used to evaluate methods of clustering parameter estimates when there is an extra variance component due to

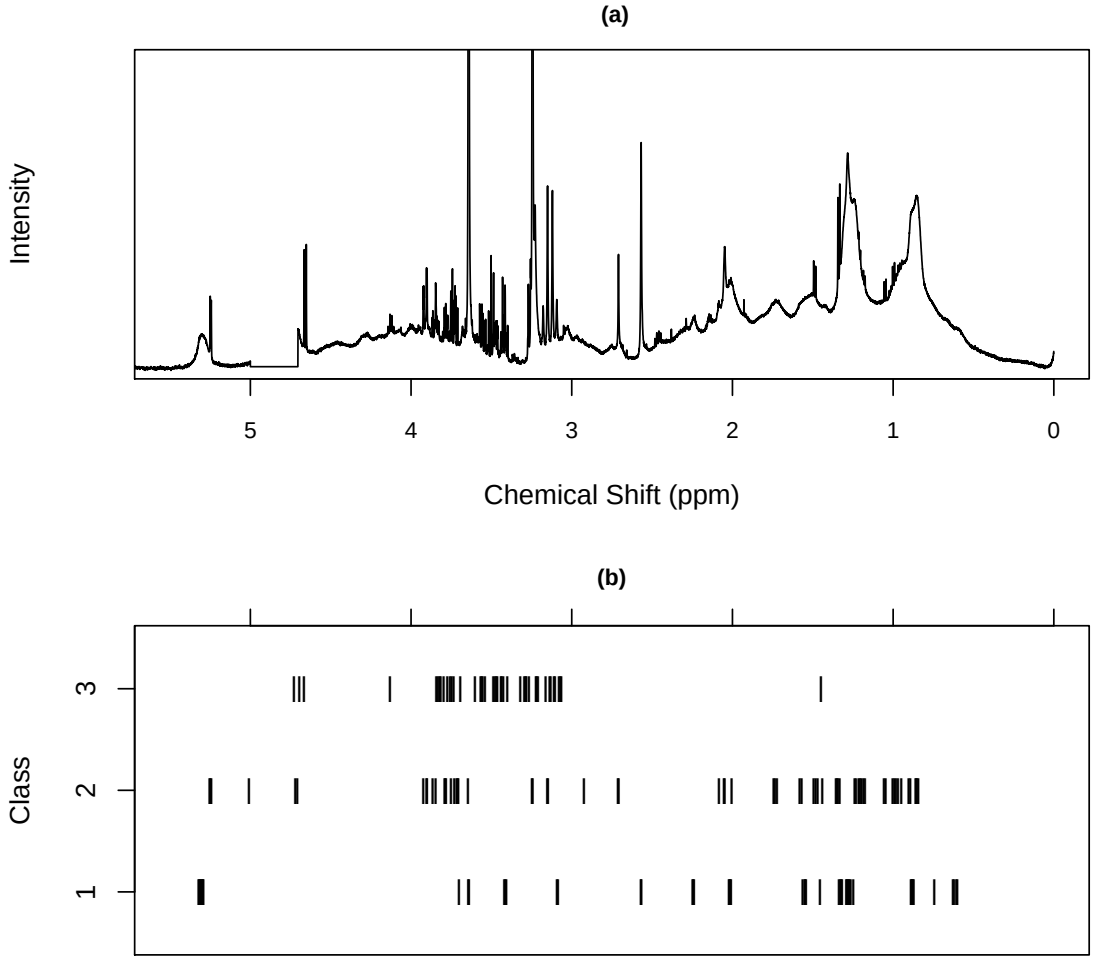

Figure 3.4: A pre-processed spectrum (a) is shown with locations for peaks in the four classes (b) so that actual locations of the class peaks can seen. Pre-processed spectrum has been truncated to reveal detail.

the fitting of these estimates. Two cases were considered: one method included the variation due to estimating the parameters as a known extra variance component when clustering, and the other method ignored this extra variation. The goal was to determine whether or not the variation of parameter estimates should be considered in the clustering process. Data for two clusters were generated using difference distances between cluster centers (1, 5, and 10) and different pairs of cluster variances ($\sigma_1^2 = 1$ with $\sigma_2^2 = 2$, $\sigma_1^2 = 1$ with $\sigma_2^2 = 0.1$, and $\sigma_1^2 = 0.1$ with $\sigma_2^2 = 0.1$). Three different χ^2 distributions were used to generate the extra variance component, denoted by s_p^2: $\chi^2_{0.25}$, $\chi^2_{0.5}$, and χ^2_1. For each combination of cluster center difference, variance pair, and extra variance distribution, values for 5000 peaks were generated and were clustered both including and ignoring the extra variance component.

Each of the 5000 peaks was randomly classified as being in cluster one or cluster two with probabilities 0.7 and 0.3, respectively, for each simulation. For a cluster center difference of one, cluster one values were generated from the distribution $N(-0.5, \sigma_1^2 + s_p^2)$, and cluster two values were generated from $N(0.5, \sigma_2^2 + s_p^2)$. For a cluster center difference of five, these generating distributions were $N(-2.5, \sigma_1^2 + s_p^2)$ and $N(2.5, \sigma_2^2, s_p^2)$ for the two clusters. For a cluster center difference of ten, the distributions $N(-5, \sigma_1^2 + s_p^2)$ and $N(5, \sigma_2^2 + s_p^2)$ were used.

Table 3.2 shows that the misclassification rates decrease as cluster centers get farther apart and that they increase as the degrees of freedom for the s_p^2 distribution increases, whether or not the extra variation is accounted for. Also, misclassification rates are smaller when the clusters have less variation, regardless of method. However, the misclassification rates for the simulations accounting for the extra variation are always less than or comparable to those in which the extra variation is ignored. This difference becomes less pronounced as the distance between cluster centers increases. Also, $\hat{\sigma}^2$ is consistently overestimated when this extra variation is ignored, as one would expect (see Appendix A.1). Again, the differences in estimates from the two

methods are more pronounced for tighter clusters, and differences decrease as the cluster centers become more spread out. For more simulation details, please see Appendix A.1.

Table 3.2: Misclassification rates for simulation scenarios with different distances between clusters, cluster-specific variances, and distributions for known peak-specific variance components

Variance Information		Including s_p^2			Ignoring s_p^2		
(σ_1^2, σ_2^2)	df[a]	D[b]= 1	D = 5	D = 10	D = 1	D = 5	D = 10
(1, 2)	0.25	0.250	0.024	<0.001	0.256	0.025	0.001
	0.5	0.258	0.029	<0.001	0.276	0.032	0.001
	1	0.263	0.045	0.002	0.292	0.046	0.003
(1, 0.01)	0.25	0.227	0.006	0	0.237	0.009	0.001
	0.5	0.244	0.014	<0.001	0.262	0.015	0.001
	1	0.272	0.023	0.001	0.517	0.024	0.002
(0.1, 0.1)	0.25	0.098	0.004	0	0.307	0.004	0
	0.5	0.146	0.006	<0.001	0.310	0.007	<0.001
	1	0.193	0.015	<0.001	0.675	0.017	0.001

5000 values were generated for each combination of parameters.
[a] $s_p^2 \sim \chi^2$ with degrees of freedom = df.
[b] D = Difference in cluster centers

3.6 Discussion

Presented is a method of clustering NMR peaks so that peaks within a cluster would have similar periodic behavior but could have heterogeneous amplitudes and phase shifts. In this method, the variation due to estimation of the period-related parameter estimates is treated as an extra, peak-specific variance component in addition to the cluster-specific variation. This clustering method is used for the analysis of NMR data in which metabolic behavior over the course of a day was observed. The general periodic behavior of the metabolites is of interest, the goal being to find groups of peaks with similar daily patterns in terms of periodicity, regardless of phase shift. It was determined that three clusters are present in the data, two with patterns repeating about once or twice per day, and one with a period of 74.9 hours, close to

three days.

Certain metabolites may have levels that rise and fall more than once per day, and these metabolites may be found in the cluster identified as having a period of about 12 hours (Class 3). This class could contain glucose, lactate, and choline. Other metabolites may have levels that rise and fall once per day, which would correspond to the pattern seen for Class 2. This group could include lipids, VLDL, LDL, HDL, and valine, and albumin. The heatmap for Class 1 shows a consistent pattern of increase throughout the day and then a dramatic decrease. The metabolites identified as being in this cluster appear to have a 24-hour pattern, but the pattern of increase until about 6:30 am, followed by an abrupt decrease, may not be modeled appropriately with a sine curve. Peaks from the substances in Class 2, representing substances cleared about once per day, could also be present in this group. Another explanation for Class 1 having such a large period, which exceeds the scope of the data, could be that subjects were going through an equilibration period caused by a change in diet. The 74.9-hour pattern could indicate that the shift toward a new equilibrium for metabolites in this cluster lasts about three days.

Previous work suggests that the glucose pattern might be different than other metabolite patterns since glucose is dependent on food intake, which supports the idea that glucose would be cleared multiple times per day and fall into Class 3. Other metabolites might increase and then be cleared once per day and fall into Classes 1 and 2 (Park et al., 2009). These classes contain lipids and lipoproteins, which are related to slower metabolic processes. The speed at which these metabolites are cleared might differentiate the substances with peaks found in Class 1 from those with peaks in Class 2, but the underlying pattern for these classes appears to be similar. This metabolite information could prove to be useful, but linking metabolite peaks to their identities more strongly could provide better information on the metabolism of the human body and aid in diagnosing and treating metabolic and eating disorders.

Also, in this study, classification is based on the most likely cluster for each feature. While in genetic studies it might be of interest to allow genes to belong to multiple clusters or pathways, allowing a feature to belong to more than one group based on periodicity does not make sense using our model. Some features may have ambiguous cluster membership, however, and these features should be further investigated.

Simulation studies were used to determine whether or not it is appropriate to cluster parameter estimates treating the variation of the estimates as an extra variance component. Misclassification rates are as good or better for the proposed method when compared to the other method for all scenarios tested. Also, estimates are often more accurate for the proposed method. Differences in these estimates and in misclassification rates are more pronounced for tighter clusters, and the two methods produce more similar results as cluster centers become farther apart.

The peaks identified as displaying periodic patterns could serve as a window into the metabolic patterns of the human body. There is a possibility that knowledge of the behavior of NMR peaks of metabolites under normal conditions could help in the diagnosis and treatment of metabolic and other disorders and aid in determining the level of success achieved by various interventions such as surgeries and dietary changes. However, the dataset used for this study is limited to only 25 measurements for each of eight people, therefore there is not an abundance of information with respect to periodicity. We chose to use a sine curve due to its simplicity and to the fact that empirical data supported this choice. Relaxing the parametric assumption of the sine curve could be possible, but observations spanning a longer period of time would be needed. A non- or semi-parametric model could provide better behavioral details for all classes, especially for the class with an apparent 75-hour periodic pattern, so this approach should be explored in the future.

Chapter 4

Using Mass Spectra to Determine Possible Ages of Metabolic System-Level Shifts

4.1 Introduction

The aging process of organisms, generally considered a change from fully functional at younger ages to a state of declining or failing function later in life, is not thoroughly understood and has been investigated for decades (Soltow et al., 2010). Studies of aging often involve non-human organisms such as worms, mice, and flies, but these organisms do not live long enough for a successful aging process similar to that of humans to be examined (Austad, 2009). An appropriate and convenient species for the study of long-term aging would be small and manageable, would live for a long time for its body size, and would have certain cognitive similarities to humans (Austad, 2009; Soltow et al., 2010). For these reasons, the common marmoset (*Callithrix jacchus*) has emerged as a useful tool in this area. This small primate is the size of a rat, lives an average of seven to ten years, and lives a maximum of 16 years. Also, it has stronger cognitive abilities than rhesus monkeys, the species most commonly used in non-human aging studies to date (Austad, 2009).

There are many hypotheses regarding the causes and effects of aging, and studies

of marmosets have been used to determine developmental stages of the animals (de Castro Leão et al., 2009), to examine neurogenesis (Leuner et al., 2007), and to characterize oxidative stress over time (Terman and Brunk, 2006). Other studies list inflammation and hormone dysregulation as possible causes for aging, but it could be a combination of these factors and more (Soltow et al., 2010). It is thought that information available using genomics, proteomics, and metabolomics could aid in the study of aging. Since the metabolome is the furthest downstream from genetic information and thus could represent gene function as well as environmental factors, the study of metabolites could be particularly useful (Soltow et al., 2010).

Due to the possibility that the metabolome could provide an abundance of information on aging, this study focuses on the relationship between marmoset age and blood plasma metabolite levels. Blood was drawn from the 76 members of a marmoset colony cross-sectionally. Metabolomic data was obtained from the samples with liquid chromatography coupled with high resolution Fourier transform mass spectrometry (LC-FTMS), using anion exchange (AE) chromatography (Soltow et al., 2011). MS is used to determine the mass-to-charge ratios (m/z) of ions and quantify those ions, with resulting intensity corresponding to the number of ions with a particular m/z value. The use of liquid chromatography results in retention times for each data point. LC-FTMS spectra thus consist of data points with values for m/z, retention time, and intensity. Using retention time information results in more accurate data, and the more precise m/z values can be used for feature identification (Yu et al., 2009).

In preliminary analyses presented in Soltow et al. (2010), marmosets are divided into "old" and "young" using a cut-off of six years of age, and metabolite characteristics for the two groups are compared. Instead of dividing marmosets into groups in this manner, our goal is to identify marmoset ages at which levels of metabolites change. These ages may not be the same for all groups of metabolites. We propose first

using a segmentation algorithm to find the locations (ages) at which mass spectra (MS) intensity changes occur. MS features with similar breakpoint locations will then be identified. The hope is that breakpoints, representing system-level shifts of the metabolic system at particular ages, can be identified. Markers associated with age-related changes could provide useful information about the aging process and could be helpful in the development of personalized medicine.

4.2 Pre-Processing of the Data

As mentioned above, blood was drawn from the 76 members of a marmoset colony, and metabolomic data was produced from resulting plasma samples with LC-FTMS, using anion exchange (AE) chromatography (see Soltow et al., 2011, for more details). Of the 76 marmosets in the colony, 61 were round to be healthy. Two spectra were available for 60 of the 61 marmosets, and the remaining marmoset had only one spectrum available. Mass spectra were processed using the method of adaptive processing of high-resolution LC-MS data (Yu et al., 2009) available in the apLCMS R package (Yu, 2011; R Development Core Team, 2008). This processing step allowed for more accurate feature detection, quantification, and alignment.

After initial processing, up to 3238 features were present for each spectrum. An intensity value of zero was interpreted as absent, so several techniques were used to address these values. Only features found in at least 75% of the spectra were kept, for a total of 1058 (32.7%) MS features. At this point, over 92% of the values were not equal to zero. Spectra from the same marmoset were then "merged," meaning the average intensity was taken if two intensities were present for a feature, and the non-zero intensity was taken if one of the values was absent. After this step, close to 97% of the values were not equal to zero. K-nearest neighbor imputation was then used to impute the remaining missing values (Hastie et al., 2011). Next, quantile

normalization was used to force all of the marmosets to have the same distribution of values (Bolstad, 2010). For each feature, a normal score transformation was then used to reduce the impact of outliers (Hothorn et al., 2006, 2008).

4.3 Segmentation and Clustering

After data pre-processing, there were 1058 MS feature intensities for each of 61 marmosets. The heatmap in Figure 4.1 shows the different features as the rows, with marmosets in age order as the columns. For some features, it looks as if intensity levels change at certain ages. We want to consider each feature individually and find locations of breakpoints that would identify marmosets of some ages as having different intensity levels than those of other ages. For example, if we found a breakpoint at 1000 days in a feature, marmosets older than 1000 days old would have a different intensity for that particular feature than younger marmosets.

4.3.1 Segmentation

Our first goal is to detect breakpoints, also called changepoints or steps, in the data. Several different types of breakpoints exist. For some data, it may be expected that a linear model would fit the data well, perhaps with a change in slope at a particular breakpoint (Seber and Lee, 2003). This model can be represented as:

$$Y = \begin{cases} \alpha_1 + \beta_1 X + \epsilon & \text{if } x \leq \gamma \\ \alpha_2 + \beta_2 X + \epsilon & \text{if } x \geq \gamma \end{cases} \tag{4.1}$$

where the α parameters give the intercepts of the lines, the β parameters give the slopes of the lines, and ϵ is the error term ($\epsilon \sim N(0, \sigma^2)$). Parameters α_1 and β_1 apply where input data point x is less than or equal to breakpoint γ, and parameters α_2 and β_2 apply otherwise. Using this model, the slope of the line changes at breakpoint

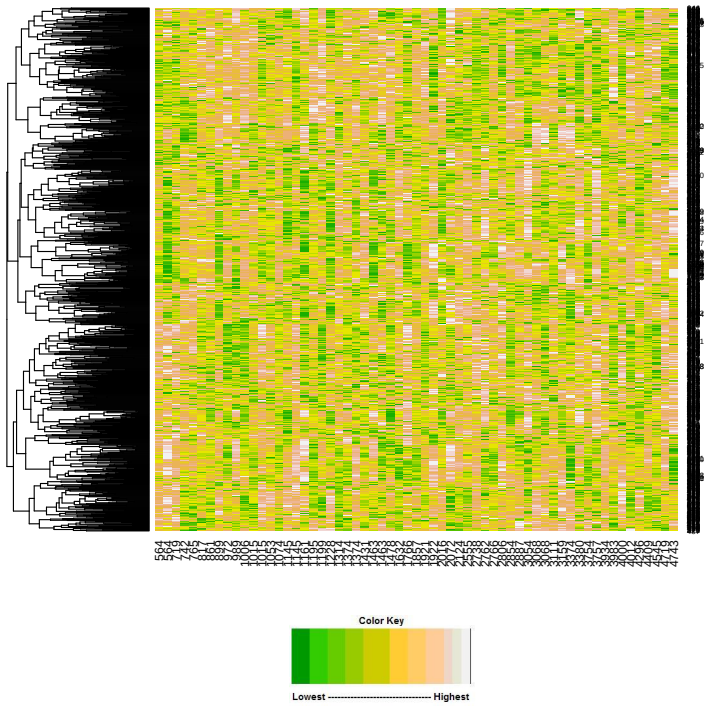

Figure 4.1: Heatmap with reordered features as rows and marmosets in age order as columns.

γ, and the two lines are forced to meet at this breakpoint.

However, for the marmoset data, we expect mean intensity levels to shift at certain times due to biological changes. Evidence of this type of shift can be seen in Figure 4.1. Thus, a slope of zero on either side of a breakpoint between changing mean intensity levels is expected, and the model becomes a piecewise constant linear model (Gustafsson, 1994; Hawkins, 1976):

$$Y = \begin{cases} \alpha_1 + \epsilon & \text{if } x \leq \gamma \\ \alpha_2 + \epsilon & \text{if } x > \gamma. \end{cases} \tag{4.2}$$

Here, the breakpoint γ indicates the location of a change in mean intensity level. More generally, if more than one breakpoint is present in the data, the model can be

written as:

$$Y = \begin{cases} \alpha_1 + \epsilon & \text{if } x \leq \gamma_1, \\ \alpha_j + \epsilon & \text{if } \gamma_{j-1} < x \leq \gamma_j \text{ and } 1 < j < J, \\ \alpha_{J+1} + \epsilon & \text{if } x > \gamma_J. \end{cases}$$

In this model, γ_j is the j^{th} breakpoint ($j = 1, 2, ..., J$), and the ϵ terms represent independent errors with mean zero. The total number of segments created by the J breakpoints is $J + 1$, and α_j represents the mean value for segment j.

In our situation, the number of breakpoints for a particular feature is unknown, and different features could have a different number of breakpoints. Aiming to find this type of breakpoint makes our goal similar to the goal of finding breakpoints in chromosomes which indicate locations of change in DNA copy numbers. This practice has been used to identify chromosomal regions of malignant tumors in which genetic material has been gained or lost, which could lead to the identification of genes involved in tumor suppression or progression (Hupé et al., 2004). Several segmentation algorithms have been used for this type of data (Yu et al., 2007) and are available as packages in R. The marmoset data is very different than the genomic data for which the algorithms were intended. For this reason, an updated version of the Gain and Loss Analysis of DNA (GLAD) method proposed in Hupé et al. (2004) was chosen for the detection of breakpoints in the marmoset data due to the availability of necessary tuning parameters. These parameters allow the user to adjust the sensitivity to changes in mean intensity levels, controlling how easily breakpoints are identified. This method, available in the GLAD R package, allows breakpoints to be detected without specifying the hypothesized number of breakpoints (Hupe, 2009).

The implementation of the GLAD method involves several steps. First, the marmoset data are smoothed, and breakpoints are detected using a piecewise constant function. Regions of constant intensity levels are identified within each feature. The number of breakpoints is then optimized, and any undesirable breakpoints representing very

minor changes in small regions are removed. The last step, which involves finding segments among all features with homogeneous intensity levels, does not help fulfill the goal of our study and will be ignored.

In the GLAD method, data can be smoothed and segmentation can be performed using several algorithms. These include extensions of the Alternative Weights Smoothing (AWS) procedure introduced by Polzehl and Spokoiny (2000), as well as an algorithm using Haar segmentation developed by Ben-Yaacov and Eldar (2008). The faster Haar segmentation was chosen for our data.

Haar segmentation involves applying an undecimated discrete wavelet transform (UDWT) using the Haar wavelet to the data. This results in a decomposition of the signal including a "smooth" subband and a set of detail subbands at different scales (Burrus, 1998; Hastie et al., 2001; Ben-Yaacov and Eldar, 2008). It is assumed that some of the levels of detail can be ignored, resulting in a representation of the data as a piecewise constant signal when reconstructed from the remaining levels. The levels of detail, or detail subbands, chosen to be retained influence how sensitive to noise the piecewise constant signal will be. If finer detail subbands are included, the algorithm becomes more sensitive to noise. This could result in very small segments of data being identified. As finer subbands are excluded, the algorithm is less sensitive to noise but is more likely to miss breakpoints separating small sections. The coarsest detail subband level should be large enough to not be overly sensitive to noise but small enough so that very minor changes are not detected.

The detail subbands are composed of wavelet coefficients. These coefficients provide information about the differences between mean intensity levels in consecutive regions and are used to determine breakpoint locations. Haar wavelet coefficients $w_{S,n}$ are

given by (Ben-Yaacov and Eldar, 2008):

$$w_{S,n} = \frac{1}{\sqrt{2^{S+1}}} \left(\sum_{k=n}^{n+(2^S-1)} y_k - \sum_{k=n-2^S}^{n-1} y_k \right)$$

where S is the detail subband and y_n is the n^{th} data point. The equation for $w_{S,n}$ can be rewritten as a difference in means of two regions:

$$w_{S,n} = \sqrt{2^{S-1}} \left(\frac{1}{2^S} \sum_{k=n}^{n+(2^S-1)} y_k - \frac{1}{2^S} \sum_{k=n-2^S}^{n-1} y_k \right)$$

As S increases, the window size of the regions being compared widens and the level of detail decreases. If there is no difference between the regions being compared, then the coefficient will be zero.

Relative maxima of the absolute values of the coefficients are found for each of the selected detail subbands, and each maximum is a breakpoint candidate. The maxima are thresholded for each subband separately using a false discovery rate (FDR) thresholding procedure (Benjamini and Hochberg, 1995). The number of coefficients kept as possible breakpoints is limited by the FDR threshold value, the larger coefficients being kept over the smaller ones.

The remaining maxima from all subbands are then grouped together to create a list of significant breakpoints in the data. Some breakpoints could surface in more than one detail subband with an offset, so redundant breakpoints should be removed. This involves taking the maxima from the level of finest detail, say S_{min}, and adding the maxima from S_{min+1} if they are not less than $2^{S_{min}-1} + 1$ measurements away from the maxima in S_{min}. For $S_{min} = 2$, maxima would be added from $S = 3$ if they were greater than $2^{2-1} + 1 = 3$ measurements away from the maxima found in $S = 2$. This process of compiling breakpoints found in the different detail subbands is repeated until maxima from all subbands, given that they meet the criteria, have been added to the breakpoint list. For more details on the Haar smoothing and segmentation,

see Ben-Yaacov and Eldar (2008).

After the GLAD method of segmentation implements Haar smoothing and segmentation to identify significant breakpoints in the data, there is a step for additional filtering of the breakpoints. Breakpoints creating small regions with small changes in intensity can be removed, as the number of breakpoints is optimized using a method based on the penalized log-likelihood. The hypothesized region size can be given, requesting that breakpoints creating regions with fewer members than the region size specified be eliminated except in cases of strong evidence. There is also a step for optimizing the number of breakpoints, which uses a penalty term to control the number of breakpoints found. Two parameters used in the penalty term, L and D, aim to eliminate unnecessary breakpoints while retaining true breakpoints. D is the parameter used in the kernel function in the penalty term, and L is the penalty term coefficient. For more information on the GLAD procedure, see Hupé et al. (2004).

After breakpoints are identified, the piecewise constant signal between two breakpoints is found by taking the average of all data points between the breakpoints. This allows the direction of change at a breakpoint, either increase or decrease, to be seen.

4.3.2 Classification

After finding a list of breakpoints at different marmoset ages for each of the MS features, it is of interest to find groups of features exhibiting similar changes with age. Although breakpoints that are identified in two features, say at 1000 and 1001 days old, may not be exactly the same, it is believed that the changes indicated by the two breakpoints could be biologically related. Kernel density estimation is used to identify clusters of intensity change locations. Density relative maxima are the locations at which change is most likely. Relative minima are used as the cut-offs for

the different clusters.

4.4 Application to Marmoset MS Data

Our data includes mass spectra (MS) from 61 marmosets of various ages. Since some ages were repeated, 0.1 was subtracted from the first marmoset with a repeated age so that the exact location of a breakpoint could be determined. The Haar segmentation algorithm was used to find age breakpoints in the individual MS features, with detail subbands being limited to levels two to five. It was determined that the penalty term coefficient L would be set to four and the penalty term kernel parameter D would be set to three since the number of breakpoints leveled off at this parameter setting 4.2. For more details on the selection of parameters, please see Appendix B.1.

The combination of parameters chosen resulted in 167 total breakpoints, with 115 features (10.9% of of 1058) having at least 1 breakpoint. A total of 74 (7.0%) features were found to have one break, 30 (2.8%) were found to have two breaks, and 11 (1.0%) were found to have three breaks. Of the 61 marmoset spectra, 39 (63.9%) were found to contain a breakpoint for at least one feature 4.1. These 39 spectra represented 38 unique marmoset ages. A total of 13 (21.3%) of these breakpoint locations occurred in only one feature, and 27 (44.3%) were found three or fewer features. On the other hand, the breakpoint at 4545 days occurred in 44 features, the breakpoint at 765 was found in 16 features, and the breakpoint at 4409 days was found in 11 features. Since 765 days and 4545 days are the minimum and maximum breakpoint ages and had more breakpoints than the other ages, a possible edge effect of the algorithm was investigated by eliminating marmosets with extreme ages. (Note that the minimum and maximum breakpoints are NOT the minimum and maximum marmoset ages, but the extremes of the breakpoints detected.) An abundance of breakpoints at the extreme breakpoint locations was no longer detected. A difference in the extreme

ages can be seen in the heatmap as well (Figure 4.1), thus it was determined that these breakpoints signified actual metabolite level changes instead of being artifacts of the algorithm used.

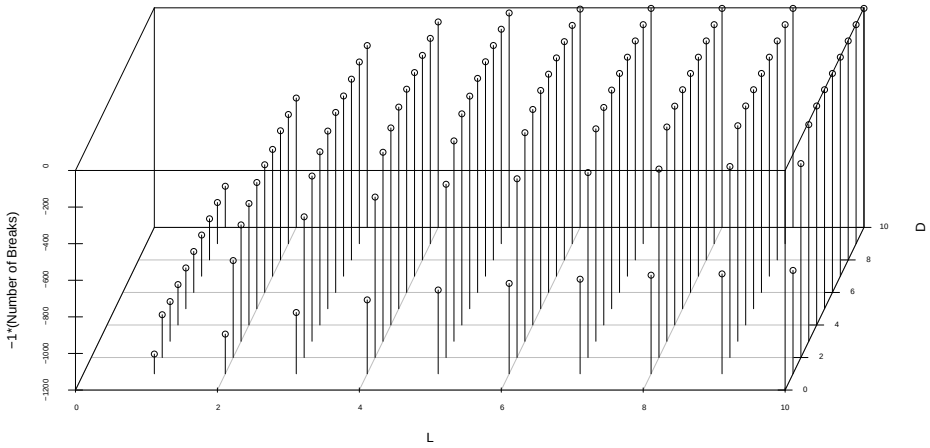

Figure 4.2: The total number of breakpoints (multiplied by -1) plotted against values for D and L.

The entire list of 167 breakpoints was considered when trying to find clusters in the data, meaning that some clusters could be identified that contained multiple breakpoints from the same feature. The kernel density of the log of the breakpoint locations was used to identify clusters. This density was estimated from 200 points using the method of Sheather and Jones (1991) to select the bandwidth, which was multiplied by two to reduce the level of detail captured by the kernel density to a reasonable level.

The relative maxima of the density were considered the most likely breakpoint locations, and the relative minima between the maxima were used as cut-off values to distinguish values in one cluster from another. Figure 4.3 shows the histogram of the log of the breakpoint locations with the kernel density. It can be seen that five relative maxima and thus five clusters were identified. The most likely breakpoints

Table 4.1: The number of features with breakpoints at the different marmoset ages. If breakpoints were found in the spectra of marmosets with non-unique ages, (1) and (2) are used to indicate the (arbitrary) ordering of these spectra.

Age in Days	Number of Features
765	16
817	2
899	1
989	1
1006	1
1015 (2)	1
1074	1
1161	2
1199	1
1228	1
1374 (1)	3
1431	1
1463 (1)	3
1463 (2)	1
1478	2
1632	2
1766	1
1857	5
1921	3
2016	5
2072	3
2124	6
2555	5
2738	3
2762	3
2766	3
2806	6
2854	2
3054	1
3374	4
3380	6
3754	1
3757	4
3934	6
3983	1
4000	2
4296	3
4409	11
4545	44

for the five clusters were found to be 771 days, 1448 days, 2006 days, 2749 days, and 4459 days, or roughly 2, 4, 5.5, 7.5, and 12.5 years. Figure 4.3 also shows how many features were found to have breakpoints at each of the breakpoint locations, indicating cluster membership for each of the breakpoint locations. A total of 19 breakpoints from 19 different features were found in Cluster 1, and 20 breakpoints from 18 features were found in Cluster 2. Cluster 3 and Cluster 4 both contain 23 breakpoints from different features. Cluster 5 contains the most breakpoints with 82, representing 72 different features.

In Figures 4.4 and 4.5, heatmaps of the features identified as having breakpoints in each of the five clusters are shown along with plots showing which breakpoints were identified as belonging to a each cluster. On the left side are the heatmaps, and on the right are plots of the breakpoint locations for the features in each cluster. Features (rows) are in the same order in the heatmaps and the breakpoint location plots so

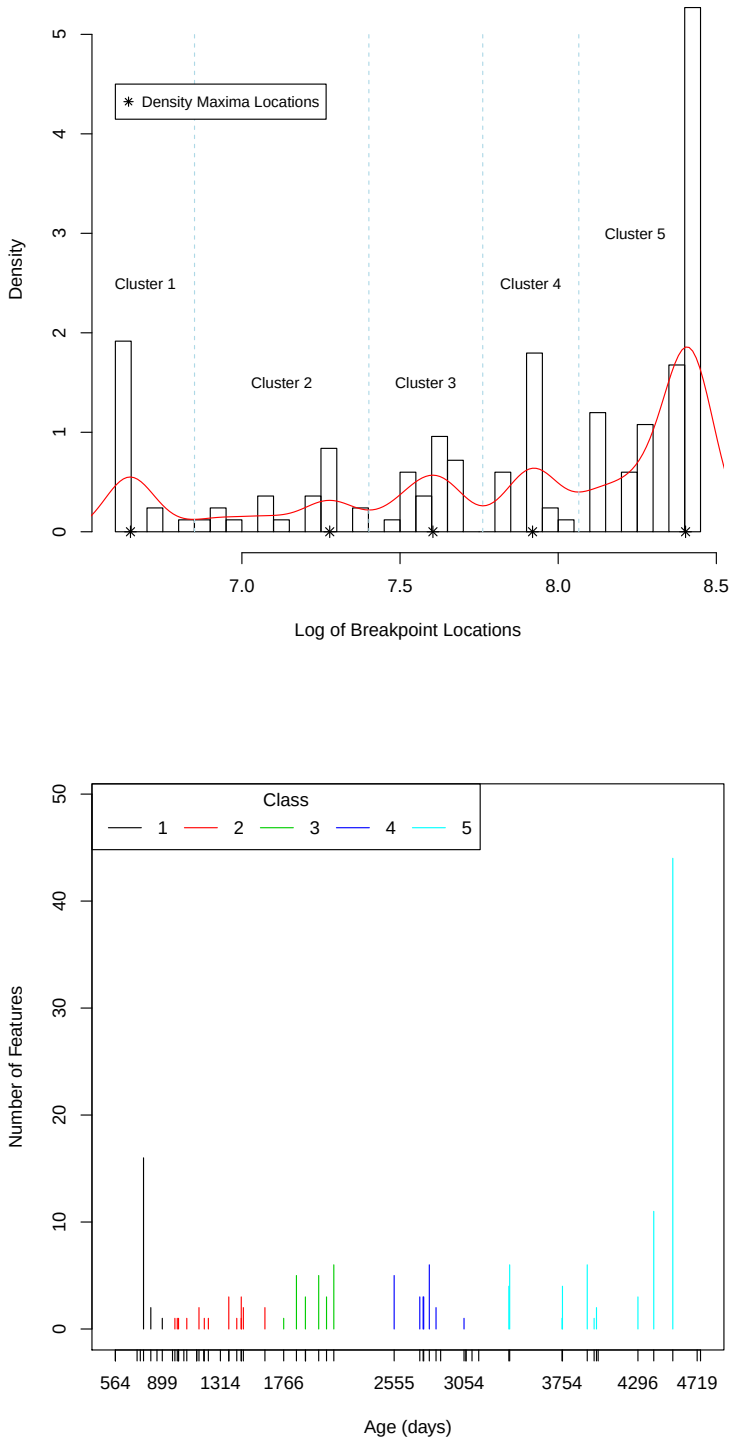

Figure 4.3: Kernel density plot showing five clusters of breakpoints identified along with plot of number of features with each of the breakpoints in each cluster.

that patterns can be compared. The breakpoints classified in the particular cluster of interest are shown in red. You can see that some of the features have more than one breakpoint in a given cluster and that features can show up in more than one cluster if they have more than one breakpoint.

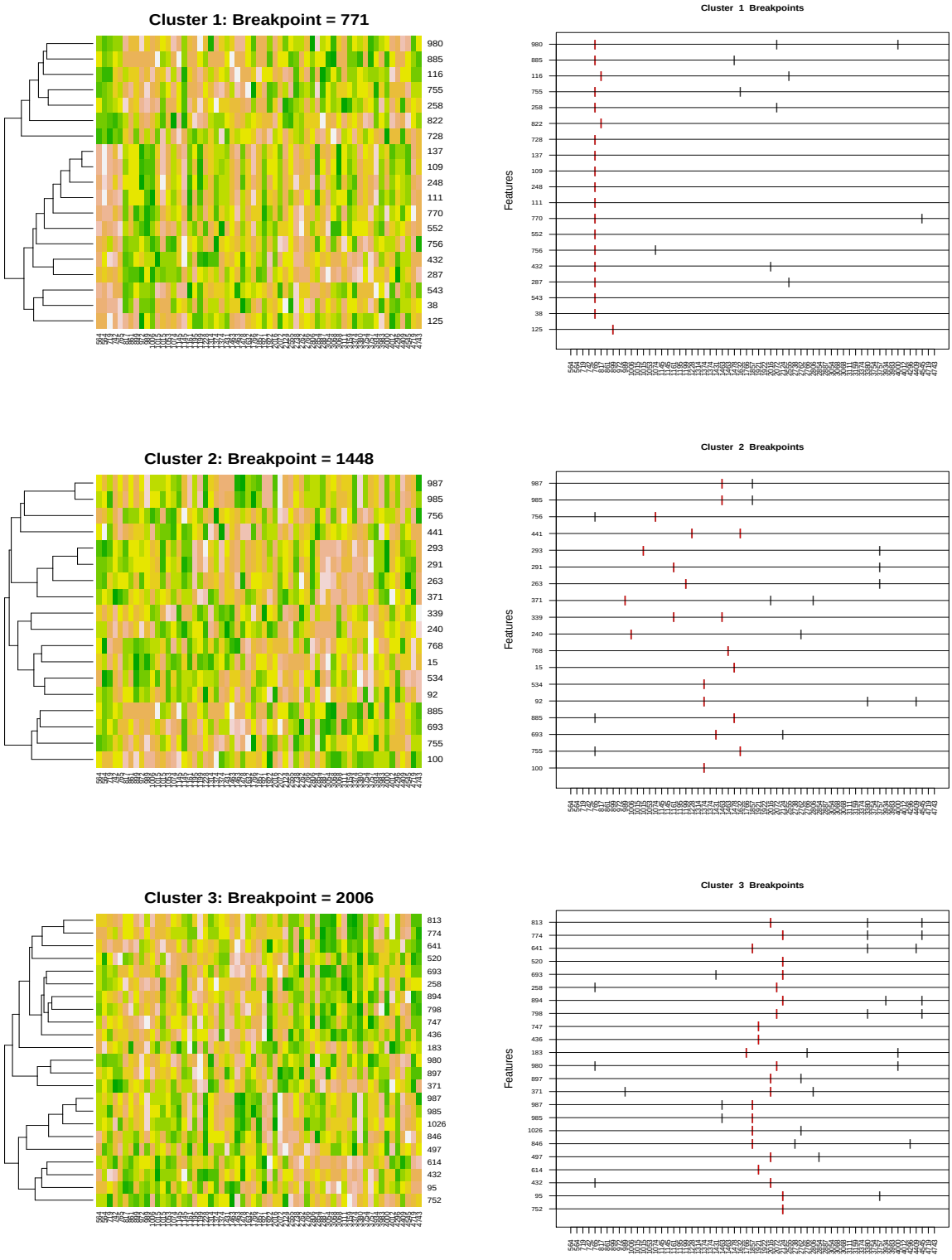

Figure 4.4: Heatmaps with breakpoint location information, Clusters 1-3.

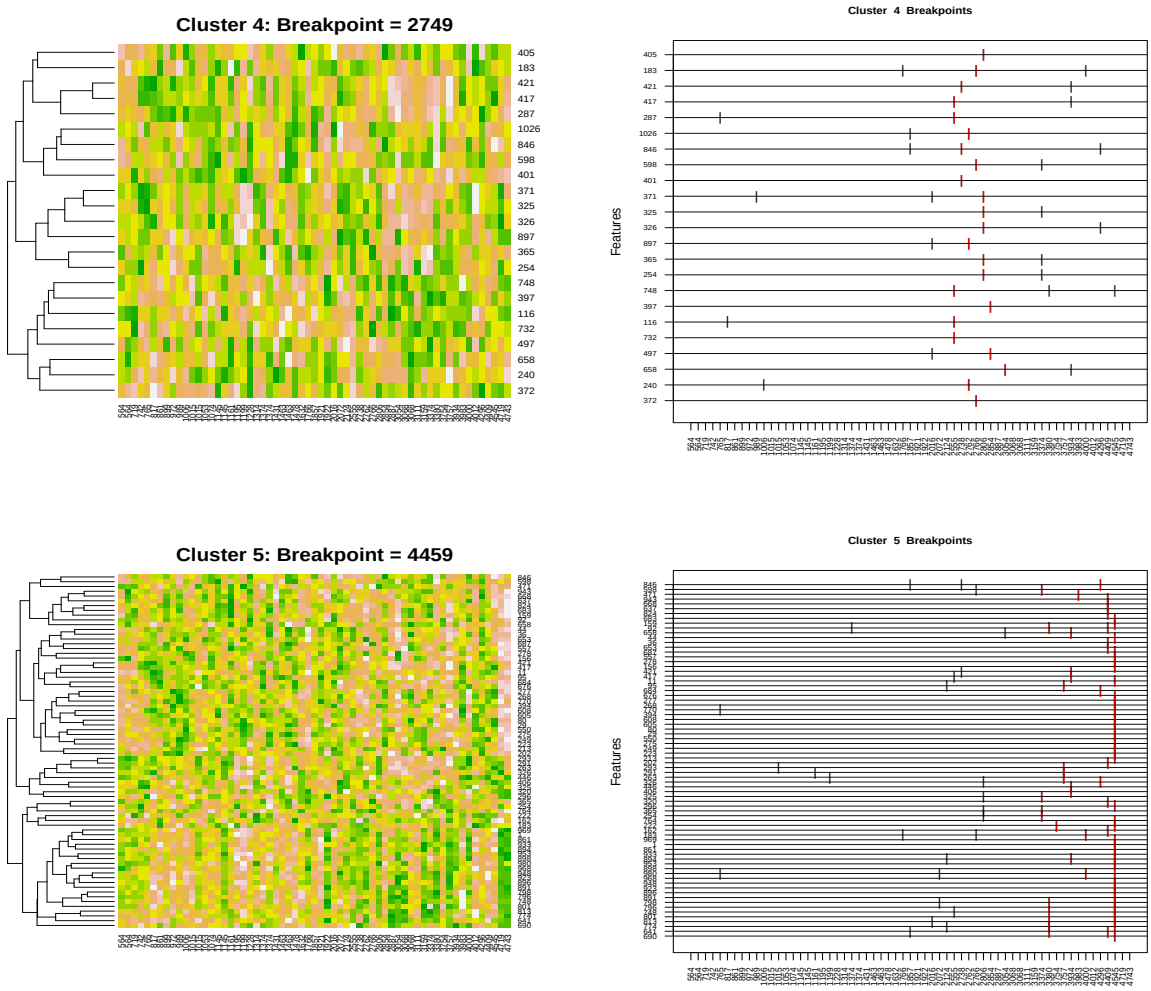

Figure 4.5: Heatmaps with breakpoint location information, Clusters 4-5.

While features are listed in Figures 4.4 and 4.5 using indices which range from 1 to 1058, these features are actually representative of mass-to-charge ratios (m/z) of the mass spectra. Figure 4.6 shows the average intensity across all marmosets for the different m/z values. Below the plot, features identified as having breakpoints in each of the five clusters are shown. The m/z values for the features found in each cluster were entered in the Madison-Qingdao Metabolomics Consortium Database (MMCD) in order to find metabolite matches (Cui et al., 2008). A subset of the results found can be seen in Tables 4.2 and 4.3, and full tables are available in Appendix B.2. Cluster 1

likely contains arginine, creatine and stearoylcarnitine appear to have been found in Cluster 2, and Cluster 3 could contain either aldosterone or cortisone, two steroids, as well as stearoylcarnitine. Emetine, vitamin D2, dipeptides, and tripeptides were likely to have been found in Cluster 4, and the Cluster 5 appears to contain many amino acids, dipeptides, tripeptides, and lysophospholipids. Not all m/z values in the clusters were found to have MMCD matches. Of the 115 features found to have breakpoints, only 51 (44.3%) had at least one match in the database. Many of those with at least one match had many possible matches (Tables 4.2 and 4.3, and Appendix B.2), making identification more difficult. From a biological standpoint, some of the possibilities were more probable than others, and these have been indicated in bold in Tables 4.2 and 4.3.

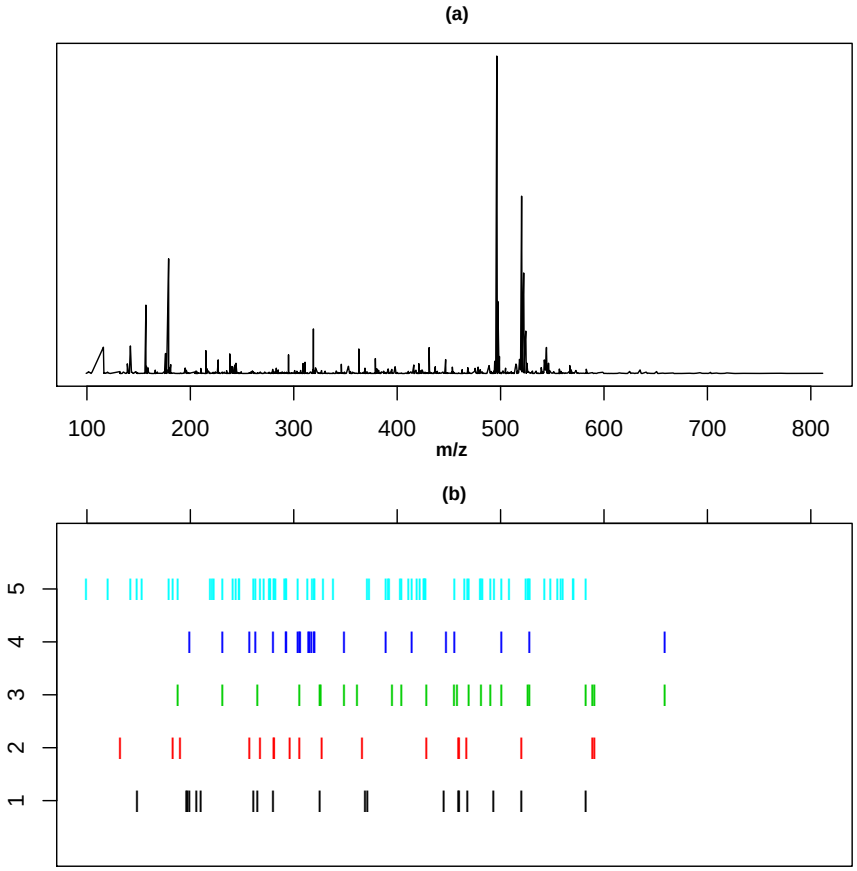

Figure 4.6: Average mass spectrum shown with locations of breakpoints from the five clusters.

Table 4.2: MMCD Matches for Clusters, Part 1. A tolerance of 10 ppm was used, and metabolites matched were of the [M+H]+ and [M+Na]+ varieties. Likely candidates for identification are shown in bold.

Cluster	Input Mass	Type	Database Mass	Formula	Name
1	197.1022449	[M+Na]+	174.1116757	C6H14N4O2	**L-Arginine;(S)-2-Amino-5-guanidinovaleric acid**
	197.1022449	[M+Na]+	174.1116757	C6H14N4O2	D-Arginine;D-2-Amino-5-guanidinovaleric acid
	459.3476132	[M+Na]+	436.35526	C27H48O4	5-b-Cholestane-3a-7-tetraol
	459.3476132	[M+Na]+	436.35526	C27H48O4	Cholestane-3,7,12,25-tetrol
	459.3476132	[M+Na]+	436.35526	C27H48O4	3alpha,7alpha,12alpha,26-Tetrahydroxy-5beta-cholestane; 5beta-Cholestane-3alpha,7alpha,12alpha,26-tetraol; 5beta-Cholestane-3alpha,7alpha,12alpha,26-tetrol
	459.3476132	[M+Na]+	436.35526	C27H48O4	5b-Cholestane-3a,7a,12a,23-Tetrol
2	132.0759201	[M+H]+	131.0694765	C4H9N3O2	**Creatine;alpha-Methylguanidino acetic acid;Methylglycocyamine**
	132.0759201	[M+H]+	131.0694765	C4H9N3O2	3-Guanidinopropanoate
	428.3692035	[M+H]+	427.3661591	C25H49NO4	**Stearoylcarnitine**
3	231.095881	[M+H]+	230.0902716	C9H14N2O5	TETRAHYDRODEOXYURIDINE
	231.095881	[M+H]+	230.0902716	C9H14N2O5	**Aspartyl-L-proline**
	231.095881	[M+H]+	230.0902716	C9H14N2O5	3,4-DIHYDRO-1H-PYRIMIDIN-2-ONE NUCLEOSIDE
	361.1976803	[M+Na]+	338.2093241	C19H30O5	Shiromodiol diacetate
	361.1976803	[M+Na]+	338.2093241	C19H30O5	Idebenone; drug
	361.1976803	[M+Na]+	338.2057365	C20H28F2O2	4,4-Difluoro-17beta-hydroxy-17alpha-methyl-androst-5-en-3-one
	361.1976803	[M+H]+	360.193674	C21H28O5	**Aldosterone;11beta,21-Dihydroxy-3,20-dioxo-4-pregnen-18-al**; drug
	361.1976803	[M+H]+	360.193674	C21H28O5	Cinerin II
	361.1976803	[M+H]+	360.193674	C21H28O5	**Cortisone;17alpha,21-Dihydroxy-4-pregnene-3,11,20-trione; Kendall's compound E;Reichstein's substance Fa**; drug
	361.1976803	[M+H]+	360.193674	C21H28O5	Prednisolone; drug
	361.1976803	[M+H]+	360.193674	C21H28O5	Tricyclodehydroisohumulone
	428.3692035	[M+H]+	427.3661591	C25H49NO4	**Stearoylcarnitine**
	481.3068156	[M+H]+	480.2988078	C29H40N2O4	Emetine
4	231.095881	[M+H]+	230.0902716	C9H14N2O5	TETRAHYDRODEOXYURIDINE
	231.095881	[M+H]+	230.0902716	C9H14N2O5	**Aspartyl-L-proline**
	231.095881	[M+H]+	230.0902716	C9H14N2O5	3,4-DIHYDRO-1H-PYRIMIDIN-2-ONE NUCLEOSIDE
	317.1130363	[M+Na]+	294.1255944	C19H18O3	METHYL (2Z)-3-METHOXY-2-2-[(E)-2-PHENYLVINYL]PHENYLACRYLATE
	317.1130363	[M+Na]+	294.1255944	C19H18O3	(2-Butylbenzofuran-3-yl)(4-hydroxyphenyl)ketone; 2-Butyl-3-(4-hydroxybenzoyl)benzofuran
	317.1130363	[M+Na]+	294.1215717	C14H18N2O5	Aspartame; drug
	317.1130363	[M+Na]+	294.1215717	C14H18N2O5	**Glutamylphenylalanine**
	317.1130363	[M+Na]+	294.1215717	C14H18N2O5	2-(BETA-D-GLUCOPYRANOSYL)-5-METHYL-1-BENZIMIDAZOLE

Table 4.3: MMCD Matches for Clusters, Part 2. A tolerance of 10 ppm was used, and metabolites matched were of the [M+H]+ and [M+Na]+ varieties. Likely candidates for identification are shown in bold.

Cluster	Input Mass	Type	Database Mass	Formula	Name
5	148.0593598	[M+H]+	147.0531578	C5H9NO4	3-METHYL-ASPARTIC ACID
	148.0593598	[M+H]+	147.0531578	C5H9NO4	L-Glutamate;L-Glutamic acid;L-Glutaminic acid
	148.0593598	[M+H]+	147.0531578	C5H9NO4	N-ACETYL-SERINE
	148.0593598	[M+H]+	147.0531578	C5H9NO4	**Glutamate;Glutaminic acid;2-Aminoglutaric acid**
	148.0593598	[M+H]+	147.0531578	C5H9NO4	N-Methyl-D-aspartic acid;N-Methyl-D-aspartate;NMDA
	148.0593598	[M+H]+	147.0531578	C5H9NO4	O-Acetyl-L-serine;O3-Acetyl-L-serine
	148.0593598	[M+H]+	147.0531578	C5H9NO4	2-Oxo-4-hydroxy-5-aminovalerate
	148.0593598	[M+H]+	147.0531578	C5H9NO4	L-threo-3-Methylaspartate
	148.0593598	[M+H]+	147.0531578	C5H9NO4	N-HYDROXY-N-ISOPROPYLOXAMIC ACID
	148.0593598	[M+H]+	147.0531578	C5H9NO4	N-(Carboxymethyl)-D-alanine
	148.0593598	[M+H]+	147.0531578	C5H9NO4	L-4-Hydroxyglutamate semialdehyde
	148.0593598	[M+H]+	147.0531578	C5H9NO4	Isoglutamate;Isoglutamic acid;3-Aminopentanedioic acid
	153.0760536	[M+H]+	152.0684735	C5H12O5	**Xylitol**
	153.0760536	[M+H]+	152.0684735	C5H12O5	**D-Ribitol;D-Adonitol**
	153.0760536	[M+H]+	152.0684735	C5H12O5	L-Arabitol;L-Arabinol;L-Arabinitol;L-Lyxitol
	153.0760536	[M+H]+	152.0684735	C5H12O5	D-Arabitol;D-Arabinitol;D-Arabinol;D-Lyxitol
	183.076994	[M+H]+	182.0691422	C8H10N2O3	2-AMINO-4-OXO-4(1H-PYRROL-1-YL)BUTANOIC ACID
	183.076994	[M+H]+	182.0707955	C6H15O4P	**TRIETHYL PHOSPHATE**
	183.076994	[M+H]+	182.0707955	C6H15O4P	Diisopropyl phosphate
	219.0959433	[M+H]+	218.0902716	C8H14N2O5	**L-Ala-gamma-D-Glu**
	219.0959433	[M+H]+	218.0902716	C8H14N2O5	gamma-L-Glutamyl-D-alanine
	231.095881	[M+H]+	230.0902716	C9H14N2O5	TETRAHYDRODEOXYURIDINE
	231.095881	[M+H]+	230.0902716	C9H14N2O5	**Aspartyl-L-proline**
	231.095881	[M+H]+	230.0902716	C9H14N2O5	3,4-DIHYDRO-1H-PYRIMIDIN-2-ONE NUCLEOSIDE
	241.029228	[M+Na]+	218.0401503	C12H10O2S	1,1'-BIPHENYL-2-SULFINIC ACID
	241.029228	[M+Na]+	218.0401503	C12H10O2S	cis-1,2-Dihydroxy-1,2-dihydrodibenzothiophene
	241.029228	[M+H]+	240.0238483	C6H12N2O4S2	L-Cystine;L-Dicysteine;L-alpha-Diamino-beta-dithiolactic acid
	241.029228	[M+H]+	240.0238483	C6H12N2O4S2	**Cystine;Dicysteine;alpha-Diamino-beta-dithiolactic acid**
	247.105723	[M+Na]+	224.1160924	C11H16N2O3	(6,10-DIOXO-OCTAHYDRO-PYRIDAZINO[1,2-A][1,2]DIAZEPIN-1-YL)-ACETALDEHYDE FRAGMENT
	247.105723	[M+Na]+	224.1160924	C11H16N2O3	2-(4-AMINO-PHENYL)-5-HYDROXYMETHYL-PYRROLIDINE-3,4-DIOL
	247.105723	[M+H]+	246.1004423	C13H14N2O3	Mephobarbital; drug
	247.105723	[M+H]+	246.1004423	C13H14N2O3	5-AMINO-4-OXO-1,2,4,5,6,7-HEXAHYDRO-AZEPINO[3,2,1-HI] INDOLE-2-CARBOXYLIC ACID
	247.105723	[M+H]+	246.1004423	C13H14N2O3	**Alanine, N-indol-3-ylacetyl- (6CI) Indole-3-acetylalanine N-(3-Indolylacetyl)-L-alanine**
	247.105723	[M+H]+	246.1004423	C13H14N2O3	N-Acetyl-D-tryptophan
	276.1167155	[M+H]+	275.1117353	C10H17N3O6	Norophthalmic acid
	277.1005288	[M+H]+	276.0939004	C22H12	Benzo[ghi]perylene;1,12-Benzoperylene
	277.1005288	[M+Na]+	254.1088985	C12H18N2O2S	Thiamylal;5-Allyl-5-(1-methylbutyl)-2-thiobarbituric acid; drug
	277.1005288	[M+H]+	276.0957509	C10H16N2O7	**Glu-Glu**
	277.1005288	[M+H]+	276.0957509	C10H16N2O7	gamma-glutamyl-D-glutamate

4.5 Discussion

Since marmosets could provide useful information about aging, it is of interest to determine marmoset ages at which metabolite levels change. In this study, the Gain and Loss Analysis of DNA (GLAD) method (Hupé et al., 2004), developed for the analysis of DNA copy numbers in tumors, is applied to marmoset mass spectra to identify breakpoints indicating changes in mean metabolite levels for a subset of features.

It was determined that the breakpoints found could be divided into five clusters, with the most likely breakpoint locations for the groups at 773 days, 1444 days, 2008 days, 2759 days, and 4496 days. This means that some metabolite levels in the marmoset colony change at about 2, 4, 5.5, 7.5, and 12.5 years old. Most features having any breakpoints have breaks at only one of these locations, but some features have two or three breakpoints.

Some of the breakpoints found could be due to changes in environment. For instance, at age two, marmosets are often removed from their original groups and form new breeding pairs. This could be accompanied by a change in diet or behavior that could influence different metabolites. Other possible causes of metabolite level shifts include changes in food or drug therapies and change in activity level. Age-related changes could include those related to the onset of renal disease and osteoporosis between seven and eight years of age. Also, there could be metabolite changes in females due to the onset of menopause. The dataset used here does not allow for the detection of sex-specific breakpoints, but this possibility should be addressed in the future.

Cluster 1 likely contains arginine, which is related to nitric oxide metabolism and is associated with hypertension (Taddei et al., 1996). Creatine and stearoylcarnitine appear to have been found in Cluster 2, which could signify changes in muscle mass or

activity level at around age four. Cluster 3 contains breakpoints that may be related to a change in steroid levels. Emetine, vitamin D2, dipeptides, and tripeptides were likely to have been found in Cluster 4, perhaps due to changes in food habits and drug therapy. Cluster 5 appears to contain many amino acids, dipeptides, tripeptides, and lysophospholipids, so breakpoints in this cluster seem to be related to protein and amino acid metabolism. Many of the other metabolites identified in the five clusters are artifacts from food and environmental sources, such as flame retardants and plasticizers absorbed after exposure to commercial products.

Being able to better identify the metabolites found in the different clusters would aid in separating age-related breakpoints from those that could be explained by environmental or other changes. While metabolite databases include more information now than in recent years, they are by no means complete (Soltow et al., 2010; Wishart et al., 2009). In our study, only 51 of the 115 MS features (44.3%) found to have breakpoints have at least one match in the MMCD database. Identification of all of the components of the metabolome would dramatically increase the amount of information available from metabolomic studies. Clusters of metabolites with similar age-related changes could provide clues as to which metabolites are biologically related, perhaps being associated with the same biological process. Differentiating those metabolites that increase at a particular age from those that decrease could also provide useful information about the aging process.

In this study, the complete list of breakpoints was considered when determining cluster membership, regardless of feature. For some features with two or three breakpoints, the breakpoints were identified as being in different clusters. For others, however, breakpoints on the same feature were classified together. It could be the case that these breakpoints should belong to the same cluster but that artifacts in the data allowed for more than one breakpoint to be identified. It could also mean that some clusters have been missed. Improved methods for classification could help determine

which of these might be true. One method of interest is using the EM algorithm assuming a mixture of truncated normal distributions to identify clusters in the data. This proposed method is discussed in more detail in the "Future Work" section below.

In addition to improving the classification method, other adjustments may also be made in the future. Several marmosets in the colony analyzed in this study have non-unique ages. For every repeated age, one marmoset's age was adjusted to make the ages unique for the purposes of this study. There is difficulty in interpreting a breakpoint found at either of the repeated ages but not the other. If there is a true breakpoint at a certain repeated age, one would expect it to surface for both marmosets of this age. Having the marmosets analyzed side-by-side when they are actually the same age may make a breakpoint more ambiguous. For this study, very few marmoset ages are repeated, and very few breakpoints are discovered at the repeated ages, so this is not thought to have caused problems. In the future, this issue will be resolved. Additional data could also improve this analysis in the future. While a cross-sectional dataset is used here, longitudinal data may soon be available. Collecting data from the same marmosets over time would allow between-subject variation to be controlled. Many improvements are possible, but this study provides a solid step in the direction of identification of metabolite groups that could provide useful information about aging and possibly lead to advances in personalized medicine.

4.6 Future Work

Another approach of interest in analyzing the marmoset data is using the Expectation-Maximization (EM) Algorithm for finite mixtures (Dempster et al., 1977). It is assumed that breakpoints exist at which metabolic systems experience shifts, resulting in changes in metabolite levels. The breakpoints detected in the marmoset data are

assumed to come from a mixture of normal distributions, with means of the mixture components representing the true breakpoints. The specific mixture component from which a particular breakpoint comes is treated as unknown. A set of starting values for the parameters is first selected. Parameters include means, variances, and mixing proportions for the different components, with mixing proportions summing to one and being between zero and one inclusive. In the expectation step (*E step*), the expected value of the log-likelihood of the complete data is computed using the current parameter values. Next, in the maximization step (*M step*), maximum likelihood estimates of the parameters are found. The E step and M step are repeated, using updated values of the parameters, until convergence.

As you can see in Figure 4.3, many breakpoints are found near the extremes in the marmoset data. In a Gaussian mixture model, one would expect values to taper off rather than peak in these areas. Therefore, a Gaussian mixture model may not be appropriate, as components may not have a normal distribution. Instead, a mixture model with truncated normal components may better fit the data. A truncated normal distribution results when a normal distribution is truncated on the left, on the right, or on both the left and right. Left-truncated and right-truncated distributions are referred to as singly truncated, and distributions with truncation on both the left and right are referred to as doubly truncated.

Let X be the random variable of interest, and let T_a represent the left truncation point and T_b represent the right truncation point. As explained in Cohen (1991), the pdf $f_{T_{ab}}$ for a doubly truncated distribution is:

$$f_{T_{ab}}(x;\mu,\sigma) = \begin{cases} \frac{1}{\sigma\sqrt{2\pi}(F(T_b)-F(T_a))} exp(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2) & \text{if } T_a \leq x \leq T_b \\ 0 & \text{elsewhere} \end{cases} \tag{4.3}$$

where μ and σ are the mean and standard deviation of the non-truncated distri-

bution, and $F(T_a)$ and $F(T_b)$ represent the CDF for points T_a and T_b ($F(x) = \int_{-\infty}^{x} f(y;\mu,\sigma)dy$). In this pdf, the normal pdf is being adjusted for the truncation by dividing by the proportion of the distribution that remains after truncation. For a right-truncated distribution, this would amount to dividing by $F(T_b)$, and for a left-truncated distribution, one would divide by $1 - F(T_a)$.

Using a truncated distribution is appealing for the marmoset breakpoints. This could be accomplished by assuming that all mixture components are doubly truncated. However, the truncation may be on the left for the smallest breakpoint component and on the right for the largest breakpoint component, with no truncation for other components. To account for this, we could allow different components to have different truncation patterns, with the leftmost component being left-truncated, the rightmost component being right-truncated, and any other components being free of truncation.

If truncation exists in the marmoset data, using the EM algorithm for Gaussian mixtures will result in incorrect estimates for breakpoint locations. For example, suppose we have a two-component mixture model with equal mixing proportions for the two components. Data points from the two components $Z1$ and $Z2$ are generated using $Z1 \sim N(-2, 1)$ and $Z2 \sim N(2, 1)$, and the distributions are truncated at -3 and 3. A visual representation of this truncated mixture model is given in Figure 4.7. Although means -2 and 2 are used to generate the data, using the EM algorithm and assuming a finite mixture model with two components results in means of -1.7 and 1.7 due to truncation (Figure 4.7). Additionally, standard deviation estimates from the EM algorithm for both components are about 0.8 instead of the value of 1 used to generate the data. While these may be accurate representations of the parameters for the truncated distribution components, we are interested in the parameters from the corresponding non-truncated distribution in the marmoset data.

Many have expressed interest in using truncated distributions, and as a result, several

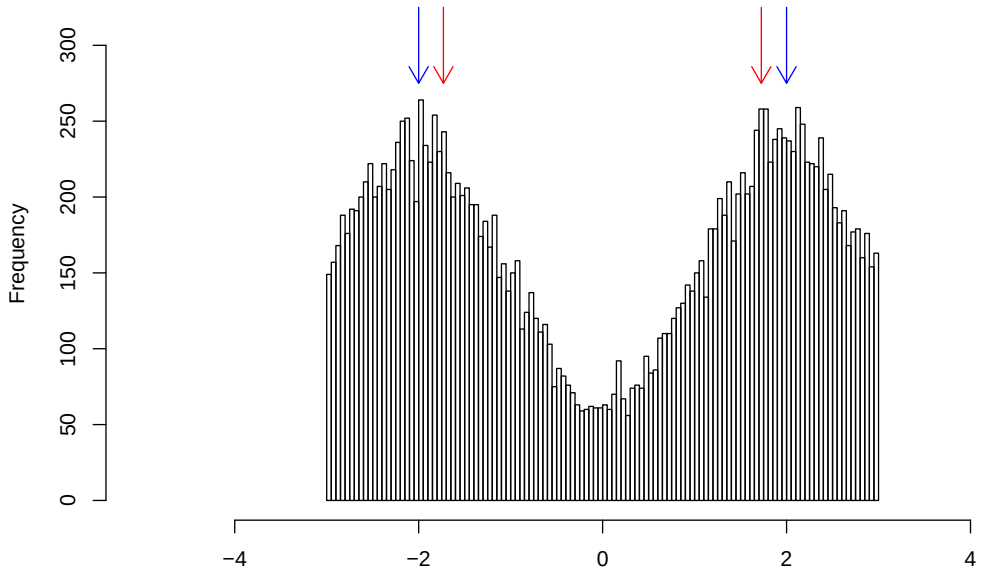

Figure 4.7: Example of truncated mixture distribution. Means used to generate the data are indicated by blue arrows, and means resulting from EM algorithm for a Gaussian mixture model are indicated by red arrows.

R functions for the analysis of truncated data exist. In the "Multi-state Markov and hidden Markov models in continuous time" package (Jackson, 2011), functions allow one to input the mean and standard deviation of a distribution before truncation and find the density, distribution function, and quantile function for the truncated distribution. Random numbers from a truncated distribution can also be produced. Functions in the "Truncated normal distribution" package (Trautmann et al., 2010) allow one to find the same values, and this package also has functions for finding the expected value and variance of the truncated distribution. The "Truncated Multivariate Normal and Student t Distribution" package (Wilhelm and Manjunath, 2010) adds to these functions a means for finding the maximum likelihood estimates from the corresponding non-truncated distribution when truncation points are known. In the future, we would like to use similar methodology to find non-truncated maximum likelihood estimates for truncated mixture models. Preliminary attempts reveal that the underlying data may not be normal, with a disproportionately high number of breakpoints at the minimum and maximum ages at which breakpoints were found. We plan on using simulation studies to investigate the robustness of the method when the underlying distributions of one or more mixture components deviate from normality.

Bibliography

Alm, E., Torgrip, R. J. O., Aberg, K. M., Schuppe-Koistinen, I., and Lindberg, J. (2009). A solution to the 1D NMR alignment problem using an extended generalized fuzzy Hough transform and mode support. *Analytical and Bioanalytical Chemistry*, 395(1):213–223.

Antti, H., Bollard, M., Ebbels, T., Keun, H., Lindon, J., Nicholson, J., and Holmes, E. (2002). Batch statistical processing of H-1 NMR-derived urinary spectral data. *Journal of Chemometrics*, 16(8-10, Sp. Iss. SI):461–468.

Austad, S. N. (2009). Comparative Biology of Aging. *Journals of Gerontology Series A-Biological Sciences and Medical Sciences*, 64(2):199–201.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17(3):166–173.

Beckonert, O., Bollard, M., Ebbels, T., Keun, H., Antti, H., Holmes, E., Lindon, J., and Nicholson, J. (2003). NMR-based metabonomic toxicity classification: hierarchical cluster analysis and k-nearest-neighbour approaches. *Analytica Chimica Acta*, 490(1-2):3–15.

Beckonert, O., Keun, H. C., Ebbels, T. M. D., Bundy, J. G., Holmes, E., Lindon, J. C., and Nicholson, J. K. (2007). Metabolic profiling, metabolomic and metabonomic procedures for NMR spectroscopy of urine, plasma, serum and tissue extracts. *Nature Protocols*, 2(11):2692–2703.

Ben-Yaacov, E. and Eldar, Y. C. (2008). A fast and flexible method for the segmentation of aCGH data. *Bioinformatics*, 24(16):I139–I145.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B-Methodological*, 57(1):289–300.

Bertram, H., Knudsen, K., Serena, A., Malmendal, A., Nielsen, N., Frette, X., and Andersen, H. (2006). NMR-based metabonomic studies reveal changes in the biochemical profile of plasma and urine from pigs fed high-fibre rye bread. *British Journal of Nutrition*, 95(5):955–962.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Clarendon Press, Oxford.

Blanco, R. A., Ziegler, T. R., Carlson, B. A., Cheng, P.-Y., Park, Y., Cotsonis, G. A., Accardi, C. J., and Jones, D. P. (2007). Diurnal variation in glutathione and cysteine redox states in human plasmas. *American Journal of Clinical Nutrition*, 86(4):1016–1023.

Blask, D. E. (2009). Melatonin, sleep disturbance and cancer risk. *Sleep Medicine Reviews*, 13(4):257–264.

Bollard, M., Keun, H., Beckonert, O., Ebbels, T., Antti, H., Nicholls, A., Shockcor, J., Cantor, G., Stevens, G., Lindon, J., Holmes, E., and Nicholson, J. (2005a). Comparative metabonomics of differential hydrazine toxicity in the rat and mouse. *Toxicology and Applied Pharmacology*, 204(2):135–151.

Bollard, M., Stanley, E., Lindon, J., Nicholson, J., and Holmes, E. (2005b). NMR-based metabonomic approaches for evaluating physiological influences on biofluid composition. *NMR in Biomedicine*, 18(3):143–162.

Bolstad, B. M. (2010). *preprocessCore: A collection of pre-processing functions*. R package version 1.12.0.

Burrus, C. (1998). *Introduction to Wavelets and Wavelet Transforms*. Prentice Hall, Englewood Cliffs.

Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Thomson Learning, Stamford.

Cohen, A. (1991). *Truncated and censored samples: theory and applications*. M. Dekker, New York.

Craig, A., Cloareo, O., Holmes, E., Nicholson, J., and Lindon, J. (2006). Scaling and normalization effects in NMR spectroscopic metabonomic data sets. *Analytical Chemistry*, 78(7):2262–2267.

Cui, Q., Lewis, I. A., Hegeman, A. D., Anderson, M. E., Li, J., Schulte, C. F., Westler, W. M., Eghbalnia, H. R., Sussman, M. R., and Markley, J. L. (2008). Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology*, 26(2):162–164.

De Bacquer, D., Van Risseghem, M., Clays, E., Kittel, F., De Backer, G., and Braeckman, L. (2009). Rotating shift work and the metabolic syndrome: a prospective study. *International Journal of Epidemiology*, 38(3):848–854.

de Castro Leão, A., Duarte Dória Neto, A., and Bernardete Cordeiro de Sousa, M. (2009). New developmental stages for common marmosets (*Callithrix jacchus*) using mass and age variables obtained by K-means algorithm and self-organizing maps (SOM). *Computers in Biology and Medicine*, 39:853–859.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from incomplete

data via EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological*, 39(1):1–38.

Denkert, C., Budczies, J., Kind, T., Weichert, W., Tablack, P., Sehouli, J., Niesporek, S., Koensgen, D., Dietel, M., and Fiehn, O. (2006). Mass spectrometry-based metabolic profiling reveals different metabolite patterns in invasive ovarian carcinomas and ovarian borderline tumors. *Cancer Research*, 66(22):10795–10804.

Denkert, C., Budczies, J., Weichert, W., Wohlgemuth, G., Scholz, M., Kind, T., Niesporek, S., Noske, A., Buckendahl, A., Dietel, M., and Fiehn, O. (2008). Metabolite profiling of human colon carcinoma - deregulation of TCA cycle and amino acid turnover. *Molecular Cancer*, 7(72).

Dumas, M., Maibaum, E., Teague, C., Ueshima, H., Zhou, B., Lindon, J., Nicholson, J., Stamler, J., Elliott, P., Chan, Q., and Holmes, E. (2006). Assessment of analytical reproducibility of H-1 NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. *Analytical Chemistry*, 78(7):2199–2208.

Dunn, W., Broadhurst, D., Begley, P., Zelena, E., Francis-McIntyre, S., Anderson, N., Brown, M., Knowles, J., Halsall, A., Haselden, J., Nicholls, A., Wilson, I., Kell, D., Goodacre, R., and The Human Serum Metabolome (HUSERMET) Consortium (2011). Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature Protocols*, 6(7):1060–1083.

Ebbels, T. M. D., Keun, H. C., Beckonert, O. P., Bollard, M. E., Lindon, J. C., Holmes, E., and Nicholson, J. K. (2007). Prediction and classification of drug toxicity using probabilistic modeling of temporal metabolic data: The Consortium

on Metabonomic Toxicology screening approach. *Journal of Proteome Research*, 6(11):4407–4422.

Eckel-Passow, J. E., Oberg, A. L., Therneau, T. M., and Bergen, III, H. R. (2009). An insight into high-resolution mass-spectrometry data. *Biostatistics*, 10(3):481–500.

Eisen, M., Spellman, P., Brown, P., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, 95(19):14863–14868.

Fernald, G., Capriotti, E., Daneshjou, R., Karczewski, K., and Altman, R. (2011). Bioinformatics challenges for personalized medicine. *Bioinformatics*, 27(13):1741–1748.

Fiehn, O. (2002). Metabolomics - the link between genotypes and phenotypes. *Plant Molecular Biology*, 48(1-2):155–171.

Forshed, J., Schuppe-Koistinen, I., and Jacobsson, S. (2003). Peak alignment of NMR signals by means of a genetic algorithm. *Analytica Chimica Acta*, 487(2):189–199.

Fraley, C. and Raftery, A. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *Computer Journal*, 41(8):578–588.

Gustafsson, F. (1994). Segmentation of signals using piecewise constant linear regression models. Technical report, Linköping University.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning*. Springer, Berlin.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G. (2011). *impute: Imputation for microarray data*. R package version 1.26.0.

Hawkins, D. (1976). Point estimation of parameters of piecewise regression-models. *Journal of the Royal Statistical Society Series C-Applied Statistics*, 25(1):51–57.

Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2006). A lego system for conditional inference. *The American Statistician*, 60(3):257–263.

Hothorn, T., Hornik, K., van de Wiel, M. A., and Zeileis, A. (2008). Implementing a class of permutation tests: The coin package. *Journal of Statistical Software*, 28(8):1–23.

Hupe, P. (2009). *GLAD: Gain and Loss Analysis of DNA*. R package version 2.6.0.

Hupé, P., Stransky, N., Thiery, J., Radvanyi, F., and Barillot, E. (2004). Analysis of array CGH data: from signal ratio to gain and loss of DNA regions. *Bioinformatics*, 20(18):3413–3422.

Jackson, C. H. (2011). Multi-state models for panel data: The msm package for R. *Journal of Statistical Software*, 38(8):1–29.

Keun, H., Ebbels, T., Antti, H., Bollard, M., Beckonert, O., Schlotterbeck, G., Senn, H., Niederhauser, U., Holmes, E., Lindon, J., and Nicholson, J. (2002). Analytical reproducibility in H-1 NMR-based metabonomic urinalysis. *Chemical Research in Toxicology*, 15(11):1380–1386.

Kim, S. B. and Park, Y. (2005). Matlab code for Spectra Alignment using Beamsearch and Genetic Algorithm, Normalization based on TSM, and Water Reduction. Clinical Biomarkers Lab. in Emory University Medical School.

Lee, G. and Woodruff, D. (2004). Beam search for peak alignment of NMR signals. *Analytica Chimica Acta*, 513(2):413–416.

Leikin, J. (2003). *American Medical Association Complete Medical Encyclopedia*. Crown Publishers, New York.

Leuner, B., Kozorovitskiy, Y., Gross, C. G., and Gould, E. (2007). Diminished adult

neurogenesis in the marmoset brain precedes old age. *Proceedings of the National Academy of Sciences of the United States of America*, 104(43):17169–17173.

McMurry, J. (1996). *Organic Chemistry*. Brooks Cole, Pacific Grove.

Nicholson, J., Foxall, P., Spraul, M., Farrant, R., and Lindon, J. (1995). 750-MHZ H-1 and H-1-C-13 NMR-Spectroscopy of Human Blood-Plasma. *Analytical Chemistry*, 67(5):793–811.

Nicholson, J., Lindon, J., and Holmes, E. (1999). 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11):1181–1189.

Odunsi, K., Wollman, R., Ambrosone, C., Hutson, A., McCann, S., Tammela, J., Geisler, J., Miller, G., Sellers, T., Cliby, W., Qian, F., Keitz, B., Intengan, M., Lele, S., and Alderfer, J. (2005). Detection of epithelial ovarian cancer using H-1-NMR-based metabonomics. *International Journal of Cancer*, 113(5):782–788.

Park, Y., Kim, S. B., Wang, B., Blanco, R. A., Le, N.-A., Wu, S., Accardi, C. J., Alexander, R. W., Ziegler, T. R., and Jones, D. P. (2009). Individual variation in macronutrient regulation measured by proton magnetic resonance spectroscopy of human plasma. *American Journal of Physiology-Regulatory Integrative and Comparative Physiology*, 297(1):R202–R209.

Pearson, G. A. (1991). *Shimming an NMR Magnet*. http://nmr.chem.uiowa.edu/manuals/Shimming-GAP-NMR-magnet.pdf.

Polzehl, J. and Spokoiny, V. (2000). Adaptive weights smoothing with applications to image restoration. *Journal of the Royal Statistical Society Series B-Statistical Methodology*, 62(Part 2):335–354.

Qin, L. and Self, S. (2006). The clustering of regression models method with applications in gene expression data. *Biometrics*, 62(2):526–533.

Quinn, B. G. and Thomson, P. J. (1991). Estimating the frequency of a periodic function. *Biometrika*, 78(1):65–74.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sanders, J. (1993). *Modern NMR Spectroscopy*. Oxford University Press, Oxford Oxfordshire.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Seber, G. A. F. and Lee, A. J. (2003). *Linear regression analysis*. Wiley-Interscience, Hoboken, N.J.

Sheather, S. and Jones, M. (1991). A reliable data-based bandwidth selection method for kernel density-estimation. *Journal of the Royal Statistical Society Series B-Methodological*, 53(3):683–690.

Simko, F. and Pechanova, O. (2009). Potential roles of melatonin and chronotherapy among the new trends in hypertension treatment. *Journal of Pineal Research*, 47(2):127–133.

Soltow, Q. A., Jones, D. P., and Promislow, D. E. L. (2010). A Network Perspective on Metabolism and Aging. *Integrative and Comparative Biology*, 50(5):844–854.

Soltow, Q. A., Strobel, F. H., Mansfield, K. G., Wachtman, L., Park, Y., and Jones, D. P. (2011). High-performance metabolic profiling with dual chromatography-Fourier-transform mass spectrometry (DC-FTMS) for study of the exposome. In Press.

Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. (1998). Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Molecular Biology of the Cell*, 9(12):3273–3297.

Stevens, R. G. (2009). Light-at-night, circadian disruption and breast cancer: assessment of existing evidence. *International Journal of Epidemiology*, 38(4):963–970.

Strang, G. (1996). *Wavelets and Filter Banks*. Wellesley-Cambridge Press, Wellesley.

Swokowski, E. (2009). *Algebra and Trigonometry with Analytic Geometry, Classic Edition*. Brooks Cole, Pacific Grove.

Taddei, S., Virdis, A., Mattei, P., Ghiadoni, L., Sudano, I., and Salvetti, A. (1996). Defective L-arginine-nitric oxide pathway in offspring of essential hypertensive patients. *Circulation*, 94(6):1298–1303.

Terman, A. and Brunk, U. (2006). Oxidative stress, accumulation of biological 'garbage', and aging. *Antioxidants & Redox Signaling*, 8(1-2):197–204.

Trainesel, D., Catherinot, V., and Delsuc, M.-A. (2007). Modeling of NMR processing, toward efficient unattended processing of NMR experiments. *Journal of Magnetic Resonance*, 188(1):56–67.

Trautmann, H., Steuer, D., Mersmann, O., and Bornkamp, B. (2010). *truncnorm: Truncated normal distribution*. R package version 1.0-4.

Viant, M., Lyeth, B., Miller, M., and Berman, R. (2005). An NMR metabolomic investigation of early metabolic disturbances following traumatic brain injury in a mammalian model. *NMR in Biomedicine*, 18(8):507–516.

Wilhelm, S. and Manjunath, B. G. (2010). *tmvtnorm: Truncated Multivariate Normal Distribution*. R package version 1.1-5.

Wishart, D. S., Knox, C., Guo, A. C., Eisner, R., Young, N., Gautam, B., Hau, D. D., Psychogios, N., Dong, E., Bouatra, S., Mandal, R., Sinelnikov, I., Xia, J., Jia, L., Cruz, J. A., Lim, E., Sobsey, C. A., Shrivastava, S., Huang, P., Liu, P., Fang, L., Peng, J., Fradette, R., Cheng, D., Tzur, D., Clements, M., Lewis, A., De Souza, A., Zuniga, A., Dawe, M., Xiong, Y., Clive, D., Greiner, R., Nazyrova, A., Shaykhutdinov, R., Li, L., Vogel, H. J., and Forsythe, I. (2009). HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Research*, 37(Sp. Iss. SI):D603–D610.

Yu, T. (2010). An exploratory data analysis method to reveal modular latent structures in high-throughput data. *BMC Bioinformatics*, 11(440).

Yu, T. (2011). *apLCMS: Adaptive processing of LC-MS data*. R package version 4.5.0.

Yu, T., Park, Y., Johnson, J. M., and Jones, D. P. (2009). apLCMS-adaptive processing of high-resolution LC/MS data. *Bioinformatics*, 25(15):1930–1936.

Yu, T., Ye, H., Sun, W., Li, K.-C., Chen, Z., Jacobs, S., Bailey, D. K., Wong, D. T., and Zhou, X. (2007). A forward-backward fragment assembling algorithm for the identification of genomic amplification and deletion breakpoints using high-density single nucleotide polymorphism (SNP) array. *BMC Bioinformatics*, 8(145).

Appendices

Appendix A

Appendix for Chapter 3

A.1 Full Simulation Results

Table A.1: Simulation Results (1): $\sigma^2 = (1, 2)$; $\pi = (0.7, 0.3)$

	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\pi}$	MR[a]	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\pi}$	MR[a]
Extra Variation	Including				Ignoring			
1.1	$s_p^2 \sim \chi^2_{0.25}$				$s_p^2 \sim \chi^2_{0.25}$			
$\mu : (-0.5, 0.5)$	(-0.43, 0.43)	(1.03, 2.33)	(0.75, 0.25)	0.250	(-0.40, 0.29)	(1.13, 2.91)	(0.72, 0.28)	0.256
$\mu : (-2.5, 2.5)$	(-2.52, 2.42)	(0.96, 2.09)	(0.70, 0.30)	0.024	(-2.53, 2.44)	(1.16, 2.19)	(0.70, 0.30)	0.025
$\mu : (-5, 5)$	(-5.03, 4.97)	(1.03, 2.00)	(0.71, 0.29)	<0.001	(-5.03, 4.96)	(1.28, 2.29)	(0.71, 0.29)	0.001
1.2	$s_p^2 \sim \chi^2_{0.5}$				$s_p^2 \sim \chi^2_{0.5}$			
$\mu : (-0.5, 0.5)$	(-0.48, 0.77)	(1.08, 1.86)	(0.78, 0.22)	0.258	(-0.35, 0.06)	(1.23, 3.32)	(0.65, 0.35)	0.276
$\mu : (-2.5, 2.5)$	(-2.49, 2.47)	(1.00, 1.92)	(0.70, 0.30)	0.029	(-2.50, 2.48)	(1.45, 2.30)	(0.70, 0.30)	0.032
$\mu : (-5, 5)$	(-5.00, 5.03)	(0.93, 1.87)	(0.70, 0.30)	<0.001	(-5.01, 5.04)	(1.45, 2.40)	(0.70, 0.30)	0.001
1.3	$s_p^2 \sim \chi^2_1$				$s_p^2 \sim \chi^2_1$			
$\mu : (-0.5, 0.5)$	(-0.56, 0.48)	(0.91, 2.09)	(0.63, 0.37)	0.263	(-0.32, 0.08)	(1.54, 5.06)	(0.70, 0.30)	0.292
$\mu : (-2.5, 2.5)$	(-2.49, 2.55)	(1.03, 2.06)	(0.70, 0.30)	0.045	(-2.51, 2.56)	(1.86, 2.94)	(0.70, 0.30)	0.046
$\mu : (-5, 5)$	(-5.02, 4.93)	(0.96, 1.84)	(0.71, 0.29)	0.002	(-5.05, 4.87)	(1.85, 2.86)	(0.71, 0.29)	0.003

[a] MR = Misclassification rate

[b] 5000 values were generated for each simulation.

Table A.2: Simulation Results (2): $\sigma^2 = (1, 0.1)$; $\pi = (0.7, 0.3)$

	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\pi}$	MR[a]	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\pi}$	MR[a]
Extra Variation	Including				Ignoring			
2.1	$s_p^2 \sim \chi^2_{0.25}$				$s_p^2 \sim \chi^2_{0.25}$			
$\mu : (-0.5, 0.5)$	(-0.54, 0.49)	(1.01, 0.11)	(0.68, 0.32)	0.227	(-0.42, 0.48)	(1.31, 0.12)	(0.76, 0.24)	0.237
$\mu : (-2.5, 2.5)$	(-2.51, 2.50)	(1.04, 0.10)	(0.70, 0.30)	0.006	(-2.49, 2.52)	(1.36, 0.25)	(0.70, 0.30)	0.009
$\mu : (-5, 5)$	(-5.01, 4.97)	(0.98, 0.01)	(0.71, 0.29)	0	(-5.01, 4.98)	(1.30, 0.33)	(0.71, 0.29)	0.001
2.2	$s_p^2 \sim \chi^2_{0.5}$				$s_p^2 \sim \chi^2_{0.5}$			
$\mu : (-0.5, 0.5)$	(-0.53, 0.53)	(1.03, 0.11)	(0.67, 0.33)	0.244	(-0.38, 0.45)	(1.73, 0.16)	(0.77, 0.23)	0.262
$\mu : (-2.5, 2.5)$	(-2.51, 2.50)	(1.03, 0.11)	(0.70, 0.30)	0.014	(-2.49, 2.53)	(1.45, 0.46)	(0.70, 0.30)	0.015
$\mu : (-5, 5)$	(-4.99, 5.00)	(1.02, 0.10)	(0.69, 0.31)	<0.001	(-4.99, 5.01)	(1.56, 0.62)	(0.70, 0.30)	0.001
2.3	$s_p^2 \sim \chi^2_1$				$s_p^2 \sim \chi^2_1$			
$\mu : (-0.5, 0.5)$	(-0.53, 0.47)	(0.99, 0.10)	(0.67, 0.33)	0.272	(-0.43, -0.02)	(3.05, 0.90)	(0.46, 0.54)	0.517
$\mu : (-2.5, 2.5)$	(-2.50, 2.50)	(1.02, 0.10)	(0.69, 0.31)	0.023	(-2.53, 2.56)	(1.86, 0.84)	(0.69, 0.31)	0.024
$\mu : (-5, 5)$	(-5.01, 4.99)	(0.95, 0.10)	(0.70, 0.30)	0.001	(-5.00, 5.02)	(1.92, 1.06)	(0.70, 0.30)	0.002

[a] MR = Misclassification rate

[b] 5000 values were generated for each simulation.

Table A.3: Simulation Results (3): $\sigma^2 = (0.1, 0.1)$; $\pi = (0.7, 0.3)$

	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\pi}$	MR[a]	$\hat{\mu}$	$\hat{\sigma}^2$	$\hat{\pi}$	MR[a]
Extra Variation	Including				Ignoring			
3.1	$s_p^2 \sim \chi^2_{0.25}$				$s_p^2 \sim \chi^2_{0.25}$			
$\mu : (-0.5, 0.5)$	(-0.51, 0.48)	(0.10, 0.10)	(0.69, 0.31)	0.098	(-0.21, -0.05)	(0.39, 2.77)	(0.93, 0.07)	0.307
$\mu : (-2.5, 2.5)$	(-2.50, 2.51)	(0.10, 0.09)	(0.71, 0.29)	0.004	(-2.51, 2.51)	(0.26, 0.30)	(0.71, 0.29)	0.004
$\mu : (-5, 5)$	(-5.00, 4.99)	(0.10, 0.10)	(0.70, 0.30)	0	(-5.01, 4.99)	(0.34, 0.33)	(0.70, 0.30)	0
3.2	$s_p^2 \sim \chi^2_{0.5}$				$s_p^2 \sim \chi^2_{0.5}$			
$\mu : (-0.5, 0.5)$	(-0.47, 0.56)	(0.11, 0.08)	(0.75, 0.25)	0.146	(-0.23, -0.14)	(0.46, 3.04)	(0.87, 0.13)	0.310
$\mu : (-2.5, 2.5)$	(-2.51, 2.49)	(0.10, 0.10)	(0.70, 0.30)	0.006	(-2.52, 2.50)	(0.54, 0.56)	(0.70, 0.30)	0.007
$\mu : (-5, 5)$	(-5.01, 4.96)	(0.10, 0.11)	(0.72, 0.28)	<0.001	(-5.02, 4.98)	(0.63, 0.53)	(0.72, 0.28)	<0.001
3.3	$s_p^2 \sim \chi^2_1$				$s_p^2 \sim \chi^2_1$			
$\mu : (-0.5, 0.5)$	(-0.51, 0.42)	(0.09, 0.12)	(0.69, 0.31)	0.193	(-0.30, -0.23)	(3.67, 0.59)	(0.22, 0.78)	0.675
$\mu : (-2.5, 2.5)$	(-2.50, 2.48)	(0.10, 0.10)	(0.70, 0.30)	0.015	(-2.55, 2.48)	(0.90, 0.94)	(0.70, 0.30)	0.017
$\mu : (-5, 5)$	(-5.00, 5.00)	(0.09, 0.09)	(0.70, 0.30)	<0.001	(-4.99, 4.99)	(1.05, 1.07)	(0.70, 0.30)	0.001

[a] MR = Misclassification rate

[b] 5000 values were generated for each simulation.

A.2 Derivation of Maximum Likelihood Estimates for μ_k, σ_k^2, and π_k

As noted in 3.3.2, the expected log-likelihood at each iteration j+1 (*E-Step*) is given by (equation (3.6) above):

$$E(\ell(\theta;\mathbf{T})|\mathbf{Z},\hat{\theta}^{(j)}) = \sum_{p=1}^{P}\sum_{k=1}^{K}[(\hat{\eta}_{pk}log\phi_{\theta_k}(\hat{\beta}_p)) + (\hat{\eta}_{pk}log(\pi_k))],$$

where $\mathbf{T}$ is the complete data and includes the unknown classes, $\mathbf{Z}$ is the observed data, and $\hat{\theta}^{(j)}$ represents the parameter estimates at iteration j. The estimate $\hat{\eta}_{pk}$ is found by replacing π_k and θ_k with their current estimates $\hat{\pi}_k$ and $\hat{\theta}_k$ in the following equation (equation (3.4) above):

$$\eta_{pk} = \frac{\pi_k\phi_{\theta_k}(\hat{\beta}_p)}{\sum_{k=1}^{K}\pi_k\phi_{\theta_k}(\hat{\beta}_p)},$$

where $\phi_{\theta_k}(\hat{\beta}_p)$ is the normal density for $\hat{\beta}_p$ with mean μ_{β_k} and variance $\sigma_k^2 + s_p^2$. Specifically,

$$\phi_{\theta_k}(\hat{\beta}_p) = \frac{1}{\sqrt{2\pi(\sigma_k^2 + s_p^2)}}e^{-((\hat{\beta}_p-\mu_{\beta_k})^2)/(2(\sigma_k^2+s_p^2))}.$$

Maximum likelihood estimates of $\hat{\mu}_{\beta_k}$, $\hat{\sigma}_k^2$, and $\hat{\pi}_k$ are then found for each iteration (*M-Step*) by taking the derivative of $E(\ell(\theta;\mathbf{T})|\mathbf{Z},\hat{\theta}^{(j)})$ with respect to each of the parameters, setting each to zero, and solving for the respective parameters. Derivations for these equations follow.

Derivation for $\hat{\mu}_{\beta_k}$:

$$\begin{aligned}
\frac{\partial(E(\ell(\theta;\mathbf{T})|\mathbf{Z},\hat{\theta}^{(j)}))}{\partial\mu_{\beta_k}} &= \frac{\partial}{\partial\mu_{\beta_k}}\sum_{p=1}^{P}\sum_{k=1}^{K}(\hat{\eta}_{pk}log\phi_{\theta_k}(\hat{\beta}_p)) \\
&= \frac{\partial}{\partial\mu_{\beta_k}}\left(\sum_{p=1}^{P}\hat{\eta}_{pk}\left(log(2\pi)^{-\frac{1}{2}} + log(\sigma_k^2+s_p^2)^{-\frac{1}{2}} - \frac{(\hat{\beta}_p-\mu_{\beta_k})^2}{2(\sigma_k^2+s_p^2)}\right)\right) \\
&= -\sum_{p=1}^{P}\frac{\hat{\eta}_{pk}(\hat{\beta}_p-\mu_{\beta_k})}{(\sigma_k^2+s_p^2)} \\
&= -\sum_{p=1}^{P}\frac{\hat{\eta}_{pk}\hat{\beta}_p}{(\sigma_k^2+s_p^2)} + \sum_{p=1}^{P}\frac{\hat{\eta}_{pk}\mu_{\beta_k}}{(\sigma_k^2+s_p^2)} \\
Set &= 0: \\
\hat{\mu}_{\beta_k} &= \frac{\sum_{p=1}^{P}\hat{\eta}_{pk}\hat{\beta}_p/(\sigma_k^2+s_p^2)}{\sum_{p=1}^{P}\hat{\eta}_{pk}/(\sigma_k^2+s_p^2)}
\end{aligned}$$

Derivation for $\hat{\sigma}_k^2$:

$$\begin{aligned}
\frac{\partial(E(\ell(\theta;\mathbf{T})|\mathbf{Z},\hat{\theta}^{(j)}))}{\partial\sigma_k^2} &= \frac{\partial}{\partial\sigma_k^2}\sum_{p=1}^{P}\sum_{k=1}^{K}(\hat{\eta}_{pk}log\phi_{\theta_k}(\hat{\beta}_p)) \\
&= \frac{\partial}{\partial\sigma_k^2}\left(\sum_{p=1}^{P}\hat{\eta}_{pk}\left(log(2\pi)^{-\frac{1}{2}} + log(\sigma_k^2+s_p^2)^{-\frac{1}{2}} - \frac{(\hat{\beta}_p-\mu_{\beta_k})^2}{2(\sigma_k^2+s_p^2)}\right)\right) \\
&= \sum_{p=1}^{P}-\frac{1}{2}\hat{\eta}_{pk}(\sigma_k^2+s_p^2)^{-1} - \sum_{p=1}^{P}-\frac{1}{2}\hat{\eta}_{pk}(\hat{\beta}_p-\mu_{\beta_k})^2(\sigma_k^2+s_p^2)^{-2} \\
&= \sum_{p=1}^{P}\hat{\eta}_{pk}(\sigma_k^2+s_p^2)^{-1} - \sum_{p=1}^{P}\hat{\eta}_{pk}(\hat{\beta}_p-\mu_{\beta_k})^2(\sigma_k^2+s_p^2)^{-2} \\
Set &= 0: \\
0 &= \sum_{p=1}^{P}\frac{\hat{\eta}_{pk}}{\hat{\sigma}_k^2+s_p^2} - \sum_{p=1}^{P}\frac{\hat{\eta}_{pk}(\hat{\beta}_p-\mu_{\beta_k})^2}{\hat{\sigma}_k^2+s_p^2}
\end{aligned}$$

For $\hat{\sigma}_k^2$, the equation was not solved explicitly for $\hat{\sigma}_k^2$ since the expressions were complicated by the s_p^2 value. Instead, an iterative procedure is used to find estimates

for $\hat{\mu}_{\beta_k}$ and $\hat{\sigma}_k^2$(3.3.2).

Derivation for $\hat{\pi}_k$:

$$\begin{aligned}
\frac{\partial(E(\ell(\theta;\mathbf{T})|\mathbf{Z},\hat{\theta}^{(j)}))}{\partial\pi_k} &= \frac{\partial}{\partial\pi_k}\sum_{p=1}^{P}\sum_{k=1}^{K}\hat{\eta}_{pk}log(\pi_k) \\
&= \sum_{p=1}^{P}\sum_{j=1}^{K}(-\hat{\eta}_{pj}\pi_k + \hat{\eta}_{pk}) \\
&= -\sum_{p=1}^{P}\pi_k + \sum_{p=1}^{P}\hat{\eta}_{pk} \\
&= -P\pi_k + \sum_{p=1}^{P}\hat{\eta}_{pk} \\
Set &= 0: \\
\hat{\pi}_k &= \frac{\sum_{p=1}^{P}\hat{\eta}_{pk}}{P}
\end{aligned}$$

In the derivation for $\hat{\pi}_k$, π_k was expressed as $\frac{e^{\lambda_k}}{\sum_{j=1}^{K}e^{\lambda_j}}$ (Hastie et al. (2001); Bishop (1995)). This enforced necessary constraints on π_k ($\pi_k \geq 0$, $\sum_{k=1}^{K}\pi_k = 1$). The index j was used to represent all values 1 to K, while k was used as the index of the particular π value of interest.

Appendix B

Appendix for Chapter 4

B.1 Selection of parameters for the implementation of GLAD

For Haar segmentation, detail subbands were limited to levels two to five. The finest level of detail was set to two since a segment was expected to span at least $x = 4$ units, so applying the rule of thumb in Ben-Yaacov and Eldar (2008), we find that the finest level of detail equals $log_2 4 = 2$. The level of coarsest detail was set to five, at which point the number of breakpoints stabilized so that including coarser detail subbands produced the same number of breakpoints. The false discovery rate (FDR) threshold used was 0.0001, meaning a very small proportion of false positives was allowed. Also, it was indicated that the region size expected was four. This does allow for regions smaller than four in some cases, but it decreases the number of very small regions that are produced by segmentation. Again, regions spanning fewer than four marmoset ages were generally not of interest unless there was strong evidence in the data that such small regions existed.

Several combinations of the penalty parameters L and D were investigated, where D is the parameter used in the kernel function in the penalty term, and L is the penalty term coefficient. Both D and L were allowed to take values from one to ten, and segmentation was performed for every combination of values. A subset of results is shown in Table B.1. In order to determine appropriate values for these parameters, the number of total breakpoints was plotted against values of parameters D and L. In Figure B.1, the number of total breaks is plotted against parameter D separately for the different L values. This relationship is reversed for Figure B.2, in which the number of breaks is plotted against L for the different D values. Additionally, Figure 4.2 shows a plot of the total number of breaks by both D and L. For the different values of L, the number of breakpoints appeared to level off after $D = 3$ (Figure B.1). Similarly, the L plots show that the number of breakpoints stabilize after $L = 4$ for

the different values of D (Figure B.2).

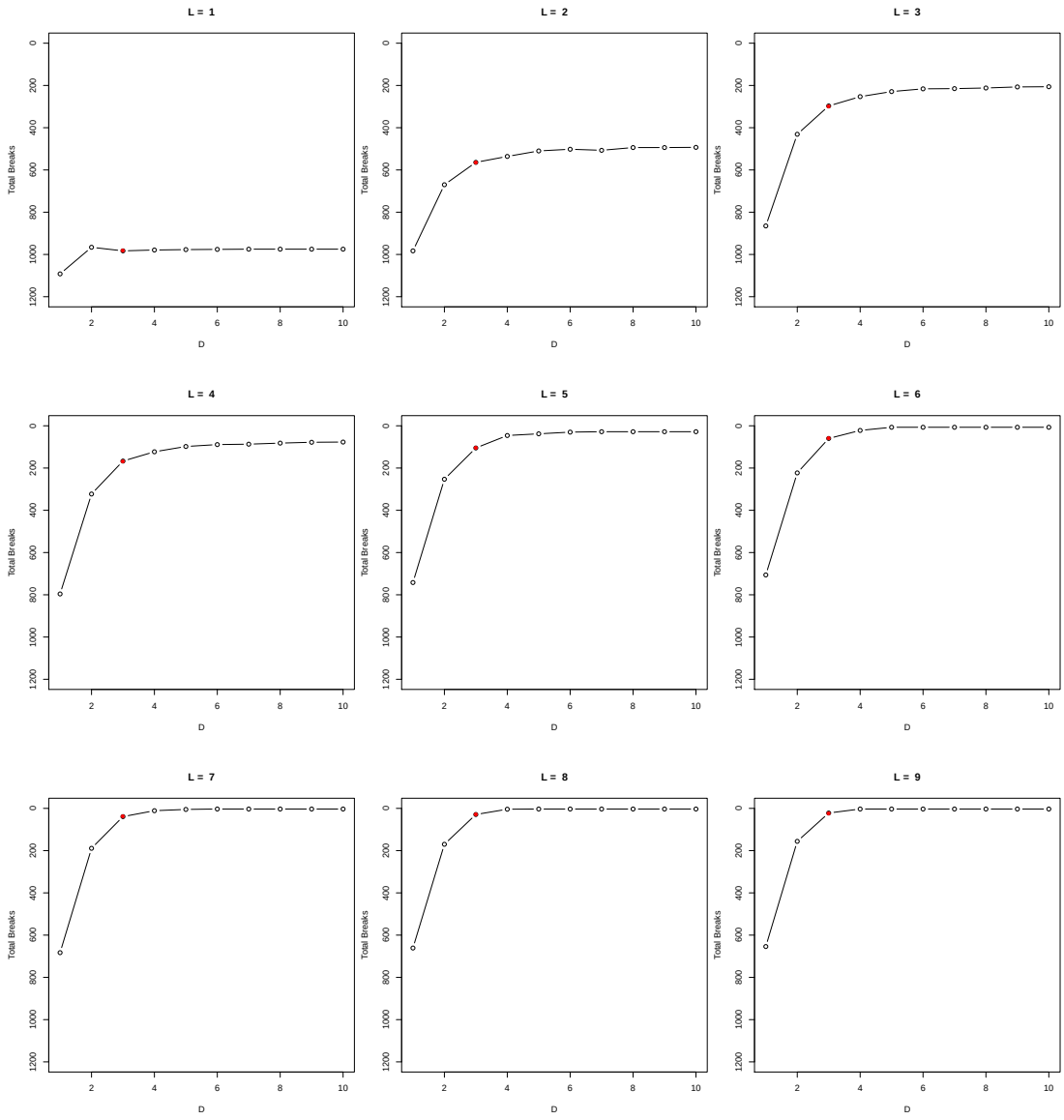

Figure B.1: Number of breakpoints plotted against penalty term kernel parameter D for various penalty term coefficient L values. Chosen value $D = 3$ is shown in red.

Table B.1: Results using various values of penalty term coefficient L and penalty term kernel parameter D.

L	D	Tot. Num. of Breaks	Num. Feat. with Breaks
1	1	1092	713
1	2	966	641
1	3	983	642
1	4	979	641
1	5	977	639
1	6	976	638
1	7	975	638
1	8	975	638
1	9	975	638
1	10	975	638
2	1	983	653
2	2	670	459
2	3	564	377
2	4	536	352
2	5	510	330
2	6	502	323
2	7	507	322
2	8	494	318
2	9	494	318
2	10	493	317
3	1	865	588
3	2	431	310
3	3	297	200
3	4	253	161
3	5	229	143
3	6	216	132
3	7	215	131
3	8	212	130
3	9	207	127
3	10	206	126
4	1	796	556
4	2	323	235
4	3	167	115
4	4	123	78
4	5	98	58
4	6	89	50
4	7	87	49
4	8	82	46
4	9	78	43
4	10	77	42
5	1	742	526
5	2	253	194
5	3	105	72
5	4	46	31
5	5	38	23
5	6	30	16
5	7	28	15
5	8	28	15
5	9	28	15
5	10	28	15
6	1	706	506
6	2	223	174
6	3	60	43
6	4	22	13
6	5	7	4
6	6	7	4
6	7	7	4
6	8	7	4
6	9	7	4
6	10	7	4
7	1	683	492
7	2	189	150
7	3	39	30
7	4	11	8
7	5	5	3
7	6	3	2
7	7	3	2
7	8	3	2
7	9	3	2
7	10	3	2
8	1	661	480
8	2	170	137
8	3	29	24
8	4	4	3
8	5	3	2
8	6	3	2
8	7	3	2
8	8	3	2
8	9	3	2
8	10	3	2
9	1	654	473
9	2	156	127
9	3	22	19
9	4	3	2
9	5	3	2
9	6	3	2
9	7	3	2
9	8	3	2
9	9	3	2
9	10	3	2
10	1	635	462
10	2	140	117
10	3	16	14
10	4	3	2
10	5	3	2
10	6	3	2
10	7	3	2
10	8	3	2
10	9	3	2
10	10	3	2

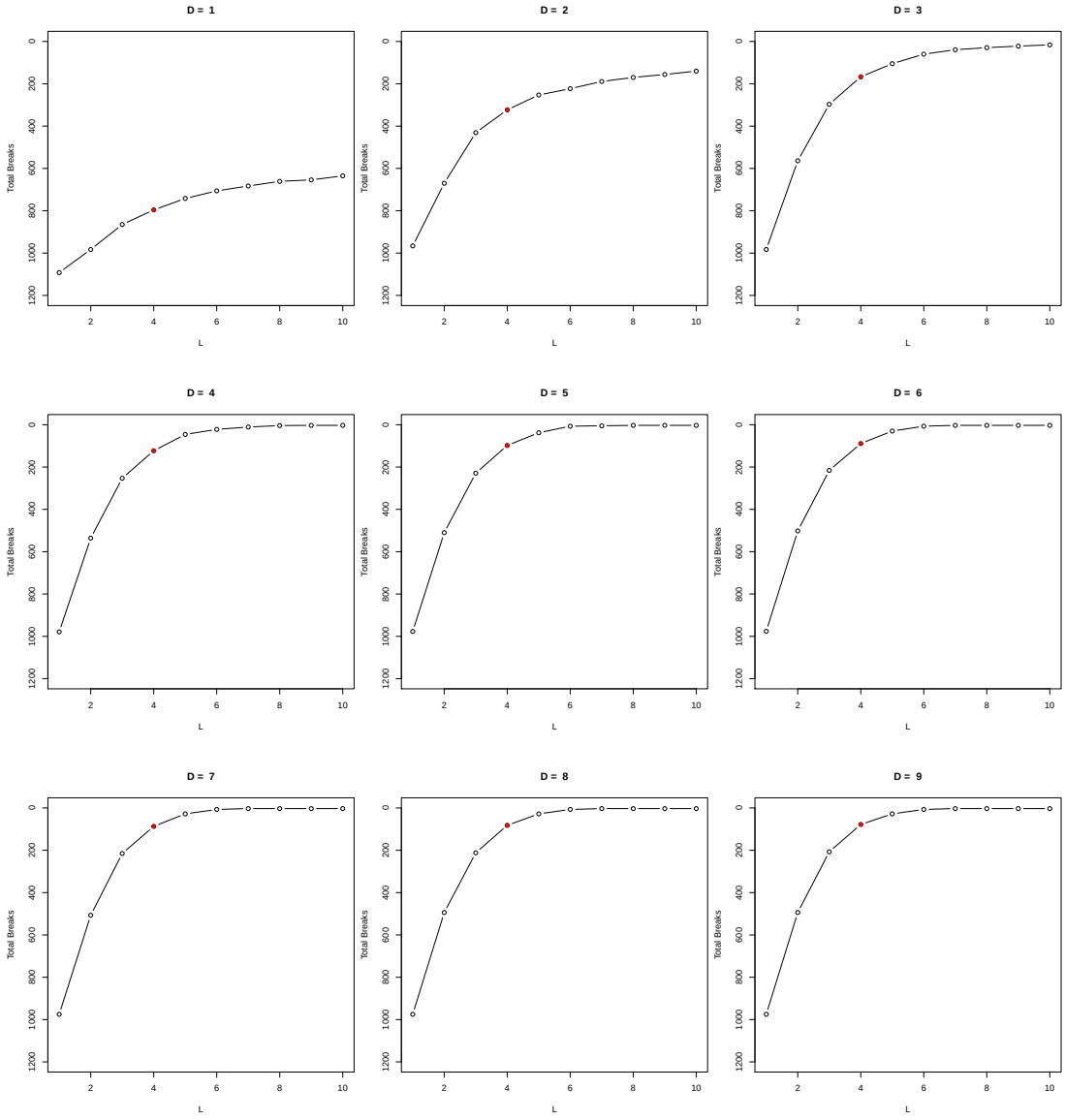

Figure B.2: Number of breakpoints plotted against penalty term coefficient L values for various values of penalty term kernel parameter D. Chosen value $L = 4$ is shown in red.

B.2 Metabolite Matches from the Madison-Qingdao Metabolomics Consortium Database (MMCD).

For all searches, a tolerance of 10 parts per million (ppm) was used. Metabolites matched were for the [M+H]+ species (no adducts) and the sodium adduct [M+Na]+.

Table B.2: Cluster 1 MMCD Matches

Input Mass	Type	Database Mass	Formula	Name
196.098814	[M+H]+	195.0929141	C7H17NO3S	ETHYL DIMETHYL AMMONIO PROPANE SULFONATE
197.1022449	[M+Na]+	174.1116757	C6H14N4O2	L-Arginine;(S)-2-Amino-5-guanidinovaleric acid
197.1022449	[M+Na]+	174.1116757	C6H14N4O2	D-Arginine;D-2-Amino-5-guanidinovaleric acid
199.1680346	[M+H]+	198.16198	C12H22O2	(-)-Menthyl acetate;l-Menthyl acetate
199.1680346	[M+H]+	198.16198	C12H22O2	Citronellyl acetate;3,7-Dimethyl-6-octen-1-yl acetate
199.1680346	[M+H]+	198.16198	C12H22O2	5-Dodecenoic acid
210.1145263	[M+H]+	209.1051934	C11H15NO3	Propoxur;Aprocarb;2-Isopropoxyphenyl N-methylcarbamate; drug
210.1145263	[M+H]+	209.1051934	C11H15NO3	Tyr-OEt
210.1145263	[M+H]+	209.1051934	C11H15NO3	p-Lactophenetide;4'-Ethoxylactanilide
210.1145263	[M+H]+	209.1051934	C11H15NO3	N,O-DIMETHYL-L-TYROSINE
210.1145263	[M+H]+	209.1051934	C11H15NO3	3-HYDROXY-4-AMINO-5-PHENYLPENTANOIC ACID
261.0901498	[M+H]+	260.0830777	C10H16N2O4S	d-biotin d-sulfoxide
261.0901498	[M+Na]+	238.0987278	C8H18N2O4S	1-Piperazineethanesulfonic acid, 4-(2-hydroxyethyl)-
264.9879506	[M+H]+	263.9800049	C4H10O9P2	3-HYDROXY-2-OXO-4-PHOPHONOXY- BUTYL)-PHOSPHONIC ACID
280.0764361	[M+H]+	279.0695568	C17H10FNO2	3-(3-FLUORO-4-HYDROXYPHENYL)-7-HYDROXY-1-NAPHTHONITRILE
371.3121317	[M+H]+	370.3083098	C22H42O4	Di(2-ethylhexyl) adipate;Dioctyl adipate
459.3476132	[M+Na]+	436.35526	C27H48O4	5-b-Cholestane-3a-7-tetraol
459.3476132	[M+Na]+	436.35526	C27H48O4	Cholestane-3,7,12,25-tetrol
459.3476132	[M+Na]+	436.35526	C27H48O4	3alpha,7alpha,12alpha,26-Tetrahydroxy-5beta-cholestane; 5beta-Cholestane-3alpha,7alpha,12alpha,26-tetraol; 5beta-Cholestane-3alpha,7alpha,12alpha,26-tetrol
459.3476132	[M+Na]+	436.35526	C27H48O4	5b-Cholestane-3a,7a,12a,23-Tetrol
582.2948387	[M+H]+	581.2908219	C23H43N5O12	N3'-Acetylapramycin; drug

Table B.3: Cluster 2 MMCD Matches

Input Mass	Type	Database Mass	Formula	Name
132.0759201	[M+H]+	131.0694765	C4H9N3O2	Creatine;alpha-Methylguanidino acetic acid;Methylglycocyamine
132.0759201	[M+H]+	131.0694765	C4H9N3O2	3-Guanidinopropanoate
183.076994	[M+H]+	182.0691422	C8H10N2O3	2-AMINO-4-OXO-4(1H-PYRROL-1-YL)BUTANOIC ACID
183.076994	[M+H]+	182.0707955	C6H15O4P	TRIETHYL PHOSPHATE
183.076994	[M+H]+	182.0707955	C6H15O4P	Diisopropyl phosphate
190.0851209	[M+H]+	189.0789786	C11H11NO2	Phensuximide; drug
190.0851209	[M+H]+	189.0789786	C11H11NO2	1,3-Dimethyl-6,8-isoquinolinediol
190.0851209	[M+H]+	189.0789786	C11H11NO2	INDOLYLPROPIONIC ACID
190.0851209	[M+H]+	189.0789786	C11H11NO2	Backebergine
190.0851209	[M+Na]+	167.0946287	C9H13NO2	Metaraminol; drug
190.0851209	[M+Na]+	167.0946287	C9H13NO2	Phenylephrine; drug
190.0851209	[M+Na]+	167.0946287	C9H13NO2	(-)-Sympatol
190.0851209	[M+Na]+	167.0946287	C9H13NO2	2,5-Dihydrophenylalanine
190.0851209	[M+Na]+	167.0946287	C9H13NO2	Epinine;Deoxyepinephrine
190.0851209	[M+Na]+	167.0946287	C9H13NO2	alpha-methyldopamine
190.0851209	[M+Na]+	167.0946287	C9H13NO2	Ethinamate; drug
190.0851209	[M+Na]+	167.0946287	C9H13NO2	4-methoxytyramine
190.0851209	[M+Na]+	167.0946287	C9H13NO2	p-Synephrine
190.0851209	[M+Na]+	167.0946287	C9H13NO2	PHENYLALANINDIOL
190.0851209	[M+Na]+	167.0946287	C9H13NO2	3-Methoxytyramine
281.1544039	[M+H]+	280.1463299	C19H20O2	demethylmenaquinone
296.129168	[M+H]+	295.1208434	C18H17NO3	Pukateine
296.129168	[M+H]+	295.1208434	C18H17NO3	(+)-Mecambroline
296.129168	[M+H]+	295.1208434	C18H17NO3	Mecambrine
296.129168	[M+H]+	295.1208434	C18H17NO3	Japonine
296.129168	[M+H]+	295.1208434	C18H17NO3	Xylopine
296.129168	[M+H]+	295.1208434	C18H17NO3	Pukateine
296.129168	[M+H]+	295.1208434	C18H17NO3	(+)-Mecambroline
296.129168	[M+H]+	295.1208434	C18H17NO3	Mecambrine
296.129168	[M+H]+	295.1208434	C18H17NO3	Japonine
296.129168	[M+H]+	295.1208434	C18H17NO3	Xylopine
428.3692035	[M+H]+	427.3661591	C25H49NO4	Stearoylcarnitine
459.3476132	[M+Na]+	436.35526	C27H48O4	5-b-Cholestane-3a-7-tetraol
459.3476132	[M+Na]+	436.35526	C27H48O4	Cholestane-3,7,12,25-tetrol
459.3476132	[M+Na]+	436.35526	C27H48O4	3alpha,7alpha,12alpha,26-Tetrahydroxy-5beta-cholestane; 5beta-Cholestane-3alpha,7alpha,12alpha,26-tetraol; 5beta-Cholestane-3alpha,7alpha,12alpha,26-tetrol
459.3476132	[M+Na]+	436.35526	C27H48O4	5b-Cholestane-3a,7a,12a,23-Tetrol

Table B.4: Cluster 3 MMCD Matches

Input Mass	Type	Database Mass	Formula	Name
231.095881	[M+H]+	230.0902716	C9H14N2O5	TETRAHYDRODEOXYURIDINE
231.095881	[M+H]+	230.0902716	C9H14N2O5	Aspartyl-L-proline
231.095881	[M+H]+	230.0902716	C9H14N2O5	3,4-DIHYDRO-1H-PYRIMIDIN-2-ONE NUCLEOSIDE
264.9879506	[M+H]+	263.9800049	C4H10O9P2	3-HYDROXY-2-OXO-4-PHOPHONOXY- BUTYL)-PHOSPHONIC ACID
361.1976803	[M+Na]+	338.2093241	C19H30O5	Shiromodiol diacetate
361.1976803	[M+Na]+	338.2093241	C19H30O5	Idebenone; drug
361.1976803	[M+Na]+	338.2057365	C20H28F2O2	4,4-Difluoro-17beta-hydroxy-17alpha-methyl-androst-5-en-3-one
361.1976803	[M+H]+	360.193674	C21H28O5	Aldosterone;11beta,21-Dihydroxy-3,20-dioxo-4-pregnen-18-al; drug
361.1976803	[M+H]+	360.193674	C21H28O5	Cinerin II
361.1976803	[M+H]+	360.193674	C21H28O5	Cortisone;17alpha,21-Dihydroxy-4-pregnene-3,11,20-trione; Kendall's compound E;Reichstein's substance Fa; drug
361.1976803	[M+H]+	360.193674	C21H28O5	Prednisolone; drug
361.1976803	[M+H]+	360.193674	C21H28O5	Tricyclodehydroisohumulone
404.2032174	[M+Na]+	381.2151377	C20H31NO6	Symphytine;7-Tiglyl-9-(-)-viridiflorylretronecine
404.2032174	[M+Na]+	381.2164751	C21H27N5O2	[PHENYLALANINYL-PROLINYL]-[2-(PYRIDIN-4-YLAMINO)-ETHYL]-AMINE
428.3692035	[M+H]+	427.3661591	C25H49NO4	Stearoylcarnitine
455.13414	[M+Na]+	432.1467406	C20H24N4O5S	METHYL N-[(4-METHYLPHENYL)SULFONYL]GLYCYL-3-[AMINO(IMINO)METHYL]-D-PHENYLALANINATE
455.13414	[M+Na]+	432.1420324	C22H24O9	2-(2,4,5-Trimethoxyphenyl)-5,6,7,8-tetramethoxy-4H-1-benzopyran-4-one
455.13414	[M+Na]+	432.1420324	C22H24O9	(-)-Medicocarpin;Medicarpin 3-O-glucoside
455.13414	[M+Na]+	432.1420324	C22H24O9	2-(3,4,5-Trimethoxyphenyl)-5,6,7,8-tetramethoxy-4H-1-benzopyran-4-one
455.13414	[M+Na]+	432.1420324	C22H24O9	(-)-medicarpin-3-O-glucoside
481.3068156	[M+H]+	480.2988078	C29H40N2O4	Emetine
582.2948387	[M+H]+	581.2908219	C23H43N5O12	N3'-Acetylapramycin; drug

Table B.5: Cluster 4 MMCD Matches

Input Mass	Type	Database Mass	Formula	Name
199.1680346	[M+H]+	198.16198	C12H22O2	(-)-Menthyl acetate;l-Menthyl acetate
199.1680346	[M+H]+	198.16198	C12H22O2	Citronellyl acetate;3,7-Dimethyl-6-octen-1-yl acetate
199.1680346	[M+H]+	198.16198	C12H22O2	5-Dodecenoic acid
231.095881	[M+H]+	230.0902716	C9H14N2O5	TETRAHYDRODEOXYURIDINE
231.095881	[M+H]+	230.0902716	C9H14N2O5	Aspartyl-L-proline
231.095881	[M+H]+	230.0902716	C9H14N2O5	3,4-DIHYDRO-1H-PYRIMIDIN-2-ONE NUCLEOSIDE
263.0352379	[M+H]+	262.0253821	C6H12FO8P	2-DEOXY-2-FLUORO-ALPHA-D-GLUCOSE-1-PHOSPHATE
280.0764361	[M+H]+	279.0695568	C17H10FNO2	3-(3-FLUORO-4-HYDROXYPHENYL)-7-HYDROXY-1-NAPHTHONITRILE
304.0616617	[M+H]+	303.0534452	C10H11ClFN5O3	2-CHLORO-9-(2-DEOXY-2-FLUORO-B -D-ARABINOFURANOSYL) -9H-PURIN-6-AMINE
315.1478418	[M+H]+	314.1412613	C13H22N4O3S	Ranitidine
315.1478418	[M+Na]+	292.1575633	C19H20N2O	16-Epivellosimine
315.1478418	[M+Na]+	292.1575633	C19H20N2O	Vellosimine
317.1130363	[M+Na]+	294.1255944	C19H18O3	METHYL (2Z)-3-METHOXY-2-2-[(E)-2-PHENYLVINYL]PHENYLACRYLATE
317.1130363	[M+Na]+	294.1255944	C19H18O3	(2-Butylbenzofuran-3-yl)(4-hydroxyphenyl)ketone; 2-Butyl-3-(4-hydroxybenzoyl)benzofuran
317.1130363	[M+Na]+	294.1215717	C14H18N2O5	Aspartame; drug
317.1130363	[M+Na]+	294.1215717	C14H18N2O5	Glutamylphenylalanine
317.1130363	[M+Na]+	294.1215717	C14H18N2O5	2-(BETA-D-GLUCOPYRANOSYL)-5-METHYL-1-BENZIMIDAZOLE
320.1609073	[M+Na]+	297.172879	C19H23NO2	Ibuprofen piconol; drug
447.3430143	[M+H]+	446.33961	C28H46O4	Stylisterol B
447.3430143	[M+H]+	446.33961	C28H46O4	3-dehydroteasterone;dehydroteasterone
447.3430143	[M+H]+	446.33961	C28H46O4	Di-n-decyl phthalate;Didecyl phthalate;Didecyl 1,2-Benzenedicarboxylate
447.3430143	[M+H]+	446.33961	C28H46O4	Diisodecyl phthalate
455.2918962	[M+Na]+	432.3028305	C30H40O2	4,4'-Diaponeurosporenic acid
455.2918962	[M+H]+	454.2831577	C27H38N2O4	Verapamil; drug

Table B.6: Cluster 5 MMCD Matches, Part 1

Input Mass	Type	Database Mass	Formula	Name
120.0470801	[M+H]+	119.0404846	C4H9NOS	S-Acetylthioethanolamine
142.0287976	[M+Na]+	119.0404846	C4H9NOS	S-Acetylthioethanolamine
148.0593598	[M+H]+	147.0531578	C5H9NO4	3-METHYL-ASPARTIC ACID
148.0593598	[M+H]+	147.0531578	C5H9NO4	L-Glutamate;L-Glutamic acid;L-Glutaminic acid
148.0593598	[M+H]+	147.0531578	C5H9NO4	N-ACETYL-SERINE
148.0593598	[M+H]+	147.0531578	C5H9NO4	Glutamate;Glutaminic acid;2-Aminoglutaric acid
148.0593598	[M+H]+	147.0531578	C5H9NO4	N-Methyl-D-aspartic acid;N-Methyl-D-aspartate;NMDA
148.0593598	[M+H]+	147.0531578	C5H9NO4	O-Acetyl-L-serine;O3-Acetyl-L-serine
148.0593598	[M+H]+	147.0531578	C5H9NO4	2-Oxo-4-hydroxy-5-aminovalerate
148.0593598	[M+H]+	147.0531578	C5H9NO4	L-threo-3-Methylaspartate
148.0593598	[M+H]+	147.0531578	C5H9NO4	N-HYDROXY-N-ISOPROPYLOXAMIC ACID
148.0593598	[M+H]+	147.0531578	C5H9NO4	N-(Carboxymethyl)-D-alanine
148.0593598	[M+H]+	147.0531578	C5H9NO4	L-4-Hydroxyglutamate semialdehyde
148.0593598	[M+H]+	147.0531578	C5H9NO4	Isoglutamate;Isoglutamic acid;3-Aminopentanedioic acid
153.0760536	[M+H]+	152.0684735	C5H12O5	Xylitol
153.0760536	[M+H]+	152.0684735	C5H12O5	D-Ribitol;D-Adonitol
153.0760536	[M+H]+	152.0684735	C5H12O5	L-Arabitol;L-Arabinol;L-Arabinitol;L-Lyxitol
153.0760536	[M+H]+	152.0684735	C5H12O5	D-Arabitol;D-Arabinitol;D-Arabinol;D-Lyxitol
179.0156802	[M+H]+	178.0088501	C9H6O2S	3-Hydroxy-2-formylbenzothiophene
183.076994	[M+H]+	182.0691422	C8H10N2O3	2-AMINO-4-OXO-4(1H-PYRROL-1-YL)BUTANOIC ACID
183.076994	[M+H]+	182.0707955	C6H15O4P	TRIETHYL PHOSPHATE
183.076994	[M+H]+	182.0707955	C6H15O4P	Diisopropyl phosphate
183.076994	[M+H]+	182.0691422	C8H10N2O3	2-AMINO-4-OXO-4(1H-PYRROL-1-YL)BUTANOIC ACID
183.076994	[M+H]+	182.0707955	C6H15O4P	TRIETHYL PHOSPHATE
183.076994	[M+H]+	182.0707955	C6H15O4P	Diisopropyl phosphate
219.0959433	[M+H]+	218.0902716	C8H14N2O5	L-Ala-gamma-D-Glu
219.0959433	[M+H]+	218.0902716	C8H14N2O5	gamma-L-Glutamyl-D-alanine
231.095881	[M+H]+	230.0902716	C9H14N2O5	TETRAHYDRODEOXYURIDINE
231.095881	[M+H]+	230.0902716	C9H14N2O5	Aspartyl-L-proline
231.095881	[M+H]+	230.0902716	C9H14N2O5	3,4-DIHYDRO-1H-PYRIMIDIN-2-ONE NUCLEOSIDE
241.029228	[M+Na]+	218.0401503	C12H10O2S	1,1'-BIPHENYL-2-SULFINIC ACID
241.029228	[M+Na]+	218.0401503	C12H10O2S	cis-1,2-Dihydroxy-1,2-dihydrodibenzothiophene
241.029228	[M+H]+	240.0238483	C6H12N2O4S2	L-Cystine;L-Dicysteine;L-alpha-Diamino-beta-dithiolactic acid
241.029228	[M+H]+	240.0238483	C6H12N2O4S2	Cystine;Dicysteine;alpha-Diamino-beta-dithiolactic acid
247.105723	[M+Na]+	224.1160924	C11H16N2O3	(6,10-DIOXO-OCTAHYDRO-PYRIDAZINO[1,2-A][1,2]DIAZEPIN-1-YL)-ACETALDEHYDE FRAGMENT

Table B.7: Cluster 5 MMCD Matches, Part 2

Input Mass	Type	Database Mass	Formula	Name
247.105723	[M+Na]+	224.1160924	C11H16N2O3	2-(4-AMINO-PHENYL)-5-HYDROXYMETHYL-PYRROLIDINE-3,4-DIOL
247.105723	[M+H]+	246.1004423	C13H14N2O3	Mephobarbital; drug
247.105723	[M+H]+	246.1004423	C13H14N2O3	5-AMINO-4-OXO-1,2,4,5,6,7-HEXAHYDRO-AZEPINO[3,2,1-HI] INDOLE-2-CARBOXYLIC ACID
247.105723	[M+H]+	246.1004423	C13H14N2O3	Alanine, N-indol-3-ylacetyl- (6CI) Indole-3-acetylalanine N-(3-Indolylacetyl)-L-alanine
247.105723	[M+H]+	246.1004423	C13H14N2O3	N-Acetyl-D-tryptophan
247.105723	[M+Na]+	224.1160924	C11H16N2O3	(6,10-DIOXO-OCTAHYDRO-PYRIDAZINO[1,2-A][1,2]DIAZEPIN-1-YL)-ACETALDEHYDE FRAGMENT
247.105723	[M+Na]+	224.1160924	C11H16N2O3	2-(4-AMINO-PHENYL)-5-HYDROXYMETHYL-PYRROLIDINE-3,4-DIOL
247.105723	[M+H]+	246.1004423	C13H14N2O3	Mephobarbital; drug
247.105723	[M+H]+	246.1004423	C13H14N2O3	5-AMINO-4-OXO-1,2,4,5,6,7-HEXAHYDRO-AZEPINO[3,2,1-HI]INDOLE-2-CARBOXYLIC ACID
247.105723	[M+H]+	246.1004423	C13H14N2O3	Alanine, N-indol-3-ylacetyl- (6CI) Indole-3-acetylalanine N-(3-Indolylacetyl)-L-alanine
247.105723	[M+H]+	246.1004423	C13H14N2O3	N-Acetyl-D-tryptophan
247.1269753	[M+H]+	246.1215717	C10H18N2O5	2,4-BIS(ACETYLAMINO)-1,5-ANHYDRO-2,4-DIDEOXY-D-GLUCITOL
261.1426419	[M+H]+	260.1347193	C15H20N2S	Methaphenilene
261.1426419	[M+H]+	260.1372218	C11H20N2O5	(E)-N 6 -[3-CARBOXY-1-(HYDROXYMETHYL)PROPYLIDENE]-L-LYSINE
263.0352379	[M+H]+	262.0253821	C6H12FO8P	2-DEOXY-2-FLUORO-ALPHA-D-GLUCOSE-1-PHOSPHATE
271.0380657	[M+H]+	270.0310421	C10H10N2O5S	2-(OXALYL-AMINO)-4,5,6,7-TETRAHYDRO-THIENO[2,3-C]PYRIDINE-3-CARBOXYLIC ACID
276.1167155	[M+H]+	275.1117353	C10H17N3O6	Norophthalmic acid
277.1005288	[M+H]+	276.0939004	C22H12	Benzo[ghi]perylene;1,12-Benzoperylene
277.1005288	[M+Na]+	254.1088985	C12H18N2O2S	Thiamylal;5-Allyl-5-(1-methylbutyl)-2-thiobarbituric acid; drug
277.1005288	[M+H]+	276.0957509	C10H16N2O7	Glu-Glu
277.1005288	[M+H]+	276.0957509	C10H16N2O7	gamma-glutamyl-D-glutamate
281.1544039	[M+H]+	280.1463299	C19H20O2	demethylmenaquinone
282.2772066	[M+H]+	281.2718648	C18H35NO	(9E)-OCTADEC-9-ENAMIDE
304.0616617	[M+H]+	303.0534452	C10H11ClFN5O3	2-CHLORO-9-(2-DEOXY-2-FLUORO-B -D-ARABINOFURANOSYL)-9H-PURIN-6-AMINE
313.1412633	[M+Na]+	290.1518092	C17H22O4	Laurenobiolide
313.1412633	[M+Na]+	290.1518092	C17H22O4	epi-Tulipinolide
313.1412633	[M+Na]+	290.1518092	C17H22O4	Tulipinolide
313.1412633	[M+H]+	312.1361591	C19H20O4	Montanin A
313.1412633	[M+H]+	312.1361591	C19H20O4	Butylbenzyl phthalate;Butyl phenylmethyl 1,2-benzenedicarboxylate

Table B.8: Cluster 5 MMCD Matches, Part 3

Input Mass	Type	Database Mass	Formula	Name
313.1412633	[M+H]+	312.1361591	C19H20O4	m-(beta-Acetyl-alpha-ethyl-p-hydroxyphenethyl)benzoic acid
313.1412633	[M+H]+	312.1361591	C19H20O4	4'-Prenyloxyresveratrol
320.1609073	[M+Na]+	297.172879	C19H23NO2	Ibuprofen piconol; drug
328.282355	[M+H]+	327.2728499	C4H22B18Co	COBALT BIS(1,2-DICARBOLLIDE)
338.073982	[M+H]+	337.0667161	C13H15N5O2S2	2-[3-(5-MERCAPTO-[1,3,4]THIADIAZOL-2-YL)-UREIDO]-N-METHYL-3-PHENYL-PROPIONAMIDE
338.073982	[M+Na]+	315.084848	C18H18ClNS	Chlorprothixene; drug
338.073982	[M+H]+	337.067501	C10H16N3O8P	5-Hydroxymethyldeoxycytidylate;2'-Deoxy-5-hydroxymethylcytidine 5'-phosphate
338.073982	[M+H]+	337.067501	C10H16N3O8P	N4-METHOXY-2'-DEOXY-CYTIDINE-5'-MONOPHOSPHATE
338.073982	[M+H]+	337.067501	C10H16N3O8P	5-METHYLCYTIDINE-5'-MONOPHOSPHATE
338.073982	[M+H]+	337.067501	C10H16N3O8P	O2'-METHYLYCYTIDINE-5'-MONOPHOSPHATE
338.073982	[M+Na]+	315.0868632	C17H15O6+	Rosinidin
404.2032174	[M+Na]+	381.2151377	C20H31NO6	Symphytine;7-Tiglyl-9-(-)-viridiflorylretronecine
404.2032174	[M+Na]+	381.2164751	C21H27N5O2	[PHENYLALANINYL-PROLINYL]-[2-(PYRIDIN-4-YLAMINO)-ETHYL]-AMINE
404.2032174	[M+Na]+	381.2151377	C20H31NO6	Symphytine;7-Tiglyl-9-(-)-viridiflorylretronecine
404.2032174	[M+Na]+	381.2164751	C21H27N5O2	[PHENYLALANINYL-PROLINYL]-[2-(PYRIDIN-4-YLAMINO)-ETHYL]-AMINE
422.023739	[M+H]+	421.0170197	C9H13BeF3N2O9P2	N-METHYL O-NITROPHENYL AMINOETHYLDIPHOSPHATE BERYLLIUM TRIFLUORIDE
422.023739	[M+Na]+	399.0326183	C17H22BrNOS2	Timepidium bromide;Timepidium bromide anhydrous
426.0157679	[M+Na]+	403.0296619	C17H13N3O5S2	Sulfathalidine;Phthalylsulfathiazole;2-[[[4-[(2-Thiazolylamino)sulfonyl] phenyl]amino]carbonyl]benzoicacid; drug
426.0157679	[M+Na]+	403.0227387	C15H15ClFN3O3S2	Fluthiacet methyl
455.2918962	[M+Na]+	432.3028305	C30H40O2	4,4'-Diaponeurosporenic acid
455.2918962	[M+H]+	454.2831577	C27H38N2O4	Verapamil; drug
455.2918962	[M+Na]+	432.3028305	C30H40O2	4,4'-Diaponeurosporenic acid
455.2918962	[M+H]+	454.2831577	C27H38N2O4	Verapamil; drug
465.0523806	[M+Na]+	442.0615992	C23H18FeN2O4+2	[[2,2'-[4-CARBOXYETHYL-1,2-PHENYLENEBIS(NITRILOMETHYLIDYNE)] BIS[PHENOLATO]](2-)-N,N',O,O']-IRON
481.3068156	[M+H]+	480.2988078	C29H40N2O4	Emetine
481.3068156	[M+H]+	480.2988078	C29H40N2O4	Emetine
508.33502	[M+Na]+	485.3464863	C25H47N3O6	3-HYDROXY-6-METHYL-4-(3-METHYL-2-(3-METHYL-2-(3-METHYL-BUTYRYLAMINO)-BUTYRYLAMINO)-BUTYRYLAMINO)-HEPTANOIC ACID ETHYL ESTER
527.3014379	[M+Na]+	504.3172518	C24H40N8O4	Dipyridamole; drug
558.2932796	[M+H]+	557.2901142	C30H40FN3O6	N-[(BENZYLOXY)CARBONYL]LEUCYL-N 1 -[3-FLUORO-1-(4-HYDROXYBENZYL)-2-OXOPROPYL]LEUCINAMIDE
582.2948387	[M+H]+	581.2908219	C23H43N5O12	N3'-Acetylapramycin; drug