NONPARAMETRIC BAYESIAN MODELS IN EXPRESSION PROTEOMIC APPLICATIONS

by

Leanna Lynn House

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Merlise Clyde, Supervisor

_____
Dr. Robert L. Wolpert, Supervisor

_____
Dr. James O. Berger

_____
Dr. Michael C. Fitzgerald

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2006

ABSTRACT

(Statistics)

NONPARAMETRIC BAYESIAN MODELS IN EXPRESSION PROTEOMIC APPLICATIONS

by

Leanna Lynn House

Institute of Statistics and Decision Sciences
Duke University

Date: _____

Approved:

_____
Dr. Merlise Clyde, Supervisor

_____
Dr. Robert L. Wolpert, Supervisor

_____
Dr. James O. Berger

_____
Dr. Michael C. Fitzgerald

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2006

Abstract

Bayesian nonparametric analyses develop probability models on very high, possibly infinite, dimensional function spaces. However, with the benefits of exploring large parameter spaces comes the responsibility of controlling potentially over-parameterized models. With thoughtful prior elicitation, Bayesian methods may naturally impose model complexity restrictions depending upon whether a function is defined by a collection of random components or as one random variable. This dissertation, via the progression of three separate works, takes advantage of two ways prior distributions may penalize complex functions in nonparametric analyses of expression proteomic data .

Since all cellular functions are carried out by proteins, the primary purpose for expression proteomics is to assess from differences in protein production how an organism responds under various conditions. One common way to assess the differences is to analyze the protein content of varying biological samples using Matrix Assisted Laser Desorption/Ionization Time-of-Flight (MALDI-TOF) mass spectrometry (MS). Although, MALDI-TOF MS has many analytical benefits, inherent within the technology are sources of measurement error that make deciphering true signal from noise difficult. Thus, all expression proteomic studies that use MALDI-TOF MS data must first extract data of interest from mass spectra before making inference.

Chapters 2 and 3 are devoted solely to developing nonparametric Bayesian models to identify significant features from individual spectra. Both models estimate an unknown function f that represents true protein signal as a weighted sum of J kernel functions with either prespecified or data-determined J location

parameters. In the prespecified case, a truncated exponential prior on the coefficients regularize the proposed over-parameterized model. In the unspecified case, the function f itself is assumed to be a random variable for which a Lévy random field prior is elicited. The process prior of f penalizes complex models and is comparable to specifying a joint prior distribution on J, kernel function parameters, and the basis coefficients.

Chapter 4 expands the model presented in Chapter 3 to include multiple spectra from two sub-populations. Underlying every observed spectra, regardless of sub-population, is one mean-spectrum that is modeled similarly to f as described in Chapter 3 with a Lévy random field prior. The difference is an added dimension to the random field that represents sub-population association. With the new dimension, the proposed model extracts and aligns population significant features and compares them to make treatment group classifications.

Acknowledgements

Working to complete this dissertation during the last five years has been both a rewarding and humbling experience: rewarding because this 100-page document describes only a fraction of what I learned as a student in the Institute of Statistics and Decision Sciences, and humbling because each day I encountered new concepts (both statistical and life related) that I did not know, needed to know, or wanted to know. Further, I routinely met people who seemed to always know more. These are the people I would like acknowledge and thank sincerely for sharing some of their knowledge with me.

I would first like to thank my advisors Dr. Merlise Clyde and Dr. Robert Wolpert for their never-ending patience and guidance. Their obvious passion to teach and develop innovative research is inspirational. I will sincerely miss Robert's sense for detail and Merlise's open door policy once I leave Duke. From Duke, I would also like to thank my remaining committee members, Dr. Jim Berger and Dr. Michael Fitzgerald, for their insightful comments, as well as, Dr. David Banks, Dr. Dalene Stangl, Dr. Michael Lavine, and Dr. Sayan Mukherjee who were willing to answer questions and offer sound advice when needed.

I am extremely grateful for the friends I made will attending Duke. In particular, I would like to thank past and present officemates Chris, Carlos, Jarad, Chong, Eric, and Ana who endured my venting after tough meetings or while trying to debug programs; the staff at ISDS, Kris, Pat, Susan, Anne, and Lance for their invaluable assistance; Beth Brown and her family who taught me that choosing to do what you love is always the right choice; Karen Gonzalez who is a wonderful person that listens intently and helps whenever needed despite what

is going on in her life; my “fitness friends”, Amy, Katharine, Laura, Jen, and Megan who live, not just teach, the life lessons of balance between career, family, health, and friendship; and Jaime Palter, a friend who might as well be family, who breathes sensibility, responsibility, and brilliance.

A special thank you is reserved for Scotland Leman. Scotland’s presence in a room is always known because of his larger than life personality. Getting to know that personality and how it came to be has been (and continues to be) a wonderful adventure.

Finally, I would like to thank my family, Mom, Dad, Michael, and now, Cristina. Words cannot describe how grateful I am for your faithful support and encouragement. I love being a member of the House family, and I dedicate this dissertation to you.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Within the Bayesian paradigm, completing a nonparametric data analysis as originally defined by Dr. Jacob Wolfowitz in 1942 is impossible. The original definition referred to *distribution free*, statistical techniques that avoided potentially restrictive or erroneous assumptions concerning the probability density or mass functions of random variables; yet, Bayesian methods require a well defined likelihood distribution and probability measures on all parameters. Thus, Bayesians redefine the meaning of nonparametric analyses and develop a definition that maintains the spirit of Wolfowitz's intentions and applies to flexible, statistical models which may summarize data that cannot or do not support typical analysis assumptions. Since model flexibility may result from the addition of parameters, or equivalently, expansion of the parameter space, Bayesian *non*parametrics, is accepted ironically as "ordinary Bayesian analysis with an infinite [or extremely large] dimensional parameter space" [Wolpert, 2002; Bernardo and Smith, 1994; Müller and Quintina, 2004]. Within the applied field of expression proteomics, this dissertation develops three novel, nonparametric data analysis approaches.

1.1 Expression Proteomics and MALDI-TOF

Expression proteomics is the study of all protein forms observed within an organism as a function of time, age, state, or other external factor(s). Since, all cellular functions are carried out by proteins [Martin and Nelson, 2001] the aim is to assess how an organism responds under various conditions and to answer questions including "What proteins are present?", "How do the proteins work together in signaling pathways?", and "What protein differences or changes may drive the development, repair, breakdown, and/or death of an organism?" [Wu *et al.*, 2002]. For example, one expression proteomic, observational study discussed within this dissertation called for serum samples from lung cancer and control patients. The purpose of the study was to gain an understanding of how the body responds to lung cancer and to find protein *biomarker(s)* that may differentiate patient disease states.

The protein content of samples can be summarized by spectra acquired from Matrix Assisted Laser Desorption/ Ionization Time-of-Flight (MALDI-TOF) mass spectrometers. MALDI-TOF mass spectrometers have four primary components: a sample inlet, an ionization source, a mass analyzer, and an ion detector. On a plate, samples insert directly into an ion source, where a laser excites, vaporizes, and ionizes molecules so that they "fly" through the instrument's field free region and separate according to size and charge. At the end of the region, an ion detector measures at uniform time intervals the intensity of an electrical current that is proportional to the number of detected ions. Upon converting the time units to molecular mass divided by charge (m/z[1] in unit Da/e), MALDI-TOF

[1]Standard mass spectrometry literature does not provide units for m/z. However, mass is measured in Daltons (Da) and e represents the change an electron in coulombs.

mass spectrometers produce a spectrum or *proteomic profile* from either one or the sum of multiple laser shot(s). Figure 1.1 is a spectrum of a serum sample that was generated from the sum of ten laser shots.

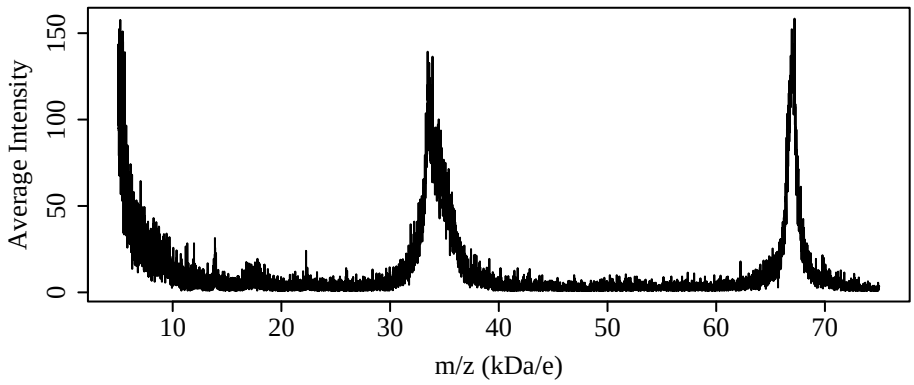


Figure 1.1: Example of a MALDI-TOF spectrum from the sum of 10 laser shots.

One spectrum may contain tens of thousands of paired {m/z, intensity} measurements, and one expression proteomic experiment may require hundreds of spectra. However, each observation will be influenced by multiple sources of measurement error and may not necessarily represent an actual protein. Thus, how to combine and compare thousands or millions of observations within and across spectra, subjects, and treatment groups deems a difficult question. To answer it, requires a thorough understanding of the data issues inherent in MALDI-TOF.

Five data quality terms tend to describe MALDI-TOF data issues: resolution, calibration/alignment, background, scaling, and noise [Morris *et al.*, 2005; Baggerly *et al.*, 2004; Coombes *et al.*, 2005a].

1. **Resolution:** For a variety of reasons including unequal kinetic energy dispersions from the ionizing laser, the spacial location of molecules in the

matrix, and the size of the protein molecules, ions of equal size and charge may not reach the TOF detector at the same time. Hence, measurements for one protein may occur over a range of TOFs and appear as peaks within a spectrum. The width of spectrum peaks in part reflects the quality of the data and is summarized by resolution. The resolution of a spectrum is typically calculated from the Full Width Half Mass procedure and represents the distance needed between two peaks in order to distinguish them [p.g. 74 Dass, 2001]. The distance will be small for spectra with narrow peaks and large for spectra with wide peaks; high resolution spectra have narrow peaks where as low resolution spectra have wide peaks.

2. **Calibration/Alignment:** The process by which time is converted to m/z is referred to as calibration. Proteins or *calibrants* with known masses are either injected into a given sample or analyzed outside the sample to collect data and fit a two or higher degree polynomial regression function that converts time to mass. The regression function is typically referred to as a calibration equation and maps time to mass within $\pm 0.1\% - \pm 0.3\%$ [Wang *et al.*, 2003; Coombes *et al.*, 2005a].

   Accurate calibration equations are needed in order to compare measurements across multiple spectra. Because molecules of equal size and charge may not reach the detector at the same time, protein peaks from different spectra may not align accurately. Although, mis-alignment can be corrected with well calibrated masses, the calibration procedure can be inexact and force researchers to use another means to match peaks across spectra. Morris *et al.* [2005] stresses that peak matching is currently more of a game than a science, and too many ad hoc decisions can lead to mismatched peaks.

3. **Scale:** The TOF detector quantifies the presence of proteins by recording an electrical current proportional to the number of ions [Dass, 2001, p.g. 85]. The proportional constant scales spectra and is extremely sensitive to small deviances in sample protocol, e.g. laser intensity, sample concentrations, and laboratory differences.

4. **Background:** Typical spectra display two types of background that propagate from different sources in MALDI-TOF. The first type is time-independent and tends to result from detector ringing- an apparent thermal noise that averages to be greater than zero. The non-zero average multiplied by the scaling factor vertically lifts spectra and appears as a constant background term. The second type is time-dependent and is due to small matrix molecules interrupting the detector. Typically, the interruption decays exponentially in time [Coombes *et al.*, 2005a].

5. **Noise:** The distribution of random error, like background, appears to depend either on time or mean intensity. However, current approaches tend to assume constant variance within and between spectra.

Because of these data issues, comparisons of raw spectra are arguably useless; feature extraction and data cleaning procedures are necessary to assure that inference from expression proteomic studies are based on biologically relevant data and not spectrum anomalies induced by measurement error [Morris *et al.*, 2005; Baggerly *et al.*, 2004]. Nonparametric Bayesian models work effectively to isolate the features of interest while accounting for the issues that arise within MALDI-TOF data.

1.2 Bayesian Nonparametrics

In Bayesian nonparametric analyses, probability distributions are developed on extremely high, possibly infinite, dimensional function spaces [Müller and Quintina, 2004] in order to explore flexible data representations. However, with the flexibility of modeling an unknown function from a parameter space with a dimension that equals or exceeds the dimension of the data, comes the responsibility of controlling potentially over-parameterized models. Bayesian methods naturally impose model complexity restrictions through prior distributions. For example, in frequentist terms, the posterior mode under a Gaussian model is the parameter estimate that minimizes a squared error loss function with the penalty set to the logarithm of the prior distribution. [p.g.34 Hastie *et al.*, 2001]. In this dissertation, I explore two ways prior distributions in nonparametric analyses may penalize complex functions depending upon whether a function is defined by a collection of random components or as one random variable.

In the case of random components, such as basis expansion models (Hastie *et al.* ch. 5 2001; Kohn *et al.* 2001) including non-adaptive kernel regression, smoothing, and wavelets, a function f is defined by a collection of p (e.g. for decimated wavelets, p equals the sample size) random basis coefficients and the prior on f is obtained from the linear combination of p distributions. The priors regularize the parameter estimates to remove or reduce the influence of insignificant components. Mixtures of normals with point mass priors at zero [Chipman *et al.*, 1997; Clyde *et al.*, 1998] and scale mixtures of normal priors [Vidakovic, 1998; Johnstone and Silverman, 1997; Brown and Griffin, 2005] have been shown to be effective for many applications. However, without transforming spectro-

metric data, such as Morris *et al.* [2006] who transform intensities to the wavelet domain, these priors do not apply well when assessing expression proteomic data. In Chapter 2, I describe a basis expansion approach for spectrum feature extraction and control the model complexity by eliciting a prior on each basis coefficient that is a scale mixture of truncated normal distributions.

Another way to assure sparse model estimation is to consider f itself as a random variable and elicit a prior directly on f. The prior is referred to as random probability measure (RPM) [Müller and Quintina, 2004]. Examples of RPMs are stochastic process (or random field) priors, including Gaussian, Dirichlet, and Lévy [e.g. O'Hagan and Kingman, 1978; Antoniak, 1974; Wolpert *et al.*, 2003], which rely on the fact that stochastic processes are random functions. Random functions, similar to stochastic processes, can be viewed as a collection of random variables; given a random function f that maps t in $\mathbb{T}$ to $X_t = f(t)$, f can be characterized by a collection of random variables indexed by $\mathbb{T}$, $\{f(t)|t \in \mathbb{T}\}$. Thus, to elicit a prior distribution on the random function is comparable to eliciting a prior on the collection of random variables, as achieved by process prior distributions. Common applications for process priors include density estimation, the mean estimation of generalized linear models, and kernel regression.

This dissertation utilizes kernel regression or locally weighted linear regression models of the form,

$$f(t) = \mathrm{E}[y|t] = \sum_{j=1}^{J} k(t, \theta_j)\eta_j$$

where $\boldsymbol{\eta}$ represents a $1 \times J$ vector that weights each respective kernel function of t with parameter $\theta \in \Theta$. Müller *et al.* [1996] propose a random measure on the joint distribution of (y, t) that results via standard Bayesian theory to posterior

estimates of θ_j, η_j, and the expected regression curve $\mathrm{E}[y|t]$. Wolpert and Ickstadt [1998a] however, utilize a Lévy process prior to assign random measures $\Gamma(d\theta)$ on Θ which weight the kernels. Wolpert and Ickstadt's approach is very similar to the basis expansion models, however, the random measure inherently elicits a joint prior distribution on all adaptive kernel parameters and coefficients. We expand on Wolpert and Ickstadt's work in Chapters 3 and 4, and, within the context of expression proteomics, and demonstrate the benefits of using weighted, adaptive kernel regression with Lévy processes. In the next section, I describe useful properties of Lévy processes.

1.3 Lévy Process Priors

Given spectra similar to the one displayed in Figure 1.1, the goal is to find a function of m/z or of time t (the raw measurements taken from the TOF detector) that estimates the expected intensity; ignoring the inclusion of parameters for the aforementioned data issues, let Y_t represent observed intensities at time t, e_t represent independent error terms with mean zero, and assume

$$Y_t = f(t) + e_t.$$

Given the nature of the spectrometric data, the function $f(t)$ should be positive with distinguishable features at a priori unknown locations and unknown amplitudes. These necessary characteristics of $f(t)$ motivate constructing the function from a pure-jump, increasing, random function or stochastic process X_t. Figure 1.2 displays a construction of f: given the number J, locations τ, and heights η of jumps in X_t, let f equal the convolution $\sum_{j=1}^{J} k(t, \tau_j)\eta_j$.

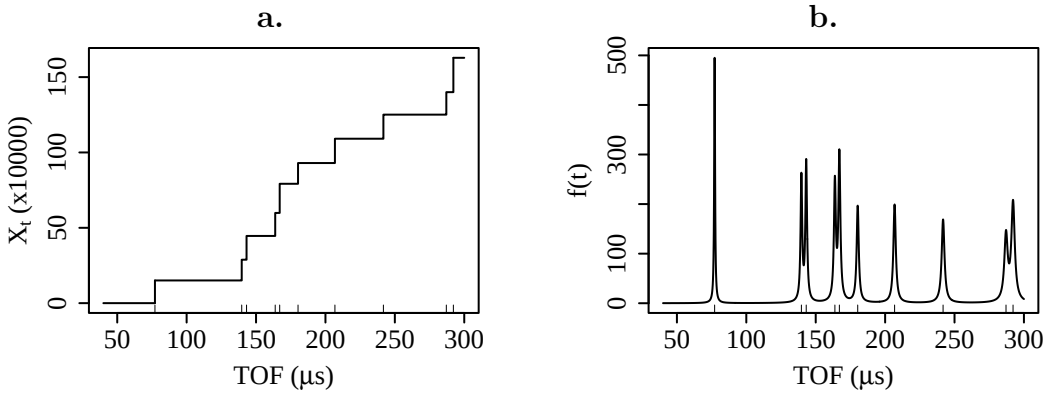


Figure 1.2: Plot a. represents a realized, pure jump, increasing stochastic process X_t. Plot b. plots $f(t)$ that generated from X_t, where $f(t) = \sum_{j=1}^{J} k(t, \tau_j)\eta_j$.

Lévy processes, as represented using the Lévy-Khintchine theorem, have characterizable pathways which, under certain conditions, are pure-jump, increasing functions. A Lévy process is a real-valued, stochastic process with stationary, independent increments (SII), defined as follows:

Definition 1.1. *Let* $X_t = \{X_{t_0}, ..., X_{t_i}, X_{t_{i+1}}, ..., X_{t_N}\}$ *where* $N < \infty$ *and* $X_{t_a} \in \mathbb{R}$ X_t *is a Lévy process if it satisfies the following:*

1. X_t *has independent increments,* $X_{t_{i+1}} - X_{t_i}$
2. X_t *has stationery increments:*
   the distribution of $X_{t_{i+1}} - X_{t_i}$ *does not depend on* t_i
3. X_t *is stochastically continuous: as* $t_{i+1} \to t_i$, $X_{t_{i+1}} \to X_{t_i}$ *in probability.*
   [Papapantoleon, 2005; Wolpert, 2002]

From the definition, each increment $X_{t_{i+1}} - X_{t_i}$ in a Lévy process can be written as the sum of n mutually exclusive increments which are, by the stationery property, identically distributed. Thus, SII processes are examples of infinitely divisible random variables,

Definition 1.2. *A random variable X is infinitely divisible when it equals in distribution the sum of n identically distributed random variables. [Papapantoleon, 2005; Wolpert, 2002]*

Since every Lévy process X_t is infinitely divisible, the celebrated Lévy-Khintchine provides the form of its characteristic function [Papapantoleon, 2005];

Theorem 1.1. Lévy-Khintchine *The law of a random variable is infinitely divisible if and only if there exists a triplet $\{m, \sigma^2, \nu\}$ with $m \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$, and ν being a measure that satisfies $\nu(0) = 0$ and $(1 \wedge \eta^2)\nu(d\eta) < \infty$, such that*

$$E[\exp\{iwX_t\}] = \exp\Big\{iw(x_0 + tm)t\frac{\sigma^2 w^2}{2} + t\int_{\mathbb{R}/0}\big[e^{iw\eta} - 1 - iw\eta\mathbf{1}_{|\eta|<1}\big]\nu(d\eta)\Big\}.$$

For X_t, the parameters x_0, m, σ^2, and $\nu(d\eta)$ refer to an initial value, a Brownian Motion(BM) Drift component, a BM diffusion constant, and a "Lévy measure" respectively. The form of the characteristic function shows that a Lévy process is the sum of two processes, BM with drift and a compensated Poisson process,

$$\begin{aligned}\mathrm{E}[\exp\{iwX_t\}] &= \exp\Big\{\big[iw(x_0 + tm)t\frac{\sigma^2 w^2}{2}\big] + \big[t\int_{\mathbb{R}}\big(e^{iw\eta} - 1 - iw\eta\mathbf{1}_{|\eta|<1}\big)\nu(d\eta)\big]\Big\} \\ &= \exp\Big\{\big[\text{BM with Drift}\big] + \big[\text{Compensated Poisson Process}\big]\Big\}.\end{aligned}$$

However, when the BM components m and σ^2 equal zero and the Lévy measure satisfies a stricter condition

$$\int_{\mathbb{R}}(1 \wedge |\eta|)\nu(d\eta) < \infty \tag{1.1}$$

the Lévy process is a pure jump, increasing process.

The expected number of jumps, $\mathrm{E}[J]$, in any finite time interval over which the process occurs $t \in [T_0, T_1]$, depends on the Lévy measure $\nu(d\eta)$. Since the Lévy measure specifies the rate for jumps of size η, $\mathrm{E}[J]$ equals

$$\mathrm{E}[J] = \iint_{[T_0,T_1]\times\mathbb{R}^+} \nu(d\eta)dt.$$

However, $\nu(d\eta)$ can be an infinite measure; the integrability constraint in Equation 1.1 only relates to the sum of jump heights and allows infinitely many jumps to occur around zero. Thus, $\mathrm{E}[J]$ need not be finite, unless a constraint is placed on the jump heights η. Constraining the jump heights to be greater than ϵ, as implemented by Wolpert *et al.* [2003] for a specific application, results in an approximate Lévy measure $\nu_+(d\eta)$ and a finite estimate for $\mathrm{E}[J]$

$$\mathrm{E}[J] = \int_{T_0}^{T_1} \int_{\epsilon}^{\infty} \nu(d\eta)dt < \infty.$$

A finite number of jumps, in the case of expression proteomics, is important for an interpretable functional estimate of f.

In expression proteomics, we assume that f is constructed similarly to what is displayed in Figure 1.2, except we represent, for now, both the kernel location and scale parameters as θ_j

$$f(t) = \sum_{j=1}^{J} k(t, \theta_j)\eta_j.$$

Because θ is two-dimensional, f is constructed from a realized Lévy random field, rather than process, and for reasons that will evolve from this thesis, we assume a Gamma random field (GaF) prior

$$\Gamma_t \sim \mathrm{GaF}(\alpha, \lambda)$$

with shape and rate parameters α and λ. The characteristic function of Γ_t given a finite measure $\pi(d\theta)$ is

$$\mathrm{E}[\exp\{iw\Gamma_t\}] \;=\; \exp\Big\{ \iint_{\mathbb{R}\times\Theta} e^{iw\eta} - 1\big]\underline{\alpha\eta^{-1}e^{-\eta\lambda}d\eta\pi(d\theta)}\Big\} \tag{1.2}$$

where the Lévy measure is underlined in equation (1.2) and equals

$$\nu(d\eta, d\theta) = \mathbf{1}_{\eta>0}\alpha\eta^{-1}e^{-\lambda\eta}d\eta\pi(d\theta). \tag{1.3}$$

From the Lévy measure we see that the Gamma random field is a generalized Poisson process and considering connections made by Wolpert and Ickstadt [1998a, 2004], $\Gamma(d\theta)$ can be written as the sum of jumps or points masses with random heights and locations (in a plane) drawn from a Poisson point process with an intensity measure $\nu(d\eta, d\theta)$

$$\Gamma(d\theta) = \sum_j \eta_j \delta_{\theta_j}(dt)$$

Further, since Equation 1.3 is an infinite measure, the approximation,

$$\nu_+(d\eta, d\theta) = \mathbf{1}_{\eta>\epsilon}\alpha\eta^{-1}e^{-\lambda\eta}d\eta\pi(d\theta)$$

results in representing $\Gamma(d\theta)$ as a finite sum,

$$\Gamma(d\theta) = \sum_j^J \eta_j \delta_{\theta_j}(dt), \tag{1.4}$$

where the number of jumps is also random. The expected value of J is easily derived provided a finite measure on θ.

From the summation in Equation 1.4, we see the connection between the adaptive kernel representation of f and Gamma random field. The weights of the

kernels are determined by the random measures assigned by the Gamma process prior,

$$f(t) = \int_{\Theta} k(t, \theta_j)\Gamma(d\theta_j) = \sum_{j=1}^{J} k(t, \theta_j)\eta_j.$$

By eliciting a Gamma process prior on Γ_t, or equivalently $f(t)$, a joint prior is specified for all of the peak identifying parameters, the number of peaks, locations, widths, and heights. We demonstrate the benefits of the joint prior in Chapters 3 and 4.

1.4 Outline of Dissertation

Chapters 2 and 3 of this thesis are devoted to developing single-spectrum feature extraction models that include measurement error and protein related parameters. The protein parameters define a latent protein signal which, in both chapters, is estimated nonparametrically. In Chapter 2 however, a basis expansion method is proposed, whereas in Chapter 3, the protein signal is assumed to be a random variable with a Lévy random field prior distribution.

Taking advantage of the benefits gained from the random variable approach, Chapter 4 completes the research for my thesis and expands the single-spectrum to a multi-spectra feature extraction model. The multi-spectra model draws information from within and between experimental treatment groups to estimate the locations of significant, possibly classifying, features. In one hierarchical model, features from every collected spectrum in an experiment are selected, scaled, aligned, and compared in order to isolate protein biomarkers.

The conclusion of my dissertation, Chapter 5, will include a brief summary of

what I accomplished as well ideas for future work. The ideas pertain to model extensions, computational improvements, and other applications for this research.

Chapter 2

Rapid Peak Identification in Matrix Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry

2.1 Introduction

Matrix Assisted and Surface-Enhanced Laser Desorption/Ionization (MALDI and SELDI) Time-of-Flight (TOF) mass spectrometers (MS) are used to estimate the protein content of biological samples. Proteomic profiles generated by these instruments may contain thousands of paired measurements which reflect the mass and quantity of all molecules present. However, many of the measurements reflect protein fragments, matrix molecules, or detector anomalies, rather than intact proteins [Campa *et al.*, 2003a]. Thus, data reduction procedures should be used to extract information of interest from spectra, i.e. the mass and quantity of proteins. Here, we present an extraction algorithm that incorporates mass spectrometry theory and utilizes an expectation maximization (EM) algorithm to estimate true protein signal.

In a spectrum, data from actual proteins or protein isotopes appear as peaks. Traditionally, spectrometrists found the peaks by hand, but the task was tedious, prone to error, and time consuming. Now, peak identification procedures are automated [e.g. Morris *et al.*, 2005; Tibshirani *et al.*, 2004; Yasui *et al.*, 2003] and tend to first, pre-process the data in order to differentiate protein signal from noise, and second, locate peaks in the protein signal. For example, before attempting to find peaks, Morris *et al.* [2005] suggests a five step procedure to de-noise, background subtract, and normalize intensities; and, Tibshirani *et al.* [2004], motivated by Yasui *et al.* [2003] and Li [2005], proposes smoothing spectra to assess/remove background, reduce random error, and/or merge protein isotopes. Although these and many other automated procedures consider known MS characteristics to isolate signal, two-step or deterministic data-cleaning procedures may alter inappropriately proteomic profiles and inhibit, rather than foster, peak detection.

Data-cleaning procedures may change peak characteristics including height, area, width, location, relation to noise, and/or shape, and make the differentiation of protein peaks from noise difficult. Thus, to avoid pre-processing and still account for measurement error within our feature extraction procedure, we develop a model-based procedure for identifying and quantifying proteins. In this setting, we estimate both MS variation in terms of background, scaling, and precision and peak characteristics, including concentration, location, and resolution. The protein parameters works within a kernel regression approach for estimating protein signal; the basis functions are weighted by the concentrations, scale to the spectrum resolution, and center initially at every data point. However, because all data points do not correspond to real proteins, we regularize the model within

a Bayesian framework. The regularizing prior is a truncated normal scale mixture prior distribution that induces L_1 shrinkage on peak concentration estimates and constrains them to be greater than a user defined minimum concentration; peaks are deleted from the model when the estimated areas are less than or equal to a minimum concentration. Ultimately, the remaining, significant model terms represent key features in a spectrum. Posterior point estimates for model parameters are obtained via Expectation - Maximization (EM) [Dempster *et al.*, 1977].

The remainder of this paper describes in detail the model, the parameter estimation process, and the implementation of our approach using simulated and real data. The paper concludes with a discussion of future work.

2.2 Statistical Model

Using MALDI-TOF characteristics, we propose a model that accounts for spectrum background, resolution, and precision in order to estimate protein peaks according to their location, width, and concentration. We assume initially that one peak is located at every observation, and rely on regularization methods to remove those that are least significant. We impose hard thresholding on shrunken parameter estimates to remove insignificant peaks. Peaks remaining in the final model are used to provide a list of the extracted features.

2.2.1 Expected Intensity

We begin by defining the response variable as Y_t, the observed intensity at time t, and map peaks according to TOF rather than m/z. Since the conversion from time to mass is one to one, we may develop a model in either domain. We choose time, the raw scale on which a TOF detector measures, to avoid error possibly induced

from mis-calibration and to model the original shape of peaks before the data are transformed by a quadratic or higher order polynomial calibration function.

We model Y_t as a Gaussian random variable

$$Y_t = \mu(t) + e_t, \quad \text{where } e_t \sim \text{No}(0, \varphi^{-1}) \tag{2.1}$$

with mean $\mu(t)$ and precision φ. We take $\mu(t)$ as the sum of a time dependent background level $\beta_0(t)$ and a scaled weighted average of J peak concentrations,

$$\mu(t) = \beta_0(t) + \sum_{j=1}^{J} k(t, \tau_j, \omega_j)\eta_j \tag{2.2}$$

where η_j, τ_j, and ω_j represent respectively the concentration, location and width of peak j, $j \in \{1, 2, ..., J\}$, and $k()$ is a basis function with a shape that mimics an isotopic peak. One major influence of peak shape is the distribution of velocities at which molecules leave the ion source [Dass 2001, p.g. 75; Coombes *et al.* 2005a]; molecules of equal mass that have more kinetic energy will arrive at the detector faster than those which have less. The distribution of arrival times forms the spectrum peaks. We opt to model the distributions as either Gaussian or Cauchy [Kempka *et al.*, 2004]

$$k(t, \tau_j, \omega_j) = \sqrt{\frac{\omega_j}{2\pi}} \exp\{-\omega_j |t - \tau_j|^2\} \tag{2.3}$$

$$k(t, \tau_j, \omega_j) = \frac{\sqrt{\omega_j}}{\pi(1 + \omega_j |t - \tau_j|^2)}. \tag{2.4}$$

where the kernel function equals the respective density function. Although the number, location, and scale kernel parameters are ultimately determined by the data, we initially assume that $\boldsymbol{\tau} = [T_0, t_2, t_3, ..., T_1]$ where one peak is centered at

every observed time t. Given $\boldsymbol{\tau}$, we derive the peak widths from the spectrum resolution.

Resolution ϱ is a unitless measure that reflects a spectrum's quality and relates to the minimum distance needed between two peaks to distinguish them. The resolution may be calculated in either the mass or time domain [Dass, 2001, p.g. 74], and accords to the Full Width Half Mass (FWHM) definition (Figure 2.1): The resolution ϱ, of a given peak j, equals the peak location τ_j divided by $\Delta\tau_j$ the width of peak j at half the height [e.g. Dass, 2001, p.g. 120],

$$\varrho = \frac{\tau_j}{\Delta\tau_j}. \tag{2.5}$$

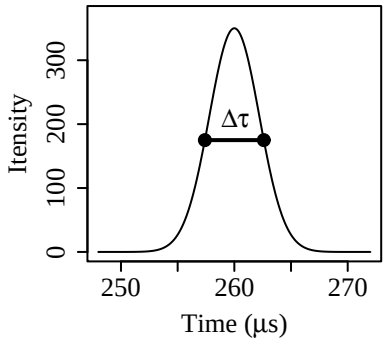


Figure 2.1: A Gaussian peak centered at τ clock-ticks with temporal resolution, $\varrho = 50$. Peak resolution equals peak location, τ, divided by $\Delta\tau$, where $\Delta\tau$ is the peak width at 50% intensity.

The FWHM definition, within the context of our kernel regression model, implies that 50% of the peak height at τ_j equals the height at $\tau_j + \Delta\tau_j/2$,

$$\frac{1}{2}\% \times \eta_j k(\tau_j, \tau_j, \omega_j) = \eta_j k(\tau_j + \frac{\Delta\tau_j}{2}, \tau_j, \omega_j)$$

so that after substituting $\Delta\tau_j = \frac{\tau_j}{\varrho}$ from Equation 2.5, ω_j solves as

$$\omega_j = \log(4)\Big(\frac{2\varrho}{\tau_j}\Big)^2 \qquad \text{and} \qquad \omega_j = \Big(\frac{2\varrho}{\tau_j}\Big)^2$$

for the Gaussian kernel and the Cauchy kernel respectively.

Expected resolutions for MALDI-TOF data may be listed in the user manuals of any mass spectrometer, [e.g. PerSeptive Biosystems, 1999], and ranges from 50-2000. The resolution reflects both the quality of the spectrum and the shape of the peaks; sharp, narrow peaks occur in high resolution, good quality spectra and short, broad peaks occur in low resolution, arguably poor quality spectra. In our model, information regarding resolution propagates into information governing peak location and width which influences peak identification. For example, given a low resolution spectrum, our model will dismiss narrow peaks as noise because they are not consistent with the overall quality of the spectrum. Methods that ignore resolution, such as local maxima methods, would likely declare the same, narrow peaks as proteins.

Also indicative of a spectrum 's quality, is the amount of background noise that may inflate molecular intensities. Typically, background in MALDI-TOF data causes an extreme intensity increase that decays in time to a non-zero constant. The time-dependent decay is a result of several small, matrix fragments leaving the ion source and inflating the intensities of comparable size proteins; the non-zero constant occurs primarily because of detector ringing. We model each background separately as either a time independent constant S or a time dependent function with parameters τ_0, ω_0, and η_0,

$$\beta_0(t) = S + \eta_0 k_0(t, \tau_0, \omega_0). \tag{2.6}$$

We estimate the background parameters by fitting a *first-peak* at $\tau_0 = T_0$ and using the same kernel framework in Equation 3.6 where η_0 represents the concentration of matrix fragments and ω_0 solves from ϱ_0, the first-peak resolution. The kernel function $k_0()$ may either match the chosen peak kernel $k()$ in Equation 2.3 or equal an exponential density function,

$$k_0(t, \tau_0, \omega_0) = \omega_0 \exp\{-\omega_0 |t - \tau_0|\}. \tag{2.7}$$

Figure (2.2) displays three Cauchy first-peaks with varying resolution values.

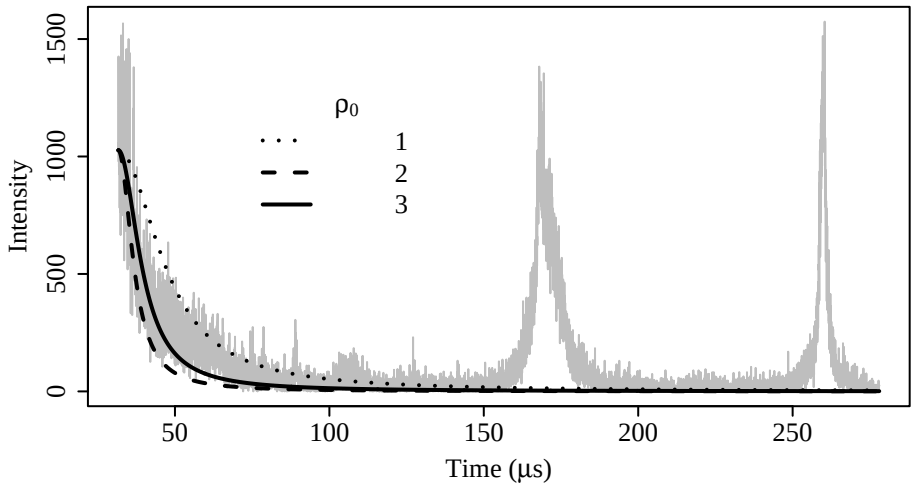


Figure 2.2: Examples of first-peak fits using three different, Cauchy first-peak resolutions.

Given $k_0()$ and substituting Equation 2.6 into Equation 3.6 with an indicator to state the basis function (i.e. if $\tau_j > T_0$ then use $k()$ or if $\tau_j = T_0$ then use $k_0()$), the expected intensity at time t becomes

$$\mu(t) = S + \sum_{j=0}^{J+1} \eta_j k_{\mathbf{1}_{\tau_j > T_0}}(t, \tau_j, \omega_j), \tag{2.8}$$

and for N observations the likelihood is

$$\mathcal{L}(\boldsymbol{Y}|S, \varphi, \boldsymbol{K}) = \mathrm{No}(\boldsymbol{K}\boldsymbol{\eta} + \mathbf{1}S, \varphi^{-1}\boldsymbol{I_N}), \tag{2.9}$$

where $\boldsymbol{Y} = \{Y(t_1), Y(t_2), ..., Y(t_N)\}$, $\boldsymbol{K}$ represents an $N \times (J+1)$ matrix with elements $[\boldsymbol{K}]_{tj} = k_{\mathbf{1}_{\tau_j > T_0}}(t, \tau_j, \omega_j)$, $\boldsymbol{\eta}' = \{\eta_1, \eta_2, ..., \eta_J\}'$ and I_N is an $N \times N$ identity matrix. Initially, $J + 1 = N$ because we assume that one peak occurs at every observation. In the next section, we describe how to regularize the model parameters to reduce the model dimension and estimate a sparse representation of $\mu(t)$.

2.2.2 Regularization

Regularization methods have been developed in both the Classical and Bayesian paradigms [e.g. Abramovich *et al.*, 1998; Vidakovic, 1998; Green, 1987] that penalize high dimensional models and result in sparse data representations. Classical approaches maximize a penalized likelihood or minimize a constrained residual sums of squares, such as the least absolute shrinkage and selection operator (LASSO)[Tibshirani, 1996]; Bayesian approaches require the specification of prior distributions. Here, we choose the Bayesian paradigm and specify a prior distribution on peak concentration $\boldsymbol{\eta}$ that is comparable to imposing an L_1 penalty. The penalty shrinks the posterior estimates of peak concentrations estimate $\hat{\boldsymbol{\eta}}$ toward ϵ, a value that interprets as the minimum concentration level needed for peak detection.

When a peak concentration reaches ϵ, $\hat{\eta_j}$ is set to zero, peak j drops effectively from the model, and J decreases. For shrinkage and selections, priors based on scale mixtures of Normal distributions have been effective [Johnstone and Silverman, 1997; Vidakovic, 1998], but the support of the Normal density does not

match the range of non-negative, protein concentrations. Thus, we model η_j as a scale mixture of truncated Normal distributions where the convolution of

$$\pi(\eta_j|\varphi,\epsilon,\sigma_j) \;=\; \begin{cases} \mathrm{No}(\epsilon,\sigma_j/\varphi) & \eta_j > \epsilon \\ 0 & \eta_j \leq \epsilon \end{cases} \tag{2.10}$$

with σ_j distributed as an Exponential distribution

$$\pi(\sigma_j|\mu) \;=\; \mathrm{Ex}(0,\mu) \tag{2.11}$$

($\mathrm{Ex}(a,b) \sim b\exp\{-b(x-a)\};\ x \in (a,\infty)$) results in the truncated Double Exponential marginal distribution ($\mathrm{DE}(a,b) \sim \frac{1}{b}\exp\{-\frac{(x-a)}{b}\}$) for η_j,

$$\eta_j|\varphi,\epsilon,\mu, \sim \mathbf{1}_{\{\eta>\epsilon\}}\mathrm{DE}\Big(\epsilon, \frac{1}{\sqrt{2\mu\varphi}}\Big) \tag{2.12}$$

[Andrews and Mallows, 1974; West, 1987], or equivalently, a shifted Exponential distribution with location and rate parameters ϵ and $\sqrt{2\mu\varphi}$. Depending upon μ, the shrinkage parameter, the Exponential prior places a large amount of prior mass near ϵ and the right tail. This distribution of weight, unlike the Normal density, naturally shrinks insignificant peaks toward ϵ and reinforces the estimation of large, significant peaks.

The specification of an Exponential prior on η_j however, makes estimating the posterior mean and variance difficult; they each depend upon the unstable matrix inversion of $\boldsymbol{K}'\boldsymbol{K}$. The full conditional posterior distribution for η_j with prior Equation 2.12 is,

$$\begin{aligned}
&\pi(\boldsymbol{\eta}|\varphi,\epsilon,\mu,\boldsymbol{Y}) \\
&\quad\propto\ \mathbf{1}_{\{\boldsymbol{\eta}>\boldsymbol{\epsilon}\}}\exp\Big\{\frac{-\varphi}{2}||\boldsymbol{Y}-S\mathbf{1}_N-\boldsymbol{K}\boldsymbol{\eta}||^2-(2\mu\varphi)^{\frac{1}{2}}\boldsymbol{\eta}'\mathbf{1}_J\Big\} \\
&\quad\propto\ \mathbf{1}_{\{\boldsymbol{\eta}>\boldsymbol{\epsilon}\}}\mathrm{No}\Big((\boldsymbol{K}'\boldsymbol{K})^{-1}\Big[\boldsymbol{K}'(\boldsymbol{Y}-S\mathbf{1}_N)+(2\mu)^{\frac{1}{2}}\varphi^{-\frac{1}{2}}\mathbf{1}_J\Big], \varphi^{-1}(\boldsymbol{K}'\boldsymbol{K})^{-1}\Big),
\end{aligned}$$

where $\mathbf{1}_{\{x>a\}} = 1$ when $x > a$, $\mathbf{1}_x = [1, 1, ..., 1]'_x$, and $\boldsymbol{\epsilon} = \epsilon \mathbf{1}_J$. Yet, with N peaks, $\boldsymbol{K}'\boldsymbol{K}$ may be nearly singular. However, if we do not analytically integrate Equations 2.10 and 2.11, the posterior mean and covariance are stabilized with the addition of $\texttt{diag}(1/\boldsymbol{\sigma}) = \texttt{diag}([1/\sigma_1, 1/\sigma_2, ..., 1/\sigma_J]')$,

$$\pi(\boldsymbol{\eta}|\varphi, \epsilon, \boldsymbol{\sigma}, \boldsymbol{Y}) \propto \mathbf{1}_{\{\boldsymbol{\eta}>\boldsymbol{\epsilon}\}} \mathrm{No}\Big(\big[\boldsymbol{K}'\boldsymbol{K} + \texttt{diag}(\frac{1}{\boldsymbol{\sigma}})\big]^{-1} \big[\boldsymbol{K}'(\boldsymbol{Y} - \mathbf{1}_N S) + \epsilon/\boldsymbol{\sigma}\big], \varphi^{-1} \big[\boldsymbol{K}'\boldsymbol{K} + \texttt{diag}(\frac{1}{\boldsymbol{\sigma}})\big]^{-1} \Big)$$

We set the mean of σ_j to an empirical Bayes estimate for μ [Vidakovic, 1998; Silverman, 1985] and use either standard reference or informative prior distributions for the remaining parameters. Given

$$\begin{aligned} \pi(S) &\propto 1 \\ \pi(\varphi) &\propto \varphi^{-1} \\ \pi(\varrho) &= \mathrm{LN}(\log(R), s_R^2) \\ \pi(\varrho_0) &= \mathrm{LN}(\log(R_0), s_{R0}^2) \end{aligned}$$

where $\{R, s_R^2\}$ and $\{R_0, s_{R0}^2\}$ represent respectively user specified location and scale parameters for a $\mathrm{LogNormal}(\mu, \sigma^2) = 1/(x\sqrt{2\pi\sigma^2}) \exp(-(x-\mu)^2/(2\sigma^2))$, the joint posterior distribution is

$$\begin{aligned} &\pi(\boldsymbol{\eta}, \boldsymbol{\sigma}, \varphi, S, \varrho, \varrho_0|\boldsymbol{Y}) \\ &\quad \propto \mathrm{No}(\boldsymbol{Y}; \boldsymbol{K}\boldsymbol{\eta} + \mathbf{1}_N S, \mathrm{I}_N \varphi^{-1}) \\ &\quad \times \mathbf{1}_{\{\boldsymbol{\eta}>\boldsymbol{\epsilon}\}} \mathrm{No}(\boldsymbol{\eta}; \mathbf{1}_J \epsilon, \texttt{diag}(\boldsymbol{\sigma}/\varphi)) \\ &\quad \times \prod_{j=0}^{M} \mathrm{Ex}(\sigma_j; \mu) \times \frac{1}{\varphi} \times \mathrm{LN}(\varrho; R, s_R^2) \times \mathrm{LN}(\varrho_0; R_0, s_{R0}^2) \end{aligned}$$

$(\mathrm{No}(x; a, b) = (2\pi b)^{.5} \exp(-(x-a)^2/(2b))$, $\mathrm{LN}(x; a, b) = (1/x)(2\pi b)^{.5} \exp(-(x-a)^2/(2b))$, and $\mathrm{Ex}(x; a) = a \exp(-xa)$). A variety of methods exist to integrate stochastically the joint posterior distribution and obtain the posterior marginal distributions of the model parameters, e.g. MCMC or importance sampling. Such approaches are advantageous for summarizing the posterior parameter distribution, but may take time to implement. Thus, we utilize the Expectation-Maximization (EM) algorithm to obtain rapidly the posterior mode estimates for peak locations, peak concentrations, and peak widths.

2.2.3 Posterior Mode Estimates

Iterating between two steps, the Expectation-step (E-step) and Maximization step (M-step), the EM algorithm finds posterior mode estimates of $\boldsymbol{\theta} = \{\boldsymbol{\eta}, S, \varrho, \varrho_0\}$ by averaging over the latent parameters, φ and $\boldsymbol{\sigma}_j$. We program the algorithm in `R` for which we supply pseudo-code at the end of this section in Figure 2.3.

During the E-step the expected value of the log complete posterior distribution of $\boldsymbol{\theta}$ and $\{\boldsymbol{\sigma}, \varphi\}$ is calculated with respect to the full posterior conditional distribution of $\{\boldsymbol{\sigma}, \varphi\}$ given current (iteration i) values of $\boldsymbol{\theta}$,

$$\begin{aligned} &\mathrm{E}[\log(f(\boldsymbol{\theta}, \boldsymbol{\sigma}, \varphi|\boldsymbol{Y}))|\boldsymbol{\theta}^{(i)}, \boldsymbol{Y}] \\ &= \iint_{\mathbb{R}^+ \times \mathbb{R}^+} \log(\pi(\boldsymbol{\theta}, \boldsymbol{\sigma}, \varphi|\boldsymbol{Y}))\pi(\boldsymbol{\sigma}, \varphi|\boldsymbol{\theta}^{(i)}, \boldsymbol{Y}) d\boldsymbol{\sigma} d\varphi. \end{aligned} \tag{2.13}$$

Ignoring terms independent of $\boldsymbol{\theta}$, Equation 2.13 equals

$$
\begin{aligned}
= \quad & -\frac{1}{2}\mathrm{E}\big[\varphi|\boldsymbol{\theta}^{(i)},\boldsymbol{Y}\big]||\boldsymbol{Y}-\boldsymbol{K}\boldsymbol{\eta}-\mathbf{1}_N S||^2 \qquad (2.14)\\
& -\frac{1}{2}(\boldsymbol{\eta}-\mathbf{1}_J\epsilon)'\mathtt{diag}\Big(\mathrm{E}\Big[\frac{\varphi}{\sigma_j}|\boldsymbol{\theta}^{(i)},\boldsymbol{Y}\Big]\Big)(\boldsymbol{\eta}-\mathbf{1}_N\epsilon)\\
& -\log(\varrho)-\frac{1}{2s_R^2}(\log(\varrho)-\log(R_0)-\log(R))^2-\frac{1}{2s_{R0}^2}(\log(\varrho_0)-\log(R_0))^2,
\end{aligned}
$$

where

$\pi(\boldsymbol{\sigma},\varphi|\boldsymbol{\theta}^{(i)},\boldsymbol{Y})$

$$
= \frac{\mathrm{No}(\boldsymbol{Y};\boldsymbol{K}\boldsymbol{\eta}+\mathbf{1}_N S,\mathrm{I}_N\varphi^{-1})\times\frac{1}{\varphi}\times\mathrm{No}(\boldsymbol{\eta};\mathbf{1}_J\epsilon,\mathtt{diag}(\boldsymbol{\sigma}/\varphi))\times\prod_{j=0}^{J}\mathrm{Ex}(\sigma_j;\mu)}{\int_\varphi \mathrm{No}(\boldsymbol{Y};\boldsymbol{K}\boldsymbol{\eta}+\mathbf{1}_N S,\mathrm{I}_N\varphi^{-1})\times\frac{1}{\varphi}\times\prod_{j=0}^{J}\mathrm{Ex}(\eta_j;\sqrt{2\varphi\mu})d\varphi},
$$

$$
\mathrm{E}\big[\frac{\varphi}{\sigma_j}|\boldsymbol{\theta}^{(i)},\boldsymbol{Y}\big]=\iint_{\mathbb{R}^+\times\mathbb{R}^+}\frac{\varphi}{\sigma_j}\pi(\boldsymbol{\sigma},\varphi|\boldsymbol{\theta}^{(i)},Y)d\boldsymbol{\sigma}d\varphi, \qquad (2.15)
$$

$$
\mathrm{E}\ \big[\varphi|\boldsymbol{\theta}^{(i)},\boldsymbol{Y}\big]=\iint_{\mathbb{R}^+\times\mathbb{R}^+}\varphi\pi(\boldsymbol{\sigma},\varphi|\boldsymbol{\theta}^{(i)},Y)d\boldsymbol{\sigma}d\varphi. \qquad (2.16)
$$

Both Equations 2.15 and 2.16 integrate to a scaled sum of two infinite series or confluent hypergeometric functions (HG1F1). When we approximated the series we found obvious inaccuracies for estimating the expected values of φ and φ/σ_j, thus we analytically integrate over $\boldsymbol{\sigma}$ and use Laplace approximation for the integration over φ.

The purpose of the M-step is to estimate values for $\boldsymbol{\theta}$ that will maximize, or at least increase, the expected log joint posterior distribution in Equation 2.14. We do so by implementing a Gauss-Seidel(GS), one-dimensional line search [Jiang,

2000] that maximizes each element of $\boldsymbol{\theta}$ individually while holding all other, current parameter estimates constant. Since convergence of the iteration scheme is not necessary for setting $\boldsymbol{\theta}$ to values that will increase Equation 2.14, the GS completes only three iterations before returning to the E-step.

In the following order, the GS algorithm estimates $\boldsymbol{\theta}$.

1. $\boldsymbol{\eta}$: Maximizing Equation 2.14 over $\boldsymbol{\eta}$ subject to $\eta_j > \epsilon$ for $j \in [1, ..., J]$ is equivalent to the quadratic programming problem,

$$\hat{\boldsymbol{\eta}} = \overset{argmin}{\boldsymbol{\eta}} \Big\{ \mathbf{1}_{\{\boldsymbol{\eta} > \boldsymbol{\epsilon}\}} \Big[\Big(\frac{1}{2}\, \boldsymbol{\eta}' \Big(\mathrm{E}[\varphi|\boldsymbol{\theta}^{(i)}, \boldsymbol{Y}] \boldsymbol{K}'\boldsymbol{K} + \mathtt{diag}(\mathrm{E}[\varphi/\boldsymbol{\sigma}|\boldsymbol{\theta}^{(i)}, \boldsymbol{Y}]) \Big) \boldsymbol{\eta} - \mathrm{E}[\varphi|\boldsymbol{\theta}^{(i)}, \boldsymbol{Y}](\boldsymbol{Y} - \mathbf{1}_N S)'\boldsymbol{K} + \mathtt{diag}(\mathrm{E}[\varphi/\boldsymbol{\sigma}|\boldsymbol{\theta}^{(i)}, \boldsymbol{Y}])(\mathbf{1}'_J \epsilon) \Big) \boldsymbol{\eta} \Big] \Big\}$$

   for which we use the function `solve.QP` within the `R` package `quadprog` that implements a dual method for constrained optimization problems introduced by Goldfarb and Idnani [1983].

2. S: A simple solution exists for the maximum estimate of S,

$$\hat{S} = \frac{(\boldsymbol{Y} - K\hat{\boldsymbol{\eta}})'\mathbf{1}_N}{N}.$$

3. $\{\varrho, \varrho_0\}$: The function `optim` in the standard `R` package implements a quasi-Newton update step to maximize Equation 2.14 over ϱ and ϱ_0. We found that updating the resolution parameters every GS iteration was time consuming and inefficient for finding peaks. Thus, we only update ϱ and ϱ_0 during the second, of the three, GS iteration.

At the completion of the GS, the parameter μ must be updated since it depends on J and $\boldsymbol{\sigma}$. Using the posterior expectation for $\boldsymbol{\sigma}$, estimated similarly to

Equations 2.15 and 2.16, we set $\hat{\mu}$ to

$$\hat{\mu} = \frac{J+1}{\sum_{j=0}^{J} \mathrm{E}[\sigma_j | \boldsymbol{\theta}^{(i)}, \boldsymbol{Y}]}.$$

Convergence of the EM algorithm is assessed by calculating the percent change between the log posterior distribution at the $I-1$ iteration and the Ith iteration. If the percent change is less than 1.0% then the EM breaks; if not, the EM algorithm continues for at most A additional iterations or until the percent change criterion is achieved.

When the EM algorithm ends, the number of significant peaks remaining in the model depends heavily upon ϵ, the chosen minimum peak concentration constraint. If ϵ is too small then several false positive peaks may occur, and if ϵ is too large then several false negatives may occur. Further, if the thresholding is invoked before significant concentrations have converged to be greater than ϵ, then again, false negatives may occur. For this reason, we do not invoke hard thresholding for $\epsilon > 0$ until half way through the I iterations. After completing at least one half of the EM iterations, the constraint is gradually increased from zero to a pre-specified ϵ (e.g. constraint $= 0.25\epsilon$ when $i \geq 0.5I$, 0.5ϵ when $i \geq 0.625I$, 0.75ϵ when $i \geq 0.75I$, and ϵ when $i \geq 0.875I$). The slow introduction of ϵ reduces the effects of hard-thresholding before parameters estimates have converged.

The algorithm's sensitivity to ϵ is not uncommon for feature extraction approaches as several tend to impose thresholds based on signal to noise ratio estimates. We however, avoid prior estimates of signal and random error and set ϵ so that ϵ divided by the observed total intensity $\sum Y_t$ equals a user defined percentage. The percentage may be adjusted to the users needs or to the sampling procedure. For example, MALDI-TOF spectra generated from fractionated

samples may only display approximately 50 distinguishable peaks [Campa *et al.*, 2003a]. Thus, ϵ should approximately equal $\sum Y_t/50$ or 2% of $\sum Y_t$. We explore the effects of varying specifications for ϵ in Section 2.3.

2.2.4 EM Implementation

Optimizing $\boldsymbol{\eta}$ when $K'K$ is initially an $N \times N$ matrix may be too large for `R` to manage. Thus, to reduce the matrix dimension, we segment the spectrum into large ($N \approx 2000$) disjoint subsets. For each subset, we apply the EM algorithm and combine the results to create informative starting locations for an analysis of the entire spectrum.

Breaking the data into subsets enables a thorough, computationally feasible, preliminary search for peaks. Given local, segmental estimates for precision and resolution, peak concentration may increase, decrease, or delete. However, to minimize the possibility of deleting true, protein peaks (false negatives) before modeling the entire spectrum, we transform all measurements by subtracting the spectrum minimum intensity and set $S = 0$ while the data are segmented; the estimation of S during the entire spectrum analysis remains as previously described. Additionally, the minimum concentration level imposed for the entire spectrum need not match that of the segments; let ϵ_s and ϵ_e represent the segment and entire spectrum constraints respectively.

We implement our two-step, EM approach to real and simulated datasets. The simulation study demonstrates our ability to find peaks with varying signal- to-noise (S/N) ratios, and compares the results to a peak finding algorithm currently available in `R`. From the simulation study, we gain user input information for modeling the real data.

```
Initialize θ, E[φ], E[φ/σ_j], E[σ_j]
i = 0
a = 0
conv = 0
while (a<20 and conv==0){
//M-step
  set ϵ
  η^(1) = θ^(i)[1]
  S^(1) = θ^(i)[2]
  ϱ^(1) = θ^(i)[3]
  ϱ_0^(1) = θ^(i)[4]
  for k in 1:3 {
       η^(k+1) = max(η|S^(k), φ^(k), ϱ_0^(k), ϱ^(k), E[φ/σ_j]^(i), E[φ]^(i))
       for j in 1:M { if (η_j^(k+1) == ϵ) delete η_j^(k+1), τ_j, ω_j}
       S^(k+1) = max(S|η^(k+1), ϱ^(k), ϱ_0^(k), E[φ/σ_j]^(i), E[φ]^(i))
       if (k==2){
          ϱ^(k+1) = max(ϱ|η^(k+1), S^(k+1), ϱ_0^(k), E[φ/σ_j]^(i), E[φ]^(i))
            ϱ_0^(k+1) = max(ϱ_0|η^(k+1), S^(k+1), ϱ^(k+1), E[φ/σ_j]^(i), E[φ]^(i))
       }
   }
   θ^(i+1) = {η^(k+1), S^(k+1), ϱ^(k+1), ϱ_0^(k+1)}

//Update Empirical Bayes Hyper-parameter
   estimate μ

//E-step
   estimate E[φ/σ_j|θ^(i+1), Y]
   estimate E[φ|θ^(i+1), Y]
   estimate E[σ_j|θ^(i+1), Y]

//Check Convergence
   if (i == I-1) logPost1 =π(θ^(i)|Y)
   if (i >= I) {
      logPost2 = π(θ|Y)
      if ((logPost2−logPost1)/logPost1<.01) conv=1
      else a+ =1
      }
   }
```

Figure 2.3: R pseudo-code for EM algorithm.

2.3 Application

We applied our feature extraction method to real data (Figure 2.8) and to simulated data (Figure 2.5). The simulated data were generated in an attempt to mimic the appearance of the real spectrum, so both datasets span approximately between 5k Da and 75k Da, or 30 microseconds (μs) to 280 μs, and were thinned by every 7th observation.

2.3.1 Simulation Study

We created varying S/N simulation scenarios by simulating a total of 100 spectra from one of four Normal error models that have the same mean or latent signal (refer to Figure 2.4), but different precisions, $\varphi = \{0.04, 0.0044, 0.0011, 0.00028\}$. The latent signal was generated from the Cauchy convolution of $J = 35$ peaks, at locations $\boldsymbol{\tau}_{true}$ (listed in Table 2.3.1), with $\{\varrho = 56, \varrho_0 = 2, S = 50, \mu = 3.68e-9, \varphi = 0.04, \epsilon = 20000\}$, and the vectors, $\{\boldsymbol{\sigma}, \boldsymbol{\eta}\}$ were drawn from the respective prior distributions. Across 25 simulated spectra, the S/N ratio for each scenario averaged respectively in the order of descending precision, 74.56, 24.85, 12.43, and 6.21. Figure 2.5 plots one spectrum from each scenario.

With the exception of ϵ_s and ϵ_e, we fit the model for each simulated dataset using all of the same user inputs: segments = [5000, 11000, 16000, 24000, 40000, 55000, 75000], $I = 10$, $A = 20$, $k()$ = Cauchy, $R = 50$, $s_R^2 = 0.5^2$, $R_0 = 3$, and $s_{R0}^2 = 0.5^2$. The scale of the resolution parameters provide an approximate geometric standard deviation of 0.5 which makes for fairly diffuse (relative to the mean), but informative distributions where 95% of the prior mass for ϱ and ϱ_0 are between respectively $[7.00, 354.95]$ and $[0.42, 21.30]$ respectively. As for ϵ_s

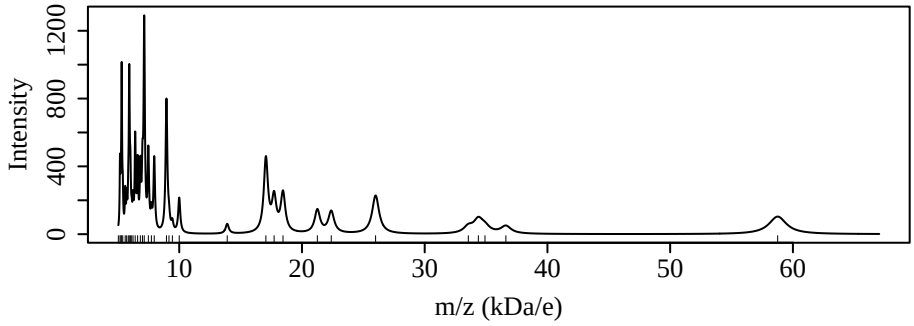


Figure 2.4: Simulated true signal from which datasets are generated.

and ϵ_e, we originally set each to $0.02 \sum Y_t$, but found the value too high when assessing the entire spectrum. Thus, we adjusted ϵ_e accordingly and set $\{\epsilon_s, \epsilon_e\}$ $= \{0.02 \sum Y_t, 0.0085 \sum Y_t\}$ for scenarios $\varphi \in \{0.04, 0.0044, 0.0011\}$ and $\{\epsilon_s, \epsilon_e\} =$ $\{0.02 \sum Y_t, 0.015 \sum Y_t\}$ when $\varphi = 0.00028$.

A summary of the EM results, by scenario, is provided in Table 2.2. The time-dependent background resolution, constant background, and precision parameters are within one standard deviation of their true value and each resolution parameter is within 3 standard deviations, but the number of peaks $\hat{J}$, with the exception of $\varphi = 0.00028$, is consistently greater than 35. We are not concerned by the over-estimation for two reasons. The first is that we place greater risk on missing peaks than flagging false positive peaks, and second, the over-estimation is minimal in comparison to local maxima approaches.

We compare our method to a peak finding algorithm of Li [2005] and available in the statistical, open source software `R` as a package called `PROcess`. The `PROcess` function `is.Peak` implements a multi-step method on background-subtracted data for finding significant features by smoothing the data twice according to the

Table 2.1: Peak locations, heights, and widths to simulate data.

j	τ	η	ω
1	8288	87283.15	0.0001826
2	8442	26980.03	0.0001760
3	8575	221014.16	0.0001706
4	8757	45359.41	0.0001636
5	9163	54269.42	0.0001494
6	9408	44489.46	0.0001417
7	9667	37901.07	0.0001342
8	9828	243056.88	0.0001299
9	10010	78389.91	0.0001252
10	10220	20132.57	0.0001201
11	10430	44701.77	0.0001153
12	10766	160725.61	0.0001082
13	11137	117713.11	0.0001011
14	11550	112407.88	0.0000940
15	11865	104475.23	0.0000891
16	12124	408963.45	0.0000853
17	12719	161810.83	0.0000775
18	13153	36802.79	0.0000725
19	13552	162076.52	0.0000683
20	15197	332156.57	0.0000543
21	15526	29380.37	0.0000520
22	15960	23881.45	0.0000492
23	16821	97049.85	0.0000443
24	22274	35659.71	0.0000253
25	26124	319530.65	0.0000184
26	26894	148826.58	0.0000173
27	27713	179595.03	0.0000163
28	30730	117703.13	0.0000133
29	31892	114780.29	0.0000123
30	35420	224042.33	0.0000100
31	42070	43754.57	0.0000071
32	42742	91832.14	0.0000069
33	43183	39489.11	0.0000067
34	44548	55174.28	0.0000063
35	60004	174267.72	0.0000035

Scenario	$\hat{J}$	$\hat{\varrho}$	$\hat{\varrho}_0$	$\hat{\varphi}$	$\hat{\mu}$	$\hat{S}$
0.0400	69.00	55.63	2.40	3.28e−02	9.40e−10	48.82
	(4.28)	(0.12)	(0.15)	(8.53e−03)	(1.98e−10)	(1.88)
0.0044	58.20	55.59	1.66	3.84e−03	7.24e−10	50.78
	(1.19)	(0.67)	(0.26)	(3.19e−04)	(1.51e−10)	(5.65)
0.0011	49.20	54.21	2.00	1.05e−03	6.88e−10	48.23
	(7.26)	(0.72)	(0.50)	(4.55e−05)	(1.62e−10)	(13.13)
0.00028	33.20	48.08	1.41	2.60e−04	2.32e−10	32.02
	(13.19)	(4.18)	(0.55)	(7.89e−06)	(7.50e−11)	(20.87)

Table 2.2: Average parameter point estimates with standard deviations in parentheses from 25 simulations per four scenarios: $\varphi = \{0.04, 0.0044, 0.0011, 0.00028\}$. The following lists the true parameters values: $J = 35$, $\varrho = 56$, $\varrho_0 = 2$, $\mu = 3.68e-09$, and $S = 50$.

parameters `span` and `sm.span` and declaring observations as peaks based on the arguments for `sm.span`, `threshold`, `SoN`, and `ratio`. In `PROcess` observed intensities are considered peaks when they are local maxima within neighborhoods of `sm.span` observations; they are greater than `threshold`; they have estimated S/N ratios that are greater than `SoN`; and their areas are greater than `ratio` times the area of the largest peak in the spectrum.

When using all default `isPeak` parameters, the function found on average 39% to 42% of the real peaks and had an extremely high false discovery rate (refer to Figure 2.6. To increase the functions sensitivity and specificity, we changed all of the arguments, except `sm.span`=11 to scenario-dependent values; respectively to scenarios $\{0.04, 0.0044, 0.0011, 0.00028\}$ we set `span`= $\{31, 81, 101, 161\}$, `ratio`= $\{0.05, 0.075, .1, .2\}$ and `SoN`=$\{2.5, 2.5, 3.5, 3.5\}$.

Table 2.3 compares the average results across the 25 simulations from our EM algorithm and the `PROcess` package. The first two rows note the number of peaks estimated ($\hat{J}$) and the proportion of true peaks found or *True Discovery*

Results	Scenario							
	$\varphi = 0.04$ ($\overline{S/N} = 74.56$)		$\varphi = 0.0044$ ($\overline{S/N} = 24.85$)		$\varphi = 0.0011$ ($\overline{S/N} = 12.43$)		$\varphi = 0.00028$ ($\overline{S/N} = 6.21$)	
	EM	PRO	EM	PRO	EM	PRO	EM	PRO
$\hat{J}$	71.04	24.16	56.32	26.56	51.48	19.04	34.76	19.96
TDR	0.94	0.57	0.80	0.43	0.75	0.35	0.54	0.29
FDR	0.04	0.18	0.10	0.43	0.17	0.35	0.35	0.49

Table 2.3: Averaged across simulations per scenario, $\hat{J}$, TDR, and FDR refer to average number of peaks estimated, the proportion of true peaks found (sensitivity), and the number of false positive peaks divided by $\hat{J}$ respectively per scenario when using either the proposed EM approach or PROcess.

Peak S/N	Scenario							
	$\varphi = 0.04$ ($\overline{S/N} = 74.56$)		$\varphi = 0.0044$ ($\overline{S/N} = 24.85$)		$\varphi = 0.0011$ ($\overline{S/N} = 12.43$)		$\varphi = 0.00028$ ($\overline{S/N} = 6.21$)	
	EM	PRO	EM	PRO	EM	PRO	EM	PRO
>20	0.97	0.50	0.96	0.52	1.00	1.00	1.00	1.00
(10,20]	0.71	0.17	0.85	0.29	0.87	0.17	0.97	1.00
(5,10]	–	–	0.54	0.59	0.82	0.20	0.68	0.16
(0,5]	–	–	0.37	0.19	0.39	0.34	0.35	0.18

Table 2.4: By scenario, the column labeled by "J" lists the number of true peaks that fall within the given signal to noise (S/N) ranges declared by the first column. The proportion of the true peaks that were found by either the EM approach or PROcess are provided in the remaining columns.

Rate (TDR). We consider a true peak "found" if an estimated location is within $\pm 0.2\% \times \tau_j$. If more than one estimated peak falls within 0.2% Da of a single true peak, then the true peak is referenced once as found; none of the simulated true peak locations $\pm 0.2\%$ overlap another. A peak is marked as a false positive if it does not fall within 0.2% of any true peak. The *False Discovery Rate* (FDR) equals the number of false positives divided by the estimated number of peaks, $\hat{J}$. For example, the first column labeled "EM" indicates that our approach, on average, found 71 peaks within a spectrum of the first scenario, 94% of the true

peaks were covered by the 71 peaks, and 4% of the 71 peaks did not fall within ±0.2% of a true peak.

We expect the model to find peaks with high S/N ratios. Defining S/N as $(f(t) - \beta_0(t))/\sqrt{\varphi}$, Table 2.3 displays the mean S/N, $\overline{S/N}$, for each scenario. The proposed model finds almost every peak in large S/N situations, and never finds less than 50% of the peaks in any S/N range. `PROcess` is consistently less sensitive. `PROcess` routinely misses small peaks that follow larger peaks and, in large S/N situations, one or more of adjacent peaks that are approximately the same height. For example, `PROcess` routinely misses true peaks occurring at approximately 14 kDa/e and within the range 33.5-36.6 kDa/e in each of the scenarios and regardless of the argument specifications (Figures 2.6 and 2.7). Although decreasing the argument `span` enabled the discovery of some peaks, the decrease also caused a dramatic increase false positive peaks(Figure 2.7).

One reason `PROcess` falsely flags several peaks is that the local maxima procedure does not account for resolution. For example, in plot d. of Figure 2.7 peaks found around 17 kDa/e meet the `PROcess` peak criteria. Given the resolution of these peaks, all of the features found after 40 kDa/e are theoretically impossible according to the definition of resolution in Equation 2.5. In our case, the resolution functions into the width of a peak and thus contributes to its identification. Thus, the EM algorithm tends to perform with a lower FDR than `PROcess`.

To further explore which peaks are found by the different approaches, Table 2.4 groups the simulated peaks according to their S/N ratio. From Table 2.4 we see that our approach is able to find peaks in very low signal to noise situations. Specifically, when $\varphi = 0.00028$ our model found almost twice as many of the 19 peaks as `PROcess`.

2.3.2 Real Data

We extract features from four real datasets (Figure 2.8) that were collected by the Duke Medical Center, Radiology Department in Durham, North Carolina for a case-control, observational study of lung cancer. In the study, serum samples from 30 diseased and healthy, Caucasian males were drawn, fractionated, and summarized by 10 replicate, MALDI-TOF spectra (MALDI-TOF Delayed Extraction Mass Spectrometer, Applied Biosystems Voyager DE). Each spectra is the sum of ten laser shots. For our purposes, Dataset 1 is one replicate chosen from one sample; Dataset 2 is the average of 10 replicates; Dataset 3 is the average for one fraction across all oncology patients; and Dataset 4 is the average for one fraction across all control patients.

We apply the model with the following user specifications to all datasets; kernel = Cauchy or Gaussian, $R = 50$, $R_0 = 2$, and segments = {5000, 10000, 15000, 26000, 40000, 60000, 75000}. When modeling a spectrum with a Cauchy kernel, the shape of the first peak mimics the kernel, but when imposing a Gaussian kernel, $k_0()$ is an exponential kernel, Equation 2.6. We learned from the simulation study that a dataset of comparable range and size may require constraints approximately equal to $\epsilon_s = 0.02 \sum Y_t$ and $\epsilon_e = 0.0085 \sum Y_t$ or $0.015 \sum Y_t$. To make sure, we explored reduced values for ϵ_s and ϵ_e for dataset 1: $0.0 \sum Y_t$, $0.0025 \sum Y_t$ and $0.005 \sum Y_t$. Most satisfied with $\epsilon_s = 0.02 \sum Y_t$ and $\epsilon_e = 0.0025 \sum Y_t$, we applied the same constraint to datasets 2-4. Table 2.5 lists the user inputs by datasets and labels the rows with a letter; per row, Table 2.6 includes the corresponding model estimates for J, ϱ, ϱ_0, φ, and S. Figures 2.9 - 2.13 graphically display the functional estimates of the protein signal.

Dataset	Row	k	R	R_0	$\epsilon_s/\sum Y_t$	$\epsilon_e/\sum Y_t$
1	a	Cauchy	50	2	0	0
	b				0.020	0.0025
	c				0.020	0.0050
	d				0.020	0.0085
2	a	Cauchy	50	2	0.020	0.0025
3	a	Cauchy	50	2	0.20	0.0025
	b	Gaussian	50			
	c	Gaussian	65			
	d	Gaussian	75			
	e	Gaussian	100			
4	a	Cauchy	50	2	0.020	0.0025

Table 2.5: User inputs for EM Algorithm to apply to four real datasets.

Figure 2.12 depicts the model results for the spectrum averaged across the diseased group, Dataset 3. Averaging spectra may either decrease resolution because of spectra misalignment or increase resolution because of the central limit theorem. Thus, for Dataset 3, we explore the use of both Gaussian and Cauchy kernels as well as varying resolution centering parameters. Table 2.5 lists the varying hyperparameter specifications and Figure 2.12 displays the results. From Figure 2.12, we see that the Cauchy kernel clearly fits better then a Gaussian kernel regardless of the resolution parameters proposed.

Both averaged and raw spectra were used because Morris *et al.* [2005] reported that more accurate peak extraction occurs when using mean, rather than individual, spectra. To compare the results from individual and mean spectrum analyses, we calculate the number of peaks found in each. Assuming peak locations are equal if they are within 0.2% Da [Campa *et al.*, 2003a], Dataset 1 and 2 when modeled with the same hyperparameter specifications share 35 peaks.

Dataset	Row	$\hat{J}$	$\hat{\varrho}$	$\hat{\varrho}_0$	$\hat{\varphi}$	$\hat{S}$
1	a	807	83.88	1.56	0.00034	2.64
	b	123	92.77	1.60	0.00032	14.50
	c	100	93.30	1.50	0.00032	15.48
	d	74	94.67	1.43	0.00032	17.89
2	a	59	94.67	1.58	0.00168	13.05
3	a	67	90.91	1.89	0.00282	12.44
	b	38	57.22	2.49	0.00118	24.11
	c	36	57.98	2.49	0.00098	25.60
	d	45	52.27	2.53	0.00121	20.99
	e	38	57.18	2.45	0.00130	22.53
4	a	102	66.29	1.78	0.01465	1.78

Table 2.6: Results from EM Algorithm application to four real datasets. The column labeled by "Row", corresponds to the same column in Table 2.5.

2.4 Discussion

To extract pertinent features from MALDI-TOF spectra we present a model-based approach that enables the simultaneous estimation of protein signal and random and/or systematic error. In modeling both, the declaration of a peak depends upon all of the characteristics of a spectrum: resolution, background, precision, peak concentration and peak location. This is a distinct difference from previous approaches that only compare peak heights to thresholds, neighboring observations, and/or noise estimates. In fact, we saw from the simulation exercises that even when attempts are made to clean the data before comparing peak heights and areas as described in `PROcess`, our modeling effort was approximately two times more sensitive than `PROcess` and had 30% lower FDR in low S/N scenarios.

Posterior point parameter estimates are achieved via an expectation maximization algorithm. Other means of parameter estimation(e.g. stochastic integration) are applicable to the proposed model, but the EM algorithm is quick, easy to im-

plement, and takes advantage of quadratic programming and optimization functions available in R. We continue our peak finding research in two sequel papers that use the results from this effort as starting values for a stochastic integration approach. Unlike the EM algorithm, peaks may die and birth during a Markov Chain Monte Carlo that will ultimately define the posterior distribution for peak concentrations, widths, and locations.

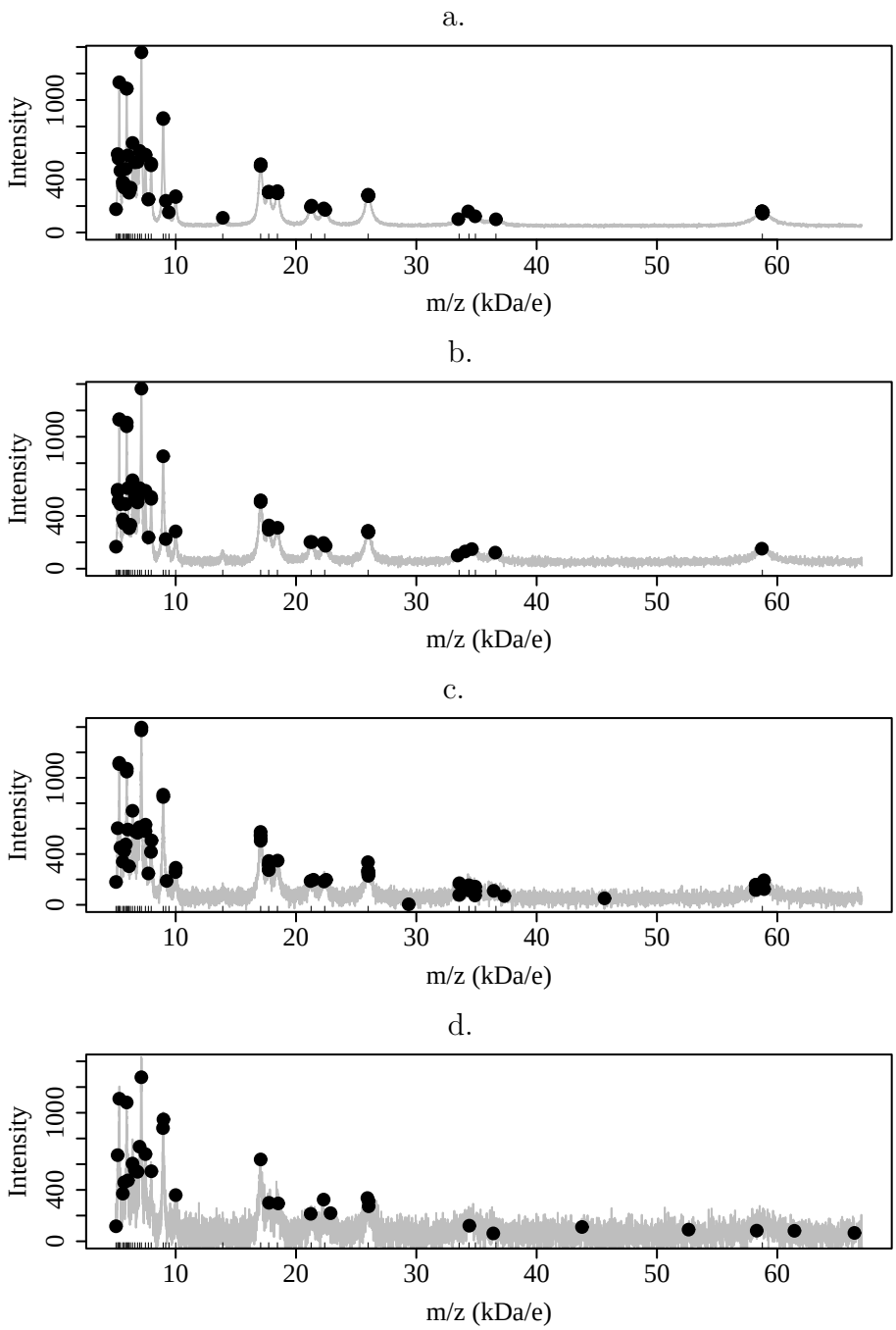


Figure 2.5: Plots a.-d. provide example feature extraction results when using the EM approach. Plots a.- d. show one spectrum from each of the simulation scenarios where $\varphi = 0.04$, 0.0044, 0.0011, and 0.00028 respectively. The circular dots reference estimated peaks locations and the dashes marks on the x-axes locate the true peak locations.

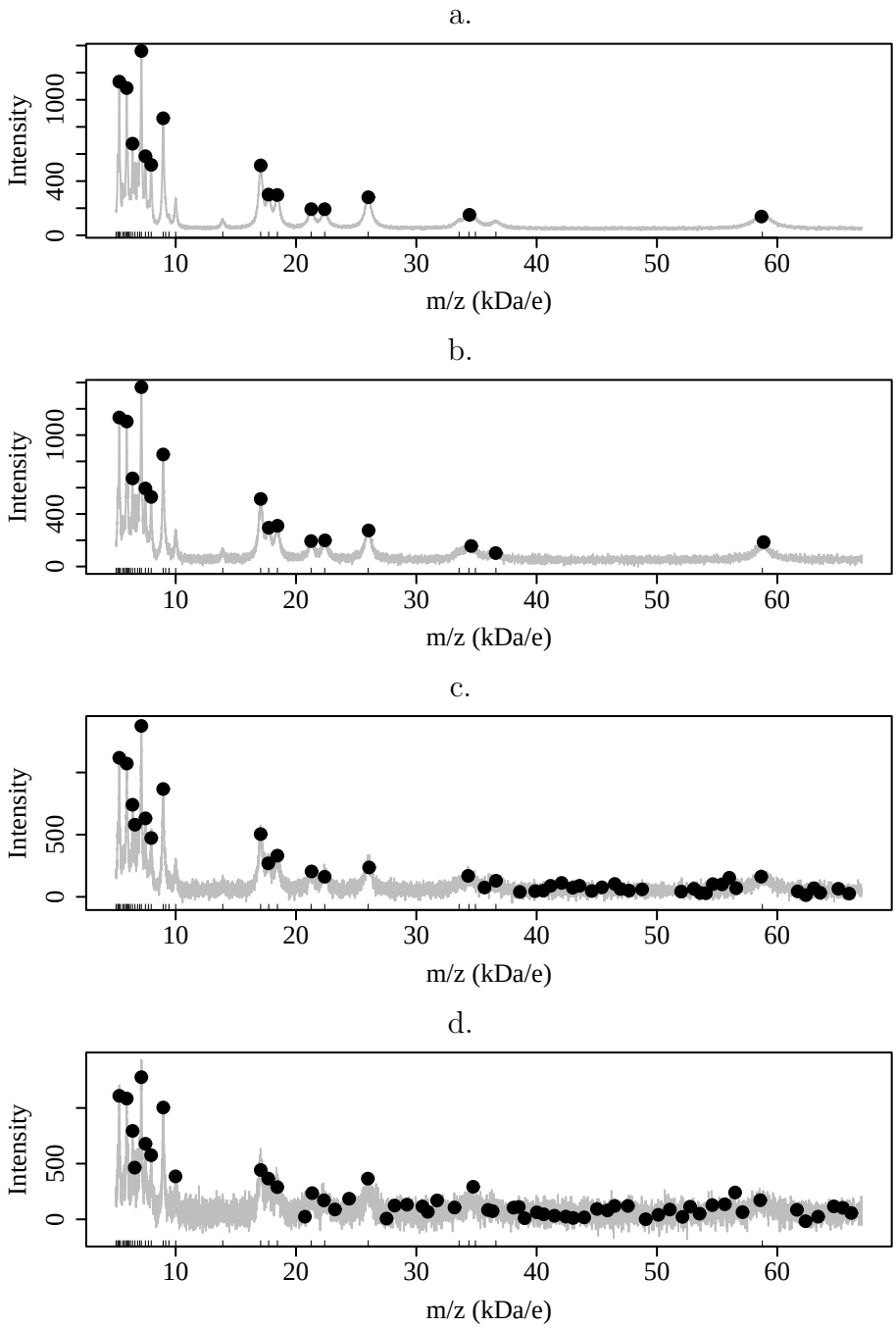


Figure 2.6: Plots a. - d. provide feature extraction results from using `PROcess` with default arguments. Plots a.- d. show one spectrum from each of the simulation scenarios where $\varphi = 0.04$, 0.0044, 0.0011, and 0.00028 respectively. The circular dots reference estimated peaks locations and the dashes marks on the x-axes locate the true peak locations.

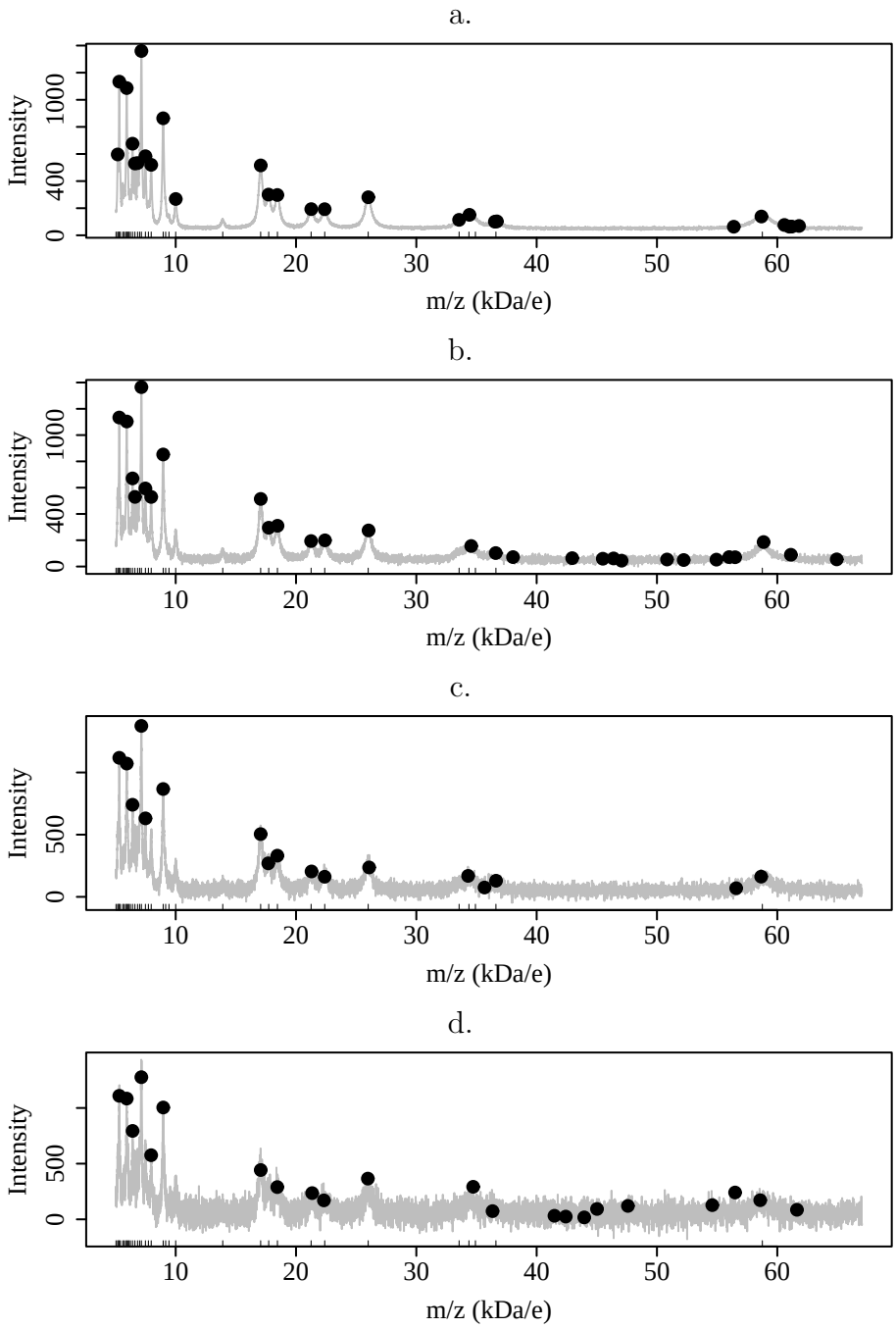


Figure 2.7: Plots a. - d. provide feature extraction results from using `PROcess` with user altered inputs, rather than default arguments. Plots a.- d. show one spectrum from each of the simulation scenarios where $\varphi = 0.04$, 0.0044, 0.0011, and 0.00028 respectively. The circular dots reference estimated peaks locations and the dashes marks on the x-axes locate the true peak locations.

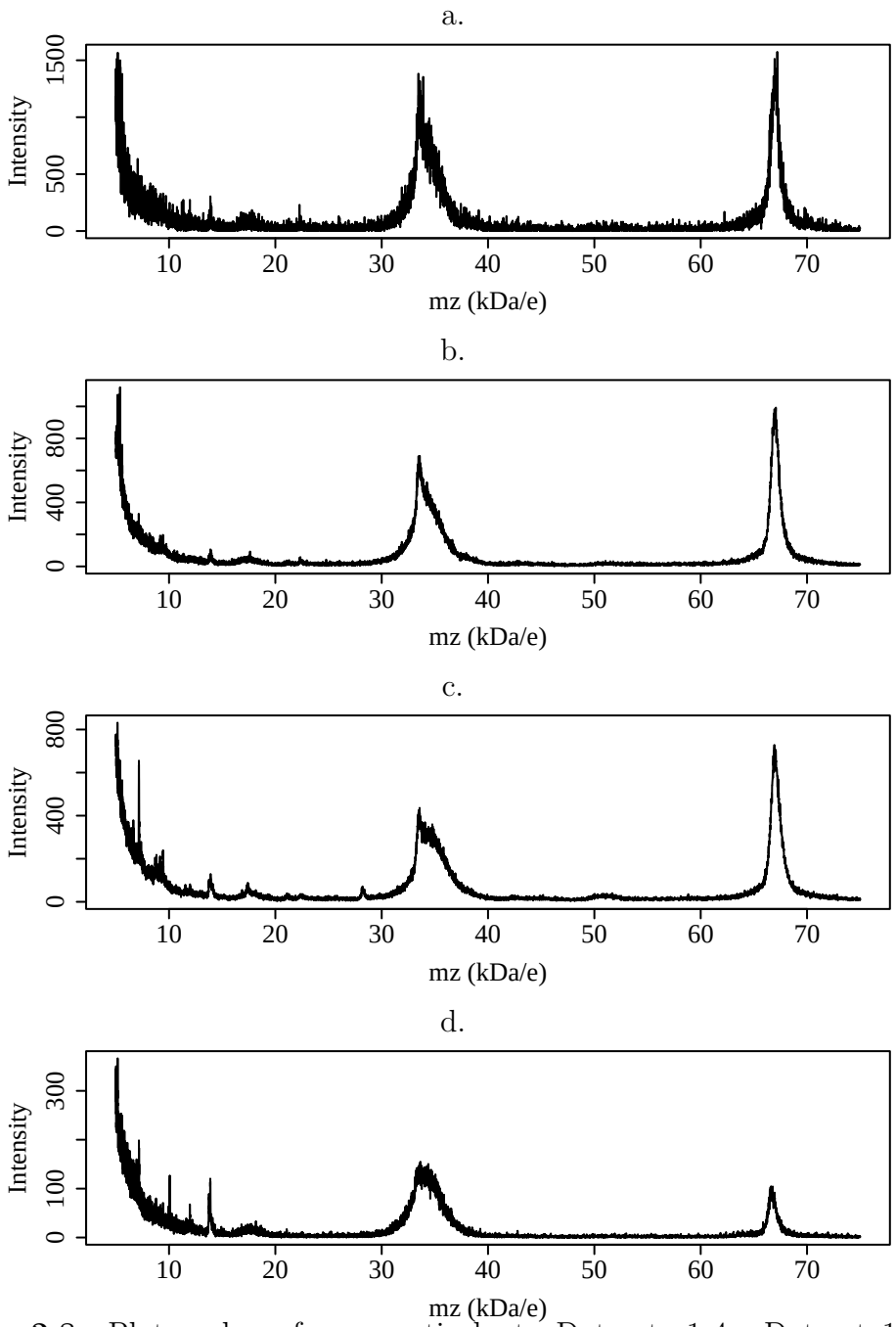


Figure 2.8: Plot a.-d. refer respectively to Datasets 1-4. Dataset 1 is one spectrum from one shot, Dataset 2 is an average spectrum for 10 shots on the same sample, Dataset 3 is an average spectrum of one shot from 16 cancer patients, and Dataset 4 is an average spectrum of one shot from 13 control patients. Each of the datasets have been shifted by minus the minimum, spectrum intensity.

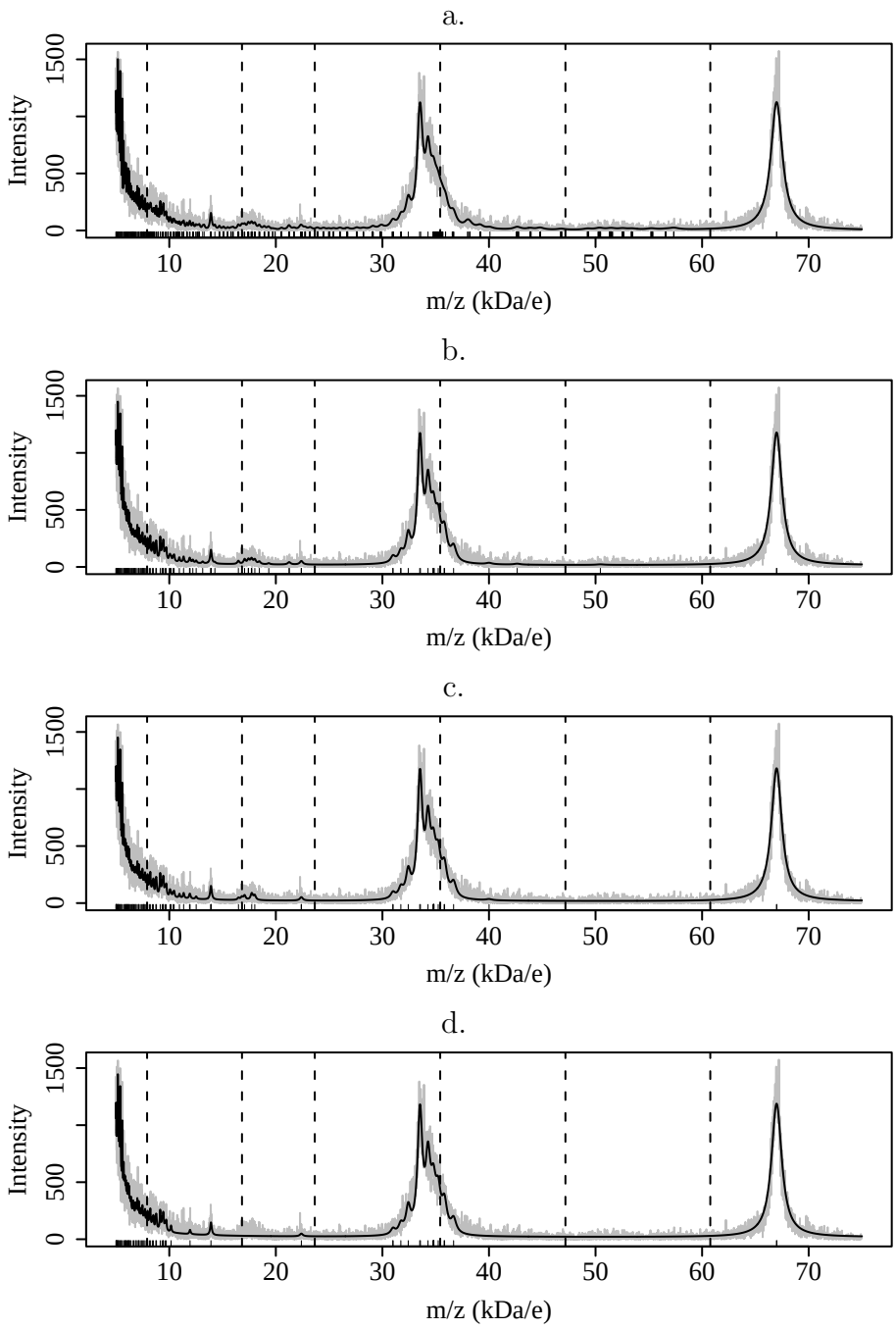


Figure 2.9: Final model fit using the proposed EM approach for dataset 4 using two different sets for ϵ_s and ϵ_e: a. $\epsilon_s = 0$ and $\epsilon_e = 0$, b. $\epsilon_s 0.02 \sum Y_t$ and $\epsilon_e = 0.0025 \sum Y_t$, c. $\epsilon_s = 0.02 \sum Y_t$ and $\epsilon_e = 0.005 \sum Y_t$, d. $\epsilon_s = 0.02 \sum Y_t$ and $\epsilon_e = 0.0085 \sum Y_t$.

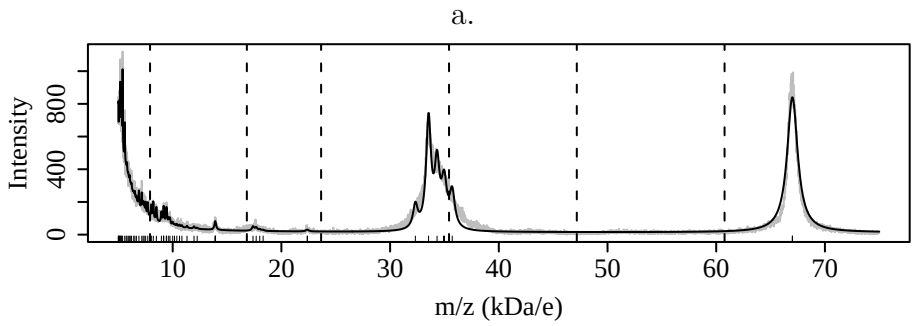


Figure 2.10: Final model fit for dataset 2 where $\epsilon_s = .02 \sum Y_t$ and $\epsilon_e = 0.0025 \sum Y_t$.

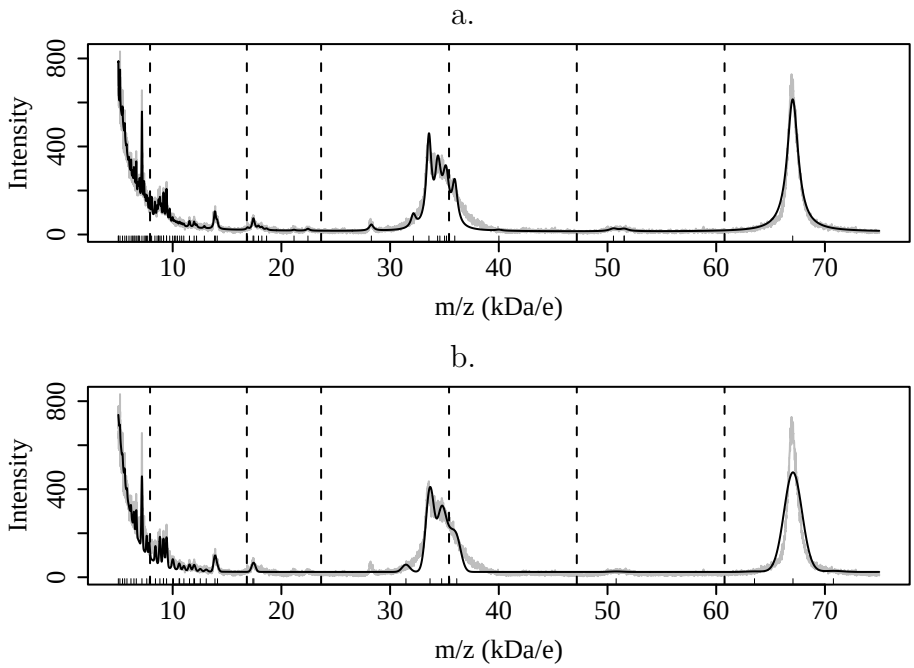


Figure 2.11: Final model fit for dataset 3 using the same values for ϵ_s and ϵ_e,$\epsilon_s = 0.02 \sum Y_t$ and $\epsilon_e = 0.0025 \sum Y_t$, but different kernels. Plots a. and b. shows the model fit with a Cauchy and Gaussian kernels respectively.

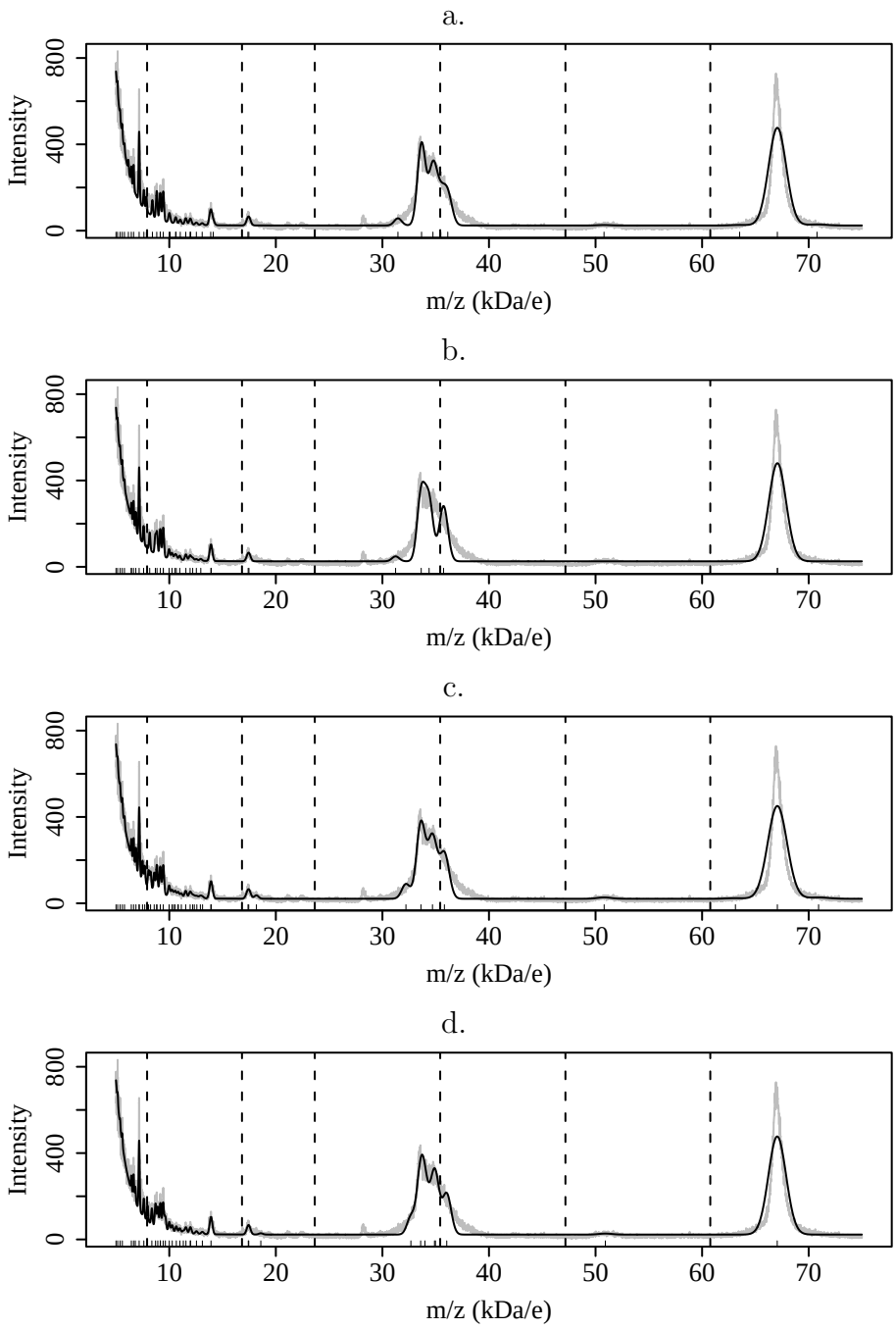


Figure 2.12: Final model fit for dataset 3 using the same values for ϵ_s and ϵ_e,$\epsilon_s = 0.02 \sum Y_t$ and $\epsilon_e = 0.0025 \sum Y_t$, but different R: Plot Plots a.- d. show Gaussian kernel model fits with R set to 50, 65, 75, and 100 respectively.

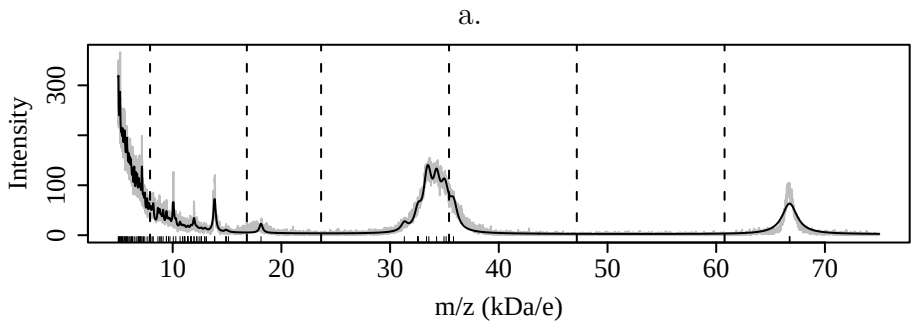


Figure 2.13: Final model fit for dataset 2 where $\epsilon_s = 0.02 \sum Y_t$ and $\epsilon_e = 0.0025 \sum Y_t$.

Chapter 3

Mass Spectrometry Feature Extraction Using a Lévy Random Field Model

3.1 Introduction

Recent innovations in protein separation methods, ionization procedures, and detection algorithms have led mass spectrometry (MS) to play a vital role in the explosive growth of proteomics [Dass, 2001, *p.* xxi]. Despite technological advances in data collection, it remains challenging to extract biologically relevant information (such as biomarkers) from MS spectral data [Coombes *et al.* 2005a; Baggerly *et al.* 2004; Dass 2001, chaps. 3, 5; Do *et al.* 2006, chaps. 14, 15].

Identifying peak locations (which represent proteins) and quantifying protein abundance is often preceded by a two or more stage analysis, involving calibration, normalization, baseline subtraction and filtering of noise [Morris *et al.*, 2005; Tibshirani *et al.*, 2004; Yasui *et al.*, 2003; Carpenter *et al.*, 2003]. A problem with such multistage analyses is that each individual step potentially introduces errors or biases that may subsequently create challenges for later stages such as classification of subjects or identification of biomarkers; methods that simultane-

ously model background, noise and features may lead to improved classification or inferences [Coombes *et al.*, 2005b]. Nonparametric models such as wavelets have proved successful in simultaneously modeling background and denoising, allowing one to extract features or regions of spectra that differentiate groups [Yasui *et al.*, 2003; Coombes *et al.*, 2005b]. While wavelets are well suited for modeling local features like spectral peaks, the coefficients and basis functions used in the representation of expected intensity have no inherent biological interpretation. In this paper, we propose a novel nonparametric model using an adaptive kernel regression model [Clyde and Wolpert, 2006] that provides the adaptivity and flexibility that make wavelet methods advantageous, but more importantly uses a model parameterization for features with direct biological interpretations.

We begin in Section 3.2 with a brief overview of MALDI-TOF mass spectrometry. In Section 3.3 we develop a statistical model for protein abundance as a function of time-of-flight using a novel nonparametric Bayesian approach. The model encompasses both signal (the protein abundance) and noise (due to artifacts of the MALDI-TOF technology), including run-to-run variability. Based on physical models for mass spectroscopy [Coombes *et al.*, 2005a], the distribution of the time of flight of a given protein may be represented by a kernel density, such as a Gaussian or Cauchy density with location parameter representing the expected time of flight and width parameter governed by both the mass of the protein and resolution of the machine (and its settings) used for MS. The unknown protein signal is then represented as a convolution of theses kernels with a distribution that characterizes protein abundance at expected times of flight. Solving this deconvolution problem provides estimates of the number of proteins, their times of flight, and abundances. As deconvolution problems typically have no unique solution, we

utilize a Bayesian approach that incorporates prior knowledge about the process which facilitates resolving the number of peaks (proteins). Prior distributions for the Bayesian model are developed in Section 3.4 from expert knowledge about the MALDI-TOF procedure and from exploratory analysis of MALDI-TOF data from related experiments. Inference about parameters of clinical interest, based on posterior distributions, are described in Section 3.5. In Section 3.6 we validate our method and compare it to the conventional peak-finding algorithm `PROcess` [Li, 2005] using simulated data. We illustrate our methodology in Section 3.7 using data from a recent lung cancer study conducted at Duke University. We conclude with a discussion and suggestions for future work in Section 3.8.

3.2 MALDI-TOF Data

In Matrix Assisted Laser Desorption Time-of-Flight Mass Spectrometry, or MALDI-TOF MS, inference about the molecular composition of a compound is based on indirect measurement of molecular masses. Molecules, initially embedded in a *matrix* of low molecular weight substance such as sinapinic acid on a metal target plate, are simultaneously dislodged (by vaporizing the substrate) and ionized (by removing one or more electrons from the molecule) by laser pulses, or *shots*. The now-charged molecules are accelerated by a strong electric field toward a detector, where the total number of molecules detected (or, more precisely, their aggregate charge) are recorded during specified time intervals (*clock ticks*, each about 4 ns long). From these, a histogram or *spectrum* is constructed of the approximate times-of-flight (TOF's) for the molecules that comprise the compound in some number of repeated laser "shots" at the same location.

Distance traveled under constant acceleration is a quadratic function of time,

leading to a simple but nonlinear relationship between TOF and the molecules' masses and ionic charge (the latter two enter only through their quotient, the *mass to charge ratio* m/z). Under ideal conditions the TOF spectrum generated by MALDI-TOF would show a narrow spike at the TOF corresponding to each molecular species present, with a height to the molecule's concentration.

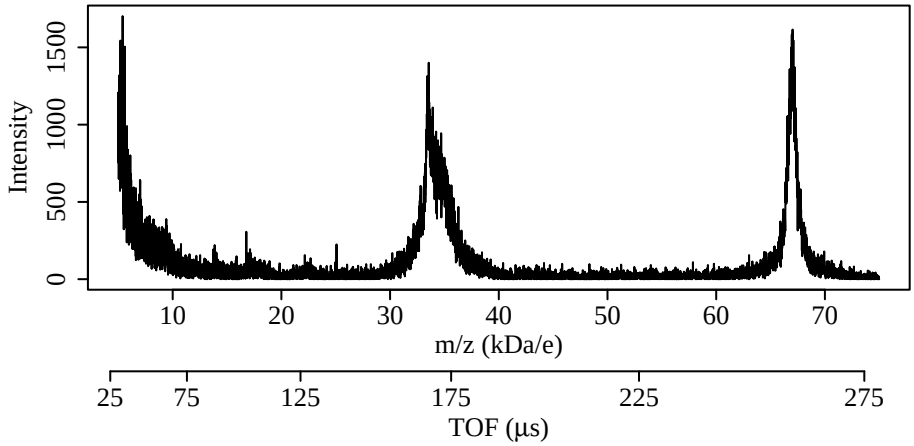


Figure 3.1: A MALDI-TOF mass spectrum, plotting intensity *vs.* both m/z and TOF, for a ten-shot spectrum.

In actual MALDI-TOF spectra (see Figure 3.1.) we observe irregular peaks rather than one-dimensional spikes because molecules of equal size and charge do not all reach the detector at the same time. The most important of the many causes of TOF dispersion is variability in the amount of ionizing laser energy received by molecules of varying location within the matrix; those further from the matrix surface or from the center of the laser pulse may receive less kinetic energy and thus have lower initial velocities than similarly-sized molecules located closer to the center, delaying their arrival at the detector. Molecules may exchange energy in collisions, and may lose or gain mass through fragmentation and agglomeration,

respectively. All these lead to TOF variation for each molecular species [Coombes *et al.*, 2005a; Zhigilei and Garrison, 1998; Franzen, 1997].

The interpretation and analysis of MALDI-TOF data are complicated by several other sources of variation described by Morris *et al.* [2005] and Coombes *et al.* [2005a]. In addition to *measurement error*, or random noise, which may mask or distort protein peaks even in a single spectrum, at least three other sources complicate the comparison or synthesis of multiple spectra: *calibration* (uncertainty in the conversion of TOF to m/z, including variable latency that affects time registration); *background* (a constant or even time-varying trend in the overall level); and *scale* (caused by many things including variability of laser intensity).

One way to accommodate these sources of variability is to construct models for peak identification and quantification that incorporate all these recognized sources of variability, as in the wavelet approach of Morris *et al.* [2005]. Our approach, described in Section 3.3, has the advantage that each of the model parameters has a direct physical interpretation.

3.3 A Model for MALDI-TOF

To eliminate variability attributable to differing numbers of laser shots and differing baselines, we model the standardized spectrum at TOF t, for some range $T_0 \leq t \leq T_1$,

$$Y_t = \frac{Y_t^{\mathsf{ob}} - \min(\mathbf{Y}^{\mathsf{ob}})}{l} \tag{3.1}$$

based on a raw spectrum $\mathbf{Y}^{\mathsf{ob}} = \{Y_t^{\mathsf{ob}}\}_{T_0 \leq t \leq T_1}$ with l laser shots. Dass [2001, *p.* 75] suggests that the initial molecular velocities will be approximately Gaussian in distribution. This and the physical modeling of the MALDI-TOF process by

Coombes *et al.* [2005a] suggest that TOFs for a single isotopic peak will also have symmetric bell-shaped distributions in the time domain, leading us (and others—see [Morris *et al.*, 2005; Kempka *et al.*, 2004; Malyarenko *et al.*, 2005]) to prefer TOF (in μs) rather than m/z (in Da/e) for spectral modeling.

3.3.1 Peak Shape

The shape of a symmetric isotopic peak may be represented by a probability density function with parameters governing the protein peak's location τ and width ω. Examples include the Gaussian

$$k(t;\tau,\omega) = \frac{1}{\sqrt{2\pi}\,\omega}\exp(-|t-\tau|^2/2\omega^2)$$

and Cauchy (sometimes called Lorentzian in the MS literature)

$$k(t;\tau,\omega) = \frac{\omega}{\pi(\omega^2+|t-\tau|^2)}, \tag{3.2}$$

as suggested by [Dass 2001, *p.* 75; Kempka *et al.* 2004; Applied Biosystems 2001, *p.* 6-30]. A protein signature associated with J peaks may now be represented as a sum

$$f(t) = \sum_{j=1}^{J} k(t;\tau_j,\omega_j)\,\eta_j, \tag{3.3}$$

where $\{\tau_j\}$, $\{\omega_j\}$ and $\{\eta_j\}$ represent the location, width, and abundance of the j^{th} peak.

3.3.2 Peak Width and Resolution

Protein peaks tend to be broader for late-arriving molecules than for earlier ones, with width nearly proportional to arrival time [Siuzdak, 2003, *p.* 44]; for this reason

it is conventional in mass spectrometry to quantify the precision (narrowness) of a kernel $k(\,\cdot\,;\tau,\omega)$ not by the width ω, but by the *resolution*

$$\rho \equiv \tau/\Delta\tau$$

where $\Delta\tau$, the so-called *full width at half mass* or FWHM, is the width of the kernel $k(\,\cdot\,;\tau,\omega)$ at half its height. For a symmetric kernel, $\Delta\tau$ is the solution of the equation

$$k(\tau \pm \tfrac{1}{2}\Delta\tau;\tau,\omega) = \tfrac{1}{2}k(\tau;\tau,\omega)$$

[*e.g.*, Dass, 2001, *p.* 120]. For the Gaussian and Cauchy kernels we have $\Delta\tau = 2\omega\sqrt{\log 4}$ and $\Delta\tau = 2\omega$, leading respectively to $\omega = \omega(\tau,\rho)$ with

$$\omega(\tau,\rho) = \frac{\tau}{2\rho\sqrt{\log 4}} \qquad \text{and} \qquad \omega(\tau,\rho) = \frac{\tau}{2\rho}. \tag{3.4}$$

Prior knowledge about precision can be used to resolve the ambiguity illustrated in Figure 3.2., where the observed spectrum may arise from either a single wide peak or a pair of near-by narrower peaks.

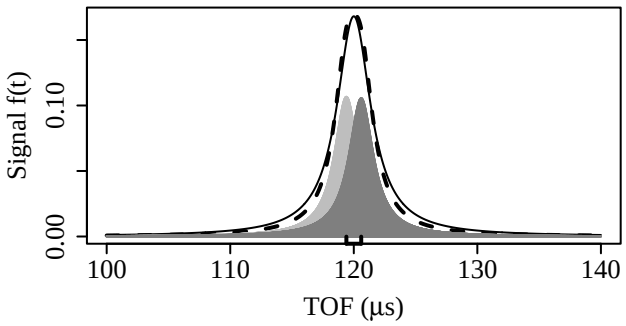


Figure 3.2: The (nearly indistinguishable) solid and dotted lines represent simulated protein signals from a sample mixture with either one wide or two narrow peaks.

3.3.3 Background Noise Sources

Even in the absence of any protein molecules (*i.e.*, with $f(t) \equiv 0$) the MALDI-TOF spectrum does not vanish. Figure 3.3 a. shows the nearly-constant level of thermal noise from a run with an empty plate, while Figure 3.3 b. shows the rapidly-decreasing signal with only the sinapinic acid matrix, showing the arrival at the detector of ionized matrix molecules (far lighter than the proteins under study, hence near the left of the spectrum). Together these contribute a background that falls off nearly exponentially to a non-zero asymptote. Exploratory analysis

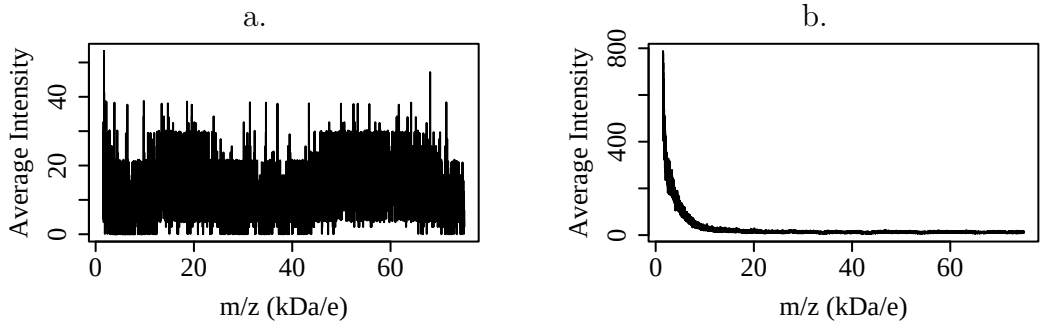


Figure 3.3: Figure 3.3 a. shows the near-uniform thermal noise spectrum (or "ringing") from an empty plate, while Figure 3.3 b. shows the rapidly-decreasing spectrum from the sinapinic acid matrix without any protein sample. Note the vertical axes are scaled differently.

suggests that the matrix molecular signal $\beta_0(t)$ can be modeled adequately as an exponential function,

$$\beta_0(t) = k_0(t; \omega_0)\, \eta_0 = \frac{\eta_0}{\omega_0} \exp\{-t/\omega_0\} \mathbf{1}_{(t>0)}, \tag{3.5}$$

with width $\omega_0 > 0$ and intensity $\eta_0 > 0$.

3.3.4 Mean Spectrum

To reflect all these features, we model the expected spectral intensity as:

$$\mu(t) = \zeta\Big\{(1-S) + S\big[f(t) + \beta_0(t)\big]\Big\} \tag{3.6}$$

for an overall scale ζ, a dimensionless number $S \in [0,1]$, the protein signal $f(t)$ from Equation (3.3), and the matrix molecular signature $\beta_0(t)$ from Equation (3.5). The term S represents the proportion of observed intensity produced by molecular signal (both matrix and protein), rather than by the ringing and thermal noise of Figure 3.3 a..

3.3.5 Likelihood

Both gamma and log-normal distributions are commonly used to model positive data like Y_t. We based our choice on the observation that the variance is proportional to the mean for gamma distributions and to the square of the mean for log-normals. Exploratory data analysis (from both a Box-Cox approach, and a regression comparison illustrated in Figure 3.4.) suggests that the variance of standardized MS data Y_t, given the mean, is nearly proportional to the first power of the mean, supporting the gamma model

$$Y_t \mid \mu(\,\cdot\,),\, \varphi \overset{ind}{\sim} \mathrm{Ga}(\varphi\mu(t), \varphi), \tag{3.7}$$

with mean $\mu(t)$ and mean : variance ratio φ. This leads to likelihood function

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^{n} \mathrm{Ga}(Y_{t_i}; \varphi\mu(t_i), \varphi) \tag{3.8}$$

for the parameter vector $\boldsymbol{\theta}$ comprising the conditional mean function $\mu(\,\cdot\,)$ (or, equivalently from Equation (3.6), all of ζ, J, $\{\tau_j, \omega_j, \eta_j\}_{1\le j\le J}$, S, ω_0, and η_0) and

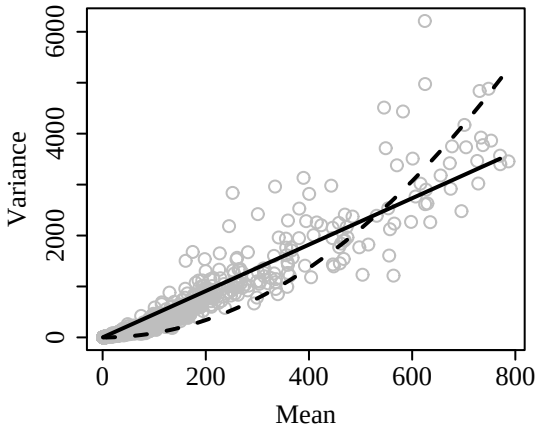


Figure 3.4: Linear and quadratic fits of mean intensity *vs.* variance of intensity for 200 μs blocks of observations from a single spectrum.

φ. Here $\mathbf{Y} = \{Y(t_i)\}_{1 \le i \le n}$ represents the vector of standardized intensities, and $\mathrm{Ga}(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \mathbf{1}_{(y>0)}$ is the probability density function at $y \in \mathbb{R}$ for the gamma $\mathrm{Ga}(\alpha, \beta)$ distribution.

Typically the likelihood function of Equation (3.8) has many modes because it is difficult to distinguish wide peaks from clusters of narrow ones, or small peaks from noise, from the data alone. Estimating $\boldsymbol{\theta}$ (and in particular J, the number of protein peaks) by direct maximization of the likelihood leads to over-fitting the data and to over-estimating J. This can be overcome by regularization [Tikhonov, 1963] or by a Bayesian approach like ours, in which prior distributions penalize overly complex models.

3.4 Prior Distributions for MALDI-TOF

We now address the problem of constructing a joint prior distribution for all the unknown parameters of the model of Section 3.3,

$$\begin{aligned} Y_t \mid \mu(\,\cdot\,), \varphi &\overset{ind}{\sim} \mathrm{Ga}(\varphi\mu(t), \varphi) \qquad (3.9) \\ \mu(t) &= \zeta\Big\{(1-S) + S\big[f(t) + \beta_0(t)\big]\Big\} \\ f(t) &= \sum_{j=1}^{J} k(t;\tau_j,\omega_j)\,\eta_j, \quad \beta_0(t) = \frac{\eta_0}{\omega_0}\exp\{-(t-T_0)/\omega_0\}\mathbf{1}_{(t>T_0)} \end{aligned}$$

3.4.1 Measurement Error φ and Overall Level ζ

The exploratory data analysis of Section 4.3.3 suggests that the sample mean of the $\{Y_t\}$ is nearly proportional to the variance, with a ratio of approximately $\varphi \approx 0.223$. We use a gamma prior distribution $\varphi \sim \mathrm{Ga}(a_\varphi = 0.5, b_\varphi = 1)$ to place about 90% of the prior mass in the interval $[0.002, 2]$.

The parameter ζ may be interpreted as the mean level or scale for Y_t, since $\mathsf{E}[f(t)] \approx 1$ (see Section 3.4.2). Since experimental levels depend on a wide range of exogenous variables and vary widely among trials, we use a rather tight data-dependent prior distribution for ζ centered at the empirical mean $\bar{Y}$,

$$\zeta \sim \mathrm{Ga}(a_\zeta, b_\zeta)$$

with a_ζ, b_ζ chosen so that $\mathsf{E}[\zeta] = \bar{Y}$ and the geometric standard deviation $(\sqrt{\mathsf{V}[Y]}/\mathsf{E}[\zeta])$ is approximately 0.10.

3.4.2 Prior Distribution for Protein Signature $f(\cdot)$

We use a negative binomial prior distribution for the number J of peaks in the protein signature

$$f(t) = \sum_{j=1}^{J} k(t; \tau_j, \omega_j)\, \eta_j$$

with mean parameter $\mu_J = \mathsf{E}[J] = 100$ and shape parameter $\alpha_J = 1$ chosen to achieve a median of $J \approx 70$ peaks with symmetric 50% and 90% ranges of approximately $30 \leq J \leq 140$ and $5 \leq J \leq 300$, respectively [Campa *et al.*, 2003b].

Conditional on J, we take the triplets $\{(\tau_j, \omega_j, \eta_j)\}_{1 \leq j \leq J}$ to be independent and identically-distributed. For peak abundances $\{\eta_j\}$ we use the truncated gamma distribution $\mathrm{Ga}(0, \lambda, \epsilon)$ with parameters λ and ϵ chosen below in Section 3.4.4. Here $\mathrm{Ga}(\alpha, \lambda, \epsilon)$ denotes the truncated gamma distribution with density function

$$\mathrm{Ga}(\eta; \alpha, \lambda, \epsilon) \equiv \frac{\lambda^\alpha}{\Gamma(\alpha, \lambda\epsilon)} \eta^{\alpha-1}\, e^{-\lambda\eta} \mathbf{1}_{(\eta > \epsilon)}, \tag{3.10}$$

where $\Gamma(\alpha, x) \equiv \int_x^\infty z^{\alpha-1} e^{-z}\, dz$ denotes the incomplete gamma function [Abramowitz and Stegun, 1964, §6.5.3]. For $\alpha, \lambda > 0$ this is the conditional distribution of a gamma-distributed $\mathrm{Ga}(\alpha, \lambda)$ random variable, given that it exceeds $\epsilon \geq 0$; it is well-defined for all $\alpha \in \mathbb{R}$ if $\epsilon > 0$.

There is little reason to give higher prior probability to one range of TOFs than another without prior knowledge of the collection of proteins present in the samples. Thus we take $\{\tau_j\}_{1 \leq j \leq J} \overset{\text{iid}}{\sim} \mathrm{Un}(T_0, T_1)$ (independently of J and $\{\lambda_j\}_{1 \leq j \leq J}$), for some interval large enough to exceed the TOF for all molecules of interest. To eliminate saturation by matrix molecules at the low end, and include

as wide a range as possible of the biologically relevant molecules, we chose the range $[5\,\text{kDa/e} \leq \text{m/z} \leq 75\,\text{kDa/e}]$, leading to TOF range $[T_0, T_1] = [32\,\mu\text{s}, 278\,\mu\text{s}]$ of length $T = T_1 - T_0 = 246\,\mu\text{s}$.

To construct a prior distribution on the widths $\{\omega_j\}_{1\leq j\leq J}$ we first use expert opinion to construct an informed prior distribution on the resolutions $\{\rho_j\}_{1\leq j\leq J}$ (see Section 3.3.2). Siuzdak [2003, *p.* 44] suggests that individual peak resolutions ρ_j *should* be nearly constant across the entire TOF range, but in practice they are observed to vary [Coombes *et al.* 2005a; Applied Biosystems 2001, *p.* 6-32]. To reflect this variation we construct a hierarchical prior probability distribution for the resolution parameters $\{\rho_j\}_{1\leq j\leq J}$. Independently of J and $\{(\lambda_j, \tau_j)\}_{1\leq j\leq J}$, we take

$$\begin{aligned} \varrho &\sim \text{LN}\big(\log(\varrho_\mu), 0.5^2\big) \\ \rho_j \mid \varrho &\overset{\text{iid}}{\sim} \text{LN}\big(\log(\varrho), 0.05^2\big), \end{aligned}$$

centered around a hyperparameter ϱ_μ taken from the literature [Applied Biosystems, 2001, Table 6-2; see also Table H-6]

$$\varrho_\mu = \begin{cases} \varrho_\mu^{(S)} = 400, \text{ for } 2\,\text{kDa/e} \leq \text{m/z} < 5\,\text{kDa/e} \\ \varrho_\mu^{(M)} = 700, \text{ for } 5\,\text{kDa/e} \leq \text{m/z} < 25\,\text{kDa/e} \\ \varrho_\mu^{(L)} = 100, \text{ for } 25\,\text{kDa/e} \leq \text{m/z} \end{cases}$$

for small (S), medium (M), and large (L) compounds. Standard deviations were chosen following discussions with collaborating spectrometrists to fit observed resolution values and to achieve overlapping ranges for adjacent molecular size ranges [Fitzgerald and Roulhac, 2006]. The relationship between width, TOF, and resolution given by Equation (3.4) now induces a prior distribution on the

width parameters, *e.g.*, $\{\omega_j\}_{1\le j\le J} \mid \boldsymbol{\tau}, \varrho \overset{ind}{\sim} \mathrm{LN}\big(\log(\tau_j/2\varrho), 0.05^2\big)$ for the Cauchy kernel.

A random variable $\eta \sim \mathrm{Ga}(\alpha, \lambda, \epsilon)$ with the truncated gamma distribution of Equation (3.10) has mean $\mathsf{E}[\eta] = (\alpha/\lambda) + \epsilon^\alpha \lambda^{\alpha-1}\, e^{-\lambda\epsilon}/\Gamma(\alpha, \lambda\epsilon)$ in general or, in our case with $\alpha = 0$,

$$\mathsf{E}[\eta] = \frac{1}{\lambda\, e^{\lambda\epsilon}\, \mathrm{E}_1(\lambda\epsilon)}$$

Exploratory data analysis and discussions with spectrometrists suggest that the smallest peak that can possibly be distinguished from noise is about 5–10% of the average signal, so we take $\epsilon/\mathsf{E}[\eta] = \lambda\epsilon\, e^{\lambda\epsilon}\, \mathrm{E}_1(\lambda\epsilon) = 0.075$, *i.e.*, $\lambda\epsilon = 0.0227$. Since $\int_{T_0}^{T_1} k(t; \tau_j, \omega_j)\, d\tau_j \approx 1$ for t well away from the boundary of $[T_0, T_1]$,

$$\mathsf{E}\big[f(t)\big] \;\approx\; \frac{\mu_J}{T\, \lambda\, e^{\lambda\epsilon}\, \mathrm{E}_1(\lambda\epsilon)} \qquad T_0 \ll t \ll T_1 \tag{3.11}$$

and so to achieve $\mathsf{E}[f(t)] = 1$ we need $\mu_J\, \epsilon = 0.075\, T$, so $\epsilon = 0.1800\, \mu\mathrm{s}$ and $\lambda = 0.1261\, \mu\mathrm{s}^{-1}$.

3.4.3 Prior Distribution for Matrix Background

Distributions for the remaining parameters, η_0 and ω_0 (which determine $\beta_0(t)$) and S, are based in part on exploratory analyses of *matrix-spectra* (from experiments with sinapinic acid matrix but no protein mixture) and *blank-spectra* (in which neither matrix nor protein mixture cover the target metal plate).

The exponential fall-off rate $\frac{1}{\omega_0}$ of $\beta_0(t)$ (see Equation (3.5)) can be estimated by logarithmic regression of an initial segment of the blank spectrum from Figure 3.3 b. From the estimate $\frac{1}{\omega_0} \approx 0.0492 \pm 0.0001367$ (mean $\pm$ one standard error, in units of $\mu\mathrm{s}^{-1}$) we infer (using the delta method) that $\log(\omega_0) \approx 11.5 \pm 0.00278$.

To accommodate possible variation between the blank spectrum experiment and the protein analysis, we use a much broader prior distribution,

$$\omega_0 \sim \mathrm{LN}(3.012, 0.5^2).$$

We use a truncated gamma model for the abundance η_0,

$$\eta_0 \sim \mathrm{Ga}(0, \lambda_0, \epsilon)$$

with λ_0 chosen to match the mean $\mathsf{E}[\eta_0] = \left\{\lambda_0\, e^{\epsilon\lambda_0}\, \mathrm{E}_1(\epsilon\lambda_0)\right\}^{-1} \approx \widehat{\eta_0}$ with the estimate $\widehat{\eta_0}$ from nonlinear regression on an initial segment of the protein data set short enough that it does not appear to include any peaks associated with proteins (we used $0 < t < 40\,\mu$s, corresponding to the range $0 < \mathrm{m/z} < 6\,\mathrm{kDa/e}$). The solution is $\lambda_0 = x/\epsilon$ for the solution x to the equation

$$x\, e^x\, \mathrm{E}_1(x) = \frac{\epsilon}{\widehat{\eta_0}}$$

(easily found using Mathematica$^{\mathrm{TM}}$ [2005] or Maple$^{\mathrm{TM}}$ [2005], for example, or the approximations in Abramowitz and Stegun [1964, §5.1.53–56]).

With the same dataset we approximate $\mathsf{E}[1 - S]$ by first estimating the noise in low intensity region, divided by the average spectral intensity $\overline{Y_t}$. Exploratory analysis suggests that the detector might be responsible for 0–46% of an observed intensity, leading us to use a beta prior distribution

$$S \sim \mathrm{Be}(\alpha_S, \beta_S)$$

with mean $\mu_S = 0.77$ and variance $\sigma_S^2 = 0.013$, *i.e.*, parameters $\alpha_S = \mu_S[\mu_S(1 - \mu_S)/\sigma_S^2 - 1] = 9.720$ and $\beta_S = (1 - \mu_S)[\mu_S(1 - \mu_S)/\sigma_S^2 - 1] = 2.903$. Notice that the signal-to-noise ratio $\frac{S}{1-S}$ has an $F_{2\beta_S}^{2\alpha_S} = F_{5.81}^{19.4}$ prior distribution with mean $\frac{\alpha_S}{\beta_S - 1} \approx 5.12$.

3.4.4 Random Field Formulation

The distribution constructed in Section 3.4.2 for J and $\{(\tau_j, \omega_j, \eta_j)\}_{1 \le j \le J}$ induces one for $f(\,\cdot\,) = \sum_{j=1}^{J} k(t; \tau_j, \omega_j)\,\eta_j$, which can be written in integral form

$$f(t) \;=\; \iint_{\mathbb{R}^2} k(t; \tau, \omega) \Gamma(d\tau, d\omega)$$

with a random measure

$$\Gamma(d\tau, d\omega) \;\equiv\; \sum_{j}^{J} \delta_{\tau_j}(d\tau) \delta_{\omega_j}(d\omega)\,\eta_j$$

that assigns masses $\{\eta_j\}_{1 \le j \le J}$ to the J discrete support points $\{(\tau_j, \omega_j)\}_{1 \le j \le J} \subset \mathbb{R}^2$. The negative binomial distribution $J \sim \mathrm{NB}(\alpha_J = 1, \mu_J = 100)$ can be written in hierarchical form as a gamma-mixture of Poisson distributions, $J \mid \lambda_J \sim \mathrm{Po}(\lambda_J)$, $\lambda_J \sim \mathrm{Ga}(\alpha_J, \beta_J)$, with $\beta_J = \alpha_J / \mu_J = 0.01$.

For disjoint sets $\{A_n\} \subset \mathbb{R}^2$ the random variables $\{\Gamma(A_n)\}$ are conditionally independent given λ_J and ϱ— that is, learning about the widths and locations of peaks in one part of the spectrum tells us nothing *a priori* about peaks and their widths in other parts of the spectrum.[1] Conditional on λ_J and ϱ, the random measure $\Gamma(d\tau, d\omega)$ is a Lévy random field, whose integrals

$$\Gamma[\phi] \;\equiv\; \iint \phi(\tau, \omega)\, \Gamma(d\tau, d\omega)$$

have characteristic functions of Lévy-Khinchine form [see Khinchine and Lévy 1936 or *p.* 74 of Rogers and Williams 1994]

$$\mathsf{E}\big[\, \exp\{is\Gamma[\phi]\} \mid \lambda_J, \varrho \big] \;=\; \exp\left\{ \iint \left(e^{i\,s\,\phi(\tau,\omega)\eta} - 1 \right) \nu(d\eta, d\tau, d\omega) \right\}$$

with finite Lévy measure

[1] Note this does not reflect the possibility of multiply-charged ions (dimers, trimers, *etc.*), although a more sophisticated version of this model could incorporate that feature.

$$
\begin{aligned}
\nu(d\eta, d\tau, d\omega) &= \lambda_J \,\mathrm{Ga}(\eta; 0, \lambda, \epsilon)\,\mathrm{Un}(\tau; T_0, T_1) \\
&\quad \times \mathrm{LN}\big(\omega; \log(\tau/c\varrho), 0.05^2\big)\, d\eta\, d\tau\, d\omega \\
&= \frac{\lambda_J\, \eta^{-1}\, e^{-\lambda\eta}\, \mathbf{1}_{(\eta>\epsilon)}}{\mathrm{E}_1(\lambda\epsilon)\, T\, \omega\, \sqrt{200\pi}} \exp\big\{ -200 \log^2(\omega\varrho c/\tau)\big\}\, d\eta\, d\tau\, d\omega
\end{aligned}
$$

where $c = 2$ for the Cauchy kernel and $c = 2\sqrt{\log 4}$ for the Gaussian.

3.5 Posterior Analysis

To support inference about protein peak locations and abundance, and about other model parameters, we construct an ergodic Markov chain in the space Θ of possible parameter vectors $\boldsymbol{\theta} = \{\zeta, J, \{\tau_j, \omega_j, \eta_j\}_{1\le j\le J}, S, (\omega_0, \eta_0), \lambda_J, \rho\}$ with the posterior distribution as its stationary distribution [Besag *et al.*, 1995; Tierney, 1994; Gelfand and Smith, 1990]. At each Markov chain step we select one of the components of $\boldsymbol{\theta}$ and either replace it with a draw from its complete conditional posterior distribution, given the other components (a Gibbs step) or, if this is impractical, propose a small change in that component which is then accepted or rejected according to the Hastings probabilities (a random-walk Metropolis-Hastings or M-H step). Note that each proposed change in J (which we always take to be an M-H step of size one) changes the *dimension* of $\boldsymbol{\theta}$ (by three), requiring some delicacy in computing the Hastings ratios; such schemes, called "reversible jump MCMC algorithms," were introduced by Green [1995]. Our approach is modeled after that of Wolpert and Ickstadt [2004], who introduced a general RJMCMC procedure that iterates though possible values of $\{\tau_j, \omega_j, \eta_j\}_{1\le j\le J}$ using just three possible moves: peak birth (incrementing J by one and introducing a new triplet $(\tau_*, \omega_*, \eta_*)$), peak death (decrementing J by one and removing a randomly-chosen

triplet $(\tau_j, \omega_j, \eta_j)$), and peak update (moving a randomly-chosen triplet $(\tau_j, \omega_j, \eta_j)$ within $\mathbb{R}^3$). To these we add two new moves, with a vast improvement in algorithmic efficiency: *splitting*, in which a single large peak is replaced by a pair of smaller ones, and *merging*, in which two nearby peaks are replaced with a single larger one.

For sufficiently large spectra or complex protein mixtures, convergence to the posterior distribution may require millions of MCMC iterations and days of computation. To reduce computation time we begin the Markov chain close to its mode, located by applying the EM algorithm [Dempster *et al.*, 1977] to a simple Gaussian approximation to our model; see House *et al.* [2006c] for details. This acceleration of convergence entails a possible cost— it may lead us to miss multiple modes, and so may overstate the posterior precision of some model parameters.

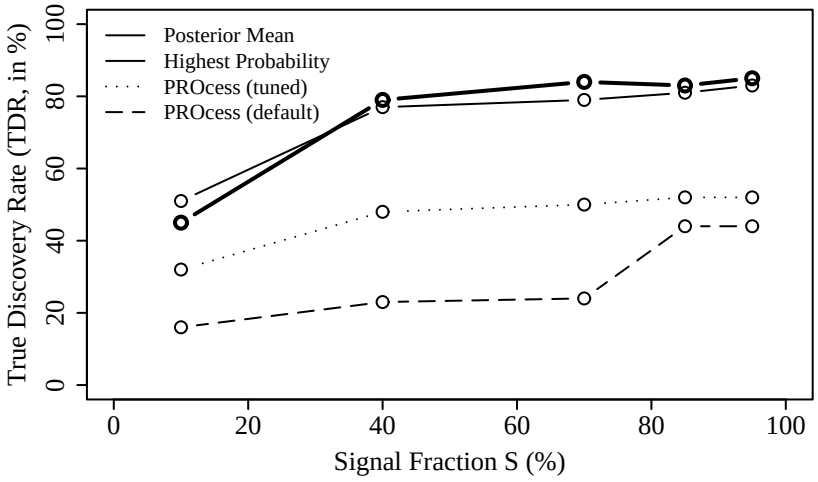
3.6 Simulation Study

In this section we describe a simulation study intended to explore how well our approach succeeds in locating true peaks within spectra of varying signal-to-noise ratios. We fit the model to simulated datasets for five values of the signal proportion: $S \in [0.10, 0.40, 0.70, 0.85, 0.95]$ (*i.e.*, signal-to-noise ratios of $[0.1, 0.7, 2.3, 5.7, 19]$). Twenty-five datasets were generated for each of these values of S, with fixed peak locations and with the remaining model parameters set to nominal values ($J = 35$, $\zeta = 130$, $\varphi = 0.50$, $(\alpha_J, \beta_J) = (1,\ 0.02)$, $\varrho = 56$, $\eta_j \equiv 3.9\,\mu$s, $\eta_0 = 35.6\,\mu$s, $\omega_0 = 46.2\,\mu$s) chosen so that the simulated spectra appeared similar to observed spectra. With a few exceptions the hyperparameter and RJMCMC specifications remained the same as those described in Section 3.4. Departures include reducing μ_J (by half) to 50 and fixing the overall spectrum resolution at $\varrho_\mu = 50$. As start-

ing values for the peak locations $\{\tau_j\}$ we took 50 equidistant peaks in the range $[32, 278]$ μs, corresponding to the range $[5\,000, 67\,000]$m/z. We used 500 000 RJM-CMC iterations, and retained the last 5 000 for further analysis. We compare our approach to a conventional peak finding algorithm using the `R` package `PROcess` [Li, 2005], which was originally designed for analysis of SELDI-TOF data, but is also applicable to MALDI-TOF. Li [2005] suggests removing background via subtracting a smoothed estimate of the spectrum local minima. Peaks are then defined as local maxima, with a signal greater than a user specified threshold and have a signal to noise ratio greater than a user specified value. Further more, an observation initially flagged as a peak may become disqualified if its area divided by the maximum peak area is less than a specified ratio. For both approaches, a discovered peak is regarded as a true peak if the TOF falls within $\pm 0.2\%$ of that of a true peak.

Figure 3.5 a,b. illustrates the fractions of true peaks found (TDR) and of spurious peak discoveries (FDR), respectively, by our approach and by `PROcess` at several signal fractions S. Both the single highest-probability simulation outcome (thin solid lines) and the posterior mean (thick solid lines) are shown, along with two estimates from `PROcess` (dashed lines showing default program settings and dotted lines lines representing carefully tuned settings). Figure 3.5. shows that we find an average of 75% of the true peaks, with a FDR in the range of 13–56%, across the range of signal fractions for both model output summaries; the posterior mean FDR remains below 16% and the mean FDR above 83% for all signal-to-noise ratios above one. The simulation shows that the kernel based approach is superior in performance to `PROcess` over a range of scenarios.

a.



b.

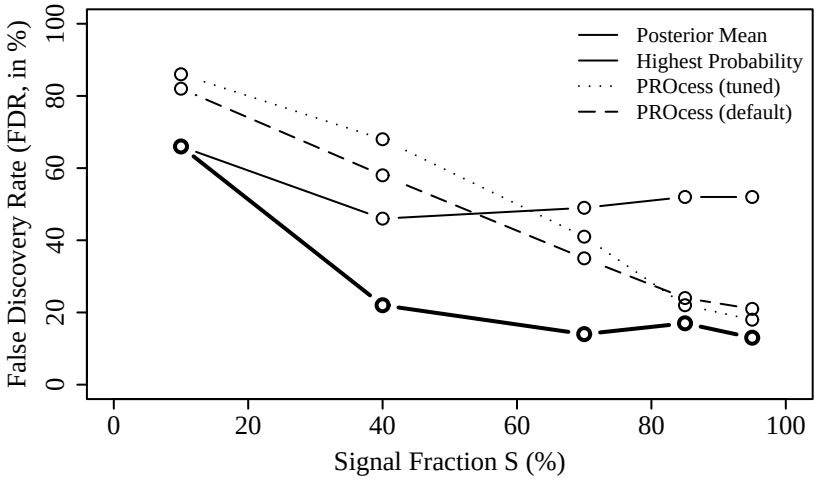


Figure 3.5: True and False Discovery Rates from simulation study. Note superiority of Posterior Mean (thick solid line) to Posterior Mode (thin solid line) and both PROcess approaches (dotted and dashed lines).

3.7 Examples

In this section, we apply our approach to three datasets and compare the results to those produced by the peak-finding algorithm `PROcess` [Li, 2005].

The first two datasets come from an observational study at the Duke Medical Center Radiology Department [Campa *et al.*, 2003b], intended to assess proteomic differences between cancerous and healthy patients. Serum from 30 diseased and control Caucasian males was collected and analyzed by a linear MALDI-TOF Delayed Extraction Mass Spectrometer, Applied Biosystems Voyager DE. Each sample was fractionated to 20 pH levels prior to the MALDI-TOF analysis to promote the ionization of a range of proteins. Ten replicate spectra were stored for each fraction, each from ten laser shots. For the present analysis we selected arbitrarily one fraction from one subject, from which we generated two datasets (see Figure 3.6.): a single spectrum from the chosen subject-fraction (DS_1), and the mean spectrum of all ten replicates (DS_{10}). The higher signal-to-noise ratio of the mean spectrum DS_{10} should make it better for supporting inference [Morris *et al.*, 2005], but the single spectrum of DS_1 offers an opportunity to show how our model accommodates noisy data.

Dataset DS_0 shows a pure matrix solution with no serum sample (it was displayed in Figure 3.3 b.). The 5–75 kDa range of each dataset (TOF 30–280 μs) was standardized as in Equation (3.1). Similar prior distributions (see Section 3.4) were used for the three datasets, with only small differences— for example, we used $\mu_J = 100$ (*i.e.*, $\beta_J = 0.01$) for DS_1 and DS_{10} but reduced it to $\mu_J = 20$ (or $\beta_J = 0.05$) for DS_0 in anticipation of a much smaller number of peaks.

Table 3.1. Table 3.2. and Figure 3.7. display posterior parameter estimates and model fits, respectively, from the last 1 200 000 draws from 4.2 million RJM-CMC iterations. Three model estimates are given in Table 3.2. for the number of proteins found: the posterior mean J^{PM}, the single highest-probability value found in the simulation J^{HP}, and the number J^{∇} of local maxima found in the

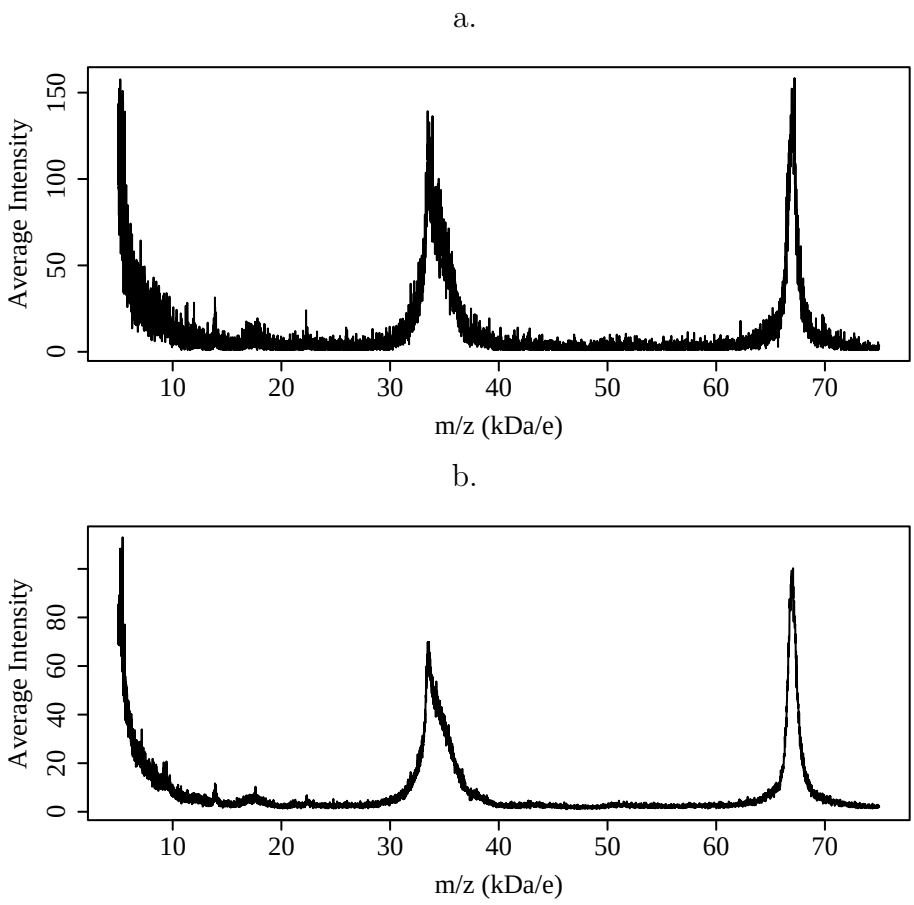


Figure 3.6: Single spectrum (a.) and mean of ten spectra (b.) from a single fraction, single subject. Note noise reduction and peak broadening in (b.)

posterior mean $\mathsf{E}[\mu(t)]$ (see Equation (3.9)). The difference $J^{\text{PM}} - J^{\nabla}$ increases with decreasing resolution as multiple nearby peaks merge [Dass, 2001, *p.* 119]. Such peak merging would be appropriate for proteins with multiple isotopes, but it will lead to distortion when two or more similar-sized but distinct proteins appear as one. Figure 3.8. shows the difference between the two model fits for the mean dataset DS_{10} at different TOF ranges.

Notice that our approach is far more satisfactory than algorithmic feature extraction methods such as that in `PRO`. When strong peaks are present, `PROcess`

Table 3.1: Posterior Model Parameter Estimates (and Standard Deviations).

Dataset	ζ	S	φ	$\varrho^{(M)}$	$\varrho^{(L)}$	η_0	ω_0	J^{PM}
Mean	5.37	0.14	4.95	102.59	108.01	180.80	16.94	154.62
	(0.11)	(0.00)	(0.08)	(1.31)	(1.04)	(4.01)	(0.22)	(7.11)
Single	11.20	0.01	0.33	60.62	95.91	105.63	18.55	103.97
	(0.86)	(0.00)	(0.01)	(5.88)	(2.49)	(8.37)	(0.55)	(7.85)
Matrix	16.84	0.23	1.52	15.32	25.56	34.12	18.40	34.99
	(0.91)	(0.04)	(0.03)	(1.76)	(2.95)	(2.74)	(0.47)	(3.69)

Table 3.2: Number of peak extracted by the model and `Process`. Three model summaries are provided: the posterior mean J^{PM}, the single highest-probability value found in the simulation J^{HP}, and the number J^{∇} of local maxima.

Dataset	J^{PM}	J^{HP}	J^{∇}	J^{PRO}
Mean	154.62	137	77	2
Single	103.97	81	57	2
Matrix	34.99	32	31	66

mistakes smaller peaks as noise; in the absence of large peaks, many noise features are falsely identified as peaks. One reason for this behavior is that local-maxima procedures do not impose uniformity of resolution across the TOF range.

Without knowledge of the true protein distribution in our serum samples, we do not know if we are over- or under- estimating the number of proteins. We do see that the method finds more peaks in the mean spectrum than in the single spectrum, and that it finds about thirty spurious peaks in the matrix spectrum. The spurious peaks mark possible fluctuations in the laser and have extremely low resolutions. Thus, the peaks found in Figure 3.7e. would be avoided in the presence of proteins because they would not comply with the posterior estimate for spectrum resolution.

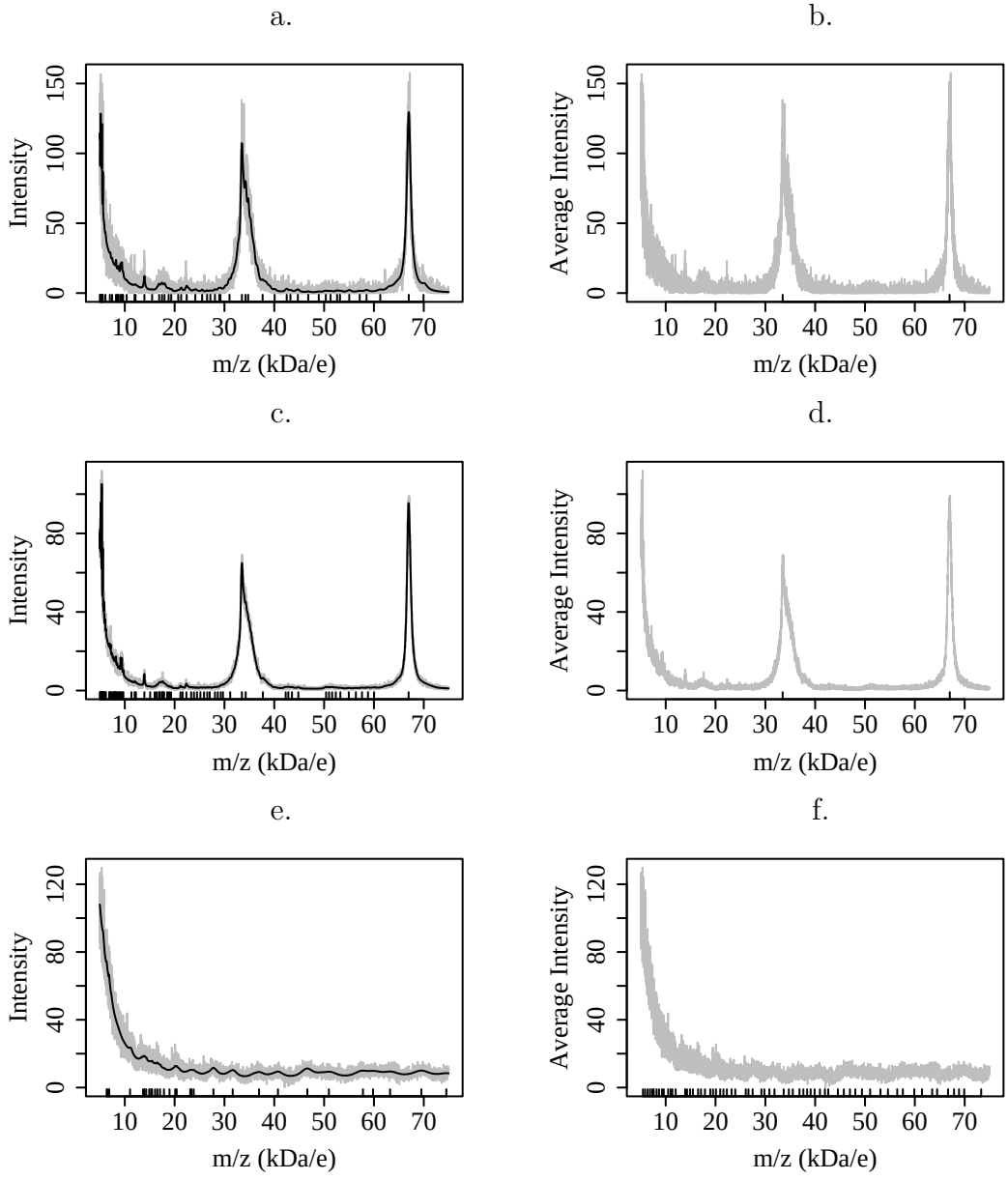


Figure 3.7: Model (left) and `PROcess` (right) peak reconstructions for single-spectrum (a., b.), mean-spectra (c., d.) and matrix-only (e., f.) experiments.

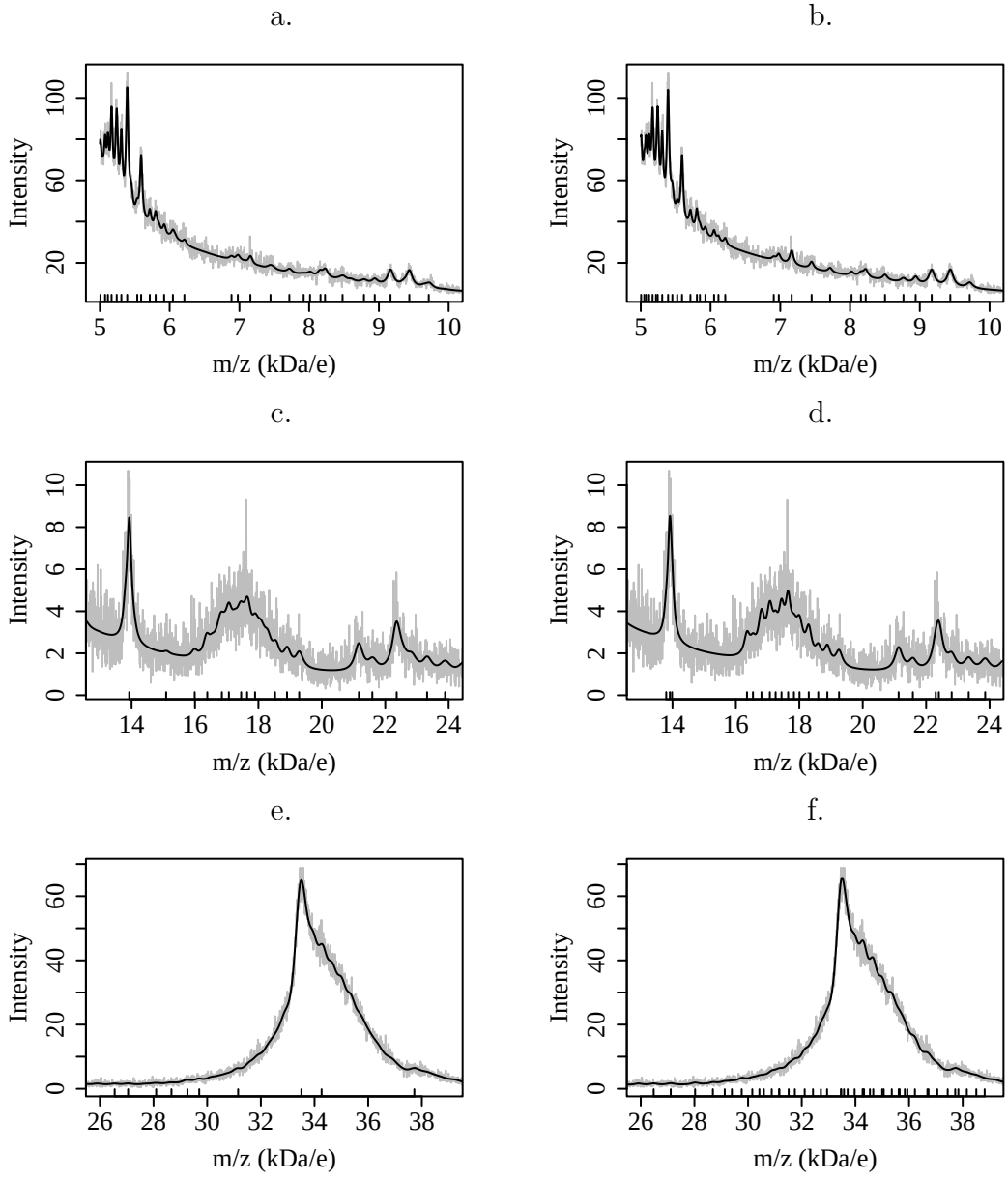


Figure 3.8: Local maxima from posterior mean (left) and highest probability model (right) estimates for various regions in mean spectra.

3.8 Discussion

We have demonstrated through simulation studies and analysis of experimental data that the our novel model based adaptive regression kernel approach provides estimates of parameters of interest and has desirable true and false discovery rates. The incorporation of physical information such as resolution and peak shape can lead to improved methods. Given the posterior samples, we are able to estimate the number of proteins, protein mass, and protein abundance, using the posterior mean or the sample with the highest posterior density. In addition to point estimates, the Bayesian model based approach can also be used to provide measures of uncertainty for any of these quantities.

While this paper describes a model for a single spectrum, modeling multiple spectra simultaneously may be carried by extending the single spectrum model to a hierarchical model to accommodate multiple spectra from different subjects, and multiple subjects within groups. Because of the variability of TOFs across shots within the same subject or experimental conditions, spectra may be slightly misaligned. Averaging spectra across shots for the same subject may lead to the broadening of peaks or possible loss of small peaks. A hierarchical model that allows the TOF parameters to vary from shot to shot, but that are centered at a subject specific expected TOF provides automatic calibration and alignment of spectra. Similar to the functional data analysis approach of Morris *et al.* [2006], the hierarchical version of our single spectrum model can be extended to identity peaks with differential abundance or presence/absence across experimental conditions.

Chapter 4

Functional Data Analysis Using a Lévy Random Field Models for Multi-Spectra Peak Identification and Classification

4.1 Introduction

Expression proteomic studies that use mass spectra and aim to differentiate experimental sub-populations (e.g. drug treatment or disease status groups) via the discovery of protein biomarkers must control for known sources of measurement error. Variation within and between spectra may cause inaccurate multi-spectrum comparisons and mask significant protein differences. Many approaches try to draw comparisons and locate biomarkers by completing three separate steps: 1. per spectrum, identify significant features or peaks that may represent actual proteins; 2. decide if the discovered peaks are present in more than one spectrum, i.e. align the peaks; and 3. compare the features according to a chosen classification scheme [e.g. Morris *et al.*, 2005; Tibshirani *et al.*, 2004; Baggerly *et al.*, 2003]. In this paper, we propose a novel expression proteomic analysis approach that addresses all three steps in a single hierarchical model.

In general, peaks observed within mass spectrometry data represent proteins-the peak areas measure the relative protein concentrations and peak locations on the x-axis identify the proteins by weight. Hence, the first two aforementioned steps, feature extraction and alignment, are necessary so that the third step, sub-population classification, can be based on biologically relevant information [Morris *et al.*, 2005; Baggerly *et al.*, 2004]. In the previous chapter, we addressed feature extraction by modeling nonparametrically Matrix Assisted Laser Desorption Ionization Time of Flight (MALDI-TOF) intensities. This paper extends the previous model to include multiple MALDI-TOF spectra in order to address alignment and classification.

We expand the single-spectrum model (SSM) to multiple spectra by assuming each observed profile, is related hierarchically to one, mean or *population* spectrum. The population peak locations, areas, and resolutions summarize a latent, population protein signal that centers the latent *observed* protein signal in each spectrum and induces a dependence structure between them; i.e. given the population protein signal, subject-specific protein parameters are generated a priori independently to produce another set of latent protein signal estimates which underly all observed spectra. The benefit of assuming this hierarchical dependence structure is two-fold. First, data across several spectra can be used to determine the existence and quantification of protein peaks common to the population; and second, all estimated peaks are aligned with the mean spectrum. The hierarchical approach also assists in classification by labeling each latent population peak with a mark that determines sub-population(s) expression. The peak marks result from assuming that the population protein signal is generated from a three dimensional Gamma random field (GaF) prior, rather than the SSM's two-dimensional

Gamma random field.

Measurement error will cause observed intensities to deviate from the latent, observed protein signal. Thus, we account for the sources of error as well as the protein signal within the likelihood distribution for all molecular intensities as discussed in Section 4.3. Subsequently, in Section s 4.4 and 4.5, we elicit parameter prior distributions and assess the posterior distributions. Sections 4.6 and 4.7 contain respectively summaries of simulation exercises and results using real data which depict the effectiveness of our approach. Before describing our approach in detail however, we motivate our research and exemplify the standard study designs to which our model applies in the next section.

4.2 Motivation

A research team from the Duke Medical Center Radiology Department conducted a retrospective, observational study to explore potential protein differences between lung cancer and control patients. The study included proteomic profiles created by a linear MALDI-TOF Delayed Extraction Mass Spectrometer (Applied Biosystems Voyager DE) of serum samples collected from 30 Caucasian, male patients who were either diagnosed or not diagnosed currently with lung cancer. Since only 2 of the 30 samples exhibited MALDI-TOF saturation or other problems, data from 16 disease patients and 12 control patients were deemed usable; i.e. $n_0 = 12$ and $n_1 = 16$.

For each patient, the serum sample was fractionated using liquid phase isoelectric focusing prior to generating a spectrum [Campa *et al.*, 2003a]. Per fraction, 10 replicate spectra were created from the sum of 10 laser shots. For this effort, we summarize the shots, replicates, and fractions by averaging all of the spectra

from one patient and subtracting the minimum average intensity. Figure 4.1 provides example spectra for three control and three diseased subjects and ; Figure 4.2 shows a heat-map of all 28 spectra used in subsequent analyses.

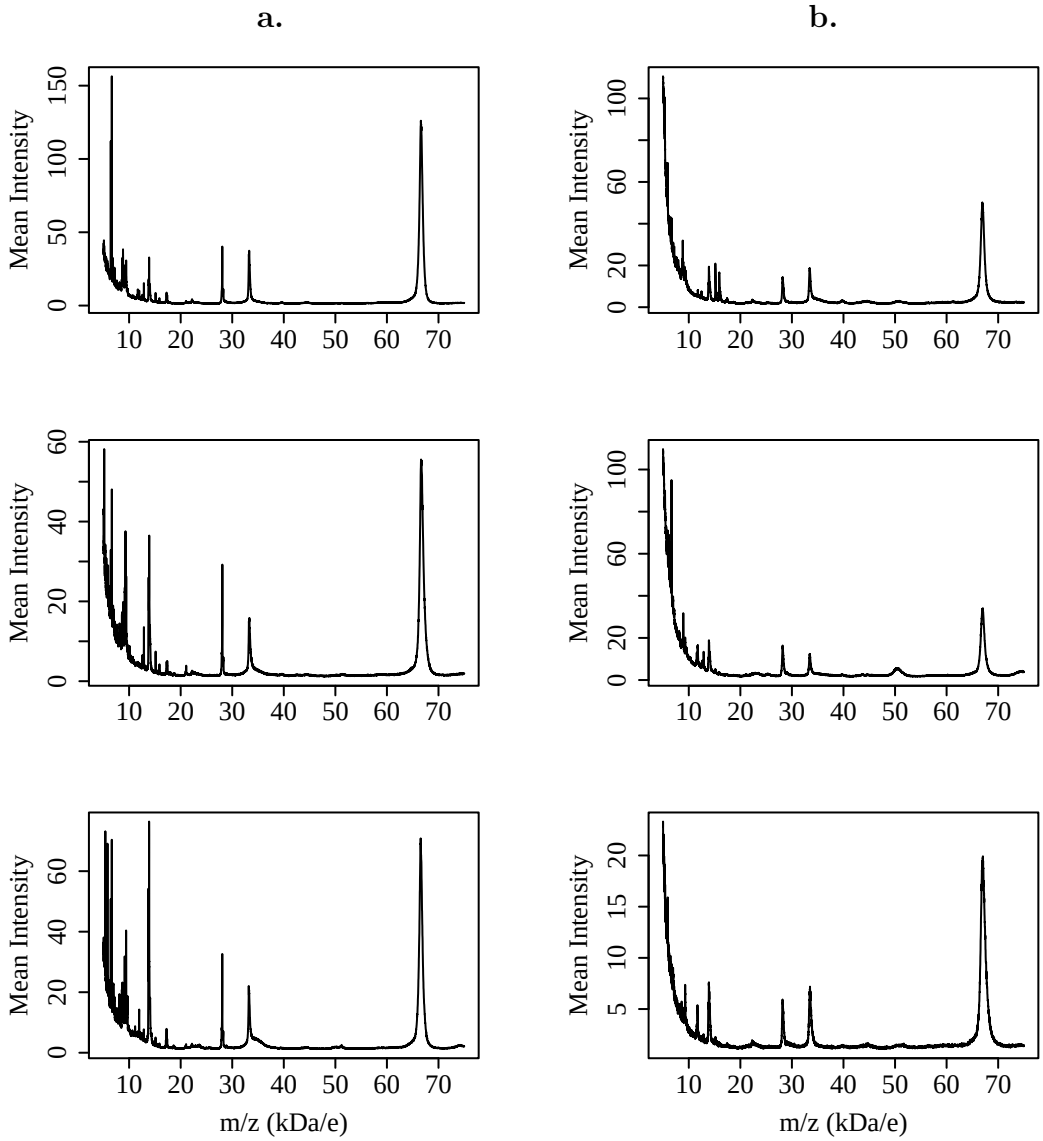


Figure 4.1: Column a. represents three example spectra from the control group and column b. represents three example spectra from the disease group. Each spectrum represents one person and is an average across 10 laser shots, 10 replicates, and 20 fractions.

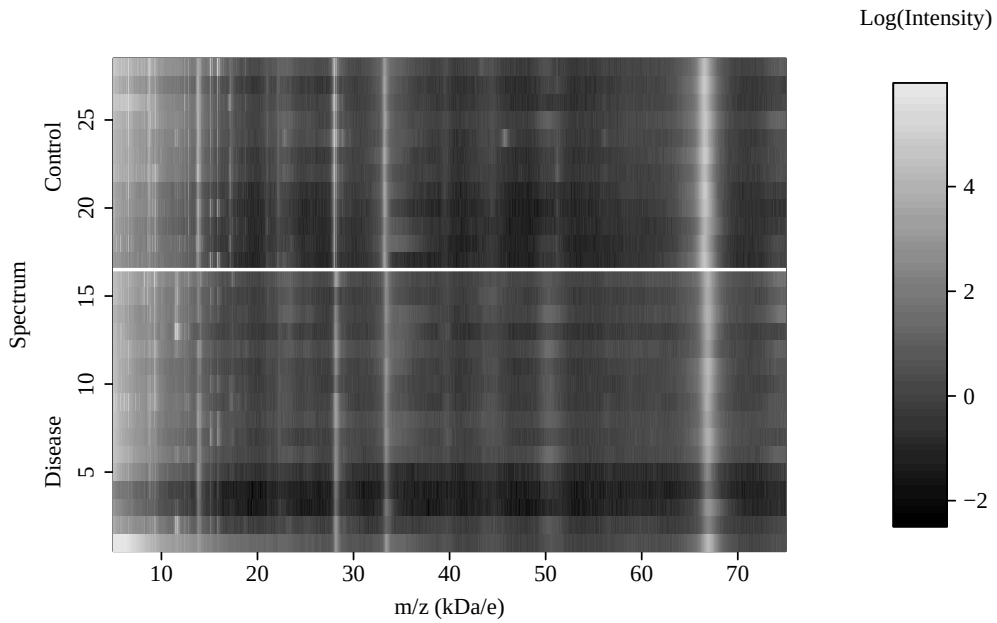


Figure 4.2: A heat-map of the log intensities analyzed in Section 4.7. Low to high intensities are represented by dark to light gray colors. The bottom 16 spectra refer to disease spectra, whereas the top 12 rows refer to control spectra.

In Figures 4.1 and 4.2, control and disease spectra peak intensities generally differ, and some peaks are found in one treatment group, but not the other. For this effort (and comparable expression proteomic studies), the primary question of interest is whether or not these differences relate significantly to disease status or result merely from measurement error and/or subject variation. In the following sections, we address the research question with an approach that models all spectra with measurement uncertainty and protein signal parameters.

4.3 Likelihood

For reasons described in (Dass 2001, p.g. 74; House *et al.* 2006b; Morris *et al.* 2005), we model mass spectra on the time, rather than mass, scale. Thus, let

$Y_{it} = Y_{it}^{\mathsf{ob}} + s$ where s is a small positive constant and Y_{it}^{ob} represents the mean intensity, as plotted in Figure 4.1, for person i at time t where $i \in [1, ..., n]$, $T_0 = 32\,\mu\text{s}$, $T_0 = 278\,\mu\text{s}$, $t \in [T_0, T_1]$, $T_1 - T_0 = T$, and N is the number of observations per spectrum. Each person i has a predetermined treatment group status d_i which may take one of two values, either "C" for control or "D" for disease. Given d_i, the expected value for Y_{it} is a function of data uncertainties inherent within MALDI-TOF and true, latent protein signal.

4.3.1 Protein Signal: Adaptive Kernel Regression

We use a kernel regression approach to model the protein signal $f_i(t, d)$ for an observed spectrum i,

$$f_i(t, d) \;=\; \sum_{j=1}^{J} \eta_{ij} k(t, \tau_{ij}, \omega_{ij}, m_{ij})$$

where the function $k()$ specifies the shape of a single isotopic peak with parameters τ_{ij}, ω_{ij}, and mij. Based on House *et al.* [2006b], we set $k()$ to a normalized Cauchy density function

$$f_i(t, d) = \sum_{j=1}^{J} \eta_{ij} \delta_1(m_{ij}) \frac{\omega_{ij}^2}{\pi(\omega_{ij}^2 + |t - \tau_{ij}|^2)} \tag{4.1}$$

and consider the unknowns, J, and $1 \times J$ vectors $\boldsymbol{\tau_i}$, $\boldsymbol{\omega_i}$, $\boldsymbol{m_i}$, and $\boldsymbol{\eta_i}$, to be the protein signal parameters. The signal parameters, in the context of function estimation, represent respectively the number, location, width, area, and existence of spectrum peaks. However, they each have biological interpretations as well. The combination of the mark and concentration vectors $\mathbf{1}_{\boldsymbol{m_i}=\mathbf{1}_J}\boldsymbol{\eta_i}$ estimates the

protein concentrations for person i given d_i; $\boldsymbol{\tau_i}$ maps to the masses of proteins; and $\boldsymbol{\omega_i}$ is a function of peak resolutions $\boldsymbol{\rho_i}$ [Dass, 2001, p.g. 74]. As a result, the signal parameterization of our model will work to identify, quantify, and determine the expression of proteins in sample i.

Typically, the quantification and precision of protein expression measurements depend upon the spectrum resolution ρ. Hence, we incorporate ρ within our estimate of protein signal via the width parameter $\boldsymbol{\omega}$. The parameter $\boldsymbol{\omega}$ is specified deterministically as a function of resolution that is calculated from the "full width half mass" (FWHM) procedure [e.g Dass, 2001, p.g. 74]. The FWHM procedure calculates resolution ρ for a a given peak by dividing its location (expected TOF) τ by one half its width $\Delta\tau$ at half the peak's height,

$$\rho \equiv \frac{\tau}{\Delta\tau}.$$

Within the context of our kernel regression, the above translates to

$$\begin{aligned}\eta_{ij}k(\tau_{ij},\tau_{ij},\omega_{ij}) &= 0.5\eta_{ij}k(\tau_{ij}+\frac{\Delta\tau_{ij}}{2},\tau_{ij},\omega_{ij}) \\ &= 0.5\eta_{ij}k(\tau_{ij}+\frac{\rho_{ij}}{2\tau_{ij}},\tau_{ij},\omega_{ij}),\end{aligned}$$

and given a kernel function, provides a means to solve for ω_{ij} as a function of τ_{ij} and ρ_j. With a Cauchy kernel function, the solution for ω_{ij} is

$$\omega_{ij} = \Big(\frac{\tau_{ij}}{2\rho_{ij}}\Big). \tag{4.2}$$

The resolution of a spectrum offers a numerical summary of its quality. The summary is important because the MALDI-TOF proteomic analysis technique is extremely sensitive and will inevitably induce both random and systematic

measurement error. The next section describes how we incorporate some of the MALDI-TOF uncertainties into the modeled expected intensity.

4.3.2 Expected Intensity

Inconsistencies in sample preparation, laser intensity, spatial matrix-molecule coordinates, and environmental variables cause positive and negative measurement errors on both protein intensities and masses; i.e. observed spectra are vertical and/or horizontal translated versions of true protein signal. To isolate the protein signal from the altered data, researchers frequently use deterministic data cleaning procedures. Our approach however, models the sources of error in the expected intensity $\mu_i(t, d)$.

Similar to the SSM, we set $\mu_i(t, d)$ to a scaled sum of detector noise and molecular signal,

$$\mu_i(t, d) = \zeta_i \Big\{ S_i \big[f_i(t, d) + \beta_{1i}(t) \big] + (1 - S_i) \Big\}.$$

where for person i, ζ_i, S_i, and $\beta_{1i}(t)$ are scaling and background parameters. Specifically, ζ_i is a scaling constant; $1 - S_i$ is between zero and one and reflects the average proportion of each intensity measurement output from detector ringing; and, $\beta_{1i}(t)$ is an exponentially decaying background level [Coombes *et al.*, 2005a]. The parameter ζ_i vertically stretches or shrinks an estimated signal while $\zeta_i S_i$ lifts the signal by a constant. The parameter $\beta_{1i}(t)$ also lifts the estimated intensities, but by a TOF-dependent quantity. Low TOFs tend to have higher intensities than large TOFs because the initial shock of the ionizing laser causes several small molecules to break from the matrix, "fly" quickly to the detector, and interfere with the abundance measures of comparable size, protein molecules [Dass, 2001].

The interference seems to decay exponentially [Coombes *et al.*, 2005a].

We model $\beta_{1i}(t)$ within the aforementioned adaptive kernel framework by estimating a *first-peak*, an imposed peak at the first, observed TOF. Let

$$\beta_{1i}(t) = \eta_{i0} k_0(t, \tau_0, \omega_{i0}) = \frac{\eta_{i0}}{\omega_{i0}} \exp(-\frac{1}{\omega_{i0}}|t - \tau_0|), \tag{4.3}$$

where η_{i0} represents the concentration of all matrix molecules, ω_{i0} is the scale parameter for the exponential decay, $\tau_0 = T_0$, and $k_0()$ is the first-peak kernel. The difference between modeling the first-peak and an actual protein peak is the pre-determined peak location τ_0 for all spectra and the specification of kernel $k_0()$. Rather than specify a Cauchy kernel for $k_0()$, we set it to the exponential density function.

Given the definitions for $\beta_{1i}(t)$ and protein signal, we may rewrite the expected intensity, Equation 4.3, as

$$\mu_i(t, d) = \zeta_i \Big\{ S_i \big[\eta_{ij} \delta_1(m_{ij}) k(t, \tau_{ij}, \omega_{ij}) + \eta_{i0} k_0(t, \tau_0, \omega_{i0}) \big] + (1 - S_i) \Big\}. \tag{4.4}$$

All of the protein and uncertainty parameters are spectrum-specific in order to account for the known, sensitive nature of mass spectrometry. In the following section, we decide to model a spectrum-specific relationship between the expected mean $\mu_i(t, d)$ and the variance of Y_{it} as well.

4.3.3 Intensity Error Model

In developing the SSM, we found that an approximate linear relationship exists between the expected mean and variance of intensities, $\mathrm{E}[variance] = \frac{1}{\varphi} mean$ (refer to section). To investigate whether the steepness of the linear slope changes between spectra (Figure 4.3) we added an interaction term to the model of *variance*

that multiples a factor variable, the spectrum identification number (id), and $mean$. Since the interaction was significant (p$-$value < 0.001), we let Y_{it} be Gamma distributed with shape and rate parameters set respectively to $\varphi_i\mu_i(t,d)$ and φ_i. Assuming Y_{it} are conditionally independent, we have the following likelihood distribution,

$$\mathcal{L}(\boldsymbol{Y}|\boldsymbol{\mu},\boldsymbol{\varphi},\boldsymbol{d}) = \prod_{i=0}^{n}\prod_{t=T_0}^{T_1} \mathrm{Ga}(Y_{it};\mu_i(t,d)\phi_i,\phi_i,0) \tag{4.5}$$

where $\mathrm{Ga}(x;a,b,e) = \frac{b^a}{\gamma(a)}x^{a-1}\exp(-xb)\mathbf{1}_{x>e}$.

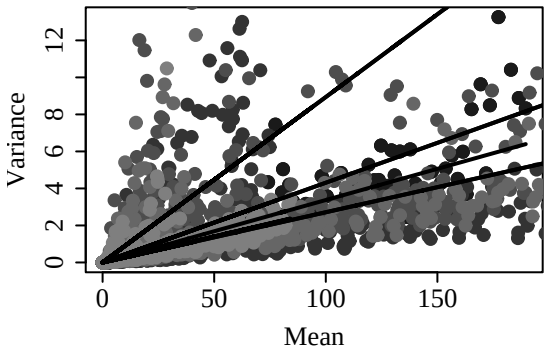


Figure 4.3: From five diseased patients, the sliding means and variances per spectrum are plotted and regressed: $variance = mean + mean \times id$.

4.4 Prior Specification

We relate hierarchically each observed spectrum to one latent, mean spectrum; each spectrum-specific, or stage-one parameter, is independent a priori of all other parameters and depends upon a corresponding latent, mean parameter; e.g. $\pi(\zeta_i|\zeta) = f(\zeta)$. The latent spectrum includes the expected protein signal $f(t)$, regardless of sub-population status, for the entire sample population as well as the

population MALDI-TOF uncertainties, $\{\zeta,\beta_0(t),S\}$. Since the mean parameters describe one spectrum, we may estimate them via the SSM, with one primary difference - the incorporation of a sub-population attribute variable m_j.

4.4.1 Latent Mean Spectrum

The functional form and parameter definitions of the latent spectrum $\mu(t)$ are identical to that of an individual spectra,

$$\mu(t) \;=\; \zeta\Big\{S\big[f(t)+\beta_0(t))\big]+(1-S)\Big\}, \tag{4.6}$$

where $f(t)$ represents the mean protein signal,

$$f(t) \;=\; \sum_{j=1}^{J}\eta_j k(t,\tau_j,\omega_j,m_j).$$

However, the prior for $f(t)$ changes. The SSM specifies a joint prior distribution on the protein parameters, J and the $1\times J$ vectors for peak locations $\boldsymbol{\tau}$, $\boldsymbol{\eta}$, and $\boldsymbol{\rho}$ by assuming $f(t)$ is a random variable and generates from a Lévy random field (Lv) prior. Specifically, a Gamma random field prior is used with shape and rate parameters $\alpha(dt)$ and λ,

$$\Gamma \sim \mathrm{GaF}(\alpha(dt),\lambda)$$

and assigns random measures to the space of the kernel parameters equal to the sum of random jump heights at random locations in the $[T_0,T_1]\times\mathbb{R}^+$ plane,

$$\Gamma(d\tau,d\rho)=\sum_{j=1}^{J}\eta_j\delta_{\tau_j}(d\tau)\delta_{\rho_j}(d\rho)$$

[Wolpert and Ickstadt, 1998b,a]. By expanding this plane to a hyperplane,

$$\Gamma(d\tau, d\rho, dm) = \sum_{j=1}^{J} \eta_j \delta_{\tau_j}(d\tau)\delta_{\rho_j}(d\rho)\delta_{m_j}(dm)$$

we change the SSM to a mean spectrum model.

The additional dimension refers to sub-population(s) association and effectively marks or labels each peak in the Gamma random field to be in one of three categories: observed only in control patients (C), observed only in disease patients (D) or shared (S) by both patient groups. We depict the incorporation of the $m-$dimension (while holding resolution constant) in Figure 4.4. The mark parameter is useful for determining the group in which each individual protein is expressed as well as grouping the proteins to determine the latent signal for each subpopulation. For example, all of the peaks marked by either C or S work to summarize the control population's expected protein signal, $f(t)|m_i \in [C, S]$ (Figure 4.5).

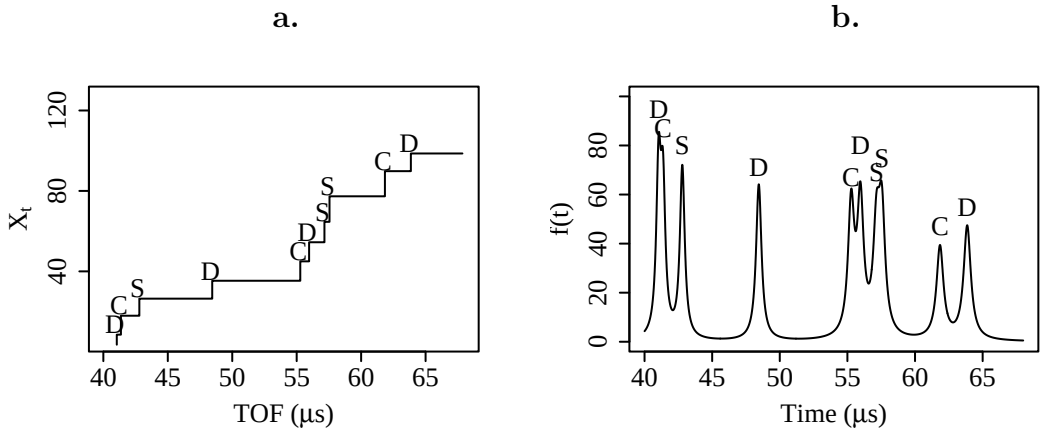


Figure 4.4: By marking each jump of the Gamma random field in plot a., the latent signal indicates the treatment groups state(s) in which each protein is expressed.

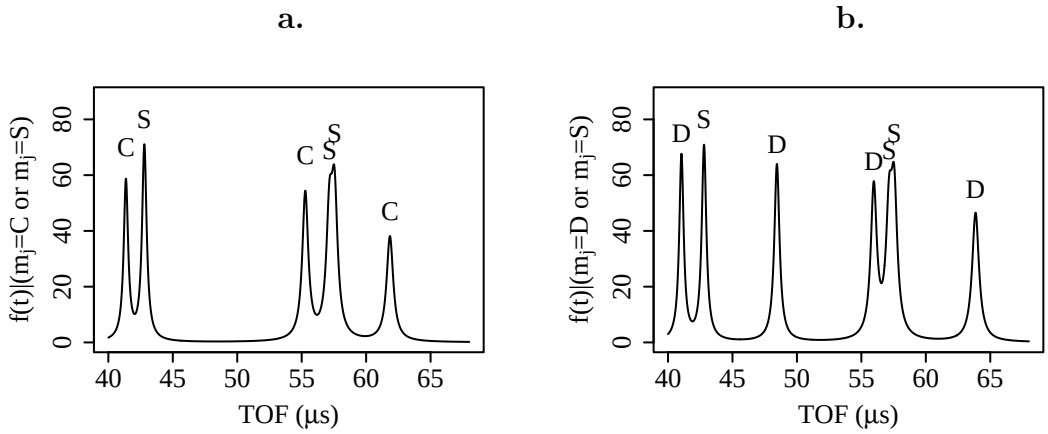


Figure 4.5: Thinning the Gamma random field displayed in Figure 4.4, results in two latent signals, one for the control group (plot a.) and one for the disease group (plot b.).

Provided a finite measure on m and the remaining kernel parameters, the Lévy measure for the three dimensional Gamma random field is

$$\nu(d\eta, d\tau, d\rho, dm) = \mathbf{1}_{\eta>0}\alpha\eta^{-1}e^{-\eta\lambda}d\eta\pi(d\tau, d\rho, dm)$$

which satisfies,

$$\int_{\mathbb{R}}(1 \wedge |\eta|)\nu(\eta) < \infty. \tag{4.7}$$

However, $\nu(d\eta, d\tau, d\rho, dm)$ is not a finite measure- without constraints, the number of jumps in a field will be infinite while still meeting the bound stated in Equation 4.7. However, in constraining η to be greater than a predetermined ϵ, the Lévy measure is approximated as

$$\nu_\epsilon(d\eta, d\tau, d\rho, dm) = \mathbf{1}_{\eta>\epsilon}\alpha\eta^{-1}e^{-\eta\lambda}d\eta\pi(d\tau, d\rho, dm), \tag{4.8}$$

and the expected value of J becomes finite.

Given the approximate Lévy measure for a Gamma random field, the peak parameter prior distributions follow naturally. Given J, Equation 4.8 presents a factorisable prior implying that individual peak concentrations η_j is proportional to

$$\begin{aligned} \pi(\eta_j|\alpha,\lambda) &\propto \mathbf{1}_{\eta_j>\epsilon}\eta_j^{-1}e^{\lambda\eta_j} && (4.9)\\ &= \mathrm{Ga}(0,\lambda,\epsilon), && (4.10)\end{aligned}$$

($\mathrm{Ga}(a,b,\epsilon) = b^a/\Gamma(a)x^{a-1}\exp(-bx)\mathbf{1}_{x>\epsilon}$) and independent a priori from the parameters $\boldsymbol{\tau}$, $\boldsymbol{\rho}$, m. For $\boldsymbol{\tau}$, we do not use prior information concerning expected proteins within a sample, thus we let each τ_j be uniform over the interval $[T_0, T_1]$,

$$\pi(\tau_j) = \mathbf{1}_{[T_0,T_1]}(t),$$

where $T_1 - T_0 = T$. For peak resolutions ρ_j, we suggest an informative, LogNormal prior based on an overall estimate of spectrum resolution, ϱ.

Typically, one or a small number of resolutions are reported per spectrum to reflect the quality of the data. However, due to the sensitive nature of MALDI-TOF, individual peak resolutions will deviate slightly from the reported value [Coombes *et al.*, 2005a]. Hence, we assume a conjugate LogNormal hierarchal model, where the individual peak resolutions have approximately 0.95 prior mass within $\pm 10\%$ of the estimated spectrum resolution, the variance for the overall spectrum resolution scales according to the n, and ϱ_μ is based on preliminary data investigations

$$\begin{aligned} \rho_j &\sim \mathrm{LogNo}(\log(\varrho),(.05)^2)\\ \varrho &\sim \mathrm{LogNo}(\log(\varrho_\mu), n(.05)^2)\end{aligned}$$

We let ϱ_μ depend upon the size of the molecules, and specify two values $\varrho_\mu^{(M)}$ and $\varrho_\mu^{(L)}$ for medium and large size molecules

$$\varrho_\mu \;=\; \begin{cases} \varrho_\mu^{(M)} = 500, & \text{if m/z} \in [5000, 25000) \\ \varrho_\mu^{(L)} = 100, & \text{if m/z} \in [25000, \infty) \end{cases}$$

The remaining parameter in the population protein signal for which we elicit a prior distribution is J. A Gamma random field (or any pure jump process) is a generalized version of a multi-dimensional Poission process. In our application, the elicited Gamma process is comparable to jumps generating from a Poission point process with intensity $\alpha T \mathrm{E}_1(\epsilon\lambda)$ because

$$\iiint \nu_\epsilon(d\eta, d\tau, d\rho, dm) = \mathbf{1}_{\eta>\epsilon} \alpha \eta^{-1} e^{-\eta\lambda} d\eta \pi(d\tau, d\rho, dm) = \alpha T \mathrm{E}_1(\epsilon\lambda)$$

where $\mathrm{E}_1()$ is the exponential integral function [Abramowitz and Stegun, 1964, *p.* 228]. However, we elicit a Gamma prior distribution for α

$$\alpha \sim \mathrm{Ga}(\alpha_a, \alpha_b)$$

which results in a Negative Binomial marginal prior distribution for J,

$$J|\alpha_a, \alpha_b, \lambda \sim \mathrm{NB}\Big(\alpha_a, \frac{\alpha_b}{T\mathrm{E}_1(\epsilon\lambda) + \alpha_b}\Big),$$

with expected value,

$$\mathrm{E}[J|\alpha_a, \alpha_b, \lambda] = \frac{\alpha_a T \mathrm{E}_1(\epsilon\lambda)}{\alpha_b}. \tag{4.11}$$

The value for ϵ is considered a minimum detection level, where $100(\epsilon/E[\eta])$ represents a minimum percent concentration level of the average signal.

To complete the latent mean spectrum model in Equation 4.6, we elicit prior distributions for the population parameters: ζ, S and $\beta_0(t)$. For ζ, a LogNormal prior is specified so that its location parameter equals the mean of all spectra and the geometric standard deviation approximately equals to 0.10;

$$\zeta \sim \text{LN}(\log\left(log(\mu_\zeta), \sigma_\zeta^2\right),$$

($\text{LN}(\mu, \sigma) = 1/(x\sigma\sqrt{2\pi})\exp\{-(\log(x)-\mu)^2/(2\sigma^2)\}$) where, $\mu_\zeta = \frac{\sum_i \overline{\boldsymbol{Y}_i}}{n}$ and $\sigma_\zeta = 0.10\mu_\zeta$. For S and $\beta_0(t)$, the distributions are similar to the SSM in House *et al.* [2006b],

$$\begin{aligned} S &\sim \text{Be}(S_a, S_b), \text{ s.t. } \text{E}[S] = .23 \text{ and } \text{V}[S] = 0.13 \\ \beta_0(t) &= \eta_0 k_0(t, \tau_0, \omega_0) \\ &\quad \eta_0 \sim \text{Ga}(0, \lambda_0, \epsilon) \\ &\quad \omega_0 \sim \text{LN}(3.012, 0.25^2), \end{aligned}$$

($\text{Be}(a, b) = \Gamma(a+b)/(\Gamma(a)\Gamma(b))x^{a-1}(1-x)^{b-1}$) with the exception of ω_0. Preliminary exponential model fits from House *et al.* [2006a] suggest that ω_0 should be centered around 11.5. Thus we place 0.95 prior mass for ω_0 between approximately 7 and 19.

Depending on the latent mean spectrum parameters, the first stage priors are centered accordingly and have either Normal or LogNormal distributions. The next section will summarize the hyper-parameter specifications for the first stage priors and conclude with the priors of the remaining parameters.

4.4.2 Remaining Prior Distributions

Two parameters in the likelihood are deterministic from the mean spectrum. The first, as previously discussed, is ω_{ij} which is a function of τ_{ij} and ρ_j; the second, is m_{ij} which equals 1 when m_j indicates that peak j is either shared or in disease group d_i,

$$m_{ij}|m_j, d_i = \delta_{[S,d_i]}(m_j).$$

With the exception of the two deterministic parameters, we elicit the remaining prior distributions while considering expert opinion, preliminary analyses using SSM, and sample size.

The spectrum-specific parameters for peak location and concentration exist only when $m_{ij} = 1$, thus for η_{ij} and τ_{ij} point-mass priors are specified,

$$\begin{aligned} \tau_{ij}|\tau_j, m_{ij} &\sim \delta_1(m_{ij})\text{LN}(\log(\tau_j), 0.002^2) \\ \eta_{ij}|\eta_j, m_{ij} &\sim \delta_1(m_{ij})\mathbf{1}_{\eta_{ij}>\epsilon}\text{LN}(\log(\eta_j), (0.11)^2). \end{aligned}$$

For $\tau_{ij}|\tau_j, m_{ij} = 1$ we elicit a LogNormal prior with $\sigma = 0.002$ because expert opinion suggests that a peak location may shift among spectra by approximately $0.15 - 0.30\%$ its mass [Campa *et al.*, 2003a]. For $\eta_{ij}|\eta_j, m_{ij} = 1$, we specify a prior distribution with the same support as η_j and determine the scale hyperparameter from preliminary SSM analyses. We implemented the SSM for all 28 spectra from one fraction (averaged across replicates) and store the model sampled with highest posterior probability. Per maximum probability model i, $i \in [1, ..., 28]$, the mean $\overline{\eta}_i$ and standard deviation s_i of peak concentrations was calculated and plotted in Figure 4.6 while noting disease state d_i. From the regression results of $s_i = \overline{\eta}_i + d_i + \overline{\eta}_i \times d_i$, we learn that d and the interaction term are insignificant. Thus, in removing those terms, $\text{E}[s_i] = 0.11\overline{\eta_i}$ or that that the geometric standard

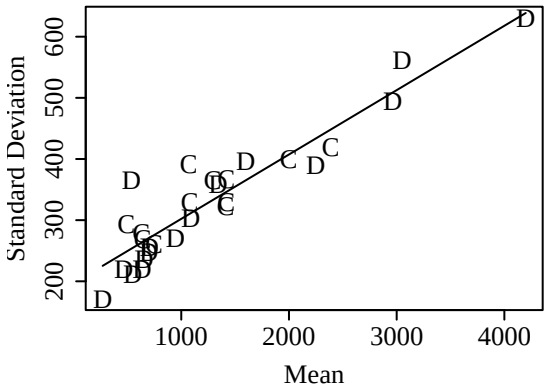


Figure 4.6: Mean and standard deviations for posterior mode estimates of protein concentration from SSM model using data from fraction 5 described in Section 4.2. The SSM analyzed spectra from both the control (C) and disease (D) sub-populations.

deviation is approximately .11. Thus, we set the scale parameter for the truncated LogNormal prior distribution of η_i to 0.11^2.

From the same SSM analyses, we estimate the scale parameter for η_{i0}. We set the scale parameter of a LogNormal prior to the standard deviation of the posterior mode estimates for the first peak concentrations,

$$\eta_{i0} \sim \text{LN}(\eta_0, 0.3).$$

Normal and LogNormal prior distributions are also assigned to the remaining spectrum specific MALDI-TOF uncertainty parameters,

$$\begin{aligned} \zeta_i|\zeta &\sim \text{LN}(\log(\zeta), 0.5^2) \\ S_i|S &\sim \text{No}(S, 0.013/n) \\ \omega_{i0} &\sim \text{LN}(\omega_0, 0.25^2/n). \end{aligned} \tag{4.12}$$

With the exception of ζ, the variance hyperparameters in Equation 4.12 differ from the respective mean-spectrum, parameter variances by a factor of $1/n$. Although, the mean parameter prior distributions are not necessarily conjugate to Normal and LogNormal distributions, they differ in scale by approximately n times which is similar to specifying to g-priors where $g = n$ [Zellner, 1986; Kass and Wasserman, 1995]. For the same reason, the prior distributions for $\{\varphi_i, \varphi\}$ and the third stage parameter ϱ depends on n

$$\begin{aligned} \varphi_i|\varphi &\sim \text{No}(\varphi, 0.5^2/n) \\ \varphi &\sim \text{Ga}(0.5, 1). \end{aligned}$$

The conclude this section, we specify a prior distribution for $[\pi_C, \pi_D, \pi_S]'$. We use a conjugate Dirichlet prior to m_j,

$$[\pi_C, \pi_D, \pi_S]' \sim \text{Di}(1, 1, 1).$$

Table 4.1 provides a comprehensive listing of all prior specifications. The hierarchical approach allows both spectra alignment and the identification of potentially classifying proteins to result naturally from estimating the model parameters. All extracted features from individual spectra $\boldsymbol{\tau_i}$ align with the mean spectrum features $\boldsymbol{\tau}$ and the marked protein concentrations indicate differences in sub-population proteomic profiles. Using the differences and Bayes rule, we may classify future patients from the posterior analysis of the model.

4.5 Posterior Analysis

We implement a reversible jump Markov chain Monte Carlo (RJMCMC) algorithm [Green, 1995] to estimate the posterior distributions for the collection of all model

parameters, $\boldsymbol{\theta}$. The RJMCMC algorithm is identical to that explained in House *et al.* [2006b] with the addition of either Metropolis-Hasting (M-H) or Gibbs steps to sample from the full conditional distributions of the spectrum specific and the sub-population indicator variables.

For cross validation, we use subset of the data $\boldsymbol{Y_i}|\boldsymbol{d_i}$ for $i \in [1, ..., n]$ to estimate the model parameters so that we may predict the sub-population status d_i^* of the remaining profiles $\boldsymbol{Y_i^*}$ for $i \in 1, .., n^*]$. The predictions are based on the posterior density, $\mathrm{P}[d_i^*|\boldsymbol{Y_i^*}, \boldsymbol{Y}, \boldsymbol{d}]$, which we obtain from the stochastic integration,

$$\pi(\boldsymbol{Y_i^*}, |d_i^* = D, \boldsymbol{Y}, \boldsymbol{d}) = \iint \pi(\boldsymbol{Y_i^*}, \boldsymbol{\theta_i^*}, \boldsymbol{\theta}|d_i^* = D, \boldsymbol{Y}, \boldsymbol{d}) d\boldsymbol{\theta^*} d\boldsymbol{\theta}$$

and applying Bayes Rule,

$$\mathrm{P}[d_i^*|\boldsymbol{Y_i^*}, \boldsymbol{Y}, \boldsymbol{d}] = \frac{\pi(Y_i^*|d_i^* = D, \boldsymbol{Y}, \boldsymbol{d})\pi(d_i^* = D|\boldsymbol{Y}, \boldsymbol{d})}{\sum_{s \in C,D} \pi(Y_i^*|d_i^* = s, \boldsymbol{Y}, \boldsymbol{d})\pi(d_i^* = s|\boldsymbol{Y}, \boldsymbol{d})} \tag{4.13}$$

[Gilks *et al.*, 2000]. Equation 4.13 simplifies to

$$\mathrm{P}[d_i^*|\boldsymbol{Y_i^*}, \boldsymbol{Y}, \boldsymbol{d}] = \frac{\pi(Y_i^*|d_i^* = D, \boldsymbol{Y}, \boldsymbol{d})}{\sum_{s \in C,D} \pi(Y_i^*|d_i^* = s, \boldsymbol{Y}, \boldsymbol{d})}$$

because d_i^* is independent of $\{\boldsymbol{Y}, \boldsymbol{d}\}$ ($\pi(d_i^* = D|\boldsymbol{Y}, \boldsymbol{d}) = \pi(d_i^* = D)$) and we set $\pi(d_i^* = D)$ to equal 0.5.

4.6 Simulation Data

Before applying our approach to real data, we analyze in this section three simulated datasets that contain sub-populations with varying degrees of distinction; datasets DS_0 DS_{30} and DS_{100} include a total of 35 peaks with disease populations that share either 0%, 30%, or 100% of the peaks in the control populations.

Given the peak locations and sub-population assignments, each of the simulated datasets were generated with the following population parameters: $T_0 = 32\,\mu$s, $T_1 = 258\,\mu$s, $l = 10$, $n_0 = 3$, $n_1 = 3$, $M = 35$, $\zeta = 13$, $\varphi = 0.5$, $S = .85$, $\alpha_a = 1$, $\alpha_b = .02$, $\varrho = 56$, $\boldsymbol{\eta} = 7 \times \mathbf{1}_{35}$, $\eta_0 = 97$, and $\omega_0 = 11.5$. We also generate $n^* = 6$ spectra for prediction.

Similar to House *et al.* [2006b], we started the RJMCMC chain with peaks in informative locations, assumed initially that the expected value for E[J]=50, and used only one overall resolution parameter, $\varrho_\mu^{(M)} = \varrho_\mu^{(L)} = 50$. We also changed the scaling parameters for η_{ij} and η_{i0} to 0.25 and 0.025 respectively in order to achieve acceptable M-H acceptance rates. For the last 5000 of 120000 iterations, we summarize the model with the highest posterior probability ($^{\text{HP}}$) and calculate the model average ($^{\nabla}$). The model average peaks locations are the critical points where first derivative of the model average equals zero and the second derivative is less than zero. The model fits resulted in high (greater then 0.80) True Discovery Rates (TDR) of the known peaks and spectra: a true-peak at location τ_{true} is discovered when there exists at least one mean spectrum, estimated peak $\hat{\tau}_j$ within the range $\tau_{true} \pm 0.003\tau_{true}$. Table 4.2 summarizes by dataset the TDR and False Discovery Rates (FDR).

The simulation exercise included the prediction of 6 spectra: 2 control, 2 with the same percent of shared peaks as the modeled data, and 2 from the disease population that either shared 5%, 60%, or 100% of the control peaks. For 100% of the cases, the spectra were accurately classified.Additionally, we see successful separation of control only, disease only, and shared signal in Figures 4.8–4.12; these figures display, per dataset, the model average estimates of the protein signal when $m_j = C$, $m_j = D$, and $m_j = S$. In conjunction with the figures, Table 4.3

documents the number of estimated peak and the number of peaks we should see per figure. Even though not all of the disease-only peaks seem to have been found in datasets DS_{30} and DS_{100} we see from Table 4.2 that the true discovery rates are adequately high. Thus, the loss of some peak identification is due to the effects of model averaging.

4.7 Real Data Application

The proposed model is applied to the real data mentioned in Section 4.2. Twenty-two ($n_0 = 9$ and $n_1 = 13$) of the 28 spectra were fit by the multi-spectra model so that we could make predictions with the remaining six. We started the RJMCMC algorithm with informative peak locations provided by the EM algorithm from House *et al.* [2006c]. Table 4.4 and Figures 4.11–4.13 summarize the last 5000 draws of the RJMCMC.

Table 4.4 provides the posterior modes for the MALDI-TOF uncertainty parameters, resolution, and the number of peaks: approximately 0.80 of an observed intensity is generated from the protein signal; the mean-variance relationship is 0.1; and the model size J^{PM} (posterior mode) is 97. Additionally, Table 4.4 includes the number of peaks within the model of the highest posterior probability (J^{HP}) and the model average (J^{∇}) which are 98 and 78 respectively. Figure 4.11 displays the model average fit for the same data displayed in section Section 4.2.

Figure 4.12 displays the latent protein signal, decomposed using the three marks, that underlies Figure 4.11; and Table 4.5 provides the posterior probabilities for

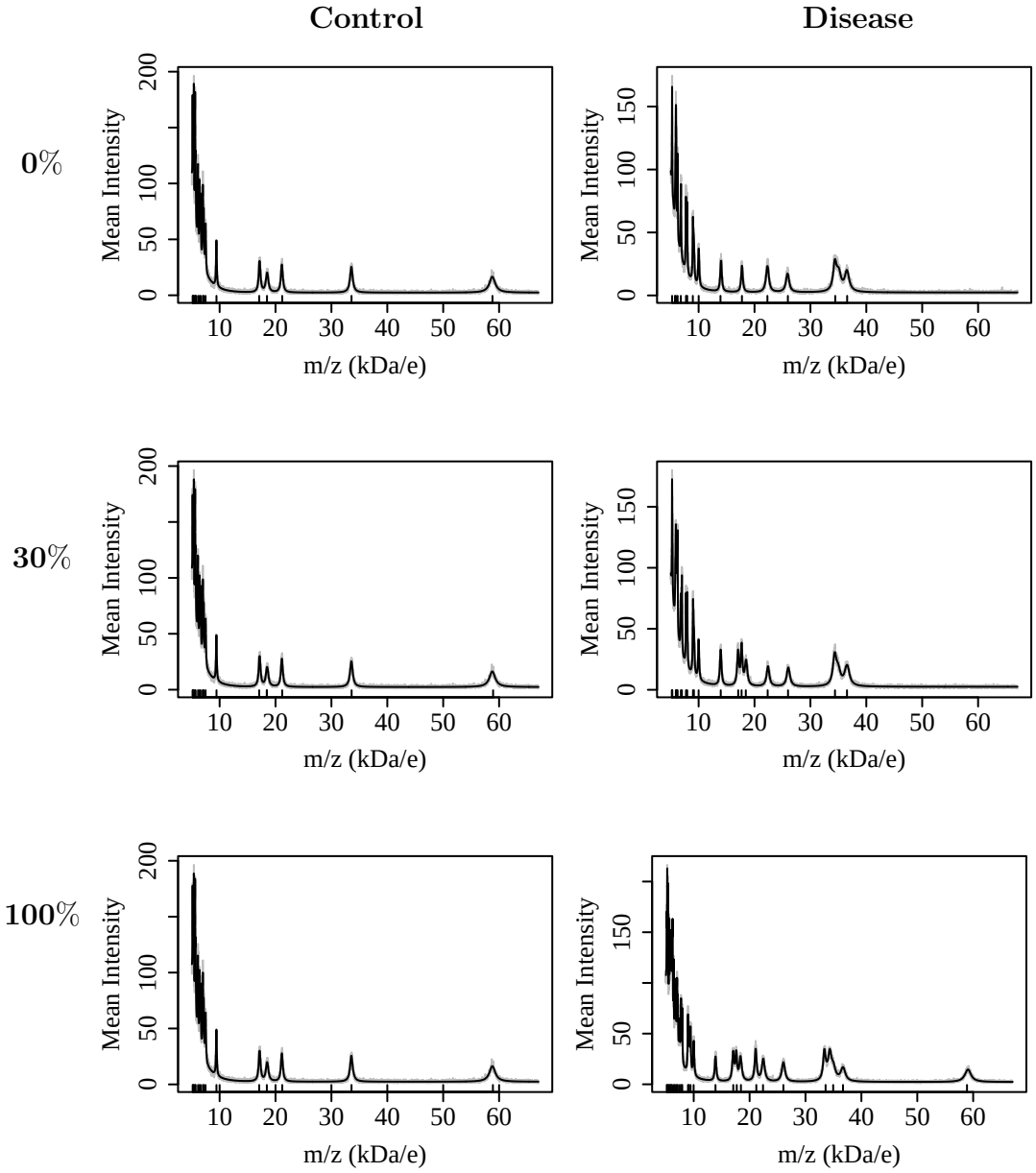


Figure 4.7: Model Average Fit for one simulated dataset per sub-population

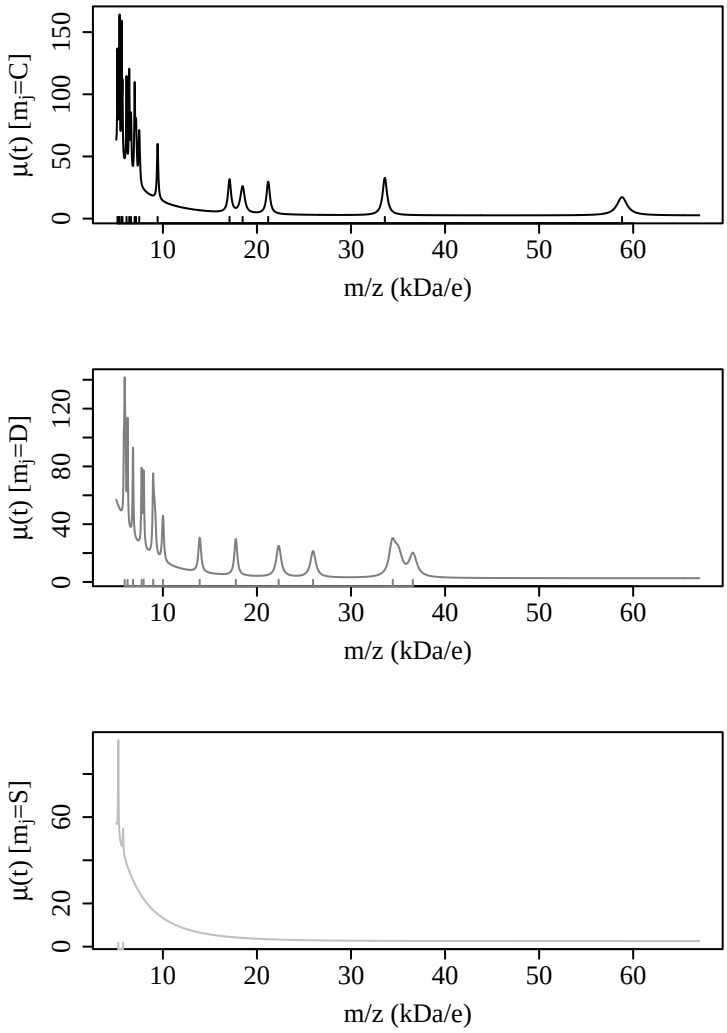


Figure 4.8: Model Average latent signal for dataset DS_0 (0%) that was observed either in control spectra only: $m_j = C$, disease spectra only: $m_j = D$, or both: $m_j = S$

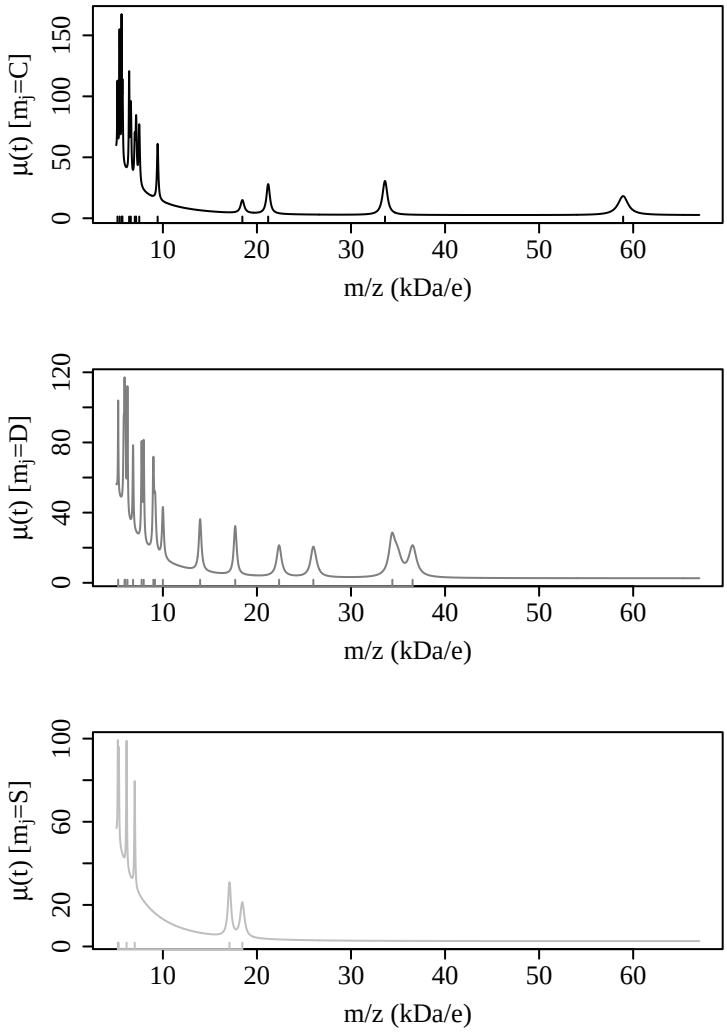


Figure 4.9: Model Average latent signal for dataset DS_{30} (30%) that was observed either in control spectra only: $m_j = C$, disease spectra only: $m_j = D$, or both: $m_j = S$

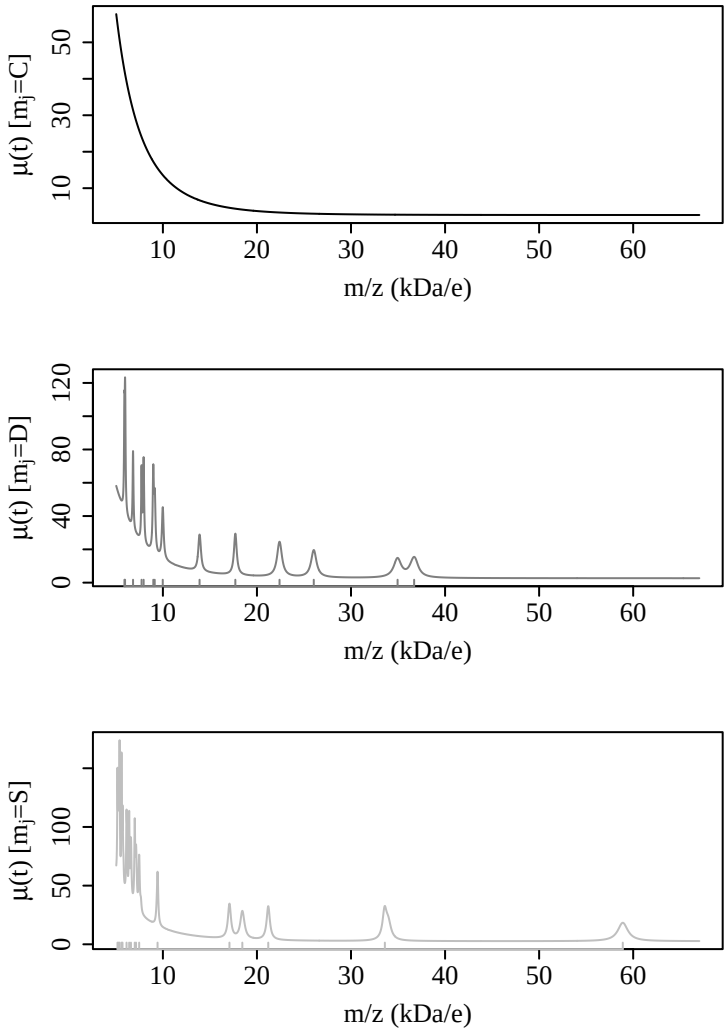


Figure 4.10: Model Average latent signal for dataset DS_{100} (100%) that was observed either in control spectra only: $m_j = C$, disease spectra only: $m_j = D$, or both:: $m_j = S$

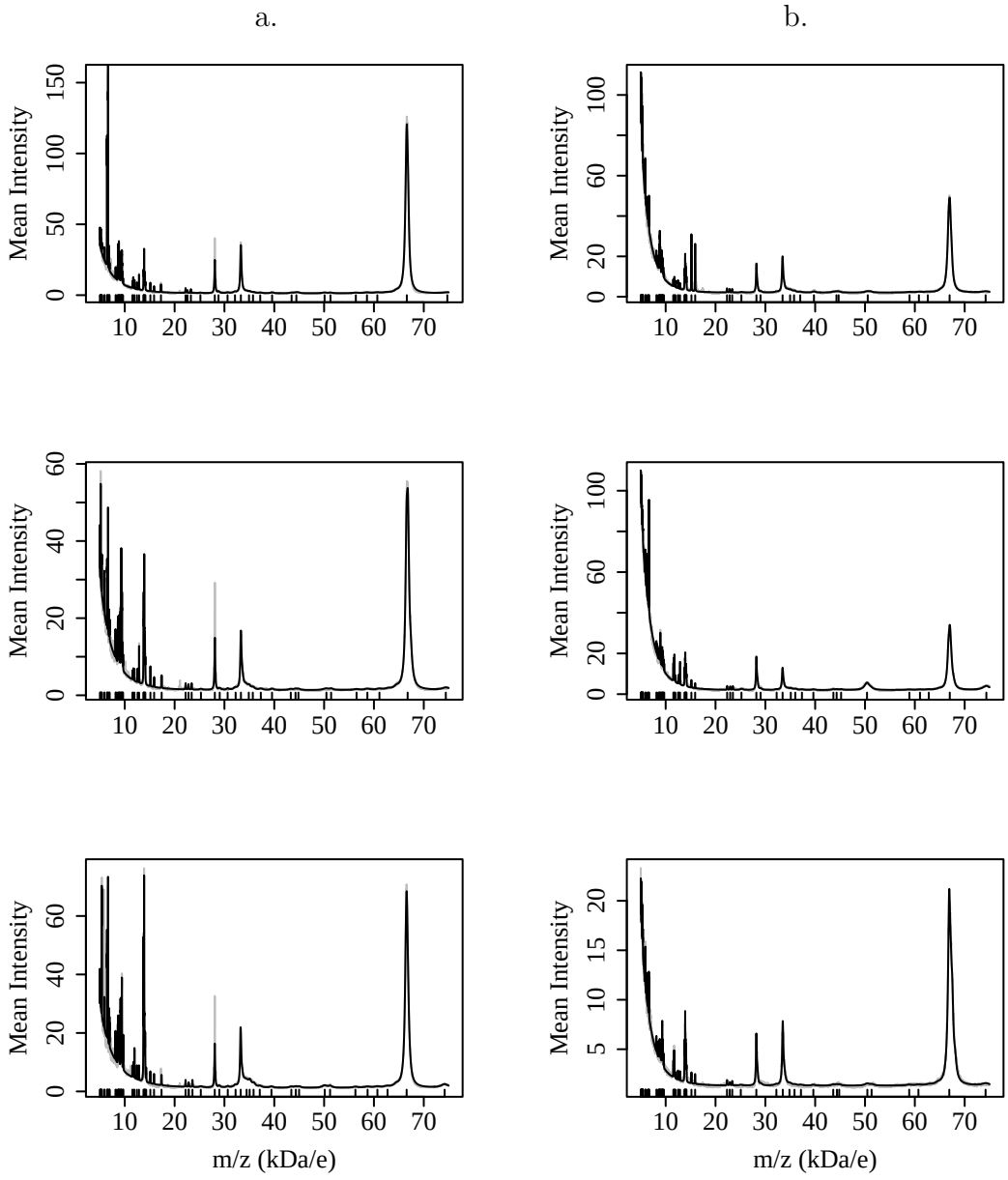
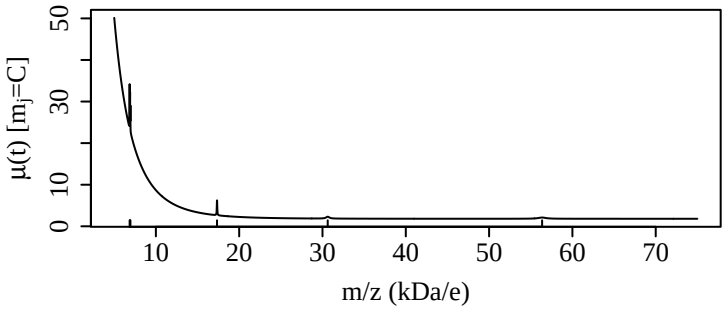


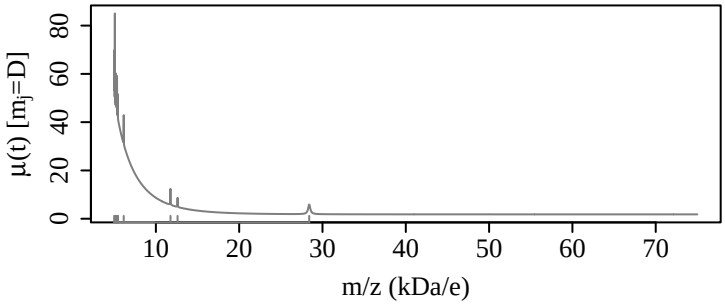
Figure 4.11: Column a. represents model average results for three example spectra from the control group. Column b. represents model average results for three example spectra from the disease group.

$m_j = C$,$m_j = D$, and $m_j = S$. Most of the extracted features are shared by both treatment groups however, there are several peaks solely expressed in the disease and control populations, specifically within the region between 10 and 20 kDa.

a.



b.



c.

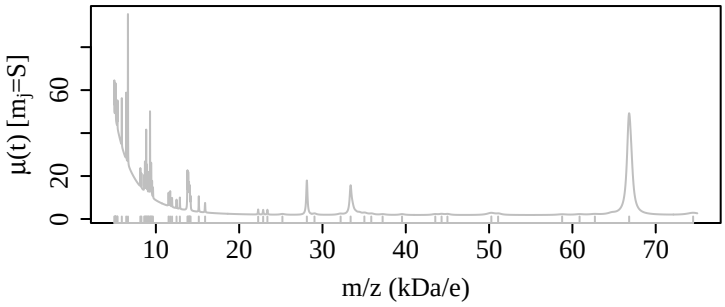


Figure 4.12: Model Average latent signal for real data that was observed either in control spectra only: $m_j = C$, disease spectra only: $m_j = D$, or both: $m_j = S$

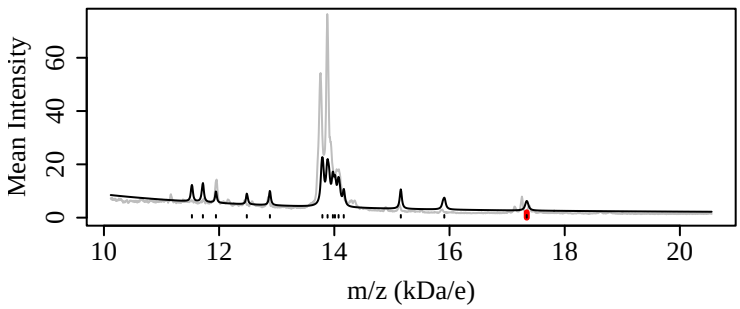
To look closely at how the latent mean-spectrum signal underlies the data, Figure 4.13 displays two disease-only peaks which occur between 10 and 20 kDs at 11,757.0 and 12,604.9 Da and one control-only peak at 17,340.9 Da. The control-only peak clearly does not exist in the disease spectrum and depicts one form differential protein expression: either a protein exists or does not exist within both sub-populations. Another way protein expression may differ is via the degree of intensity: if one peak is present in both population spectra, but on average, has different intensity levels, then the protein is differentially expressed. Our model estimates the latter form of differential expression by placing a second peak near the shared peak in the population with higher intensity. The convolution of both peaks results in a higher expected mean intensity.

Given the small differences in the populations, accurate classifications for three control and three disease spectra are still made. Again, because of computational restrictions, the posterior probability that each spectrum belonged to the control population $\mathrm{P}[d_i^* = D|\boldsymbol{Y_i^*}, \boldsymbol{Y}, \boldsymbol{d}]$ was either zero or one; $\mathrm{P}[d_i^* = D|\boldsymbol{Y_i^*}, \boldsymbol{Y}, \boldsymbol{d}]$=1 for the disease spectra and $\mathrm{P}[d_i^* = D|\boldsymbol{Y_i^*}, \boldsymbol{Y}, \boldsymbol{d}]$=0 for the control spectra.

4.8 Discussion

Our model successfully completes the three steps necessary to analyze expression proteomic data by extracting biologically relevant information from multiple MALDI-TOF mass spectra, assuring the extracted features from one spectrum align with those extracted from other spectra, and classifying profiles that were not within the training dataset. To make the classifications, the model accounts for two forms of differential expression: 1. the existence of proteins in only one

a.



b.

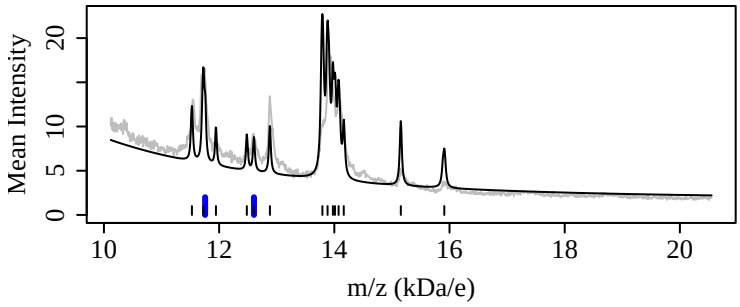


Figure 4.13: Top plot displays the model average, latent signal for a control patient $(\mu(y, d = C)|m_{ij} = 1)$ and the bottom plots the model average, latent signal for disease patients $(\mu(y, d = D)|m_{ij} = 1)$. Underlying each plot is real data from one patients belonging to the appropriate sub-population group. Thick rug marks are made at the peaks that differ between the two populations.

sub-population group, and 2. the expected intensity is increased in one sub-population over the other. A significant difference in intensity can be interpreted as differential expression after adjusting for the sample scaling and background.

We developed the model within the context of two experimental sub-populations and without subject-specific peaks (peaks that only occur within one spectrum, but not in others). However, by increasing the number of possible marks, the model may account for more experimental groups and/or estimate proteins present in one subject. For example, by including n labels U_i to signify *unique* within

spectrum i, in addition to the aforementioned labels C, D, and S, a peak may be declared as either observed in control patients, observed in disease patients, observed in all patients, or observed only in the ith patient. The benefit of adding U_i to the model is that the mean spectrum would identify the entire experimental proteome and list all present proteins, regardless of their significance in making classification. For this research however, treating subject-specific peaks as noise was justified since the primary goal was biomarker detection.

The inclusion of additional marks is also an example of the proposed model's flexibility. Some model specifications may change to accommodate the needs of datasets. For example, in some datasets, peaks may not appear symmetric due either to the ionizing process or low resolution (protein peaks merge and display an asymmetric, isotopic distribution). In which case, the kernel basis function as well as the deterministic function for ω_{ij} may change to accommodate data specific characteristics. Specifically, one would change the kernel to be a normalized summation of two basis functions, such as a Gaussian kernel function for the left peak tail summed with a Cauchy distribution kernel for the right peak tail, and still model the parameters as described in this paper.

First Stage

$$
\begin{aligned}
J|\alpha,\lambda &\sim \text{Po}(\alpha T \text{E}_1(\lambda\epsilon) \\
m_{ij}|d &= \delta_{[d,S]}(m_j) \\
\tau_{ij}|\tau_j, m_{ij} &\sim \delta_1(m_{ij})\text{LN}(\log(\tau_j), 0.002) \\
\eta_{ij}|\eta_j, m_{ij} &\sim \delta_1(m_{ij})\mathbf{1}_{\eta_j>\epsilon}\text{LN}(\log(\eta_j), 0.11) \\
\omega_{ij}|\tau_{ij} &= g(\tau_{ij}, \rho_j)^* \\
\zeta_i|\zeta &\sim \text{LN}(\log(\zeta), 0.5^2) \\
S_i|S &\sim \text{No}(S, 0.013/n) \\
\beta_{1i}(t) &= \eta_{i0} k_0(t, \tau_0, \omega_{i0}) \\
&\quad \eta_{i0} \sim \text{LN}(\log(\eta_0), 0.3) \\
&\quad \omega_{i0} \sim \text{LN}(log(\omega_0), 0.25^2/n) \\
\varphi_i|\varphi &\sim \text{No}(\varphi, 0.5^2/n)
\end{aligned}
$$

Second Stage

$$
\begin{aligned}
m_j &\sim \text{Mu}(3, \pi_C, \pi_D, \pi_S) \\
\tau_j &\sim \text{Un}(T_0, T_1) \\
\eta_j &\sim \text{Ga}(0, \lambda, \epsilon) \\
\rho_j|\varrho^{(M)}, \varrho^{(L)}\tau_j &\sim \begin{cases} \text{LN}(\log(\varrho^{(M)}), 0.05^2) & \text{if } \tau_j \in [5, 25]\text{kDa} \\ \text{LN}(\log(\varrho^{(L)}), 0.05^2) & \text{if } \tau_j > 25\text{kDa} \end{cases} \\
\zeta &\sim \text{LN}(\log\left(\frac{\sum_i \overline{\boldsymbol{Y}_i}}{n}, \sigma_\zeta^2\right), \text{ s.t. } \text{V}[\zeta]/\text{E}[\zeta] \approx 0.10^2 \\
S &\sim \text{Be}(S_a, S_b), \text{ s.t. } \text{E}[S] = .23 \text{ and } \text{V}[S] = 0.013 \\
\beta_0(t) &= \eta_0 k_0(t, \tau_0, \omega_0) \\
&\quad \eta_0 \sim \text{Ga}(0, \lambda_0, \epsilon) \\
&\quad \rho_0 \sim \text{LN}(0, \log(11.5), 0.25^2) \\
\varphi &\sim \text{Ga}(0.5, 1) \\
\alpha &\sim \text{Ga}(\alpha_a, \alpha_b)
\end{aligned}
$$

Third Stage

$$
\begin{aligned}
\varrho^{(M)} &\sim \text{LN}(\log(\varrho_\mu^{(M)}), 0.05\sqrt{n}) \\
\varrho^{(L)} &\sim \text{LN}(\log(\varrho_\mu^{(L)}), 0.05\sqrt{n}) \\
[\pi_C, \pi_D, \pi_S]' &\sim \text{Di}(1, 1, 1)
\end{aligned}
$$

Table 4.1: Hierarchical model for the modeling multiple spectra. The symbol "$*$" is a reminder that $g()$ is the function provided in Equation 4.2.

Rate	0%		30%		100%	
	J^{HP}	J^{∇}	J^{HP}	J^{∇}	J^{HP}	J^{∇}
TDR	0.94	0.86	0.97	0.80	0.86	.80
FDR	0.27	0.03	0.27	0.12	0.33	0.12

Table 4.2: Summary of true and false feature discovery rates per dataset using the posterior model ($^{\text{HP}}$) and model average ($^{\nabla}$) estimates.

%	$f(t)$		$f(t)\|m_j = C$		$f(t)\|m_j = D$		$f(t)\|m_j = S$	
	J	J^{∇}	J	J^{∇}	J	J^{∇}	J	J^{∇}
0	35	31	17	17	18	13	0	2
30	35	32	12	14	23	16	5	6
100	35	32	0	0	18	14	17	17

Table 4.3: True J and estimated J^{∇} number of peaks from the model average estimates of the mean spectrum across all marks $f(t)$, the mean spectrum for only control spectra $f(t)|m_j = C$, the mean spectrum for only disease spectra $f(t)|m_j = S$, and the mean spectrum for shared signal $f(t)|m_j = S$.

Parameter	Mean	Std. Dev
ζ	7.80	0.83
S	0.23	0.01
φ	9.39	0.08
$\varrho^{(M)}$	416.32	3.19
$\varrho^{(L)}$	105.20	0.98
η_0	147.20	9.40
ω_0	18.13	1.93
J^{PM}	97.04	0.53
J^{HP}	98.00	-
J^{∇}	78.00	-

Table 4.4: Posterior mode estimates for model parameters.

Parameter	Mean	Std. Dev
π_C	0.07	0.03
π_D	0.14	0.03
π_S	0.79	0.04

Table 4.5: Stochastic means for the parameters π_C, π_D, and π_S, which estimate the probabilities that m_j is either C, D, or S. Notice, very little information differs between the control and disease populations.

Chapter 5

Discussion

5.1 Summary

The protein signal in a spectrum appears as multiple peaks with widths that relate to the measurement resolution. Finding the peaks can be challenging since they are among tens or hundreds of thousands spectrum data points. Chapters 2 and 3 were devoted solely to developing nonparametric Bayesian models to identify significant peaks from individual spectra. Chapter 4 also found significant features in spectra, but for the purpose of classifying treatment groups as well as identifying proteins. Each of the models were based on estimating the expected intensity as a function of measurement error parameters and a latent protein signal which was estimated from four protein parameters, the number, location, width, and area of peaks. Distinct differences between the approaches however, present advantages and disadvantages.

In Chapter 2, we described a kernel regression approach to estimate protein signal. The model included initially one protein peak at every observed m/z and relied on parameter shrinkage or thresholding to reduce the model dimension. Pos-

terior model parameter estimates were obtained via an EM algorithm programmed in R. Provided the optimization packages available in R, the algorithm was fast to implement. However, the proposed model faulted in two primary ways: 1. for the sake of speed, we did not calculate variance estimates for the posterior parameter distributions, and 2. the basis expansion model constrains peaks to locate only at observed data points. The latter is particularly problematic because the true mass of a protein may exist between measurements.

The second feature extraction model proposed in Chapter 3 avoided the pitfalls of the previous approach, but possibly introduced others. Again, estimates for a latent protein signal were obtained via kernel regression, but the location and scale parameters of the kernels were adapted to the data. Thus, upon specifying a joint prior distribution on the kernel parameters and the protein concentrations, peaks were able to be placed any where within the mass domain of a given spectrum. Further, variance estimates for the model parameters were calculated from stochastically sampling the full joint posterior distribution using RJMCMC. The RJMCMC was considerably more difficult to program than the aforementioned EM algorithm, and depending upon the dataset, the Markov chain took millions of iterations to converge. However, from simulation studies in Chapter 3, we learned that the model may find true peaks in extremely noisy situations.

Taking advantage of these benefits, the final model developed in Chapter 4 extracted significant features from multiple spectra and classified them according to treatment group. In turn, the list of features did not include every peak seen in each individual spectrum, but rather, identified peaks observed in several spectra.

This dissertation was the progression of three nonparametric Bayesian models, each extending or improving the one proceeding. However, improvements and

extensions to the presented research remain.

5.2 Future Work

Improvement or extensions to my research are grouped in the following categories: Computational Efficiency, Prior elicitation and Model specifications, and Applications.

Computational Efficiency

First and foremost, the software developed to estimate the proposed models should be made available for others to use. Before doing so, the code should be streamlined so that the speed of the current RJMCMC increases. However, large computational changes to the programs developed for Chapters 3 and 4 may increase the speed of the Markov chain as well. Three possibilities are presented below.

1. Approximate the likelihood: Calculating the likelihood distribution several times per RJMCMC iteration is extremely time consuming. Developing a good way to approximate the likelihood that will decrease computation time and not effect significantly the final parameter estimates would be very useful.

2. Parallelizing the likelihood calculation: The final model in Chapter 4, increases the time needed to calculate the likelihood distribution by n, the number of spectra. Parallelizing this calculation to n nodes in a computer cluster however, would reduce the processing time back to that of a single spectrum.

3. Proposal distributions: Currently the RJMCMC explores the model space via a birth/death process; one peak is proposed to be either added, deleted,

or updated during each iteration. To add a peak, one location is proposed uniformly from the entire time domain, thus proposing a location near a true peak is small. We may increase this probability by proposing peaks in data dependent locations.

4. Tempering Scheme: In our application, feature extraction is comparable to finding modes within a spectrum. Plentiful research exists for sampling schemes to explore multi-modal distributions, and some of the ideas apply here, such as tempering. For example, the current RJMCMC relies on the birth or movement of peaks in a birth/death process to find the spectrum modes. One way to promote movement and increase the probability of birthing a peak in a meaningful place is to decrease the temperature of a spectrum. This will temporarily broaden the spectrum modes to cover more area on the x-axis and increase the probability or peaks within the birth/death chain to find them. Increasing the temperature as the RJMCMC iterates will eventually allow accurate estimates for the remaining parameters.

Prior Elicitation and Model specifications

Wasserman [1998] notes another irony of Bayesian nonparametrics in that the methods are applied traditionally when little is known about the structure of the data, yet "huge amounts of prior information" are necessary to search infinite dimensional parameter spaces. Mass spectrometric data fit the paradoxical data requirements for employing nonparametric analysis methods, and informative priors for many parameters are elicited. However, improvements can always be made on the prior distributions with either more data or better insight into the behavior of the parameters.

Specifying a meaningful value for ϵ was challenging. Thus, possibly modeling the Lévy process with a data determined ϵ is worth exploring, but care must be taken to assure identifiability. Another option might be to assume the spectrometric data are generated from a mixture of two Lévy processes; one process may model the protein signal, and the other may model the small noisy observations.

Applications

In my mind, nonparametric regression using Lévy processes is the wave of the future. In the past, Dirichlet and Gaussian process priors were used for curve estimation because sampling from each was easy within an MCMC framework. However, when Wolpert and Ickstadt [1998b] introduced an easy way to sample Lévy processes, time was the only factor keeping it from entering the nonparametric literature. I look forward to using models similar to what was developed here in other applications.

Bibliography

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Stat. Soc. B* **60**, 4, 725–749.

Abramowitz, M. and Stegun, I. A., eds. (1964). *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, vol. 55 of *Applied Mathematics Series*. National Bureau of Standards, Washington, D.C.

Andrews, D. F. and Mallows, C. L. (1974). Scale mixtures of normal distributions. *J. Roy. Stat. Soc. B* **36**, 1, 99–102.

Antoniak, C. E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* **2**, 6, 1152–1174.

Applied Biosystems (2001). *Voyager Biospectrometry Workstation with Delayed Extraction Technology User Guide Version 5.1.* Applied Biosystems, Foster City, CA.

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

Baggerly, K. A., Morris, J. S., Wang, J., Gold, J., Xiao, L.-C., and Coombes, K. R. (2003). A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. *Proteomics* **3**, 9, 1667–1672.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. John Wiley & Sons, New York, NY.

Besag, J., Green, P. J., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems (with discussion). *Stat. Sci.* **10**, 1, 3–66.

Brown, P. J. and Griffin, J. E. (2005). Alternative prior distributions for variable selection with very many more variables than observations. Tech. Rep. UKC/IMS/05/08, University of Kent, UK.

Campa, M. J., Fitzgerald, M. C., and Patz, Jr., E. F. (2003a). Exploring the proteome with MALDI-TOF (editorial). *Proteomics* **3**, 9, 1659–1660.

Campa, M. J., Wang, M. Z., Howard, B. A., Fitzgerald, M. C., and Patz, Jr., E. F. (2003b). Protein expression profiling identifies MIF and Cyclophilin A as potential molecular targets in non-small cell lung cancer. *Cancer Research* **63**, 7, 1652–1656.

Carpenter, M., Melath, S., Zhang, S., and Grizzle, W. E. (2003). Statistical process and analysis of proteomic and genomic data. In *Proceedings of the Pharmaceutical SAS Users Group, Miami, FL*, 545–548.

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997). Adaptive bayesian wavelet shrinkage. *J. Am. Stat. Assoc.* **92**, 440, 1413–1421.

Clyde, M. A., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 2, 391–401.

Clyde, M. A. and Wolpert, R. L. (2006). Nonparametric function estimation using overcomplete dictionaries. In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 8*, In Press, Oxford, UK. Oxford Univ. Press.

Coombes, K. R., Koomen, J. M., Baggerly, K. A., Morris, J. S., and Kobayashi, R. (2005a). Understanding the characteristics of mass spectrometry data through the use of simulation. *Cancer Informatics* **1**, 1, 41–52.

Coombes, K. R., Tsavachidis, S., Morris, J. S., Baggerly, K. A., Hung, M.-C., and Kuerer, H. M. (2005b). Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform. *Proteomics* **5**, 16, 4107–4117.

Dass, C. (2001). *Principles and Practice of Biological Mass Spectrometry*. John Wiley & Sons.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm (with discussion). *J. Roy. Stat. Soc. B* **39**, 1, 1–38.

Dey, D. K., Müller, P., and Sinha, D., eds. (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*, vol. 133 of *Lecture Notes in Statistics*. Springer-Verlag, New York, NY.

Do, K.-A., Müller, P., and Vannucci, M., eds. (2006). *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, Cambridge, UK.

Fitzgerald, M. C. and Roulhac, P. L. (2006). Personal communication.

Franzen, J. (1997). Improved resolution for MALDI-TOF mass spectrometers: A mathematical study. *International Journal of Mass Spectrometry and Ion Processes* **164**, 1, 19–34.

Gelfand, A. E. and Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Am. Stat. Assoc.* **85**, 410, 398–409.

Gilks, W. R., Richardon, S., and Spiegelhalter, D. J. (2000). *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC.

Goldfarb, D. and Idnani, A. U. (1983). A numerically stable dual method for solving strictly convex quadratic problems. *Mathematical Programming* **27**, 1, 1–33.

Green, P. J. (1987). Penalized likelihood for general semi-parametric regression models. *Int. Stat. Rev.* **55**, 3, 245–259.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82**, 4, 711–732.

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer-Verlag, New York, NY.

House, L. L., Clyde, M. A., and Wolpert, R. L. (2006a). Functional data analysis using Lévy random field models for multi-spectra peak identification and classification. Discussion Paper 2006-xx, Duke University ISDS.

House, L. L., Clyde, M. A., and Wolpert, R. L. (2006b). Peak identification. Discussion Paper 2006-xx, Duke University ISDS.

House, L. L., Clyde, M. A., and Wolpert, R. L. (2006c). Rapid peak identification in matrix-assisted laser desorption ionization time-of-flight mass spectrometry. Discussion Paper 2006-xx, Duke University ISDS.

Jiang, J. (2000). A nonlinear Gauss-Seidel algorithm for inference about GLMM. *Comput. Stat.* **15**, 2, 229–241.

Johnstone, I. M. and Silverman, B. W. (1997). Wavelet threshold estimators for data with correlated noise. *J. Roy. Stat. Soc. B* **59**, 2, 319–351.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* **90**, 431, 928–934.

Kempka, M., Södahl, J., Björk, A., and Roeraade, J. (2004). Improved method for peak picking in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Rapid Communications in Mass Spectrometry* **18**, 11, 1208–1212.

Khinchine, A. Y. and Lévy, P. (1936). Sur les lois stables. *Comptes rendus hebdomadaires des seances de l'Académie des sciences. Académie des science (France). Serie A. Paris* **202**, 374–376.

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Stat. Comput.* **11**, 4, 313–322.

Li, X. (2005). `PROcess`. Version 2.10 `R` package, http://cran.r-project.org.

Malyarenko, D. I., Cooke, W. E., Adam, B.-L., Malik, G., Chen, H., Tracy, E. R., Trosset, M. W., Sasinowski, M., Semmes, O. J., and Manos, D. M. (2005). Enhancement of sensitivity and resolution of Surface-Enhanced Laser Desorption/Ionization Time-of-Flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clinical Chemistry* **51**, 1, 65–74.

Martin, D. B. and Nelson, P. S. (2001). From genomics to proteomics: Techniques and applications in cancer research. *Trends in Cell Biology* **11**, 11, S60–S65.

Müller, P., Erkanli, A., and West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika* **83**, 1, 67–79.

Müller, P. and Quintina, F. A. (2004). Nonparametric bayesian analysis. *Stat. Sci.* **19**, 1, 95–110.

Monagan, M. B., Geddes, K. O., Heal, K. M., Labahn, G., Vorkoetter, S. M., McCarron, J., and DeMarco, P. (2005). *Maple 10 Programming Guide*. Maplesoft, Waterloo ON, Canada.

Morris, J. S., Brown, P. J., Baggerly, K. A., and Coombes, K. R. (2006). Analysis of mass spectrometry data using Bayesian wavelet-based functional mixed models. In Do *et al.* [2006], chap. 14, 269–292.

Morris, J. S., Coombes, K. R., Koomen, J., Baggerly, K. A., and Kobayashi, R. (2005). Feature extraction and quantification for mass spectrometry in biomedical applications using mean spectrum. *Bioinformatics* **21**, 9, 1764–1775.

O'Hagan, A. and Kingman, J. F. C. (1978). Curve fitting and optimal design for prediction. *J. Roy. Stat. Soc. B* **40**, 1, 1–24.

Papapantoleon, A. (2005). An introduction to léy processes with applications in finance. Lecture notes for mini-course at University of Piraeus and University of Leipzig.

PerSeptive Biosystems (1999). *Data Explorer Software User's Guide*. PerSeptive Biosystems, Framingham, MA.

Rogers, L. C. G. and Williams, D. (1994). *Diffusions, Markov Processes, and Martingales*, vol. 1. John Wiley & Sons, New York, NY, 2nd edn.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J. Roy. Stat. Soc. B* **47**, 1, 1–52.

Siuzdak, G. (2003). *The Expanding Role of Mass Spectrometry in Biotechnology*. MCC Press, San Diego, CA.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* **58**, 1, 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q.-T. (2004). Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics* **20**, 17, 3034–3044.

Tierney, L. (1994). Markov chains for exploring posterior distributions (with discussion). *Ann. Stat.* **22**, 4, 1701–1762.

Tikhonov, A. N. (1963). Solution of incorrectly formulated problems and the regularization method. *Soviet Mathematics Doklady* **4**, 1035–1038. English translation of *Doklady Akademii Nauk SSSR*, 151(3), 501–504.

Vidakovic, B. (1998). Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *J. Am. Stat. Assoc.* **93**, 441, 173–179.

Wang, M. Z., Howard, B. A., Campa, M. J., Patz, Jr., E. F., and Fitzgerald, M. C. (2003). Analysis of human serum proteins by liquid phase iso-electric focusing and Matrix-Assisted Laser Desorption/Ionization Mass Spectrometry. *Proteomics* **3**, 9, 1661–1666.

Wasserman, L. (1998). Asymptotic properties of nonparametric bayesian procedures. In Dey *et al.* [1998], 293–302.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika* **74**, 3, 646–648.

Wolfram Research, Inc. (2005). *Mathematica*. Wolfram Research, Inc., Champaign, IL, 5th edn.

Wolpert, R. L. (2002). Lévy processes. In A. H. El-Shaarawi and W. W. Piegorsch, eds., *Encyclopedia of Environmetrics*, vol. 2, 1161–1164. John Wiley & Sons, Chichester, NY.

Wolpert, R. L. and Ickstadt, K. (1998a). Poisson/gamma random field models for spatial statistics. *Biometrika* **85**, 2, 251–267.

Wolpert, R. L. and Ickstadt, K. (1998b). Simulation of Lévy random fields. In Dey *et al.* [1998], 227–242.

Wolpert, R. L. and Ickstadt, K. (2004). Reflecting uncertainty in inverse problems: A Bayesian solution using Lévy processes. *Inverse Problems* **20**, 6, 1759–1771.

Wolpert, R. L., Ickstadt, K., and Hansen, M. B. (2003). A nonparametric Bayesian approach to inverse problems (with discussion). In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 7*, 403–418, Oxford, UK. Oxford Univ. Press.

Wu, W., Hu, W., and Kavanagh, J. J. (2002). Proteomics in cancer research. *International Journal of Gynecological Cancer* **12**, 5, 409–423.

Yasui, Y., McLerran, D., Adam, B.-L., Winget, M., Thornquist, M., and Feng, Z. (2003). An automated peak identification/calibration procedure for high dimensional protein measures from mass spectrometers. *Journal of Biomedicine and Biotechnology* **4**, 242–248.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243.

Zhigilei, L. V. and Garrison, B. J. (1998). Velocity distributions of analyte molecules in matrix assisted laser desorption from computer simulations. *Rapid Communications in Mass Spectrometry* **12**, 1273–1277.

Biography

I was born June 17, 1976 in Ithaca, NY. I received a Bachelor of Science (B.S.) in Biometry and Statistics in 1998 and a Master of Arts in Teaching (M.A.T.) in 1999 from Cornell University, Ithaca, NY. In 2003, I earned an M.S. in statistics from the Institute of Statistics and Decision Sciences, Duke University. I have co-authored the following articles:

1. House, L. and Banks, D. (2004). Robust Multidimensional Scaling *Compstat, Proceedings in Computation Science*, **16**, 251–260.

2. Banks, D., House L., McMorris, F.R., Arabie, P., and Gaul, W. *Editors* (2004). Classification, Clustering, and Data Mining Applications, Springer-Verlag.

3. Clyde, M., House, L., Tu, C., Wolpert, R. "Bayesian Nonparametric Function Estimation Using Overcomplete Representations and Lévy Random Field Priors." *Statistische und Probabilistische Methoden der Modellwahl. Oberwolfach Report 47* (2005).

4. Clyde, M. House, L. and Wolpert, R. Nonparametric Models for Proteomic Peak Identification and Quantification, In *Bayesian Inference for Gene Expression and Proteomics*. Edited by K.-A. Do, P. Müller, and M. Vannucci (to appear)

5. House, L. and Clyde, M. (2005). Bayesian Identification of Differential Gene Expression Induced by Metals in Human Bronchial Epithelial Cells. *Bayesian Analysis* **1.1** 105–120.