MODEL-BASED PRE-PROCESSING IN PROTEIN MASS SPECTROMETRY

A Dissertation

by

JOHN C. WAGAMAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2009

Major Subject: Statistics

MODEL-BASED PRE-PROCESSING IN PROTEIN MASS SPECTROMETRY

A Dissertation

by

JOHN C. WAGAMAN

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Co-Chairs of Committee,	Jianhua Huang
	Webster West
Committee Members,	Alan Dabney
	Natarajan Sivakumar
Head of Department,	Simon Sheather

December 2009

Major Subject: Statistics

ABSTRACT

Model-Based Pre-processing in Protein Mass Spectrometry. (December 2009)

John C. Wagaman, B.S., Millersville University;

M.S., University of Central Florida

Co–Chairs of Advisory Committee: Dr. Jianhua Huang
Dr. Webster West

The discovery of proteomic information through the use of mass spectrometry (MS) has been an active area of research in the diagnosis and prognosis of many types of cancer. This process involves feature selection through peak detection but is often complicated by many forms of non-biological bias. The need to extract biologically relevant peak information from MS data has resulted in the development of statistical techniques to aid in spectra pre-processing. Baseline estimation and normalization are important pre-processing steps because the subsequent quantification of peak heights depends on this baseline estimate.

This dissertation introduces a mixture model to estimate the baseline and peak heights simultaneously through the expectation-maximization (EM) algorithm and a penalized likelihood approach. Our model-based pre-processing performs well in the presence of raw, unnormalized data, with few subjective inputs. We also propose a model-based normalization solution for use in subsequent classification procedures, where misclassification results compare favorably with existing methods of normalization. The performance of our pre-processing method is evaluated using popular matrix-assisted laser desorption and ionization (MALDI) and surface-enhanced laser desorption and ionization (SELDI) data sets as well as through simulation.

To Donna, Mark and Mary

ACKNOWLEDGMENTS

To Eustace Conway: I have only met you once for just a fleeting moment, yet I am continually enthralled by your spirit. Although you will likely never read this, your simple words, "you can", have resonated with me over the past few months. You are truly the Last American Man.

Para Michelle, Goya, Gilberto y Stephanie, Papa Iel, Taia y Luis, Juan Felipe, Isabela, Nicolas y Lucia, Tiz, Luis, Sarita y Julian, Luz, Pablo, Paula y Cristobal, Orlando, Chila, Leon, Gelli y Daniel Simon, Diana, Liria, Fernando y Sylvia, Glauco, Laura, Juan Sebastian, Julian Ricardo y Sara Manuela, Pablo, Margarita Rosa y David Felipe: Gracias por recibirme con los brazos abiertos. Nunca olvidare nuestro tiempo en Medellin, Santa Helena, El Retiro, Rionegro, Marinilla, En Penol y Amaga. Espero que nos podamos ver pronto.

To Hunter Layton: To my newest friend. I could always count on you for a beer on the front porch of 202 while listening to "Bobby" Keen and Songs from the Wood. I look forward to another Texas Pearl with you.

To Dan Glab: I am not sure if I would have gotten through 614 without you. Thanks for your words of motivation over the past few months and for your friendship over the past six years.

To John Dougherty: I came to this state without a single friend. Thank you for asking me to play kickball six years ago and for the memories we have had since then (Lowlands 2007).

To Professors Dabney and Sivakumar: Thank you for willingness to serve on my committee over the past two years. I appreciate your time and helpful advice throughout the compilation and completion of this document.

To Mark, Donna and Mary: I dedicated this document to you, and words cannot express my love for you. I'm moving closer to home, so I hope to make up for missed time over the past few years.

To Michelle: You have seen me at my best and worst over the last two years. You put up with a steady diet of stress, sudden mood changes and the Grateful Dead for me. I will see you soon, my sweet. Amor mio, me da verguenza no haberte escrito antes. Aunque siempre queria, la verdad es que no podia. Que decirte? Soy medio lento. Me llevo todos estos anos aprender tu hermosa lengua. Y todavia no alcanza. Asi que te ofrezco mi primer intento – una cancion sencilla solo dice que te quiero.

To Jianhua and Web: I do not know if I could have finished this project with anyone else. I am amazed by the patience and understanding you have displayed with me over the past few years. I walked into each of your offices and tried to quit, and when I left your offices, somehow I was convinced to keep moving forward. I hope I can be half of the professionals that you both are and I will remain forever indebted to you.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

TABLE Page

LIST OF FIGURES

FIGURE Page

FIGURE Page

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Proteomics is the study of the functions and structures of proteins. The extraction of proteomic information from biological matter is often obtained through the use of mass spectrometry. A substantial portion of the current research in protein mass spectrometry currently uses statistical techniques to identify molecular biomarkers that are associated with diseases of interest, which include breast cancer (Coombes *et al.*, 2003; Kuerer *et al.*, 2004; Li *et al.*, 2002), ovarian cancer (Petricoin *et al.*, 2002a; Rai *et al.*, 2002; Sorace and Zhan, 2003; Wu *et al.*, 2003), lung cancer (Zhukov *et al.*, 2003) and prostate cancer (Adam *et al.*, 2002; Coombes *et al.*, 2004; Qu *et al.*, 2002; Yasui *et al.*, 2003b). The applications of these techniques can be used, for example, in the monitoring of disease progression, observing the reaction to a treatment and in early cancer diagnosis (Coombes *et al.*, 2005a). The importance of proteomics in early cancer detection is especially critical since therapies can be more effective when applied earlier. However, some cancers, like early-stage ovarian cancer, lack identifying symptoms which may delay the diagnosis (Petricoin *et al.*, 2002b). The use of proteomics in early cancer detection is an attempt to provide non-invasive and diagnostic information in the discovery of useful biomarkers.

This dissertation follows the style of the *Journal of the Royal Statistical Society*.

I.1. Mass Spectrometry Overview

In mass spectrometry, a sample of tissue, serum, biological fluid, etc. is collected and input into a mass spectrometer, where a mass spectrum is produced consisting of intensity and mass-to-charge ratio pairs for the molecules in the sample. This spectrum serves as a graphical description of the contents of the sample; that is, what molecules are present and in what quantity. The process for producing a mass spectrum from a biological sample begins with a transformation of the sample input into a gas and the ionization of these gaseous molecules. While several types of ionization sources exist, our work has mainly focused on spectra resulting from matrix-assisted laser desorption and ionization (MALDI) and surface-enhanced laser desorption and ionization (SELDI) mass spectrometry (MS). In short, the sample input is embedded in a crystalline matrix, where the both the sample input and matrix are transformed and ionized with pulsed shots from a laser beam. As the ions are separated, they are directed towards the mass analyzer, where ions will be sorted according to their mass-to-charge ratio. For instance, in a time-of-flight mass analyzer, the ions are propelled through a flight tube towards the ion detector. As each ion hits the detector, its flight time is recorded; smaller ions move faster. The masses of the ions are determined from their respective flight times through a flight tube whose distance is known. The mass-to-charge ratio for each ion is its atomic mass divided by its charge (i.e., +1), and, because most ions have the same charge (+1), the m/z ratio is sometimes referred to as "mass" only. The mass and intensity information is summarized in a bivariate spectrum which plots the range of observed mass-to-charge ratios (m/z) and the respective ion intensities over this range. The peptides present in the sample are represented as anonymous peaks in intensity at particular mass-to-charge ratios, where each peptide, protein, molecule, etc. has a unique mass.

An ovarian cancer data set analyzed in Tibshirani *et al.* (2004) and Wu *et al.* (2003) has one MALDI spectrum for each of 89 subjects, each with known disease status. We want to build a classifier based on these spectra to discriminate between cancerous and healthy samples. Specifically, we want to classify each patient's spectrum based on peak heights or peak presence at particular values of the mass-to-charge ratio, which correspond to discriminating proteins, peptides, etc. Four of these spectra are plotted in Figure 1. These figures clearly exhibit some forms of non-biological bias that can complicate the extraction of biologically relevant information from the spectra. These problems are the main motivation for this work. In this chapter, we describe some of these biases and review existing methods of pre-processing spectra to correct for such biases. In Chapter II, we propose a method for modeling the intensities of the spectra. In Chapter III, we discuss baseline correction of the spectra and in Chapter IV, we consider peak detection and classification.

I.2. Pre-processing of Spectra

A significant amount of pre-processing must be done to extract biologically-meaningful patterns from the data for later use. Pre-processing consists of several steps that interact in complex ways (Baggerly *et al.*, 2003; Coombes *et al.*, 2005b), and these steps are essential, because the processed spectra should only reflect the behavior inherent in the biological specimens. The pre-processing steps summarized below are not necessarily performed in the same order and some steps may be omitted altogether. Poor choices during the pre-processing stage may result in spectra with substantial biases which may prevent the identification of important biomarkers for disease classification (Coombes *et al.*, 2005a). The goal of pre-processing spectra is to identify individual proteins in each sample through peak detection and quantification. The methods

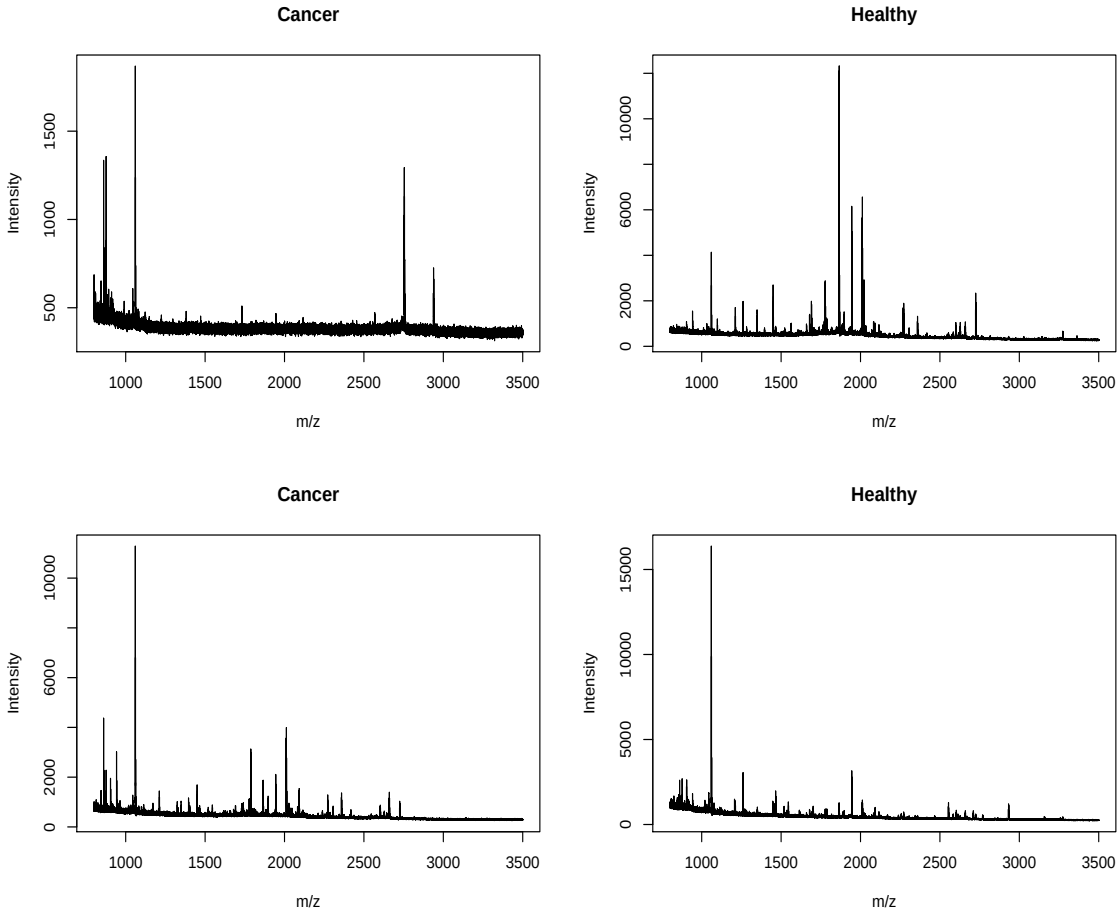

Fig. 1 Four spectra from the ovarian cancer data set of (Wu *et al.*).

used to pre-process and extract peaks from spectra vary. Successful pre-processing will yield a list of biologically relevant locations and intensities for each spectrum. We briefly outline some typical pre-processing steps below.

I.2.1. Calibration

Calibration maps the observed time of flight to the inferred mass-to-charge ratio (Coombes *et al.*, 2005a). This step aligns the spectra so that peaks in different spectra corresponding to the same protein can be identified as such (Morris *et al.*, 2005).

One method of calibrating spectra uses a windowing technique which depends on the accuracy of the mass spectrometer used. The number of peaks that fall within this window width for each location is calculated, and the mass-to-charge ratios which have the largest number of peaks in their corresponding window of potential shift are included in a set of calibrated mass-to-charge ratios (Yasui *et al.*, 2003a). Extending the process, the locations in each bin are sometimes replaced with the midpoint location of the bin and the maximum intensity in the bin. This procedure improves peak calibration and reduces the dimensionality of the data (Carpenter *et al.*, 2003). Calibration is also proposed using parametric time warping (Eilers, 2004). The intensities of two spectra can be calibrated by finding a "warping" equation for one spectrum which minimizes its differences with the other spectrum in the quadratic norm. Evidence of uncalibrated spectra is illustrated in Figure 2.

I.2.2. Baseline Correction

The baseline artifact of MS spectra is not biological in nature, rather, it often stems from a large number of matrix molecules hitting the detector early in the analysis (Coombes *et al.*, 2005a). The baseline is the signal that would be produced by a mass spectrometer in the absence of noise and a sample (Gras *et al.*, 1999). Figure 3 shows two mass spectra from healthy controls with different baselines. Poor baseline estimation may lead to increased detection of false positive peaks, as a peak may have a greater intensity due to non-biological bias in the data. Baseline estimation using a "loess" smooth is illustrated in Figure 4.

Clearly, the baseline estimation (and subtraction) procedure must be carried out for each spectra, separately. Baseline estimates should be very smooth, especially in peak regions, where poor baseline estimation may affect subsequent peak detection

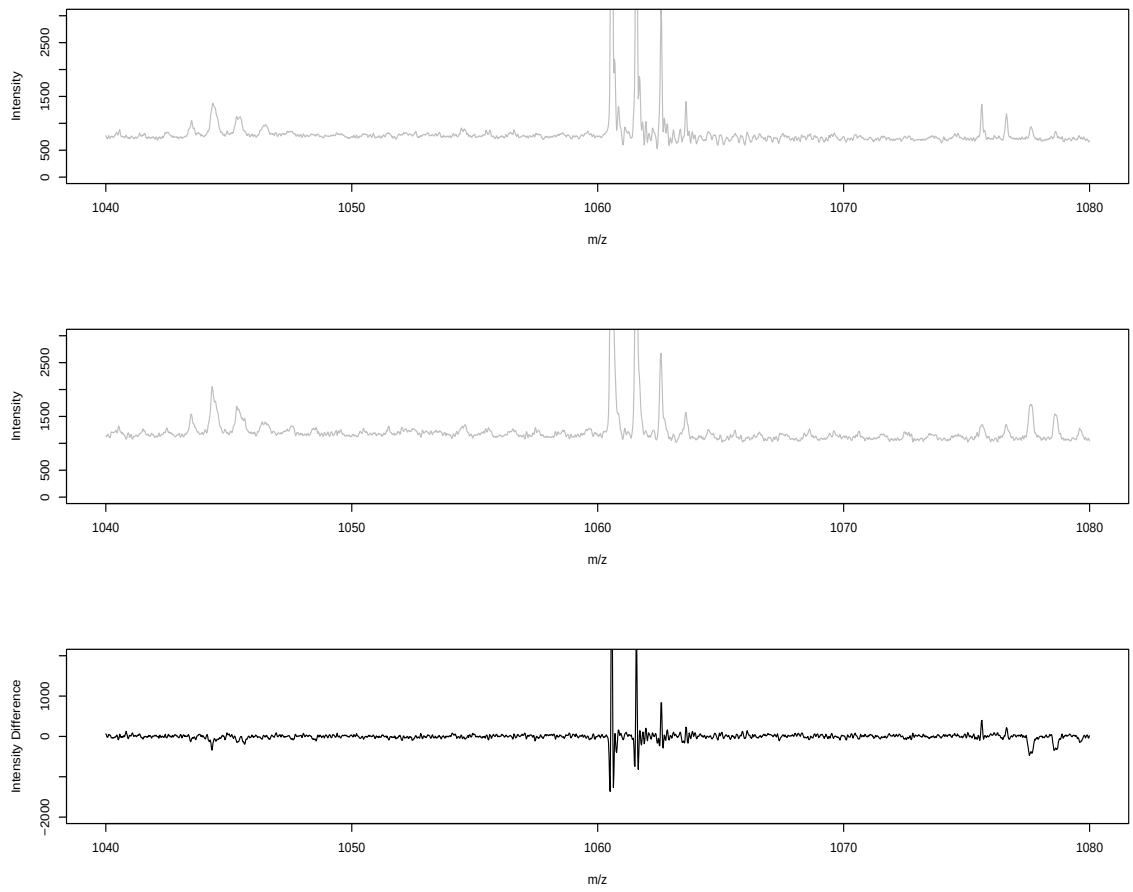

Fig. 2 Pieces of two raw spectra (top, middle) from Wu *et al.* (2003) and the difference between their baseline-corrected intensities (bottom). Note how the large peaks are slightly offset, even though they likely correspond to the same protein.

and quantification methods, as can be seen in Figure 4. We now summarize some baseline estimation procedures from previous work. A "semimonotonic" baseline is estimated by first removing a non-biological sinusoidal noise component, and then using a function of the local and monotonic minima to estimate the baseline (Baggerly *et al.*, 2003; Morris *et al.*, 2005). Since high-intensity peaks can affect baseline estimation, a simple peak finding algorithm (SPF) is used to remove peaks based on local maxima, before interpolating across the bases of removed peaks. The baseline is estimated as the local windowed minima calculated using a fixed window width, and

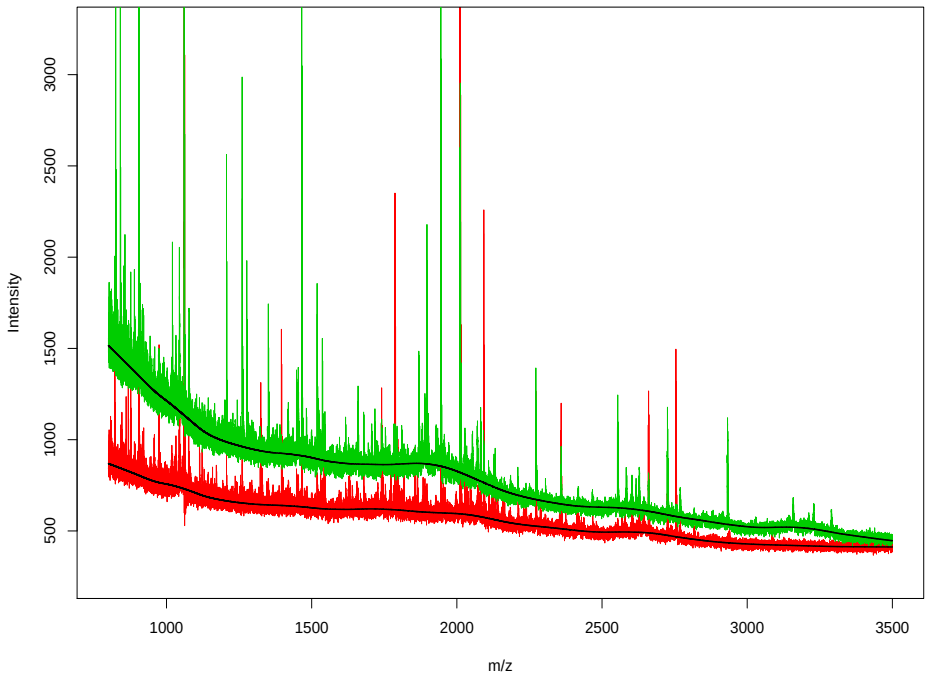

Fig. 3 Two healthy spectra (n=91380) with initial baseline estimates in black. Baselines are estimated using a "loess" smooth of 10% of the data.

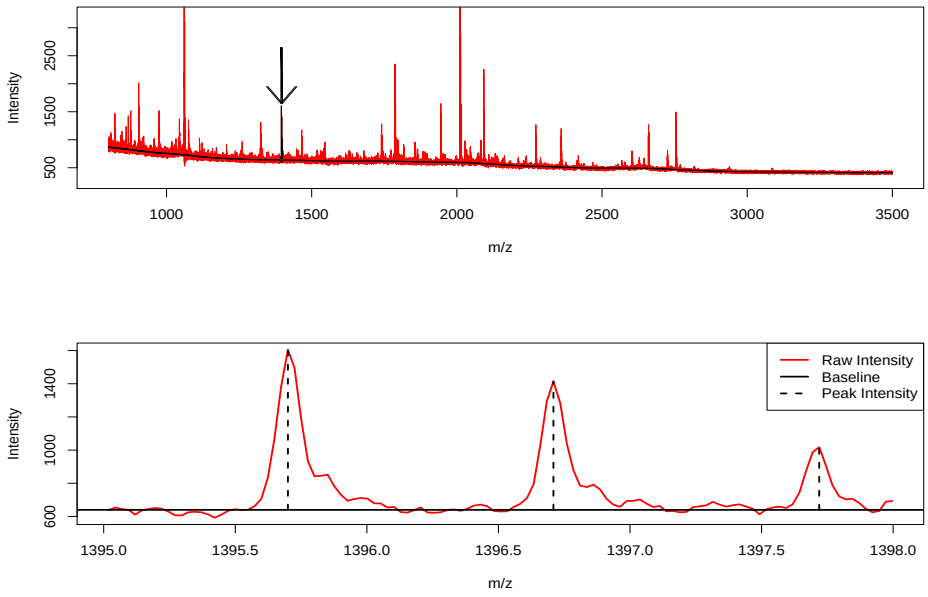

Fig. 4 Zoom-in (n=119) on baseline correction of a single healthy spectrum from MALDI ovarian cancer data set.

this baseline is removed prior to a subsequent peak detection (Coombes *et al.*, 2003). A similar method uses monotone local minima, while allowing the window size to grow smoothly across the spectrum (Kuerer *et al.*, 2004). While the local minima are easy to compute and do not require a model, one important disadvantage of using local minima is that the subsequent baseline estimates between overlapping peaks may be biased high, since the local minima will not drop to the true baseline (Dubitsky *et al.*, 2007). Further baseline estimation techniques include "loess" (Tibshirani *et al.*, 2004), local linear regression (Wagner *et al.*, 2003), local medians (Torgrip *et al.*, 2003) and cubic splines with local medians (Yu *et al.*, 2006). Locally, the baseline should be a very smooth function that is unaffected by large intensities from peaks. This suggests median-based methods and the removal of peaks in baseline estimation are more appropriate. It is unclear which method of baseline estimation is best, since this is generally an unsupervised problem.

I.2.3. Normalization

Normalization corrects for systematic differences in the total amount of protein desorbed and ionized from the sample plate (Coombes *et al.*, 2005a; Morris *et al.*, 2005), as illustrated by the large scale changes in ion intensity on the vertical axes in Figure 1. This can make the classification of spectra using ion intensities difficult. Data normalization is an important element of pattern recognition, as bias introduced by ProteinChip quality, machine performance and operator characteristics can affect the overall spectral quality (Conrads *et al.*, 2004). Several authors have normalized by dividing the intensities by the total ion current (Kozak *et al.*, 2003; Kuerer *et al.*, 2004; Li *et al.*, 2002; Morris *et al.*, 2005; Wagner *et al.*, 2003) or mean intensity (Zhu *et al.*, 2003). This method is motivated by the thought that the total ion current is

a surrogate for the total amount of protein in the sample being measured (Coombes *et al.*, 2005b). Log transformations are also used as part of a two-step normalization process. In one study, intensities were first normalized by total ion current, then followed by a logarithmic transformation (Li *et al.*, 2002). A logarithmic transformation preceded a linear transformation which mapped the 10th and 90th percentiles to 0 and 1, in another study (Tibshirani *et al.*, 2004). Normalization steps sometimes include transforming the minimum and maximum intensities in each of the spectra to 0 and 1, respectively (Baggerly *et al.*, 2004; Conrads *et al.*, 2004).

The need for normalization is best argued by observing Figure 1. Normalization is crucial for later classification purposes if the ion intensities are to be compared across samples, since the ion intensity is often used to represent the abundance of specific peptides. Clearly, comparing such abundance should be done relative to the amount of input sample, as well as controlling for other external biases.

I.2.4. Denoising

Peak detection in spectra is usually affected by random noise that is typically electronic or chemical in nature (Coombes *et al.*, 2005a), so estimation of this noise is an important step in many peak detection algorithms. Smoothing is performed in many of these cases to enhance the signal-to-noise ratio that may be used to quantify peaks. Another benefit of smoothing is a decreased number of detected false positive peaks due to large intensity values from the underlying noise component. Methods of denoising that have been proposed include supersmoothers (Tibshirani *et al.*, 2004), moving averages (Carpenter *et al.*, 2003), wavelets (Morris *et al.*, 2005) and Gaussian filters (Yu *et al.*, 2006; Zhu *et al.*, 2003). The wavelet denoising in Morris *et al.* (2005) uses a form of hard thresholding where the wavelet coefficients are computed for the

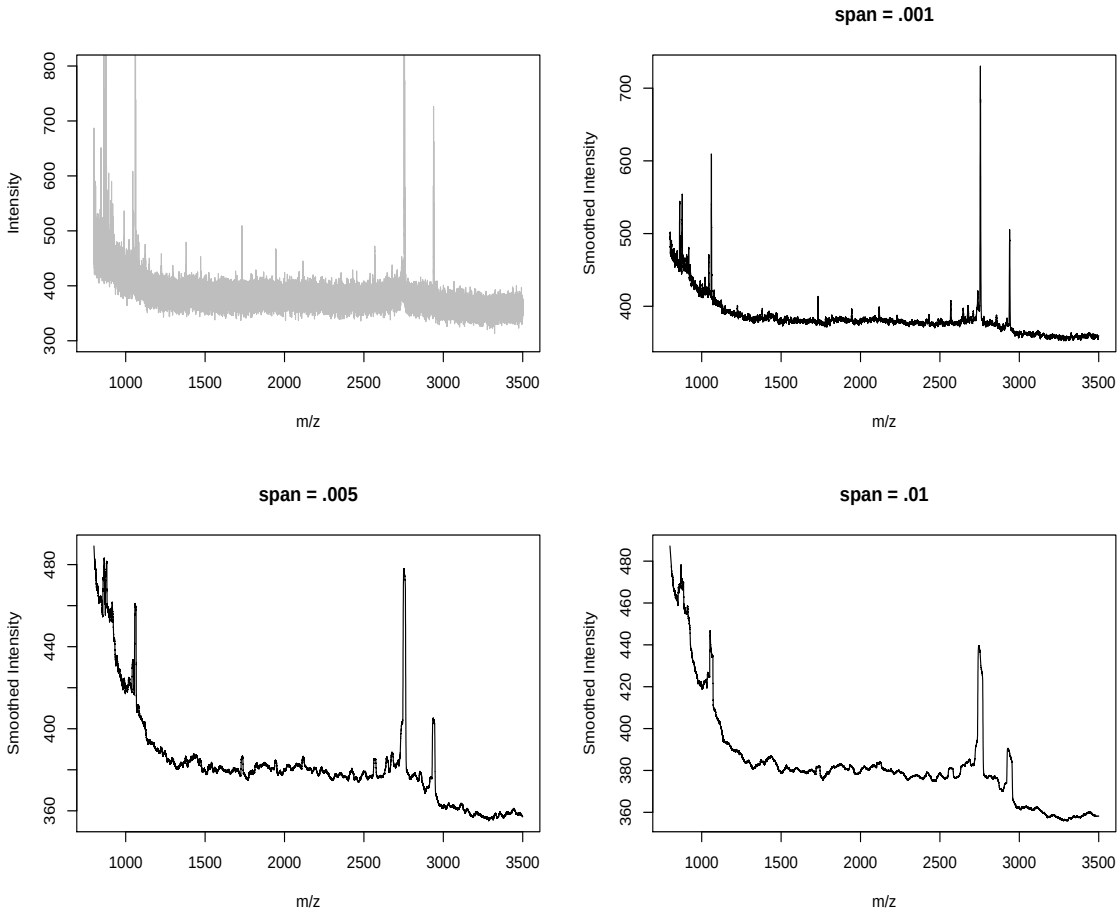

Fig. 5 Raw spectum (top-left) and spectum after smoothing using a supersmoother, using spans of .001 (top-right), .005 (bottom-left) and .01 (bottom-right).

observed spectrum, and those coefficients smaller than a threshholding parameter are set to zero. A more hardware-specific smoothing method may involve envelope extraction, since MALDI spectra obtained in reflectron mode maintain a constant distance between neighboring isotope peaks (Yu *et al.*, 2006). Figure 5 shows several degrees of smoothing for a single spectrum using a supersmoother. We must be careful not to oversmooth spectra, because important discriminatory information may be lost, as we can see several small peaks that have been razed to the baseline.

I.2.5. Peak Detection, Quantification and Matching

Peak detection and quantification is the process of identifying locations which correspond to specific proteins or peptides in the sample (Dubitsky *et al.*, 2007). Peaks in each spectrum are identified as local maxima, although, not every local maximum is a peak of interest. Some local maxima may stem from electronic or matrix noise in unsmoothed spectra. Furthermore, multiple peaks may correspond to isotopes from the same protein. Poisson distributions have been used to model the peaks in an isotopically resolved group, to determine which of the peaks is the monoisotopic peak (Breen *et al.*, 2000).

Peak matching is used to determine whether peaks that have proximate m/z values in different spectra represent the same biological feature or different biological features (Coombes *et al.*, 2005a). Two peaks that correspond to the same protein or peptide, may have (slightly) different m/z values, even after calibration. A peak alignment algorithm described in Adam *et al.* (2002) calculates a mass error score between each pair of proximate peaks, using the absolute difference in the masses, divided by the smaller mass. Each of these mass error scores is then compared to a threshold. Two peaks in different spectra are said to be "matched" if their mass error score does not exceed the threshold, and the peaks are believed to represent the presence of the same molecule in each sample.

The peak probability contrasts method (Tibshirani *et al.*, 2004) uses a peak-finding algorithm that looks for locations with intensities that are higher than those intensities at the surrounding $\pm s$ sites, and higher than the estimated average background at that location. Since peaks tend to be shorter and broader at higher values of m/z, a log-transformation of m/z is used to transform the peaks so that the peak width is roughly constant throughout the spectrum. Once peaks have been identified in all spectra, these peaks are then matched through a one-dimensional clustering

method based on location. The locations of the common peaks are defined to be the centers of common peak clusters, and peaks from individual spectra are extracted if they are within a certain distance from the cluster center.

Another method for peak detection and matching uses the mean spectrum to detect peaks in individual spectra, by first finding peaks in the mean spectrum. These peak intervals are searched in each individual spectrum, and the maximum (log) intensity is recorded for each spectrum. These maxima are retained as peaks if they exceed some minimum threshold for the signal-to-noise ratio (Morris *et al.*, 2005). If the individual spectra are well-calibrated, a single peak region in the mean spectrum indicates the presence of a set of matched peaks in various individual spectra at that locale.

A simple peak finding algorithm (SPF) compares peaks to local noise estimates using first differences between successive time points. An initial set of peaks is identified through a local maxima search based on these first differences. The median absolute value of these first differences is used as the noise estimate. Each peak whose distance from the closest local minimum is less than this noise estimate is removed from the initial set of peaks. Additional peaks whose left and right peak slopes are less than half of the noise estimate are removed from this peaklist (Coombes *et al.*, 2003; Kuerer *et al.*, 2004).

Peak width restrictions are imposed after selecting local maxima through a discrete differentiation method. Local maxima whose distance from nearest local minima does not exceed 1.5 Daltons (Da) are removed from future peak consideration. This peak width restriction is helpful in reducing the number of detected false positive peaks, since such peaks appear as narrow spikes in the spectra (Yu *et al.*, 2006). Another study combines peak detection and identification by matching peak templates from a database of known proteins to features in the data (Gras *et al.*, 1999). How-

ever, since peak identification is primarily a biological issue, it is not of particular interest to us.

I.3. Resulting Classification

After all spectra have been pre-processed and all peaks have been quantified and matched, classification based on peak information can begin. The peak information can be viewed as a set of variables (peak locations) and measurements on each of these variables (peak intensities, signal-to-noise ratios, etc.), and traditional statistical methods can be used in the disease classification. The number of peaks used in classification is a matter of feature selection; some common methods for selecting locations of significance have utilized two-sample t-statistics (Baggerly *et al.*, 2003; Wu *et al.*, 2003) and Wilcoxon tests (Sorace and Zhan, 2003; Zhukov *et al.*, 2003). A two-sample t-test is computed at each location that compares the average intensity for the cancerous spectra and the average intensity for the non-cancerous spectra. Those locations yielding large magnitudes of the test statistic are investigated as potential discriminating locations. Use of the Wilcoxon test has a similar application.

Another method for the feature selection uses split points based on peak height. At each location, a height is selected that maximally discriminates the peak intensities between the groups to be classified (Tibshirani *et al.*, 2004). Other methods for feature selection include random forests (Wu *et al.*, 2003), fitness tests based on Euclidean distances (Conrads *et al.*, 2004), Unified Maximum Separability Analysis (UMSA) (Li *et al.*, 2002), area under the curve (AUC) (Qu *et al.*, 2002; Adam *et al.*, 2002) and binomial distributions based on desired specificity (Coombes *et al.*, 2004). One previous study boasted of high sensitivity and specificity in ovarian cancer classification (Petricoin *et al.*, 2002a), but the lack of reproducibility of these results is a

cause for concern (Baggerly *et al.*, 2004).

An extensive study (Wu *et al.*, 2003) compared statistical methods on the aforementioned ovarian cancer data peaklists to evaluate the prediction error of each of these methods. These methods included linear and quadratic discriminant analysis, nearest neighbors, bagging, boosting, support vector machines and random forests. Random forests consistently performed well as the number of biomarker selections changed; the misclassification error rates were about .10, on average, depending on the number of features selected. The variables (locations) were selected by examining the t-statistics calculated at each location. Each t-statistic was computed from the normalized differences between the mean intensity for each group. The use of t-statistics and Wilcoxon tests to identify locations for feature selection has been questioned, based on the belief that cancer patients need not differ from normal patients in the same way (Coombes *et al.*, 2004).

Many of the pre-processing methods described in this chapter rely on subjective choices of the modeler, some of which may prevent reproducibility of results. While these choices cannot be altogether avoided, we present an approach where one of our focuses is to allow the data to be ultimate arbiter for model and parameter choice instead of the modeler. In the next chapter, we present our method of modeling the intensities from the spectra.

CHAPTER II

METHOD

II.1. Model

Let Y_t represent the observed intensities which are modeled as

$$Y_t = f_t + Z_t, \qquad t = 1, \ldots, n, \tag{2.1}$$

where t indexes the locations within spectra, f_t is the baseline, and Z_t is the baseline-corrected intensity. Our goal is to separate f_t and Z_t from Y_t. We assume that $f_t = f(t)$ for a smooth function f and assume that Z_t is a random variable from a mixture distribution, where the component distributions may vary based on application. While analyzing data from Wu *et al.*, (2003), we have found that the mixture of a single normal and several normal-exponential sums to suitably model the intensities in each spectrum. The details of the model that we describe in this chapter will be specific to this motivating data set with a more general discussion of other component choices to follow.

Assume that a baseline-corrected intensity, Z_t, is generated from a mixture of m components which is comprised of a single normal component and $m-1$ normal-exponential sum components. These densities may be written as $\phi(z_t; \mu, \sigma^2)$ and $\psi(z_t; \mu, \sigma^2, \theta_j, \alpha_j), j = 1, \ldots, m-1$, respectively, where μ and σ^2 denote mean and variance of the normal distribution. The parameters θ_j and α_j denote the rate and shift parameters of the exponential distribution, respectively, corresponding to the jth normal-exponential component, where $\alpha_1 < \alpha_2 < \ldots < \alpha_{m-1}$. Let the mixing probabilities be denoted as $\pi_1, \ldots, \pi_m$, where π_1 is the mixing probability for the normal

component, and $\pi_2, \ldots, \pi_m$ are the mixing probabilities for the normal-exponential components, $\psi(z_t; \mu, \sigma^2, \theta_j, \alpha_j), j = 1, \ldots, m-1$, respectively. The density function of Z_t can then be written as

$$p_\theta(z_t) = \pi_1 \phi(z_t; \mu, \sigma^2) + \sum_{j=2}^{m} \pi_j \psi(z_t; \mu, \sigma^2, \theta_{j-1}, \alpha_{j-1}),$$

where the mixing probabilities satisfy $\sum_{j=1}^{m} \pi_j = 1$.

The normal density has the form

$$\phi(z; \mu, \sigma^2) = \frac{e^{-(z-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}, \qquad -\infty < z < \infty,$$

and the normal-exponential sum convolution is

$$\psi(z; \mu, \sigma^2, \alpha, \theta) = \Phi\left\{\frac{(z-\alpha) - (\mu + \theta\sigma^2)}{\sigma}\right\} \theta e^{\left\{-\theta\left(z-\alpha-\mu-\frac{\theta\sigma^2}{2}\right)\right\}}. \tag{2.2}$$

To derive the convolution, we let $\phi(\cdot)$ and $\eta(\cdot)$ denote the normal and shifted exponential densities, respectively, and write the convolution as

$$\begin{aligned}\psi_Z(z) = \psi_{X+Y}(x+y) &= \int_{-\infty}^{z-\alpha} \phi_X(w)\eta_Y(z-w)dw \\ &= \int_{-\infty}^{z-\alpha} \frac{e^{-(w-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\, \theta e^{-\theta(z-w-\alpha)} dw.\end{aligned}$$

Note that $\eta(\cdot)$ refers to the $\alpha-$shifted exponential density, so $z - w > \alpha$ and it follows that, for fixed z, $w < z - \alpha$. We first combine terms in the exponential, and then follow by completing the square (in w) in the kernel of the normal density, which

yields

$$\begin{aligned}
\psi_Z(z) &= \int_{-\infty}^{z-\alpha} \frac{e^{-(w-\mu)^2/(2\sigma^2)}}{\sigma\sqrt{2\pi}}\,\theta e^{-\theta(z-w-\alpha)}dw \\
&= \int_{-\infty}^{z-\alpha} \frac{1}{\sigma\sqrt{2\pi}}\,\theta \exp\left(-\frac{w^2}{2\sigma^2} + \frac{2w\mu}{2\sigma^2} - \frac{\mu^2}{2\sigma^2} - \theta z + \theta w + \theta\alpha\right) dw \\
&= \int_{-\infty}^{z-\alpha} \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{w^2}{2\sigma^2} + \frac{2w(\mu+\theta\sigma^2)}{2\sigma^2} - \frac{(\mu+\theta\sigma^2)^2}{2\sigma_t^2}\right\} dw \\
&\times \theta \exp\left\{-\theta z + \theta\alpha + \frac{(\mu+\theta\sigma^2)^2}{2\sigma^2} - \frac{\mu^2}{2\sigma^2}\right\}.
\end{aligned}$$

Finally, we simplify the last two terms in the exponential and perform some rearrangement, which gives

$$\psi_Z(z) = = \Phi\left\{\frac{(z-\alpha)-(\mu+\theta\sigma^2)}{\sigma}\right\}\theta\exp\left\{-\theta\left(z-\alpha-\mu-\frac{\theta\sigma^2}{2}\right)\right\},$$

where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution.

One of the attractive properties of using the normal-exponential convolution is that it can simultaneously model the peaks from the underlying biological signal with noise from the mass spectrometer. Several graphs of this density are provided in Figure 6. Comparison of the top two pictures shows the effect of an increased noise variance, σ^2. Since the variance of the convolution is $\sigma^2 + \frac{1}{\theta^2}$, we see that a smaller proportion of this variance is contributed by the exponential variance, $\frac{1}{\theta^2}$, in the top-right graph. A change in the shift of the exponential part, α, is illustrated by comparing the top-right and bottom-left graphs. Not surprisingly, we see no change in the shape of the graphs – only a shift. Finally, the effect of a decreased θ is shown in the bottom two graphs. Since we have chosen θ to represent the rate parameter of the exponential, the mean and variance of the exponential increase for decreasing values of θ. This is evident in the bottom-right graph with increased spread and right

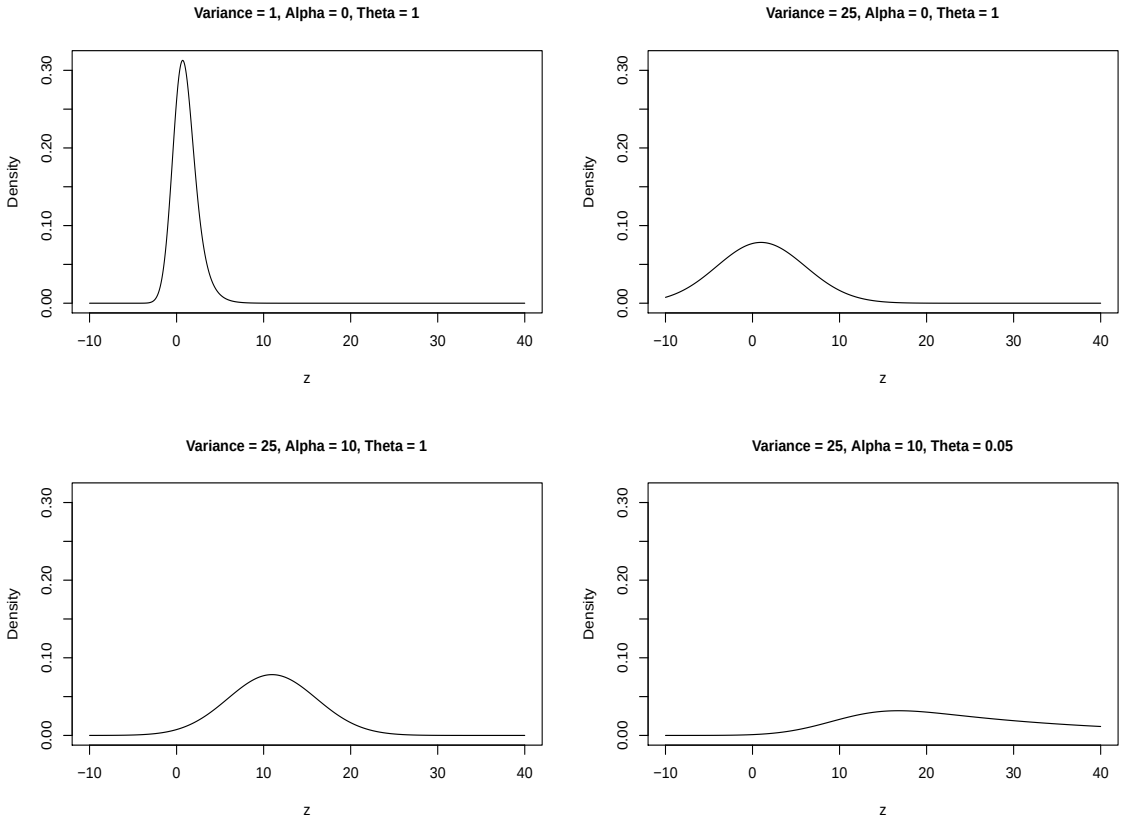

Fig. 6 Graphs of the normal-exponential sum convolution density for varying noise variance, peak minimum and mean.

skewness.

We now present some graphical support for our choice of the normal and normal-exponential sum components in the analysis of the data from our motivating MALDI data set. We first consider a smooth baseline estimate using "loess" in Figure 7, and using the baseline-corrected intensities in the entire spectrum, we compute an upper fence at $Q_3 + 3 \times IQR$ and examine the baseline-corrected intensities inside (less than) this upper fence in Figure 8. The histogram confirms that the majority of the intensities are located near the baseline through a dominating peak at 0, so we restrict $\mu = 0$ in our data generating model, and assign a zero-mean normal density to represent random error. From this figure, this choice of normal component is

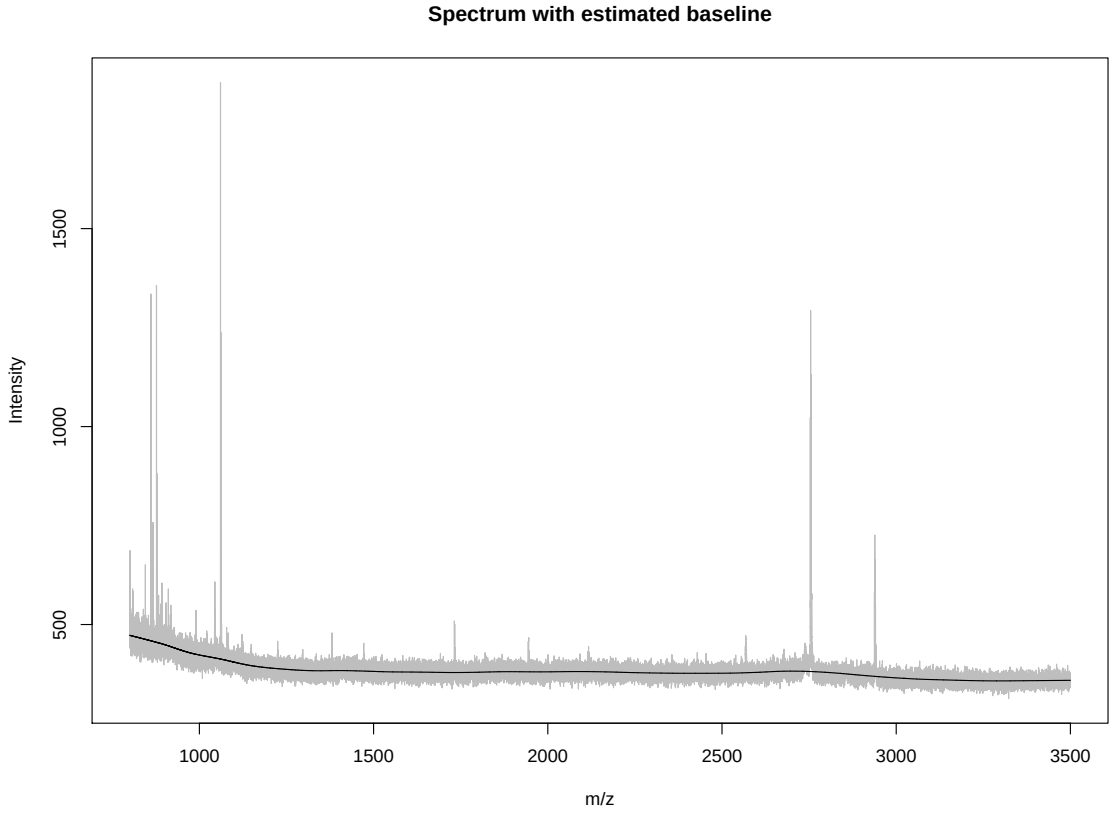

Fig. 7 Raw mass spectrum from our motivating MALDI data set with initial baseline estimate.

well-founded. The heavy right skewness of the intensities above (greater than) the upper fence in Figure 9 suggests that an exponential component or other right-skewed density may be appropriate. However, since both peak and non-peak intensities are affected by electronic noise in the mass spectrometer, modeling peak intensities as a sum of normal and exponential components seems more appropriate.

Assuming independence of the data, we can find the parameters of our mixture model by maximizing the log-likelihood of the observed data

$$\begin{aligned}\ell(\Theta) = &\sum_{t=1}^{n}\{\log \pi_1 + \log \phi(y_t - f_t; 0, \sigma^2)\} \\ &+ \sum_{t=1}^{n}\sum_{j=2}^{m}\{\log \pi_j + \log \psi(y_t - f_t; 0, \sigma^2, \theta_{j-1}, \alpha_{j-1})\}.\end{aligned} \tag{2.3}$$

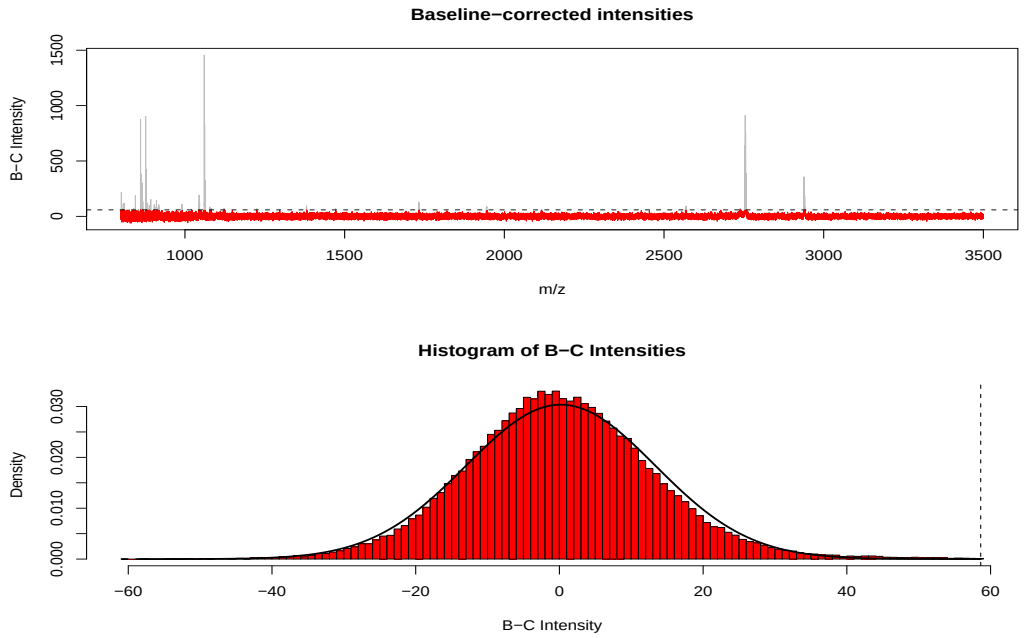

Fig. 8 Baseline-corrected intensities near zero and corresponding histogram. The dotted line represents the upper fence and the solid line represents the fitted normal density.

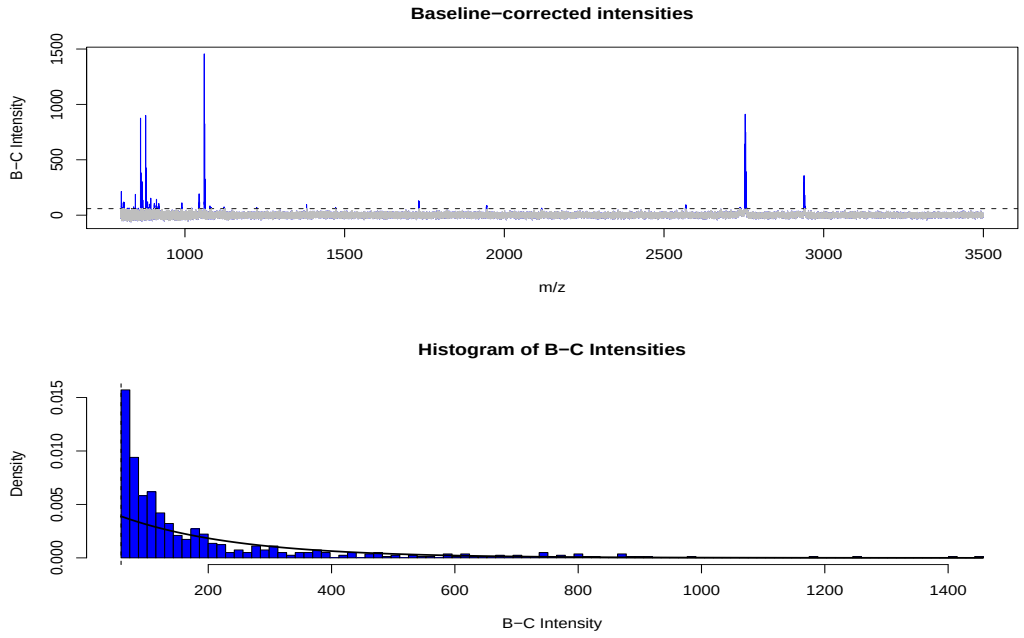

Fig. 9 Potential peak intensities and corresponding histogram. The dotted line represents the upper fence and the solid line represents an (unsuitable) exponential density.

While the assumption of independent observations may not be entirely true, using a diagonal covariance matrix for large datasets is justifiable (and convenient) as shown in Bickel. Thus, we will assume that all of the intensities are independent throughout this document.

II.2. EM Algorithm

Since direct maximization of (2.3) is difficult, we use the EM algorithm to maximize a function which minorizes the observed data log-likelihood, up to a constant. A graph of the relationship between these two functions is shown in Figures 10 and 11. To motivate our estimation method, let us introduce a latent (unobservable) indicator U_t denoting which distribution Z_t is from. Then $\pi_j = P(U_t = j)$ and Z_t is generated from $\phi(0, \sigma^2)$ if $U_t = 1$, or $\psi(0, \sigma_t^2, \theta_{j-1}, \alpha_{j-1})$ if $U_t = j$, for $2 \leq j \leq m$. The joint density of U_t and Z_t is

$$\{\pi_1 \phi(z_t; 0, \sigma^2)\}^{I(U_t=1)} \prod_{j=2}^{m} \{\pi_j \psi(z_t; 0, \sigma^2, \theta_{j-1}, \alpha_{j-1})\}^{I(U_t=j)}.$$

where $I(U_t = j)$ equals 1 if $U_t = j$, and equals 0, otherwise.

Introducing such a latent indicator is a standard method of parameter estimation for mixture models. Treating the latent variable as missing, the EM (Expectation-Maximization) algorithm can be applied to compute the maximum likelihood estimator. In the E-step of the algorithm we calculate the conditional expectation of the log-likelihood of the complete data ($\boldsymbol{Y}$, $\boldsymbol{U}$), given the observed data, $\boldsymbol{Y}$ and the current guess of parameter values, denoted Θ_i. In the M-step, we maximize the conditional expectation obtained in the E-step, and we iterate the E- and M-steps until convergence. Under mild conditions, the observed data likelihood increases after each iteration of the algorithm and the algorithm converges to a local maximum of the

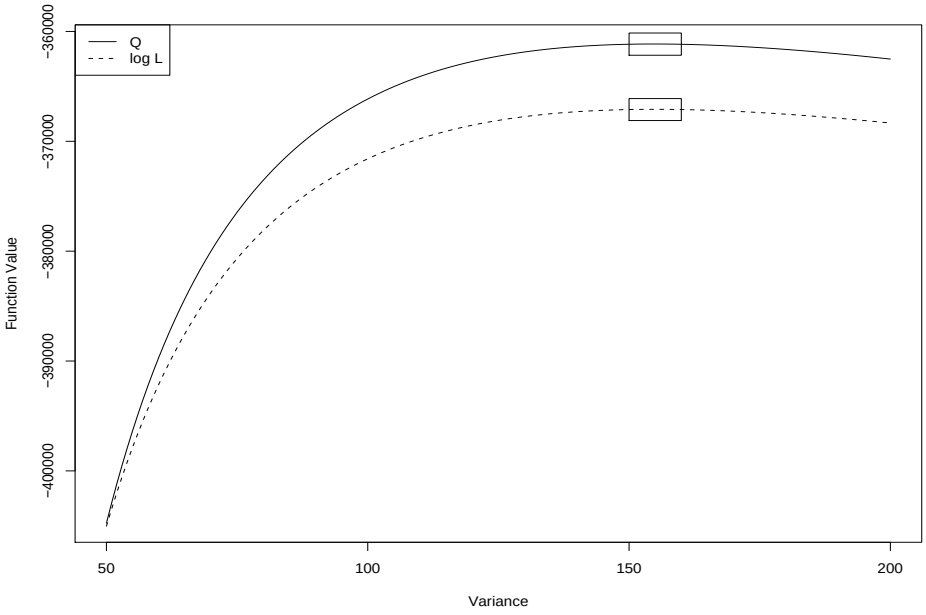

Fig. 10 Graph showing the relationship between the observed data log-likelihood and conditional expected log-likelihood for the normal component variance.

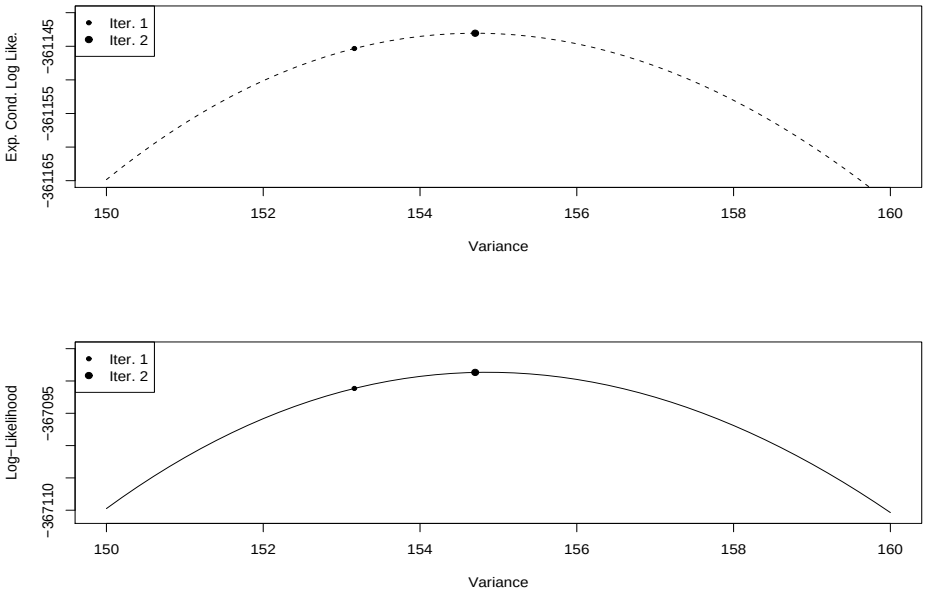

Fig. 11 Zoom in of Figure 10 showing update of the variance.

likelihood.

II.2.1. E-step

The likelihood of the complete data $(\boldsymbol{Y}, \boldsymbol{U})$ is

$$L(\Theta|\boldsymbol{Y}, \boldsymbol{U}, \mathbf{f}) = \prod_{t=1}^{n} \left[\{\pi_1 \phi(y_t - f_t; 0, \sigma^2)\}^{I(U_t=1)} \prod_{j=2}^{m} \{\pi_j \psi(y_t - f_t; 0, \sigma^2, \theta_{j-1}, \alpha_{j-1})\}^{I(U_t=j)} \right].$$

Thus, the log-likelihood is

$$\begin{aligned} \ell(\Theta) = &\sum_{t=1}^{n} I(U_t = 1) \left\{ \log \pi_1 + \log \phi(y_t - f_t; 0, \sigma^2) \right\} \\ &+ \sum_{t=1}^{n} \sum_{j=2}^{m} I(U_t = j) \left\{ \log \pi_j + \log \psi(y_t - f_t; 0, \sigma^2, \theta_{j-1}, \alpha_{j-1}) \right\}. \end{aligned} \tag{2.4}$$

Let $\Theta_i = (\pi_1^i, \sigma^i, \{\pi_j^i, \theta_{j-1}^i, \alpha_{j-1}^i\}, j = 2, \ldots, m)$ respresent current estimates of the parameters. To obtain the expected log-likelihood, we need to compute $\hat{\pi}_{j,t} = P(U_t = j|\boldsymbol{Y}, \mathbf{f}; \Theta_i)$, called "responsibilities". The term "responsibility" is used in the sense that the jth density is responsible for generating the intensity at location t, with a certain probability or level of "responsibility". The responsibilities can be computed using Bayes Theorem as follows:

$$\hat{\pi}_{1,t} = \frac{\pi_1^i \phi(Y_t - f_t; 0, \sigma^i)}{\pi_1^i \phi(Y_t - f_t; 0, \sigma^i) + \sum_{k=2}^{m} \pi_k^i \psi(Y_t - f_t; 0, \sigma^i, \theta_{k-1}^i, \alpha_{k-1}^i)}, t = 1, \ldots, n \tag{2.5}$$

and

$$\hat{\pi}_{j,t} = \frac{\pi_j^i \psi(Y_t - f_t; 0, \sigma^i, \theta_{j-1}^i, \alpha_{j-1}^i)}{\pi_1^i \phi(Y_t - f_t; 0, \sigma^i) + \sum_{k=2}^{m} \pi_k^i \psi(Y_t - f_t; 0, \sigma^i, \theta_{k-1}^i, \alpha_{k-1}^i)}, j = 2, \ldots, m, \ t = 1, \ldots, n. \tag{2.6}$$

It immediately follows that the component responsibilities for a given location sum to 1. With these responsibilities, we obtain from (2.4) that

$$\begin{aligned} Q(\Theta|\Theta_i) &\equiv E(\ell(\Theta)|\boldsymbol{Y}, \mathbf{f}; \Theta_i) \\ &= \sum_{t=1}^{n} \hat{\pi}_{1,t} \left\{ \log \pi_1 + \log \phi(y_t - f_t; 0, \sigma^2) \right\} \\ &+ \sum_{t=1}^{n} \sum_{j=2}^{m} \hat{\pi}_{j,t} \left\{ \log \pi_j + \log \psi(y_t - f_t; 0, \sigma^2, \theta_{j-1}, \alpha_{j-1}) \right\}. \end{aligned} \tag{2.7}$$

II.2.2. M-step

To perform the M-step, we need to maximize the expected log-likelihood function given in (2.7) over each of the parameters. Closed-form solutions for the updates may exist depending on the choices of the component distributions. The normal-exponential convolution prevents us from obtaining closed-form updates for most of the parameters in the model. Using the fact that $\sum_{j=1}^{n} \pi_j = 1$, the updates for the mixing probabilities are

$$\hat{\pi}_j = \sum_{t=1}^{n} \frac{\hat{\pi}_{j,t}}{n}.$$

II.3. Number of Components and Initial Values

We now consider the issue of determining the number of mixture components and initial values for our mixture model. Estimating the number of mixture components has been considered by Furman and Lindsay (1994), Karlis and Xekalaki (1999) and Titterington *et al.* (1985), although there does not appear to be much work in the area of mixtures from components of different densities, as we are proposing. It is more difficult to apply the moment-based procedure in Furman and Lindsay (1994) to our problem, since each additional normal-exponential convolution component requires three additional moment equations to estimate the additional π, α and θ. We fit each

spectrum over a range of m, and use AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) to determine the optimal number of components, by choosing the m for which these criteria are minimized.

In addition to the number of components, the EM Algorithm also requires initial estimates of the parameter values to be optimized over. A study of initial values for the EM algorithm with normal mixtures was compared in Karlis and Xekalaki (2003). Many of the methods reviewed used a grid of starting values for the initial parameter estimates, so that at least one of the sets of starting values will lead to the global maximum. Among those that consistently performed well, the moment estimation proposed in Furman and Lindsay (1994), becomes unwieldy even for $m = 3$, since we must solve a moment equation for each initial estimate of π_2, π_3, σ^2, α_1, α_2, θ_1 and θ_2. Other methods that performed well used random or user-defined starting points.

We now describe our method for selecting initial values, where we make use of prior information about the spectra. Assuming that our baseline is known, we believe that intensities below this baseline are solely attributed to the electronic noise in the machine and, thus, are generated only from the normal error component. Under this assumption, it is sensible to restrict $\mu = 0$, so we estimate the noise variance using these negative baseline-corrected intensities as $\hat{\sigma}^2 = \sum_{z_-} z_-^2/n_-$, where n_- represents number of intensities below the baseline. Once we have estimated the variance, we compute an upper fence at 3σ and initially classify any baseline-corrected (b-c) intensities less than this upper fence as members of the error component and $\pi_1 = \sum_{z_t} I(Z_t \leq 3\sigma)/n$. If $m = 2$, then $\pi_2 = 1 - \pi_1$, and we then initially classify any intensities greater than this upper fence as members of the peak component and use these intensities to compute α_1 and θ_1. If $m > 2$, we use a k-means algorithm to find $k = m-1$ means among the baseline-corrected intensities greater than 3σ. We use the intensities that are classified to cluster j $\{j : 1 \leq j \leq m-1\}$ to compute π_{j+1}, α_j, θ_j,

where the clusters are labeled in increasing order to correspond with the $m-1$ means in increasing order. A histogram of baseline-corrected intensities for a single spectrum with initial component classification is illustrated in Figures 12 and 13 using $m = 4$. The α and θ chosen for each initial density are those that maximize the log-likelihood of the baseline-corrected intensities assigned to the respective component.

There is an important distinction in using only the negative baseline-corrected intensities to compute the variance, while assigning all of the intensities less than the upper fence to the normal component. The negative b-c intensities are observations believed to be generated by only the normal component; that is, we do not expect to see (m)any peaks with negative intensities, assuming that we have estimated our baseline well. Including intensities above the baseline in this variance estimate, may include many small intensities generated from a peak density in addition to those from the error component. The use of the upper fence serves the same purpose for estimating the parameters in the peak densities, as we are very unlikely to see electronic noise from the machine generate intensities larger than 3σ.

For each spectra, we estimate the baseline with a "loess" smooth using the closest 2% of the data, which provides for a very smooth baseline. Since the AIC and BIC may not be monotone in the number of mixture components selected, we fit $m = 2, \ldots, 12$ components to each spectrum using the k-means to initially determine the locations of peak clusters and their initial values, as described above. We then begin iterating the E- and M-steps according to the EM Algorithm, until the relative increase in the log-likelihood after a single iteration is less than 10^{-8}. This information is summarized for a single spectrum from our motivating MALDI dataset, in Tables 1 and 2. The number of components selected by these criteria will be used in the next chapter.

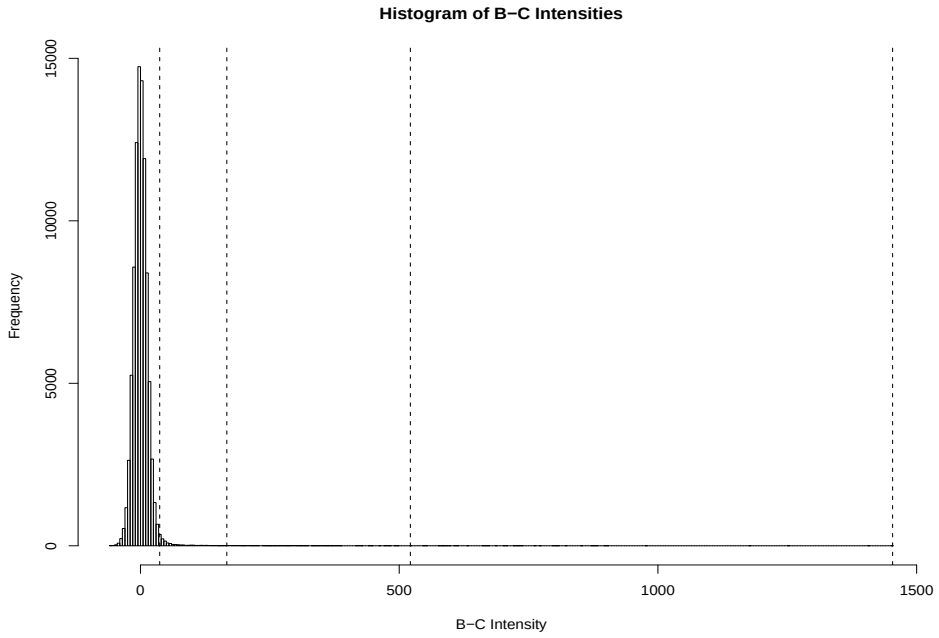

Fig. 12 Histogram and initial component assignment ($m = 4$) for a single spectrum. Vertical lines indicate a change in component.

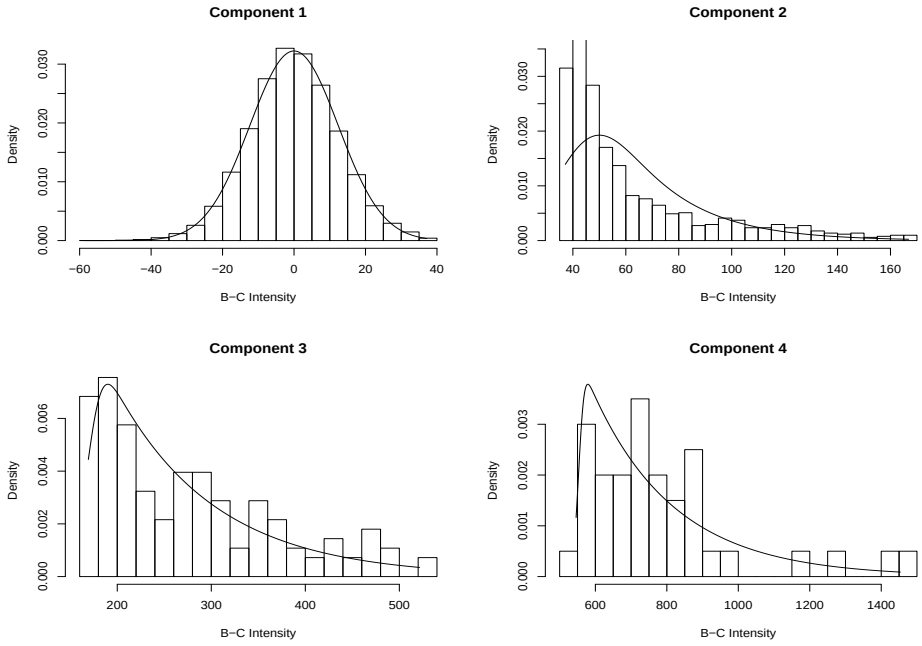

Fig. 13 Histograms and initial fitted mixture densities for the components in Figure 12.

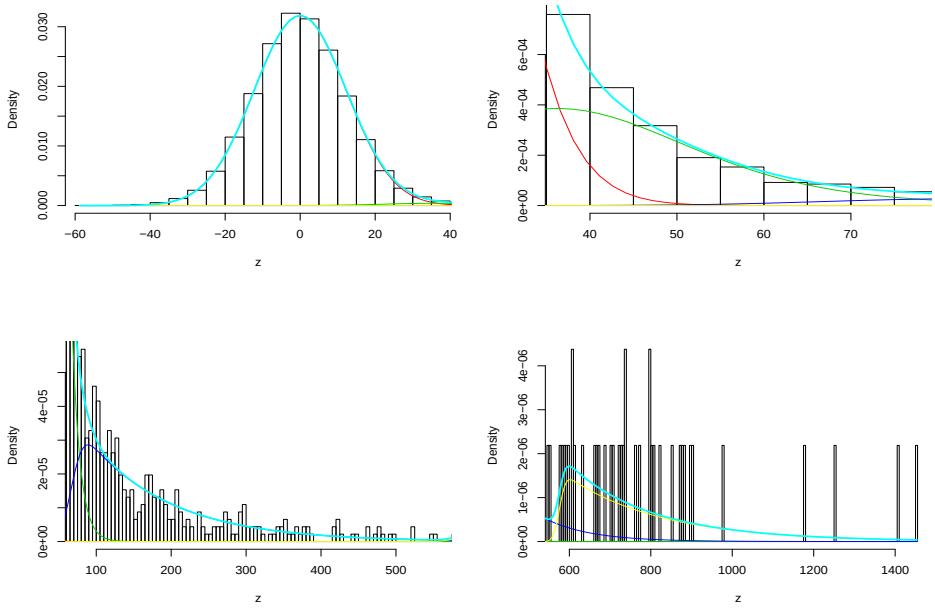

Fig. 14 Histogram and fitted mixture density ($m = 4$) for a single spectrum, with α unrestricted.

II.4. Goodness-of-Fit

In Figures 14 and 15, we provide histograms of the baseline-corrected intensities for a single spectrum with the fitted mixture densities for unrestricted α and $\alpha = 0$, respectively. In Figure 14, the top-left, top-right, bottom-left and bottom-right plots are histograms corresponding to intensities likely generated from components 1, 2, 3 and 4, respectively. In Figure 15, the top-left figure shows a histogram of all baseline-corrected intensities with fitted mixture density in cyan. The remaining three plots are enlarged areas of the top-left plot corresponding to intensities likely generated from components 1 (red), 2 (green) and 3 (blue), in the top-right, bottom-left and bottom-right graphs, respectively. The log-likelihood, AIC and BIC for this spectrum are displayed in Tables 1 and 2 for a range of m. Tables 1 and 2 suggest that, from a modeling perspective, there is some benefit to allow $\alpha > 0$, insofar as the AIC and BIC are concerned. This is generally the case for other spectra in our MALDI data,

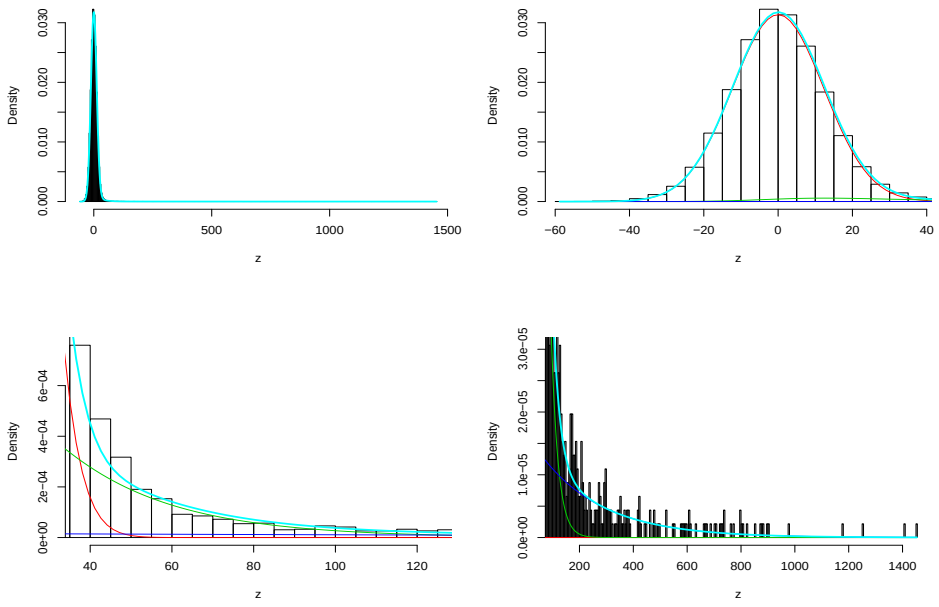

Fig. 15 Histogram and fitted mixture density ($m = 3$) for a single spectrum, with α=0.

as well. However, if we consider the fact that peaks appear to rise from the baseline, restricting $\alpha = 0$ is an intuitive selection. In Table 3, we provide the parameter estimates for the fitted mixture models depicted in Figures 14 and 15. In Tables 4 and 5 we provide the number of mixture components selected for all of the spectra with $\alpha \neq 0$ and $\alpha = 0$, respectively.

To investigate the efficacy of the model fit, we compared the empirical distribution function to the cumulative distribution function from the fitted mixture model to the baseline-corrected intensities. The empirical distribution function is calculated as $\hat{F}(z_t) = \frac{1}{n}\sum_{z_t} I(Z_t \leq z_t)/n$, and the cumulative distribution function is calculated as $F(z_t) = \int_{-\infty}^{z_t} p_\theta(z_t)dz_t$. Plots comparing the empirical and cumulative distribution functions for two spectra appear in Figure 16, along with graphs of the spectra and their noise bounds. The noise bounds provided in this figure are computed as $f_t \pm 3\sqrt{\sigma_t^2}$, where f_t is the baseline estimate at t_i and σ_t^2 is the normal component variance at t_i. It is interesting to note that the adquate fit of the model in the top-left

Table 1. AIC and BIC values for a single spectrum from Wu *et al.* with $\alpha \neq 0$.

m	Log-likelihood	AIC	BIC	Iterations
2	-367305.3	734618.6	734656.3	41
3	-366996.6	734007.8	734073.8	102
4	-366978.5	733977.0	**734071.3**	146
5	-366967.9	**733961.9**	734084.4	310
6	-366967.2	733966.4	734117.2	306
7	-366966.8	733971.6	734150.6	288
8	-366963.8	733971.6	734178.9	150
9	-366961.4	733972.8	734208.4	68
10	-366960.7	733977.5	734241.3	49

Table 2. AIC and BIC values for a single spectrum from Wu *et al.* with $\alpha = 0$.

m	Log-likelihood	AIC	BIC	Iterations
2	-367305.3	734616.6	734644.9	12
3	-367030.6	**734071.2**	**734118.3**	68
4	-367030.6	734075.2	734141.1	64
5	-367030.6	734079.2	734164.0	61
6	-367028.5	734079.1	734182.7	101

graph, is, in part, due to the relatively small change in the variance of the noise across the spectrum. The larger change in noise variance in the bottom-right graph results in a serious compromise of the model fit, as evidenced in the bottom-right graph. Figure 16 suggests that the use of a nonconstant estimate of the error variance would be justified.

While the nonconstant variance issue is not of particular interest to us at this point, we attempted to address it in the spirit of model fitting. We divided each spectrum into ten pieces so that there were roughly the same number of points in each

Table 3. Parameters for the models shown in Figures 14 ($m = 4$) and 15 ($m = 3$).

	π_1	σ^2	π_2	α_1	$1/\theta_1$	π_3	α_2	$1/\theta_2$	π_4	α_3	$1/\theta_3$
$m = 4$	.9811	150.8	.0145	28.5	10.6	.0040	67.6	111.5	.0004	576.2	232.7
$m = 3$	.9741	152.5	.0232	0	37.2	.0027	0	286.7	–	–	–

Table 4. Optimal number of components selected by AIC and BIC with $\alpha \neq 0$.

	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$	$m=9$	$m=10$	$m=11$	$m=12$
AIC	0	0	0	2	2	10	13	22	16	15	9
BIC	0	0	7	10	17	25	24	5	1	0	0

Table 5. Optimal number of components selected by AIC and BIC with $\alpha = 0$.

	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$
AIC	0	5	78	5	1
BIC	0	14	75	0	0

section. Note that there is a larger concentration of points at small values of m/z than large values of m/z, which is particularly convenient since the variance appears to change more at those small values. We also refit each mass spectrum assuming that $\alpha = 0$ and $\alpha \neq 0$ for $m = 2, \ldots, 8$, as we had done previously where we had assumed that the error variance was constant. We want to point out that we only allowed the variance to change across the range of m/z. The number of components in each of these ten sections remain the same, as well the associated parameter values. Interestingly, the optimal number of components selected using AIC and BIC were reduced under the assumption of nonconstant variance. This may suggest that some of the peak components were modeling parts of the spectrum with larger error variance rather than actual peaks. A summary of the optimal number of components selected under the nonconstant variance assumption is provided in Tables 6 and 7 for $\alpha \neq 0$ and $\alpha = 0$, respectively. Graphs of the two spectra from Figure 16 are plotted again in Figure 17 under the nonconstant variance assumption with details from the model fit.

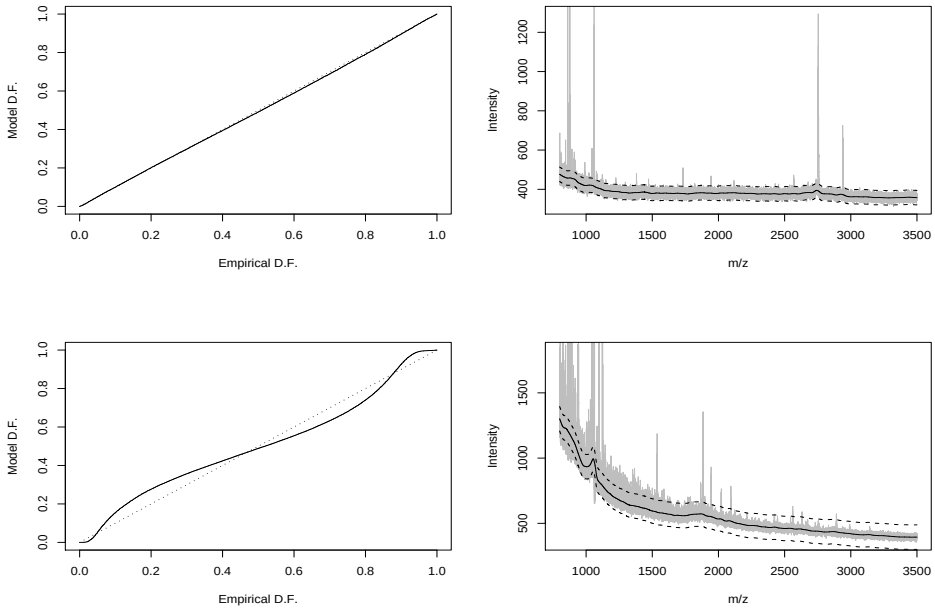

Fig. 16 Graphs of empirical versus cumulative distribution function and mass spectra with noise bands for two spectra. Note the relationship between the nonconstant variance and the discordance of the distribution functions.

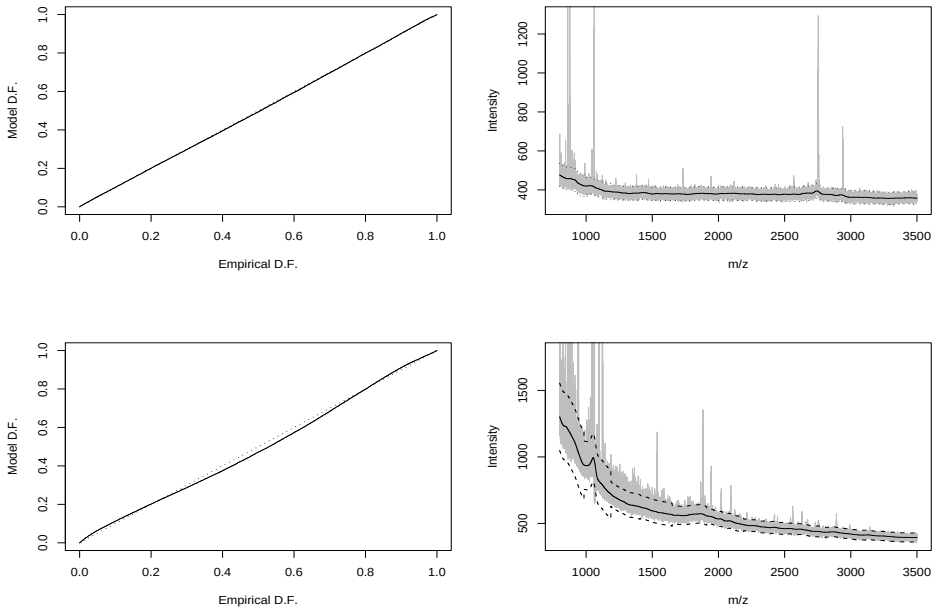

Fig. 17 Graphs of empirical versus cumulative distribution function for the two spectra from Figure 16 using nonconstant variance.

Table 6. Optimal number of components selected by AIC and BIC with $\alpha \neq 0$ and nonconstant variance. The number of components and associated parameter values are constant across the range of m/z.

	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
AIC	0	0	9	26	13	28	13
BIC	0	3	38	32	13	3	0

Table 7. Optimal number of components selected by AIC and BIC with $\alpha = 0$ and nonconstant variance. The number of components and associated parameter values are constant across the range of m/z.

	$m=2$	$m=3$	$m=4$	$m=5$	$m=6$	$m=7$	$m=8$
AIC	0	1	67	10	8	2	1
BIC	0	6	80	2	1	0	0

II.5. Other Components

To this point, we have restricted our consideration for a suitable model choice to a combination of normal and normal-exponential densities, however other component choices could be considered, depending on application. We considered the normal-exponential not only because of its good fit, but also because it is practically sensible to model peaks that are truly recorded with noise. For this reason, we feel that our model choice has intuitive appeal, as well. Nevertheless, lognormal and Weibull densities possess similar positive skewness properties and may yield similar fits when compared with our normal-exponential component for our data. However, in the case of the lognormal, it may seem unnatural to model error intensities with logarithms of peak intensities. The Weibull density may involve unnecessary extra parameters and complication, when we may be able to model the spectra effectively using a more parsimonious approach. Careful inspection of Figure 15 shows that our normal-exponential densities fit well, even with $\alpha_1 = \alpha_2 = 0$. This is evidence that more complicated models may not be necessary. It also may be more logical from a clas-

sification standpoint to model intensities *across* spectra, rather than within spectra, as it is our hope that the intensities at a certain location for healthy spectra may be modeled differently than the intensities from cancerous spectra. While this approach may be preferred with an eye towards classification purposes, we still use the peak densities as the main part of our peak detection algorithm after pre-processing and consider classification issues later.

In this chapter, we have described our method for modeling the baseline-corrected spectra, assuming that the baseline was known or fixed in advance. In the next chapter we address this assumption and consider ways of estimating this baseline in our model, within our likelihood framework.

CHAPTER III

BASELINE ESTIMATION

III.1. Roughness Penalty

In Chapter II we assumed that the baseline component was known or fixed. We will now extend this model to allow for criticism of the baseline component in addition to the parameters of the mixture density. However, analytical maximization of the expected conditional log-likelihood, Q, given in (2.7) for each $f_t, t = 1, \ldots, n$, results in a baseline estimate which simply interpolates the intensities closest to the baseline. Since we believe that our baseline component is a very smooth function, we want to maximize (2.7) subject to some smoothness constraint.

There are many choices of smoothness constraints. One example from Pawitan (2001) shows that if the intensity locations, $t_i, i = 1, \ldots, n$, are equally spaced, we can impose such a constraint by restricting the squared second differences of the baseline estimates

$$\sum_{t=2}^{n-1} \left[(f_{t+1} - f_t) - (f_t - f_{t-1}) \right]^2 = \sum_{t=2}^{n-1} \left(f_{t+1} - 2f_t + f_{t-1} \right)^2 \tag{3.1}$$

to be sufficiently small. We can represent the expression in (3.1) using a second-difference penalty matrix, K, where

$$K = \begin{pmatrix} 1 & -2 & 1 & & & & 0 \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ 0 & & & & 1 & -2 & 1 \end{pmatrix},$$

by writing (3.1) as $\mathbf{f}^T K \mathbf{f}$, where $\mathbf{f}$ denotes the $n \times 1$ vector of f_t and $\mathbf{f}^T$ denotes its matrix transpose. The expression $\lambda \mathbf{f}^T K \mathbf{f}$ is often referred to as a roughness penalty in a penalized log-likelihood function, where it is subtracted from the log-likelihood of the data. We include this roughness penalty with our complete data log-likelihood in (2.4) to yield a penalized complete data log-likelihood

$$\ell_p = \ell - \lambda \, \mathbf{f}^T K \mathbf{f}. \tag{3.2}$$

and penalized expected conditional log-likelihood

$$Q_p = Q - \lambda \, \mathbf{f}^T K \mathbf{f}. \tag{3.3}$$

We revise our EM algorithm to include an update of the baseline in the M-step of each iteration, by finding the $\mathbf{f}$ that maximizes (3.3). Adding the roughness penalty term to our likelihood developed in Chapter II does not change the convergence properties of the EM algorithm.

The smoothness parameter, λ, is positive and controls the balance between the data modeling and the degree of smoothing. Larger values of λ correspond to

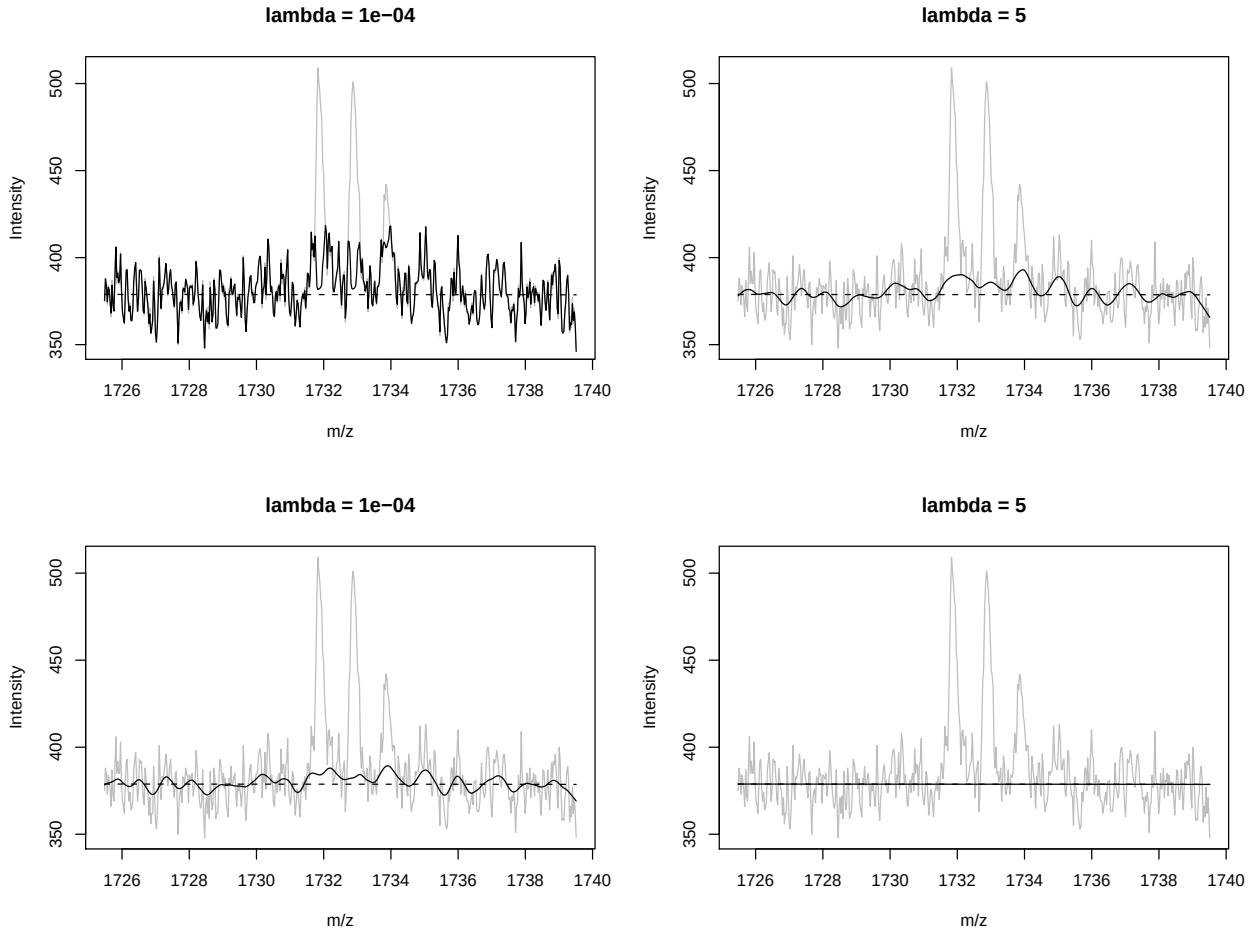

Fig. 18 Baseline updates from the maximization of (3.3). The top graphs use the penalty matrix considered in Pawitan (2001) and the bottom graphs use the penalty matrix considered in Green and Silverman (1994).

smoother baseline estimates; $\lambda = 0$ corresponds to a baseline with no smoothing. It should also be noted that the choice of penalty matrix also affects the smoothness of the resulting baseline. The interplay between the smoothing parameter, λ, and the choice of penalty matrix, K, is illustrated in Figure 18. Here, we compared the baseline updates for a piece of spectrum ($n = 500$) from Wu *et al.* (2003) for two different values of the smoothing parameter (λ) and two different penalty matrices (K). The top-left and top-right graphs show resulting baselines from the maximization of

(3.3) for $\lambda = .0001$ and $\lambda = 5$, respectively, using the penalty matrix from [25] shown previously. The bottom-left and bottom-right graphs show the resulting baselines for updating the same equation with the same values for λ, but for a penalty matrix described in Green and Silverman (1994) that can adjust for locations that are not equally spaced. After dividing the matrix coefficients by $2\sigma^2$, the first few rows and columns of this penalty matrix are

$$K = \begin{pmatrix} 243 & -551 & 390 & -103 & 27 & -8 & 2 & \cdots \\ -551 & 1491 & -1433 & 619 & -162 & 45 & -12 & \cdots \\ 390 & -1433 & 2108 & -1571 & 648 & -180 & 48 & \cdots \\ -103 & 619 & -1571 & 2065 & -1524 & 651 & -174 & \cdots \\ 27 & -162 & 648 & -1524 & 2068 & -1583 & 667 & \cdots \\ -8 & 45 & -180 & 651 & -1583 & 2157 & -1614 & \cdots \\ 2 & -12 & 48 & -174 & 667 & -1614 & 2167 & \cdots \\ -1 & 3 & -13 & 47 & -179 & 675 & -1617 & \cdots \\ 0 & -1 & 3 & -13 & 48 & -181 & 676 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

Both penalty matrices have the property that $\mathbf{Kf} = 0$ for any straight line, $\mathbf{f}$. Most of the visual difference in the smooths between the baselines using different penalty matrices is due to the magnitude of the matrix coefficients, however we can see that the bandwidths of the two matrices are slightly different. While the bandwidth of the matrix above is larger than its counterpart for equally-spaced data, the coefficients do decrease to zero for matrix positions further from the diagonal. We omit the details of the computation of this matrix, but point out that it requires an inversion of a banded matrix of the same size.

Since MS data typically contain intensity measurements at tens of thousands of

locations, we must maximize the likelihood over an equal number of corresponding baseline locations. This results in very large and complicated optimization issues at each iteration. In order to avoid some of these issues, we consider updating the baseline in a piecemeal fashion. There are several main reasons to motivate this approach. First, the computation of the penalty matrix may require the storage and inversion of a large matrix, if we do not consider the spectra in smaller pieces. Also, it should be much easier and take less time to update a 100,000-location baseline by updating the baseline in, say, 100 adjacent sections than to update the entire baseline as one large piece. We also illustrated in Chapter II that the variance of the normal error component appeared to be nonconstant across the range of m/z for the MALDI data. Perhaps the biggest advantage of updating the baseline in smaller pieces is a natural solution to this nonconstant variance issue, since we can use a different choice of smoothing parameter for each spectrum piece with a local error variance estimate.

The relationship between penalty parameter and initial estimate is evident from Figure 19. In each of the graphs, a section of 200 points plotted with the initial baseline estimate as a dotted black line from a 'loess' smooth using .00025, .00075, .0025 and .025 of the entire data set, respectively. Each baseline is updated by maximizing (3.3) with $\lambda = 10^{-6}, 10^{-4}, 10^{-2}, 10^{0}$ and these resulting updates are plotted in solid blue, green, red and black, respectively. Since the data in our motivating data set are not exactly equal-spaced, the baseline iterations pictured in Figure 19 use the penalty matrix described in Green and Silverman (1994), so these values of λ are comparable with the bottom two graphs in Figure 18. Note that the initial estimate in the top-left graph has the shortest span and is heavily influenced by the periodicity of the noise. We can see the baseline estimates from the same value of the penalty parameter for different initial baseline estimates have varying degrees of smoothness, particularly for the most wiggly baselines. Thus, the optimal choice of smoothing pa-

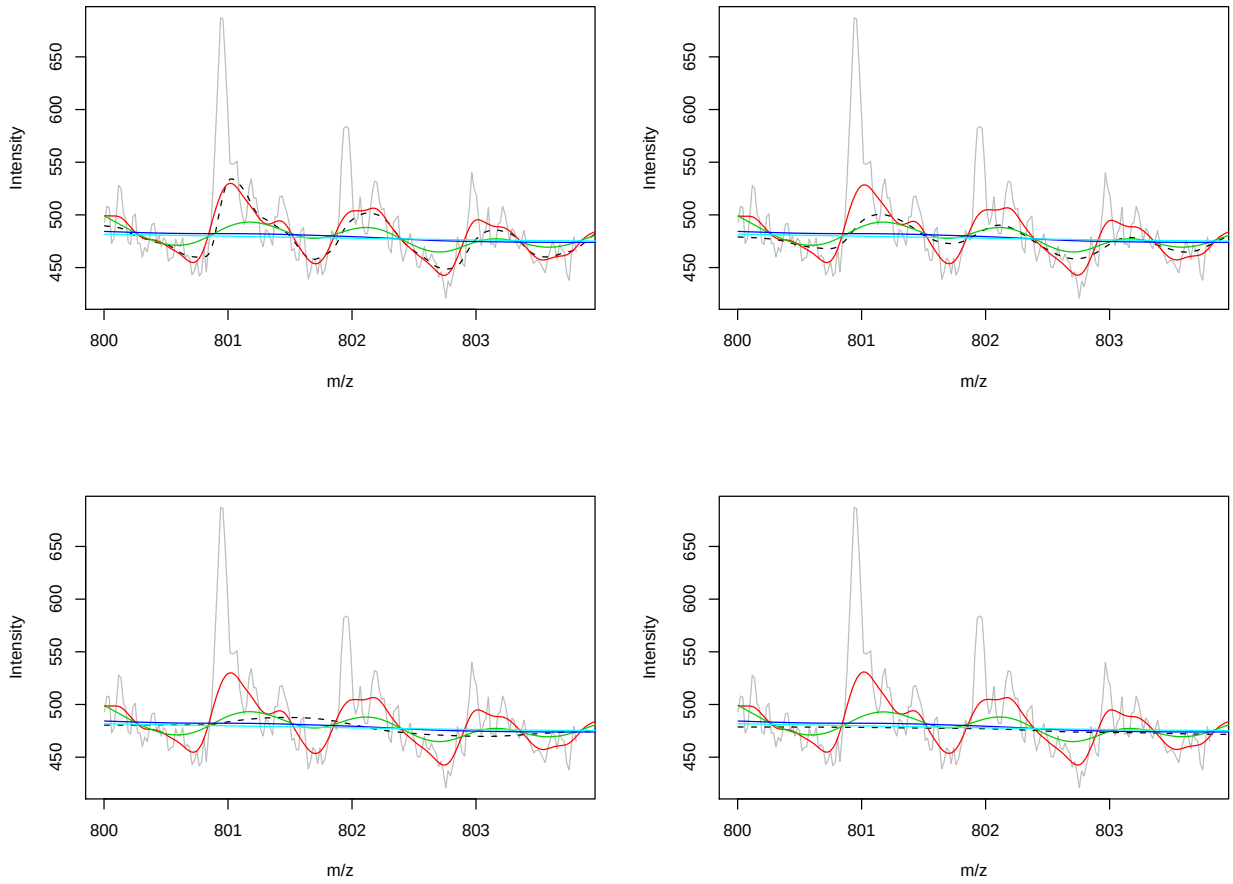

Fig. 19 Baseline updates from the maximization of (3.3). The black dotted lines indicate initial baseline estimates using "loess" smooths of differing spans. Colored lines represent resulting baseline updates for varying $\lambda = 10^{-6}, 10^{-4}, 10^{-2}, 10^{0}$.

rameter may depend upon the initial baseline estimate, but only slightly. This may result from the fact that the maximization of the conditional expected log-likelihood is done using responsibilities based on the initial baseline estimate which may weight the data points from each component differently. In the next section, we investigate the performance of our baseline estimation method in a simulation study.

III.2. Simulation Study

Since the issue of baseline estimation is generally an unsupervised issue, we conducted a simulation study to examine the effect of starting values and smoothing parameters on the estimation of the baseline. The purpose of this simulation study is to evaluate the performance of our baseline estimation procedure in a setting where the true baseline is known. To generate each spectrum, we used an exponential decay function to generate the baseline at equally-spaced design point locations. We chose the exponential decay in accordance with Chapter I, since the belief is that the baseline should roughly resemble this and our motivating data set seems to support this, as well. We should point out that for each spectrum, a different function of exponential decay was generated, since the baselines will be different for each spectrum.

We then added a zero-mean, normally distributed error component with constant variance about each baseline estimate to simulate machine noise. To add peaks, we used initial estimates of the mixing probabilities to randomly select a proportion of the locations where an exponentially distributed peak was added. For instance, in a data set of 400 points we used the mixing probabilities $(\pi_1, \pi_2) = (.95, .05)$, we randomly selected 20 locations (without replacement) to add an exponentially distributed peak to the baseline plus noise estimates. The mixing probabilities were selected based on results from Chapter II, since the data points in the spectra from our motivating data set were mostly machine noise near the baseline.

For each generated spectrum, we considered five different initial baseline estimates. Four different initial baselines used "loess" smooths and for these estimates, we used spans of .10, .20, .40 and .80, which produced baselines that were very wiggly to very smooth, respectively. We also considered the true baseline estimate as an initial estimate, as well, for each generated data set. For each of these initial baselines,

we fit the parameters of our mixture model and updated the baseline using one of eleven different values of λ. We then iteratively update the mixture parameters and baseline according to the EM Algorithm as set out in Chapter II, until the relative increase in the incomplete data log-likelihood was less than 10^{-7} or 100 iterations had been performed. We repeated this process for eleven different values of the smoothing parameter for each initial baseline. In total, 55 different models were fit to each of 100 generated spectra.

After fitting all 100 spectra, we computed the mean square error of the difference between the true baseline and fitted baseline after convergence as well as the original "loess" baseline. We show one of these data sets in the graphs of Figure 20. The solid green line shows the true simulated baseline and the black dotted lines show the initial baseline estimates using "loess". The initial baseline estimates use smoothing spans of .80, .40, .20 and .10, for the top-left, top-right, bottom-left and bottom-right graphs, respectively. The solid lines in red, blue and violet show the baseline estimates for selected $\lambda = 10^5, 10^6, 10^7$, respectively, after running the EM algorithm until convergence. Note that for this simulation study we have used the equally-spaced penalty matrix described in Pawitan (2001), so relative smoothness is comparable with the top two graphs in Figure 19. We also show a summary table of the mean square errors for the simulation study in Table 8. The numbers in the table represent the average mean square errors of the 100 simulated data sets for each combination of λ and initial baseline estimate.

We can learn several things from Figure 20 and Table 8. From the display in Figure 20, it appears as though there is very little difference between the converged baselines for same penalty parameter for different initial smooths. Recall that Figure 19 displayed a more visible difference between the baselines, but this picture depicted unconverged baselines after only a single iteration. From the average mean squared

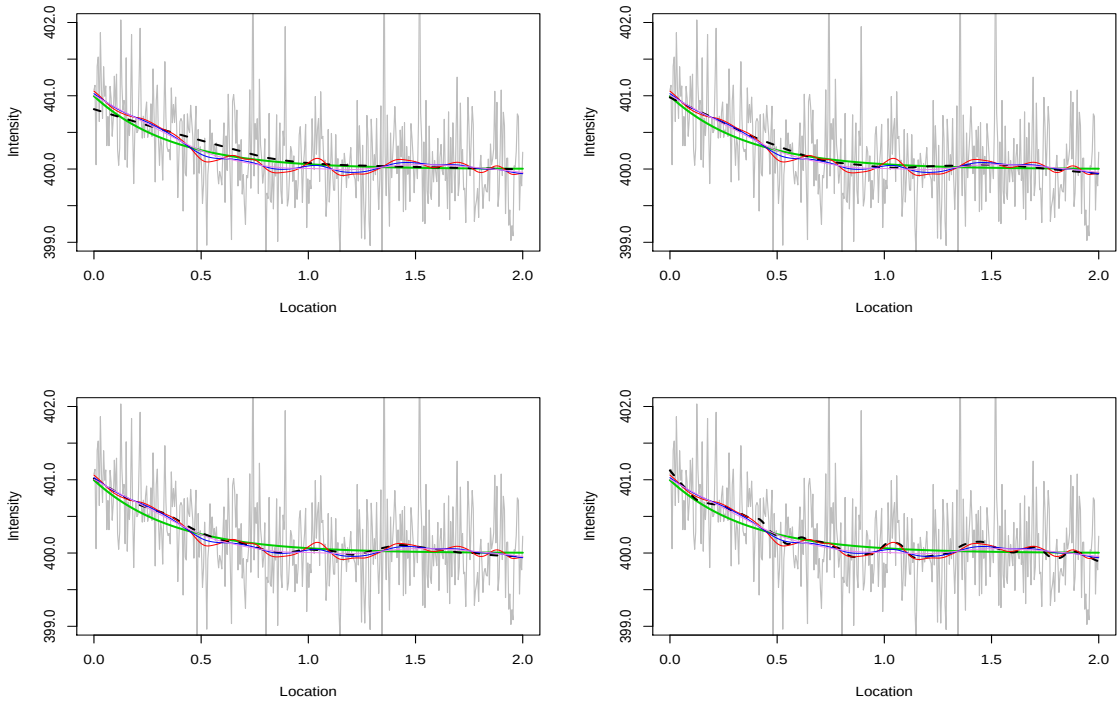

Fig. 20 Simulated baseline (solid, green line) with four different initial estimates (dotted, black lines). Red, blue and violet lines show the baseline estimates using $\lambda = 10^5, 10^6, 10^7$.

errors in Table 8, it appears that the initial baseline estimate makes only a very small difference in the accuracy of the resulting estimate of the converged baseline when compared with the true baseline. From the average mean square errors in Table 8, it appears that our method can produce a better baseline estimate than even competitive initial estimates from "loess" as far as the mean square error is concerned. Moreover, the superior performance of our approach is not confined only to the models where the initial baseline estimates were competitive. We can easily see that our method performs well in cases where the initial baseline is overly smooth or wiggly.

Table 8. Average mean square errors for the simulation study described in Section III.2 for combinations of initial baseline and smoothing parameter.

Start — Lambda	Start	1×10^5	5×10^5	1×10^6	5×10^6	1×10^7	5×10^7	1×10^8	1×10^9
True **f**	.0000	.0093	.0061	.0051	.0038	.0034	.0033	.0035	.0079
Loess, .80	.0076	.0095	.0063	.0054	.0040	.0036	.0034	.0038	.0082
Loess, .40	.0044	.0095	.0064	.0055	.0041	.0037	.0034	.0037	.0082
Loess, .20	.0058	.0096	.0065	.0057	.0042	.0036	.0034	.0037	.0081
Loess, .10	.0105	.0097	.0066	.0058	.0042	.0037	.0034	.0037	.0079

III.3. Choice of Smoothing Parameter

In Section III.1 we showed that the value of the penalty parameter that produces a desired degree of baseline smoothness is dependent upon the choice of penalty matrix. In the previous section, we illustrated some promise in using our method to estimate the baseline in a controlled simulation where the baseline was assumed to be some sort of exponential decay function. In practice, our baseline estimation is generally an unsupervised issue, so it is desirable to automatically compute a data-driven estimate of the smoothing parameter in some optimal way.

III.3.1. Generalized Cross-Validation

One popular method for choosing the smoothing parameter in nonparametric smoothing problems uses a cross-validation score, where the optimal value of the parameter is chosen to be the minimizer of this cross-validation score. The idea of cross-validation is as follows. Consider a single spectrum and a fixed value of λ, for which we will find the cross-validation score. A single data point is omitted from the spectrum, say (t_i, y_i) and the $\mathbf{f}$ that minimizes (3.3) is determined, with the ith point omitted. Denote this $\mathbf{f}$ as $\hat{f}^{(-i)}$. This baseline estimate at t_i is $\hat{f}^{(-i)}(t_i)$ and is used as an "unbiased" predictor of $f_t = f(t_i)$ with smoothing parameter λ. This procedure

is carried out for each of the remaining $n-1$ data points, yielding the predictions $\hat{f}^{(-i)}(t_i), i = 1, \ldots, n$. The cross-validation score for this single value of λ is computed as

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{f}^{(-i)}(t_i))^2. \tag{3.4}$$

The expression in (3.4) is of limited utility to us. Clearly, calculation of the cross-validation score using (3.4) directly is prohibitive, since for each λ, we must solve n smoothing problems to yield the estimates $\hat{f}^{(-i)}(t_i), i = 1, \ldots, n$. This is especially important in the context of MS data. There are algorithms where the cross-validation score can be computed without using Equation 3.4 directly, however our mixture distribution representation may complicate matters.

In an attempt to circumvent this problem to use existing methods, we initially consider computing a cross-validation score by using only points that are believed to be generated from the normally distributed error component. Let $\mathcal{E} = \{t : \hat{\pi}_{1,t} = \max_j \hat{\pi}_{j,t}\}$. Under this assumption, we compute the generalized cross-validation (GCV) score for each value of λ as

$$GCV(\lambda) = \frac{\sum_{t\in\mathcal{E}}(y_t - \hat{f}_t)^2}{(n_{\mathcal{E}} - df_{\mathcal{E}})^2}, \tag{3.5}$$

where $n_{\mathcal{E}}$ is the cardinality of $\mathcal{E}$, $\hat{\mathbf{f}}$ denotes the minimizer of the penalized sum of squares over the $n_{\mathcal{E}}$ points in the error component and

$$df_{\mathcal{E}} = \text{ trace}\{(I + \lambda K)^{-1}\}. \tag{3.6}$$

Note that the K in (3.6) is computed using only the locations in $\mathcal{E}$.

We used our simulation study to evaluate our GCV-based method for selecting the smoothing parameter. For each of the 500 combinations of simulated data sets and initial baseline estimates, we computed the GCV criterion for each of the values

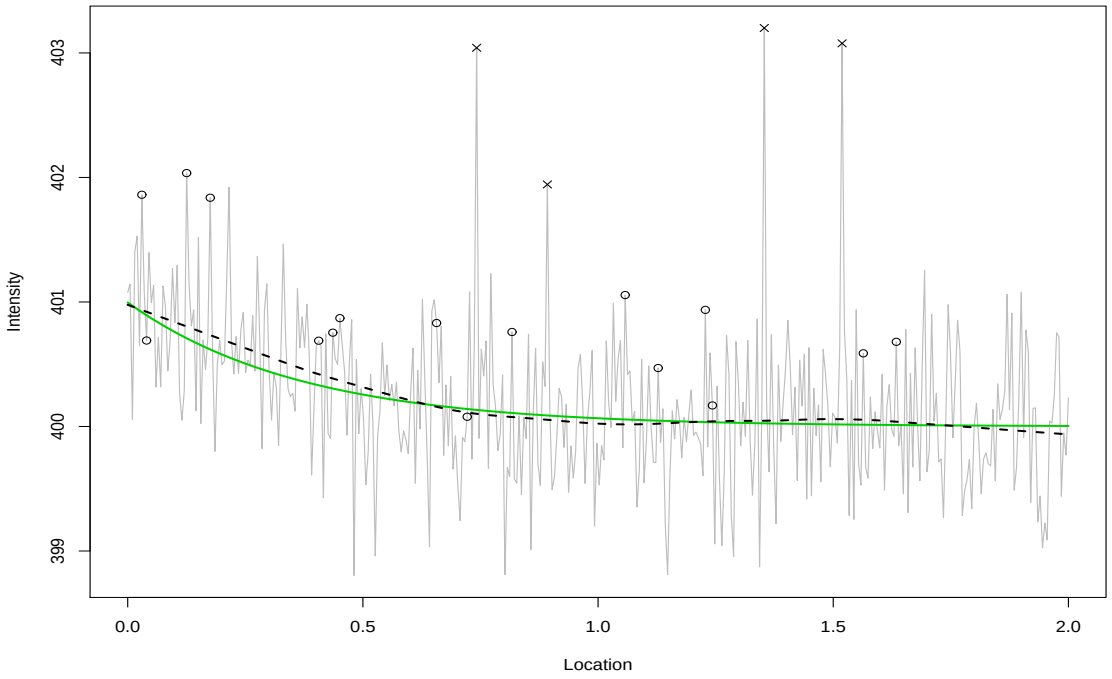

Fig. 21 Simulated data set with points correctly omitted ('×') and points erroneously included ('∘') in the GCV calculation.

of λ in the grid. We computed the true GCV criterion which used all 400 points from the generated noise. In practice, this entire set of points would not be known to the modeler. We also computed a more honest estimate of the criterion by using the responsibilities to determine which points were noise and which were peaks. These responsibilities were calculated based on the initial parameter values.

In Figure 21 we show a simulated baseline with some points marked with × or ∘. The points denoted by × were correctly excluded from the honest GCV criterion calculation; they were deemed to be peaks, when they were in fact peaks. Each point denoted with ∘ had responsibilities which indicated that it was likely generated from the error distribution, when, in fact, it had been generated as a peak. Not surprisingly, most of these points tended be on the higher side of the baseline. It would appear that any difference between the smoothing parameters of the true and honest-

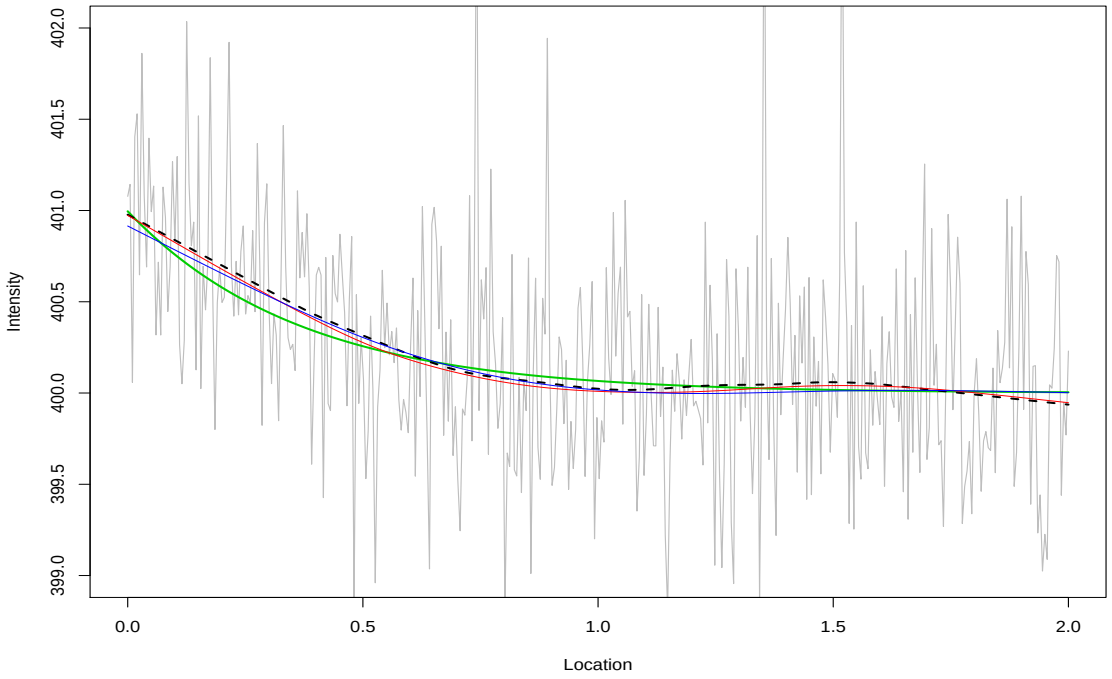

Fig. 22 Simulated data set with true baseline (green), best initial baseline (dotted black), converged baseline selected by GCV (red) and best converged baseline (blue) from the grid of λ.

based GCV criterion estimates stems from very small peaks estimated to be within the confines of the error component.

The performance from automatic smoothing parameter selections is evaluated and displayed in Table 9. The left-most column shows the average mean squared error between the initial baselines using "loess" and the true simulated baseline. To produce the results in the second column from the left, we compared the mean squared errors for each of the converged baselines over the grid of λ and retained the minimizing mean squared error for each simulation and starting value as "best". The values in this column are then the average of these mean squared errors. The column labeled "GCV Choice" displays the average mean squared errors from converged baselines where the smoothing parameter was selected using normal component points as determined by the responsibilities. For example, this would include all of the points in

Table 9. Average mean square errors for the simulation study described in Section III.2 for initial baseline, best baseline, modified GCV-selected baseline and full GCV-selected baseline.

Start — Lambda	Start	Best	GCV Choice	GCV Full
True **f**	.0000	.0032	.0043	.0040
Loess, .80	.0076	.0032	.0046	.0042
Loess, .40	.0044	.0031	.0046	.0042
Loess, .20	.0058	.0031	.0046	.0042
Loess, .10	.0105	.0031	.0047	.0042

Figure 21 without a $\times$. The column labeled "GCV Full" displays the average mean squared errors from converged baselines where λ was selected using all of the points generated as noise. This would include all locations, but the intensities at locations denoted with a $\times$ or $\circ$ would be replaced with their corresponding error intensity.

The results in Table 9 further suggest that there is some promise in using the penalized likelihood idea to estimate the baseline. Through observation of the first two columns we can see that there is at least one value of the smoothing parameter for which the penalized likelihood approach yields a baseline that is considerably closer to the true baseline in mean squared error than the competitive original baselines estimated via "loess". We can also see that use of the automatic smoothing parameter selection results in an improved baseline estimate in most cases, with the full GCV performing slightly better. However, while both GCV methods performed very well and improved upon the initial fit, they did not always choose the value of λ from the grid with minimizing mean squared error.

From Figure 22, we can get a good idea of how comparable the performances are between the "best" choice of λ and GCV-motivated selection. The green line represents the true simulated baseline for this data set, while the dotted black line provides a very good initial estimate using a 'loess' smooth with a span of .40. The red line shows the converged baseline from the GCV-selected value of $\lambda = 5 \times 10^7$,

while the blue line shows the converged baseline from the mean squared error minimizer using $\lambda = 1 \times 10^8$. It would appear that the GCV-based selected λ yields a considerable improvement over initial baselines, especially when the initial baseline is much too wiggly or smooth. There is still improvement when compared to the best initial baselines, but, as expected, these returns diminish with better choices of the initial smoothing span.

Finally, we were also interested to see whether the automatic selection of the smoothing parameter had a tendency to underestimate or overestimate the value of λ that yielded the best performance in terms of mean squared error. We also did a similar comparison to examine the honest GCV-based criterion relative to the true GCV-based criterion. Graphs comparing these results appear in Figure 23. Note that we have jittered the values of λ here, since the points would otherwise fall into stacks corresponding to the few values of λ used in the study. In the left graph, we can see that the honest GCV estimate does fairly well in estimating the best value of λ, however, there appears to be a slight tendency to overestimate this smoothness as evidenced by a cluster of points in the lower right of this graph. On a more positive note, we can see from the right graph that the honest and full GCV estimates are much more comparable. It would appear that any shortcomings in the estimation of the best value λ would be attributed to the generalized cross-validation in general; our extension of the GCV using responsibilities yields very similar estimates to the full GCV.

From a peak detection standpoint, estimating the optimal value of the smoothing parameter with such precision may not be necessary, since there is a considerable range of values of λ which produce smooth baselines that are practically similar. While the optimal choice of λ is somewhat dependent on the initial estimate, the fruits of our baseline correction method via roughness penalty can be easily seen where even rough

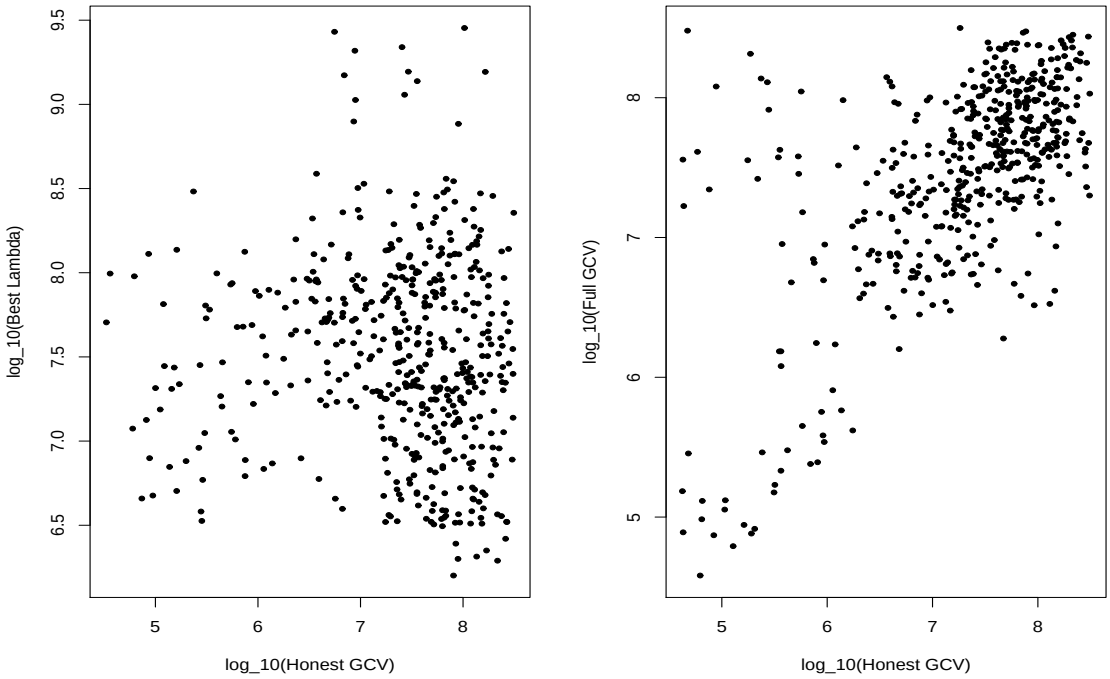

Fig. 23 Scatterplot of best value of λ versus honest GCV-selected λ (left) and full GCV-selected λ versus honest GCV-selected λ.

and wiggly initial baselines can be smoothed to a satisfactory degree provided that λ is in an appropriate range. When considering future peak detection, the estimate of λ is not as important as the resulting baseline, since satisfactory baseline estimates can be obtained whether the initial baseline is too rough or not.

Also, note that the mean squared errors for the relatively accurate initial baselines are approximately the same when compared to updated versions of these baselines using the GCV-chosen value of λ. In other words, there is a limited benefit from updating these baselines with certain values of the smoothing parameter, insofar as the mean squared error is concerned. However, from a practical standpoint, visual inspection of each baseline may be unreasonable, so updating the baseline with these values λ will yield smooth baselines, regardless of initial estimate. At some point, one may need to decide to balance the computation time with model performance.

We briefly touch on this concept in the next section.

To apply these results to real data, we employ a grid search to find the λ which minimizes (3.5) for a piece of spectrum from our MALDI dataset. A graph of the generalized cross-validation score for varying λ appears in Figure 24. We then use this GCV-minimizing λ in the numerical maximization of (3.3) for the baseline update. A graph of the resulting baseline appears in Figure 25.

Clearly, the smoothing parameter that minimizes (3.5) does not yield a very smooth baseline. In this case, the baseline models the somewhat periodic behavior of the noise, which we include in the normal error distribution about our proposed smooth baseline. Thus, the λ that yields a sufficiently smooth baseline for our purposes must be found in an alternative way.

III.3.2. Restricted Maximum Likelihood

From Figure 25, it appears that the optimal value of the smoothing parameter is affected by correlated observations, as evidenced by the periodic behavior of the associated baseline estimate. In a paper by Krivobokova and Kauermann (2007), the authors show that use of a restricted maximum likelihood (REML) criterion can provide a more reasonable estimate of the smoothing parameter, in the presence of correlation. Additionally, they found that the use of this REML estimate is preferred even when the correlation is misspecified, as may be the case with our motivating data set. Recall that we have assumed that the intensities at neighboring locations are uncorrelated.

We investigated the effect of this REML estimate with our MALDI data. We chose a section of spectrum and calculated $-2 \times REML$ over a grid of λ. Note that this criterion uses the assumption of normally distributed errors, so we again use only the points which belong to $\mathcal{E}$, as previously defined. A graph of $-2 \times REML$ for

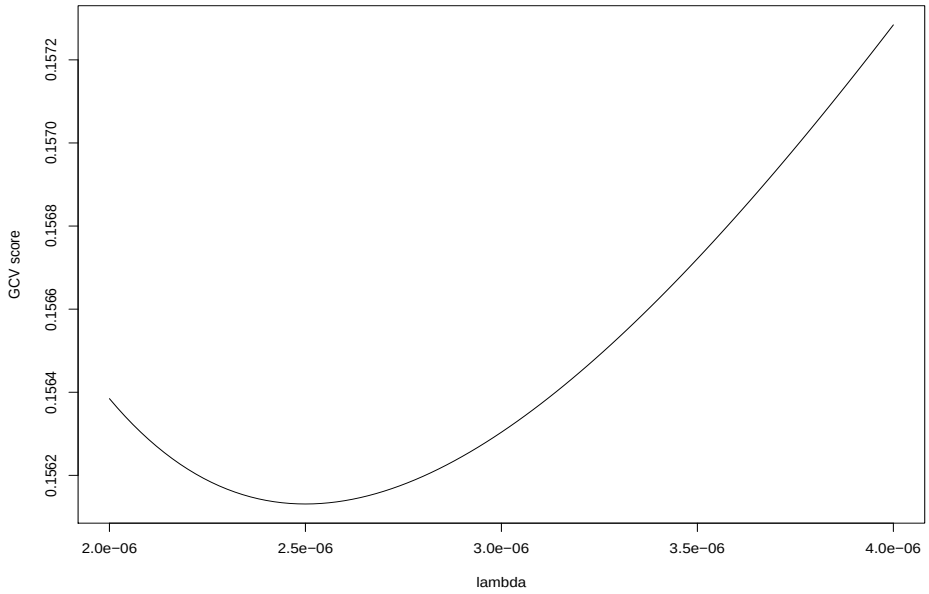

Fig. 24 Generalized cross-validation score for varying λ.

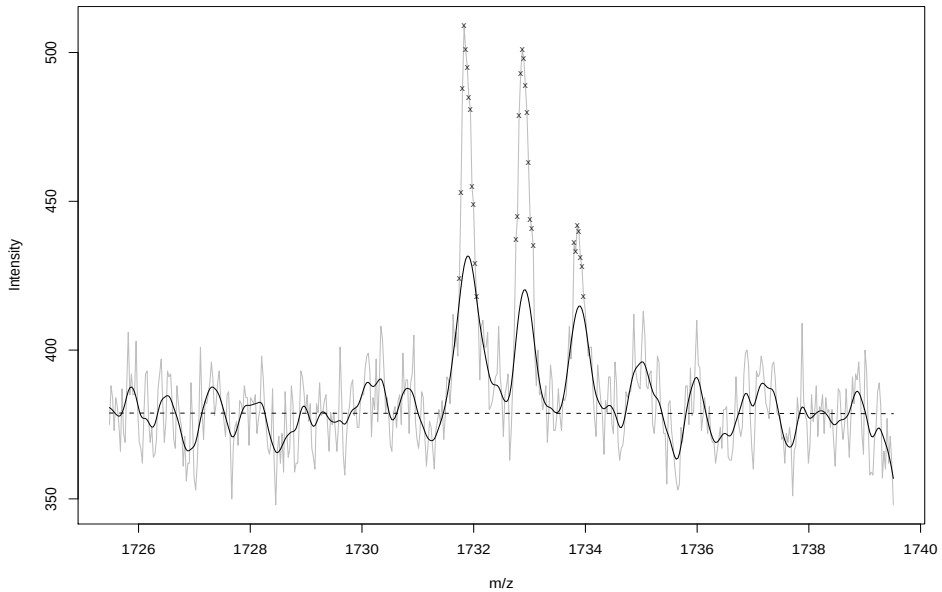

Fig. 25 Resulting baseline update from maximization of (3.3) with $\lambda = 2.5 \times 10^{-6}$ (solid line). Location and intensity pairs denoted with '$\times$' are omitted from inclusion in the generalized cross-validation score. The dashed line represents the initial baseline estimate.

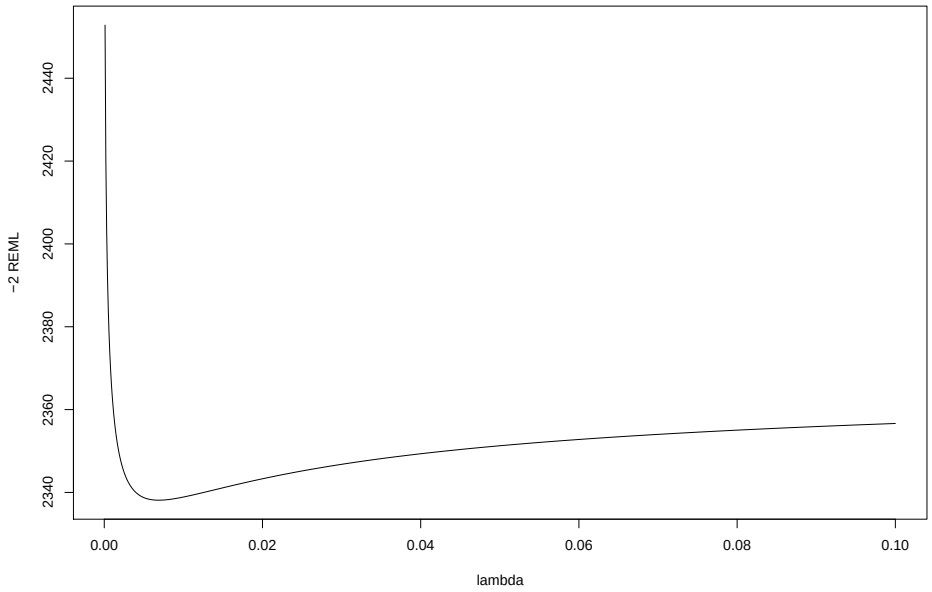

Fig. 26 -2 × REML for varying λ.

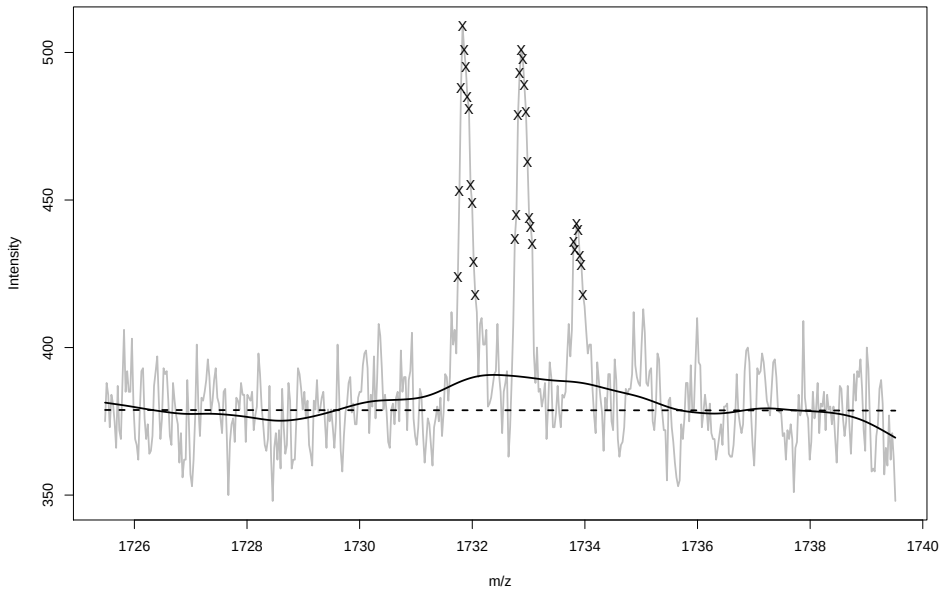

Fig. 27 Resulting baseline update from maximization of (3.3) with $\lambda = .00685$ (solid line). Location and intensity pairs denoted with '×' are omitted from inclusion in the REML criterion. The dashed line represents the initial baseline estimate.

varying λ appears in Figure 26. The minimizing value of this curve is at $\lambda = .00685$, which is considerably larger than the λ yielded by the GCV criterion.

After finding the minimizer of $-2 \times REML$, we use this λ in the maximization of (3.3) to obtain our updated baseline, which appears in Figure 27. The initial smooth using 'loess' on the entire data set is represented as a dashed line. The dotted line displays the baseline update using only the points belonging to $\mathcal{E}$. The solid line displays the baseline update using the same value of λ, using all of the points, including peak locations. We can see that use of the REML criterion has certainly produced an improved choice of λ, compared to that of GCV. However, we feel that the resulting baseline still produces baseline estimates that are too wiggly and biased high in the peak regions, which requires us consider the estimation of the optimal value of the smoothing parameter in yet another way.

III.4. Baseline Correction for Ovarian Cancer Data Set

Recall from Chapter II that we initially estimated a very smooth 'loess' baseline for the ovarian cancer data set. We initially attempted to estimate the baseline using all of the points in the spectrum. This involves the optimization of more than 90,000 baseline locations to find a maximum in the expected conditional log-likelihood, for a given value of λ. The baseline correction is the most computationally intensive aspect of our model-based pre-processing approach.

Our simulation study suggests that our improvement in baseline estimation may be mitigated if we have selected a good initial baseline estimate. Of course, since this is an unsupervised problem, we cannot be exactly sure how well our initial baseline estimates the true baseline. Nonetheless, we were interested to see what improvement was gained in attempting to update this initial baseline. We have seen from Section

III.3 that an automatic and aesthetically-pleasing choice of smoothing parameter may be obstructed by correlated errors, whereas the automatic choice of smoothing parameter worked considerably better in the absence of correlated errors. We now present an alternative method for the selection of this value of λ.

In addition to our local baseline smoothness expectations, we believe that the baseline is a non-increasing function similar to an exponential decay. We use this picture to develop a restriction on the first differences of the resulting baseline updates to yield an estimate of λ as follows. We update the baseline as an additional part of the M-step and begin by dividing each spectra into pieces of roughly equal size. There are several reasons that we divide each spectrum into pieces. From a practical standpoint, we must be wary of the curse of dimensionality. The storage and computation time required to optimize over one piece of baseline with a large number of points will be much greater than if the spectrum is divided into adjacent sections and optimized piecemeal.

Theoretically, the optimal values of the smoothing parameter for each section may, in fact, be different. If the baseline is an exponential decay function, the baseline may be more curved at smaller values of the mass-to-charge ratio, and, thus, a relatively smaller value of λ would be more apropos. Likewise, the baseline function at larger values of the mass-to-charge ratio may be flatter, which would suggest that a larger value of the smoothing parameter should be applied. Another important point of support for the use of piecewise baseline estimation is that the noise variance estimate tends to be much larger at smaller values of the mass-to-charge ratio and smaller for larger values of the mass-to-charge ratio. This change in variance may result in a change of the value smoothing parameter, as well. We now describe our initial approach for baseline correction.

For a single piece of spectrum, we employ a systematic grid search over λ and

check the first differences of each baseline estimate. When the maximum of the first differences of the baseline estimate is negative, this indicates that the resulting baseline is nonincreasing. We begin with a relatively small value of λ and observe the resulting update from the maximization of (3.3). If any of the first differences from the resulting update are positive, λ is increased by a factor of ten and (3.3) is subsequently maximized with the new value of λ. If the maximum of the resulting first differences is negative, we retain the resulting update, and proceed to update the section of the spectrum.

Ideally, we hope to find the smallest λ which yields a non-increasing baseline. The purpose of this approach is two-fold. Our idea to restrict the baseline to be non-increasing ensures that the baseline resembles something that is practically useful. By choosing the *smallest* such λ, we allow the data to have more influence on the resulting baseline, than our subjective restriction. However, given that the number of locations that must be optimized over is very large, maximizing (3.3) for a piece consisting of a few thousand points over a large grid of λ is somewhat impractical. Thus, it should be noted that it is computationally expensive to estimate this λ with great precision, especially as the number of points in each piece grows large. Furthermore, if we were able to estimate this λ with a great deal of precision, the resulting baseline update may be very similar to other baselines resulting from a large range of λ.

In Figure 28, we present two pieces of a single spectrum and a visualization of the baseline update process for each piece. These pieces are two of the 100 pieces from this spectrum, and each of these pieces has approximately 900-1,000 points in each section. On the left, we can see visible short-run increases in the red and green baselines updates, from $\lambda = .001$ and .005, respectively. When we increase the smoothing parameter to $\lambda = .01$, the resulting baseline (in blue) still has a small positive increase

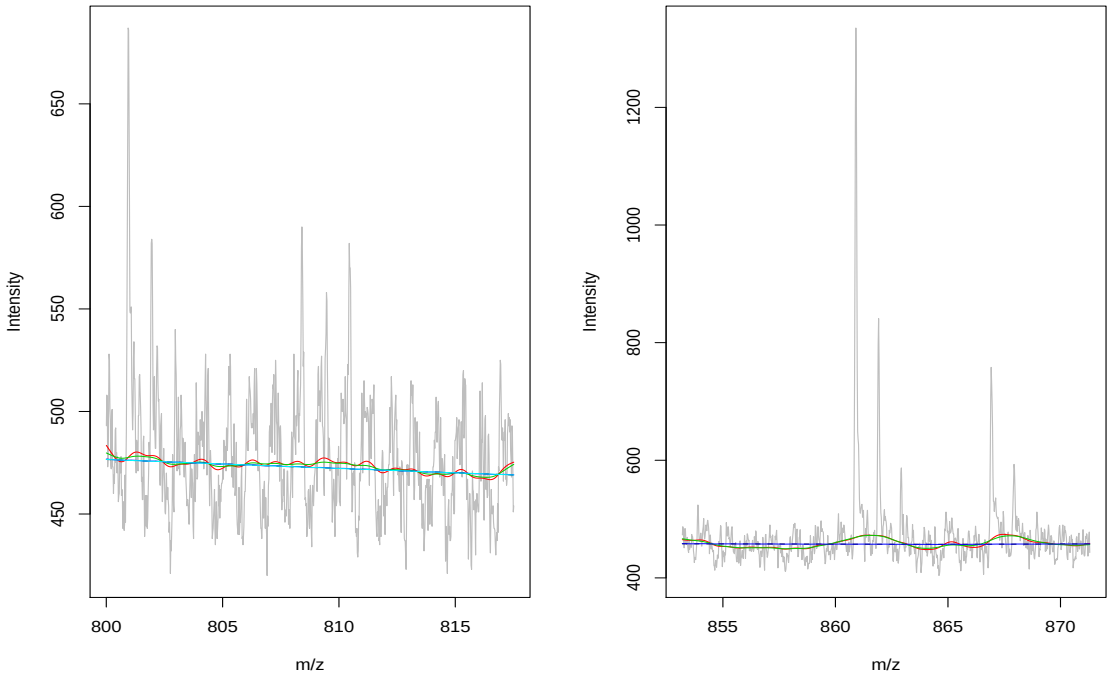

Fig. 28 Series of proposed baseline estimates for two sections of a single spectrum. On the left, the red, green and blue baselines have at least one positive first difference. On the right, the smoothness of a long-run increasing initial baseline prevents us from obtaining an update which is non-increasing over the entire interval.

of .006. When we increased the smoothing parameter to $\lambda = .05$, the maximum first difference decreased to -.007, but the resulting changes in corresponding baseline locations from the previous estimate using $\lambda = .01$ were less than a single intensity unit. This is somewhat evident by the fact that the cyan line completely covers the blue line in the left graph. The right piece shows a spectrum with a slow, long-run increase that is difficult to discern by eye. The original baseline estimate is also slowly increasing over this range and a choice of relatively large λ's revert to the original baseline, using our restriction.

Figure 28 is indicative of many of the baseline updates and it raises some interesting talking points. Due to the large amount of data, we may have divided the

spectrum into too many pieces in an attempt to procure a reasonable number of points in each section to optimize over. One problem with this approach is that the baseline often exhibits little more than straight line behavior in such small sections, which may partially explain why the change between the updated and initial baselines is quite small. Another drawback to having sections that are too small is illustrated in the right graph of Figure 28. The data suggests that the baseline have a very slow increase over this range. However, if this section were larger and extended more on the side of larger mass-to-charge ratios, the spectrum may begin to decrease and a horizontal or even decreasing baseline may then suffice. To make matters worse, even with the large number of sections, the computational time to update the baseline is still quite large, especially considering that few significant changes were garnered through the baseline update.

To this point, we have considered the number of points to be somewhat of a detriment to our baseline update procedure. From the previous study, we have learned that the number of points in each section must be a manageable number to optimize over, while spanning enough of the spectrum to illustrate some useful behavior in the baseline. In an attempt to estimate the baseline under these revised restrictions, we selected 10,000 points from each spectrum to use in baseline estimation and divided each spectrum into ten sections. We selected the 10,000 points according to a sequence; for instance, from a spectrum with 91,380 locations, we selected the first ten locations indexed at positions 1, 10, 19, 28, 37, 46, 55, 64, 74 and 83 of the original raw spectrum. We then used the responsibility calculations based on the initial baseline and initial parameter estimates to retain points from the normal error component for inclusion into the baseline update.

For each section, we start with a relatively small value of λ and optimize the baseline using R's `optim()` function for multivariate optimization. As we mentioned from

previous experience, using the maximum first difference as a criterion for whether to retain the revised baseline is too stringent. In fact, some large sections of spectra will prohibit this or will lead to a poorly estimated baseline. So, as an alternative restriction, we require that the penalty term, $\mathbf{f}^T\mathbf{K}\mathbf{f}$, be smaller than the original baseline over this region at the selected points. If the penalty term from the revised baseline is larger than that of the original baseline, the value of the smoothing parameter is increased by a factor of ten. This procedure continues until the penalty term from a resulting baseline is less than the penalty term of the original baseline.

In Figure 29, we present an illustration of this initial update process. The dotted black line represents the initial baseline estimate from 'loess'. The solid red line shows a subsequent baseline estimate using an insufficiently small value of the smoothing parameter, since the resulting penalty term is larger than that of the initial estimate. The solid black line shows a new baseline estimate from an increase of λ by a factor of ten, where the resulting penalty term is smaller than the initial baseline estimate. Thus, this baseline update is retained as an update to the initial baseline. We want to point out the visual appeal of this revised baseline in that the baseline estimates at the peaks on the ends of this piece of spectrum do not appear to be affected by the high intensities, whereas the initial baseline estimate is higher in both of these peak regions. Another example of the improved estimate in the peak region is provided in Figure 30.

One issue that must be addressed is the possible discontinuity of the baseline estimate at the ends of each baseline piece. If it all possible, we would prefer to have the baseline be a continuous function in these regions. In areas where there are very few peaks and mostly just electronic noise, the baseline estimates tend to be close. In an attempt to make the baseline continuous, we look for the closest point where the baseline estimate in the left (right) piece of the spectrum is equal to the first (last)

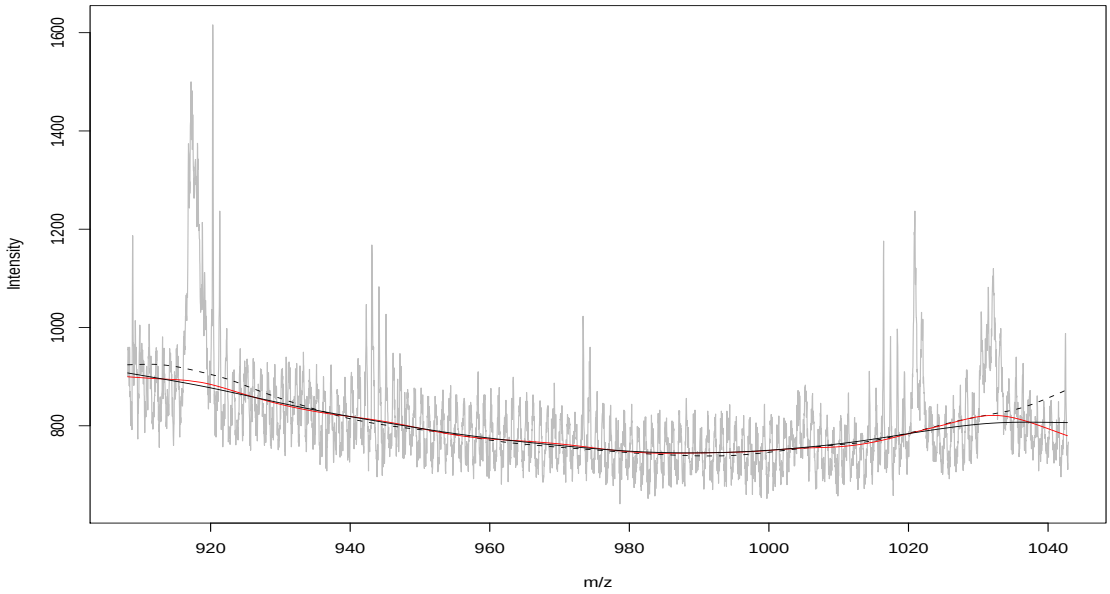

Fig. 29 Initial baseline estimate (dotted line) with baseline updates with values of λ that yield a larger penalty term than the initial estimate and smaller penalty term than the initial estimate, respectively, in red and black solid lines.

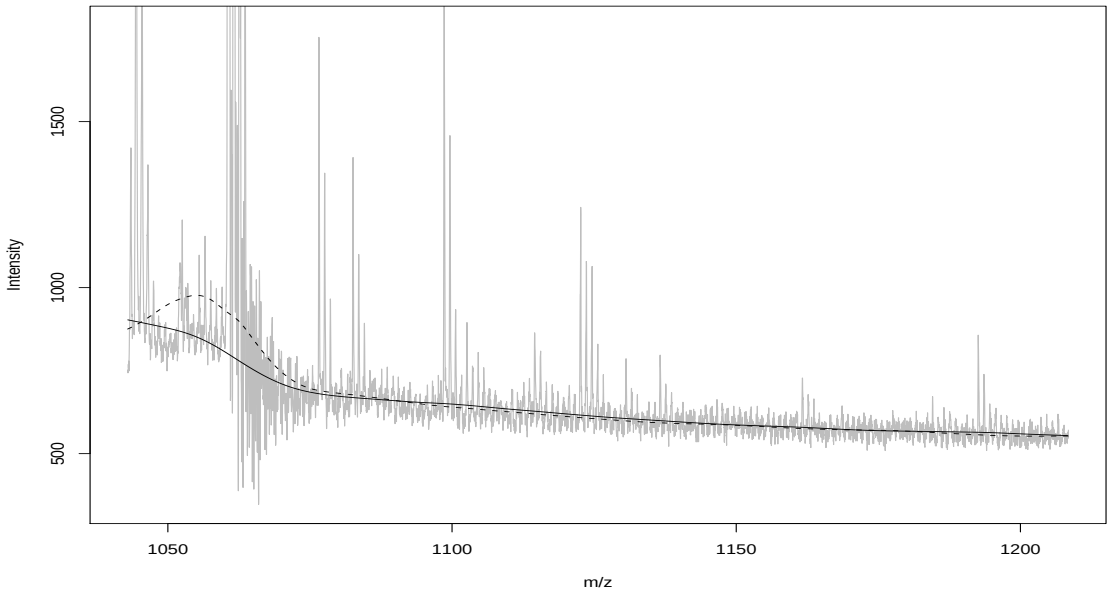

Fig. 30 Initial baseline estimate and updated baseline estimate denoted by dotted and solid line, respectively.

baseline intensity in the right (left) piece of the spectrum. If these points are within 1% of each other, we revise the baseline to be horizontal in this region. We illustrate this idea in Figures 31 and 32. We should point out that not every section break can be adjoined in this manner. However, this appears to be confined to regions where peaks are absent and the baseline discontinuity is relatively small. Figure 32 provides evidence that we should take care to select breaks in the spectra sufficiently far away from significant peaks.

For subsequent iterations of the baseline, we reevaluate the selected points based on updated responsibilities. This may result in only a few points being removed or added to the baseline optimization routine, if any, but this will depend on how close the current baseline is to the previous estimate. Once the set of points has been redetermined, the baseline is updated again. After the initial baseline update, subsequent baseline estimates do not change much, so we only optimize the baseline every ten iterations. However, during each iteration we do optimize the shift in the baseline for each spectrum. This may be important since the large increases in the penalized likelihood are generally attained by making the baseline smoother with, possibly, little regard to the shift of the baseline. Shifting the baseline up or down by only a few units may increase (or decrease) the likelihood after optimization, but such a change may not result in a large change in the observed data likelihood or conditional log-likelihood which may not be significant in the optimization routine.

In the simulation study, we have shown that baseline estimates depend somewhat on the initial estimate of the baseline, to a certain extent, but mostly on the smoothing parameter. In particular, we have used a 'loess' smooth with a relatively large smoothing span as our initial baseline estimate for the spectra from our MALDI data. This initial estimate is non-increasing over much of the interval, so piecemeal baseline update restriction does not yield much change in the resulting baseline in

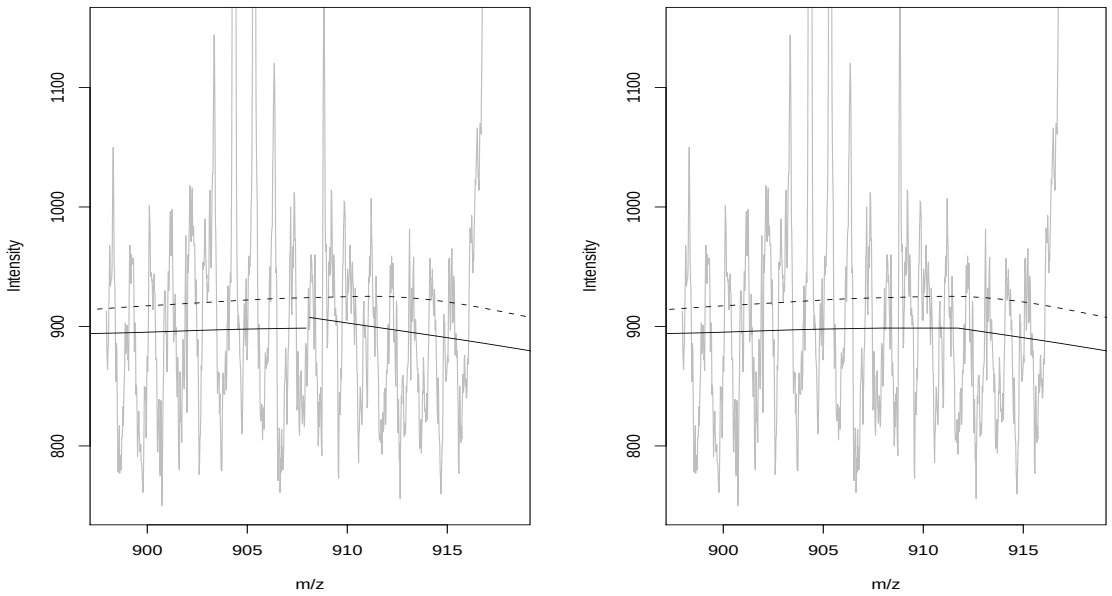

Fig. 31 Region that adjoins two sections of a single spectrum with two baseline estimates (left) and the resulting baseline estimate (right) to force baseline continuity.

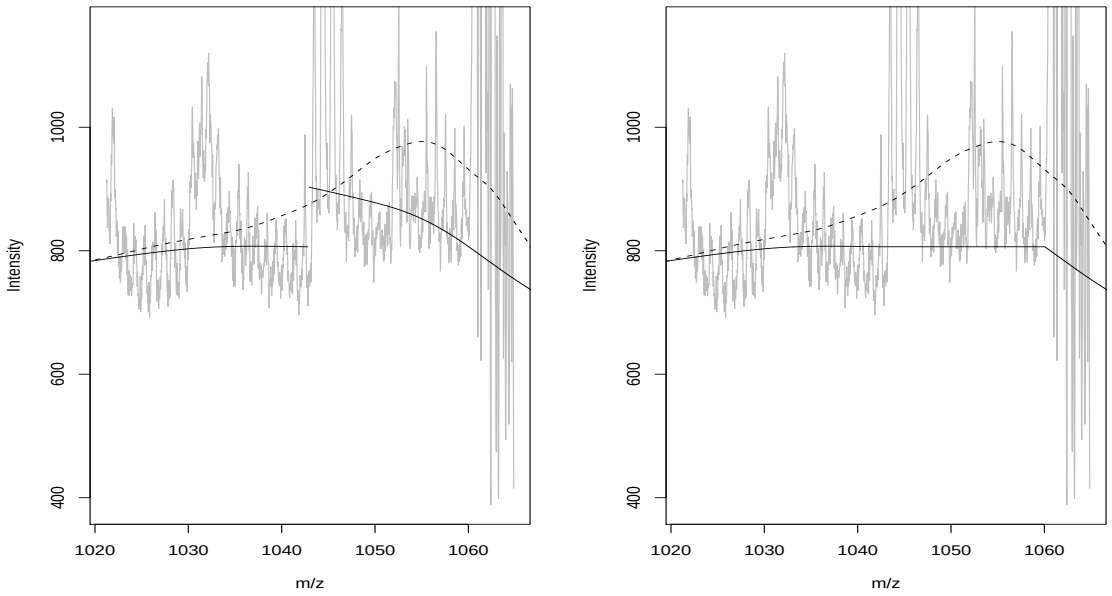

Fig. 32 Region that adjoins two sections of a single spectrum with two baseline estimates (left) and the resulting baseline estimate (right) to force baseline continuity.

this case. However, from our simulation study we recall that an initial baseline that is quite rough (top-left) can be smoothed to the same degree that an initially smooth baseline is smoothed (lower-right). Since our initial baseline estimate is very smooth to begin with, the improvements in our baseline updates have diminished returns.

In this chapter, we have considered the problem of baseline estimation and improvements through the use of a penalized likelihood. While the implementation of the penalized likelihood criterion does not yield significant improvement in the case where the baseline is already a smooth function, using relatively large values of the smoothing parameter will produce baselines that are visually appealing regardless of the initial estimate. We may also be able to reduce the dimension of the spectra with little loss of model performance, while significantly decreasing the memory and time required. A suitable baseline may arise from the combination of the data thinning with the non-increasing baseline restriction, with a maximum value of λ as a cap for baseline sections that are nearly horizontal or even slightly increasing. Another future consideration may be to model the correlation structure of the data, if it is not in fact independent. In the next chapter, we consider peak detection after the baseline and mixture model have been fit.

CHAPTER IV

PEAK DETECTION AND CLASSIFICATION

In Chapter II, we modeled the intensities as machine error from a normal distribution or a peak from one of possibly several normal-exponential convolutions, assuming the baseline was known or fixed. In Chapter III, we extended our model to include an update of the original baseline estimate. We now consider the issue of peak detection, assuming that we have suitably modeled the intensities and estimated the baseline. There are several issues one must balance when considering peak detection from mass spectra which may affect subsequent classification procedures.

IV.1. Peak Detection

Peak detection algorithms from previous work often involve some sort of local maximum search. However, this definition of "local" is not exactly constant from modeler to modeler, and the size of the neighborhood in which a maximum is sought is an important consideration since it effects the size of the resulting set of possible peaks. The advantage to selecting a small peak set is the benefit of a smaller number of potentially discriminating m/z locations to analyze in subsequent classification procedures. Increasing the size of the neighborhood has some benefit in that it will often filter out many of the smaller peaks from isotopically resolved groups, as, from a biological perspective, the monoisotopic peak is usually the peak of interest from such a group and it is often the largest within these groups. However, a small peak set may omit very small discriminating peaks if a peak is in some relatively larger

proximity of a larger peak. A natural solution to this is to make the neighborhood smaller in which a peak must be the maximum. This, in turn, increases the size of the candidate peak set. Practically speaking, one must balance the computation time and increased detection of false positive peaks in comparing larger candidate peak sets with the chance that small and possibly discriminating peaks may be overlooked during peak detection. We feel it is better to include more peaks since we have the benefit of ample computing resources necessary to possibly compare a number of biologically irrelevant peak locations.

In addition to the size of the neighborhood, minimum peak height restrictions are sometimes enforced as part of a local maximum search, so that many spurious and non-biological peaks from the machine noise are excluded from potential biomarker consideration. The same arguments from the previous paragraph apply in the context of selecting a minimum peak height for such a restriction; peak height restrictions using a smaller minimum will yield a larger peak set. A comforting notion is that the choices of the size of the neighborhood and minimum peak height need not be unguided and arbitrary. There are several factors that should be considered when making such a choice.

As mentioned previously, isotopically resolved groups may be present in the spectra. These peaks are often spaced approximately one Dalton apart, so adapting the neighborhood size so that it exceeds one Dalton should eliminate the detection of peaks which have been claimed to be biologically irrelevant as isotopes of other compounds which are present in the spectrum. Another method for selecting the neighborhood size may be based on the mass accuracy of the mass spectrometer used. This mass accuracy may range from 0.1% to as high as 0.5% of the mass-to-charge value, however, the use of mass accuracy is often used in subsequent peak matching algorithms, where peaks are matched across spectra if they are within some neighborhood

based on this mass accuracy. Using this mass accuracy in the local maximum search within each spectrum prior to peak matching will remove some of the smaller peaks that may be later discarded in a subsequent peak matching algorithm.

Minimum peak height restrictions are often based on some characterization of the machine noise component, since many spurious peaks are often the result of relatively large machine noise intensities. After baseline correction and normalization, this noise has been estimated by computing a local estimate of median absolute deviation of the pre-processed intensities and then restricting a peak intensity to be larger than three median absolute deviations above the baseline. Since we have gone to the trouble of fitting our baseline and mixture model in our likelihood framework, we preferred a model-based approach to find a suitable peak height restriction. By using the EM Algorithm to model the intensities, we can use the responsibilities in our approach for peak detection with little effort.

To illustrate how we can use the responsibilities to enforce this restriction, recall that each intensity has m responsibilities corresponding to its conditional density value from each component. For the initial phase of peak detection, we need only concern ourselves with responsibilities from the normal machine error component. We display a histogram of baseline-corrected intensities whose maximum responsibilities correspond to the error component in Figure 33. The violet bars of the histogram correspond to the violet points in the associated spectrum section below. The red and green curves on the histogram denote the weighted normal component density and weighted peak convolution density with smallest mean, respectively. The dotted line at the right of the histogram indicates the point at which these densities intersect; that is, points where intensities whose highest responsibilities change from the error component to the first ordered peak component. Note that this dotted line appears in the spectrum below, just above the machine noise. The points in this section of spec-

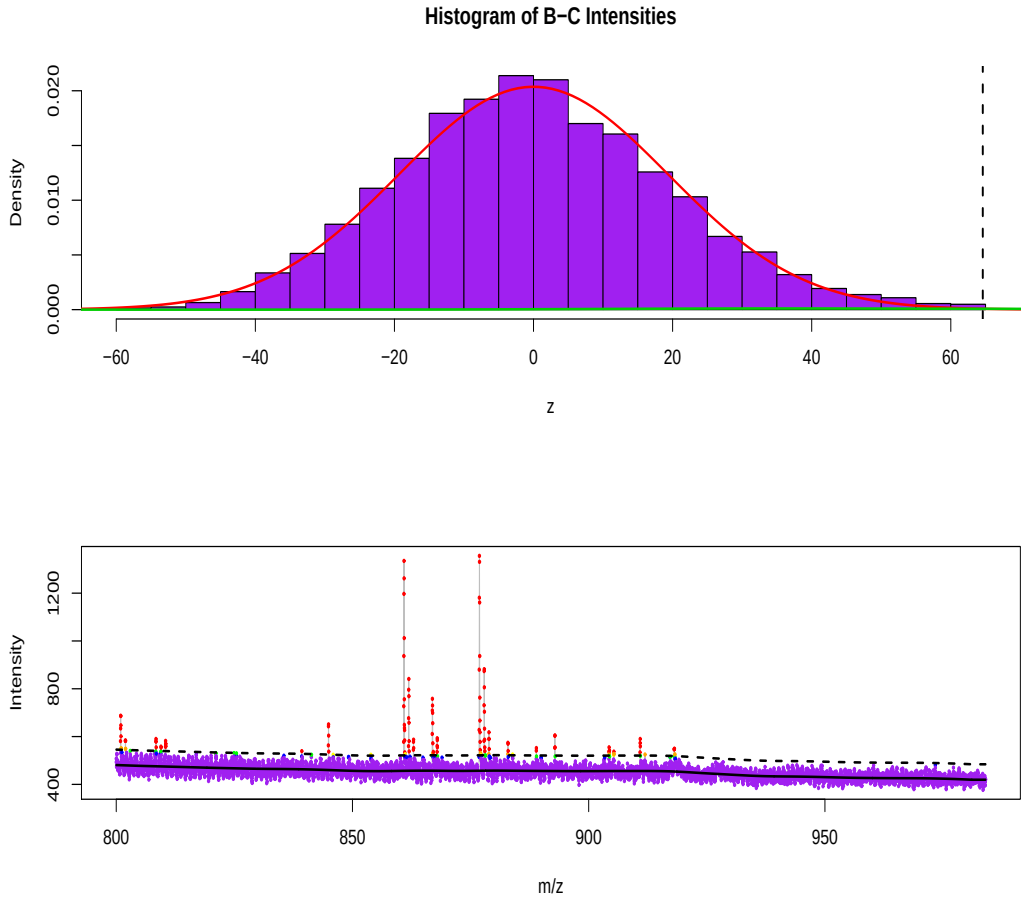

Fig. 33 Histogram of baseline-corrected intensities from the error component and spectrum of intensities with points color-coded by responsibility from error component.

trum are color-coded by their responsibility from the error density. The color of the points are scaled so that violet points denote intensities with highest responsibilities from the error density, while red points denote those intensities with responsibilities from the error density near zero.

From Figure 33 it is clear that peaks can easily be identified as intensities with very low responsibilities from the normal error component. Since peak detection is an integral part of many classification techniques for mass spectrometry, we investigated to what extent responsibilities could serve as a viable tool for classification, as well. We can see from Figure 33 that using just the responsibilities from the error

Table 10. Average number of peaks found in each mass spectrum using local maximum search with varying neighborhood size based on mass accuracy and error responsibility restriction on peak height in the MALDI ovarian cancer data.

Mass Accuracy	.001	.002	.003	.004	.005
Cancer	470.0	299.3	230.6	191.8	165.7
Healthy	499.1	317.0	242.7	200.8	172.9

component may suffice to a certain degree since the responsibilities from the error component decrease for intensities with increased distance from the noise, as evidenced by the color changes. Even with the aid of responsibilities, we must still use a local maximum search to avoid the run-up and run-down of points for each peak.

The purpose of peak detection is usually geared towards later classification. While some of these peaks may be in fact spurious, we feel that a classification procedure should sort which of these are spurious and which are meaningful. To this end, we feel that finding more peaks is an advantage. In Table 10, we show the effect of how a change in neighborhood size affects the initial peak harvesting counts from each spectrum. To facilitate subsequent peak matching, we use mass accuracies to determine the size of the neighborhood, since we may use these mass accuracies to match peaks across spectra at a later point in time. One idea is to retain intensities in an initial set of peaks for each spectrum if the intensity is the maximum in a window of size $[m/(1+a), m(1+a)]$, where m represents the m/z value of the peak and a represents the mass accuracy (i.e., .001 for .1% mass accuracy). To impose a peak height restriction, the responsibility from the machine noise component of the local maximum intensity must be the smallest of the responsibilities for that location. As expected, we see that as the neighborhood size grows larger, the number of peaks retained is smaller.

In Figure 34, we illustrate how the neighborhood size affects the size of the

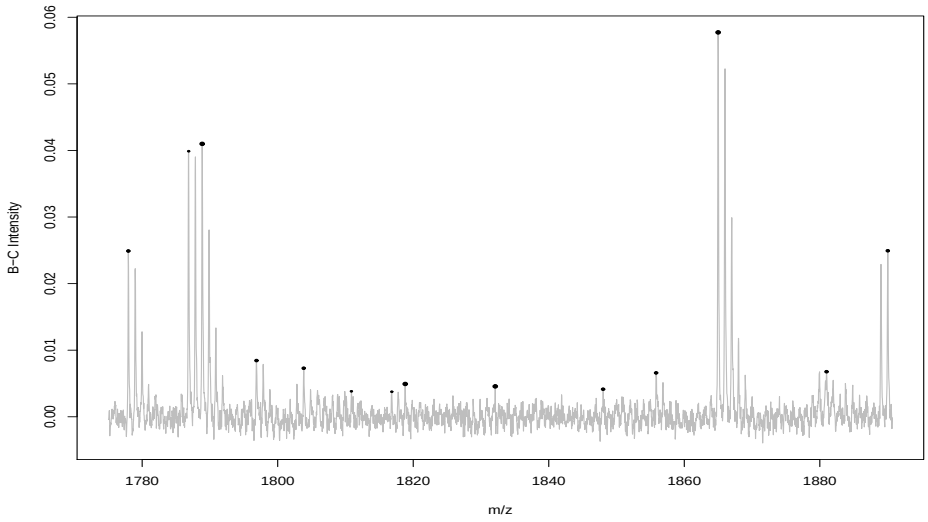

Fig. 34 Local maxima in one section of spectrum from the MALDI ovarian cancer data. The biggest dots are the largest intensities in a window of .5% of the m/z of the peak. The medium-sized and biggest dots are local maxima in a window of .3%, while all of the dots are local maxima in a window of .1%.

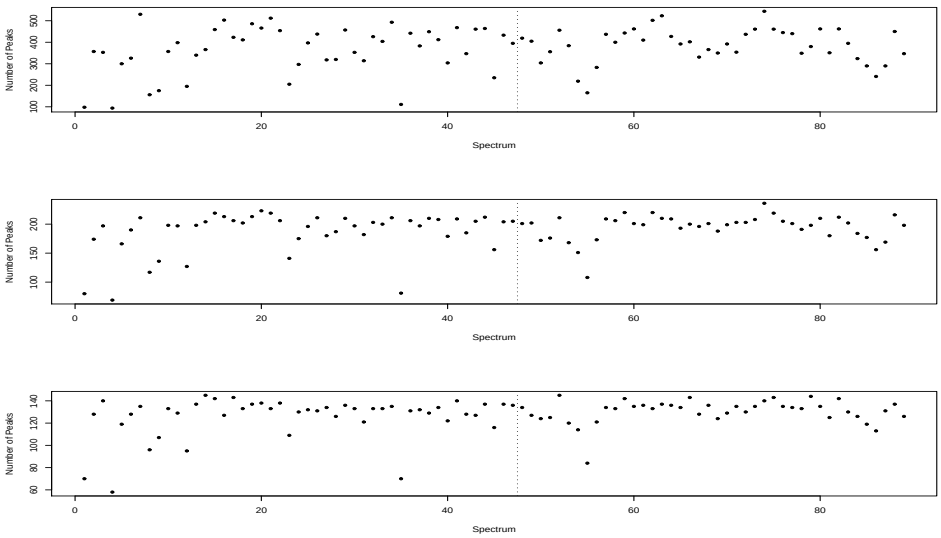

Fig. 35 Number of peaks per spectrum in initial local maxima search from the MALDI ovarian cancer data. The top, middle and bottom plots display the peak counts for each spectrum using mass accuracies of .1%, .3% and .5%, respectively. The vertical dotted line separates cancerous and healthy spectra.

resulting set of candidate peaks. Peaks are denoted by solid black dots of three different sizes. The largest dots correspond to local maxima in a neighborhood with mass accuracy of .5%. The local maxima resulting from the smallest neighborhood search with mass accuracy .1%, includes all of the large, medium and small dots. In Figure 35, we show how these neighborhood sizes affect the peak counts per spectrum; as we mentioned before, a larger neighborhood yields a smaller candidate peak set. The neighborhood size also affects the range of the number of peaks found across spectra. Note that spectra 1, 4 and 35 show the smallest number of candidate peaks as the mass accuracy changes, but the peak counts from these spectra appear as more significant outliers as the mass accuracy increases.

IV.2. Spectral Alignment and Peak Matching

Up to this point, it was not necessary to address potential peak calibration issues since we had restricted our analysis to within each spectrum. In our peak detection discussion and in Chapter I, we mentioned the need to align the spectra in such a way that peaks corresponding to the same peptide can be appropriately matched. We considered two approaches to aligning the mass spectra and matching peaks. Our first approach was to do a point-by-point mass spectral alignment for both peaks and non-peaks, in the hopes of avoiding a potentially messy peak matching routine. We found that our point-by-point spectral alignment was not sufficient for this data; we briefly summarize our findings, before considering a more traditional peak matching approach.

IV.2.1. Point-by-Point Mass Spectral Alignment

In our analysis of our MALDI data, the number of points in each spectrum differed slightly, but the mass-to-charge ratios were slightly offset in that the intensity at the kth ordered mass-to-charge location was not always closest to the corresponding mass-to-charge location in another spectrum. We attempted to circumvent this issue and, hopefully a peak matching issue, by initially aligning the peaks and non-peaks of each spectra, prior to peak detection, in a way that minimizes the variance of the mass-to-charge ratios that are matched together.

This approach begins with the first (or smallest) m/z value in each spectrum matched together, initially. Now consider the group of second m/z values in each spectrum as candidates to be matched to this initial set of m/z values. Suppose that the smallest of these second m/z values comes from spectrum s. We compare the variance of the initial set of matched m/z values with a revised set where the first m/z of spectrum s is replaced with the second m/z of spectrum s. If the variance of the former set is smaller, the initial set of m/z remains matched together, and we then proceed to the set of the next-largest m/z values from each spectrum. If the variance of the latter set is smaller, we omit the first m/z value from spectrum s and match the second m/z from spectrum s to the initial set. We then repeat the above process by considering the smallest of the remaining unmatched m/z values.

To illustrate this concept we show a dot plot of the mass-to-charge values for small sections of selected spectra in Figure 36. In the top graph, points of the same color are in the same position in each raw spectrum; that is, each point of the same color is the kth smallest m/z location in its spectrum. We can see that several of these spectra are clearly shifted, in this regard. In the bottom graph, we display the same points color-coded according to a new alignment based on a minimum variance criterion. Locations in the same color are "matched" together.

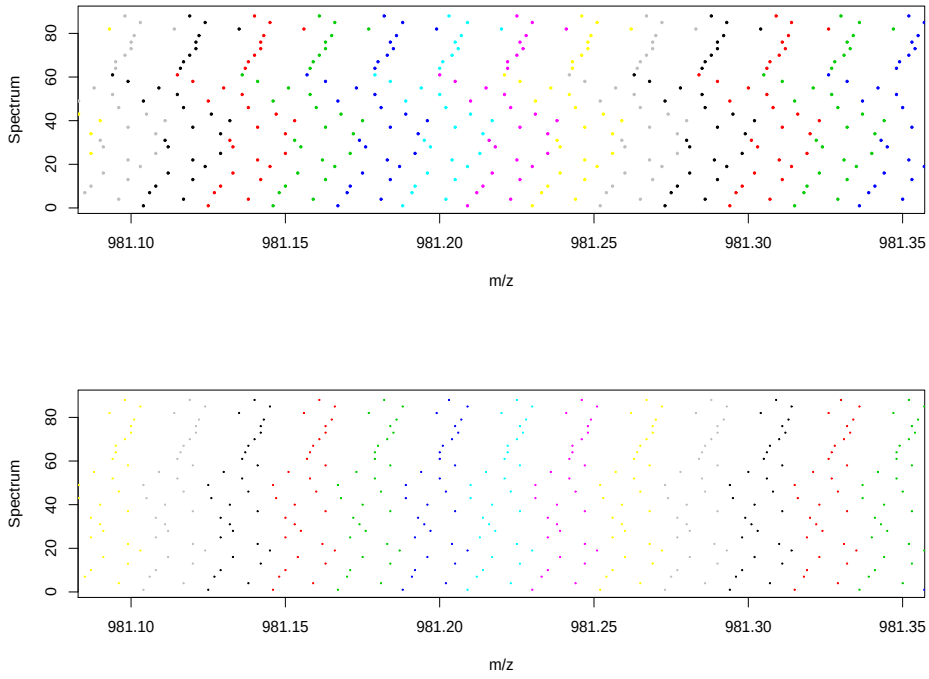

Fig. 36 Dot plots of m/z values for selected misaligned spectra. In the top graph, each point of the same color is the kth smallest m/z location in its spectrum. In the bottom graph, we display the same points color-coded according to a new proposed alignment. Locations in the same color are "matched" together.

Once these points are aligned, we compared regions of high peak responsibilities across spectra. To visualize our idea, we display overlapped baseline-corrected spectra zoomed to a single matched location that is identified as a peak for both healthy and cancerous spectra in Figure 37. Our baseline correction followed the method outlined at the end of Chapter III. The dots indicate the matched intensities according to the previously described spectral alignment; red dots signify cancerous spectra and blue dots denote healthy ones. We want to point out that we have compared some non-maxima and maxima at this location. In Figure 38, we present side-by-side dot plot of spectrum index versus baseline-corrected intensities on the left, along with a corresponding plot displaying the error component responsibilities versus the spec-

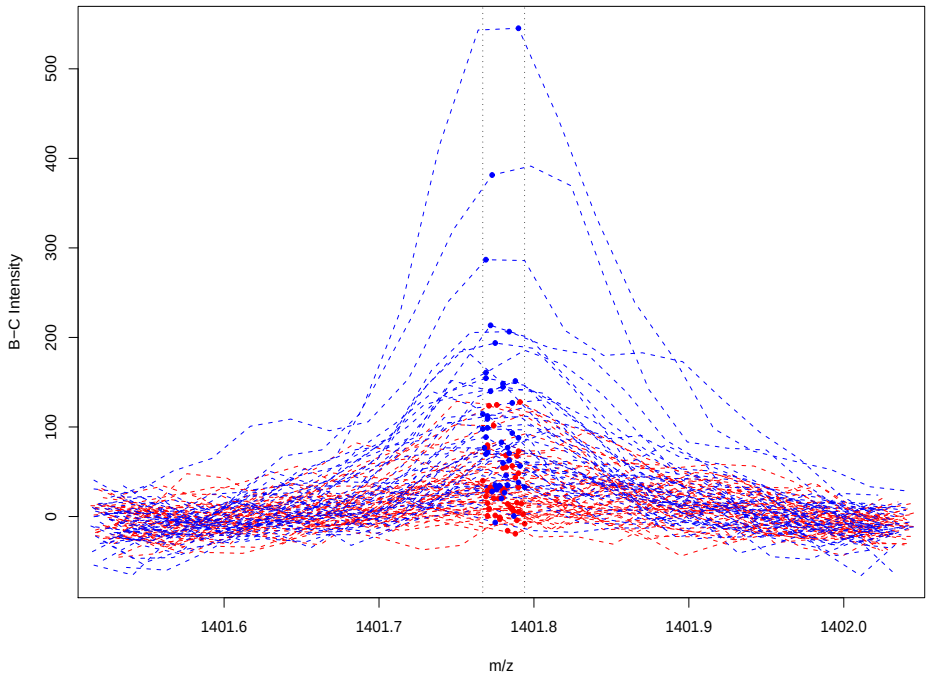

Fig. 37 Cancerous and healthy spectra in the vicinity of a peak. Red dots indicate cancerous spectra; blue dots denote healthy spectra.

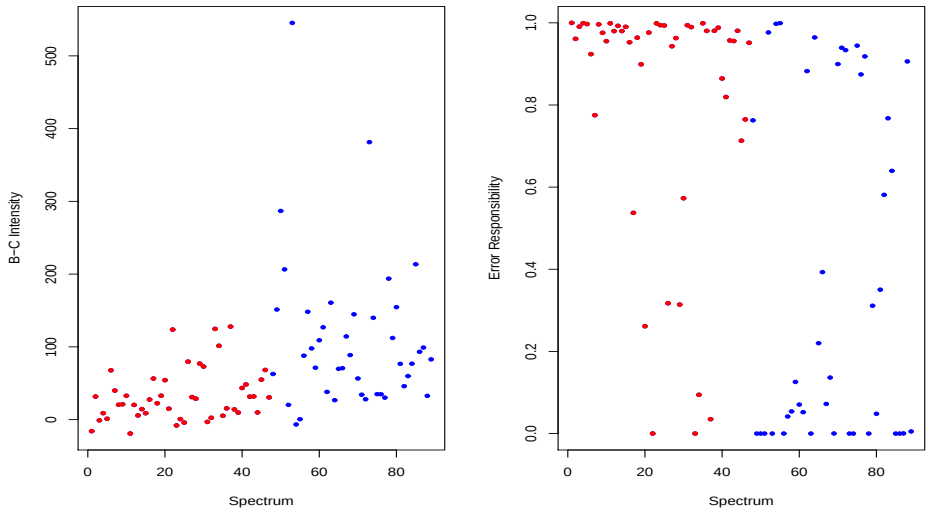

Fig. 38 Dot plots of spectrum index versus baseline-corrected intensity (left) and error component responsibility (right). There appears to be a clearer visual separation between the cancerous and healthy error responsibilities.

trum index on the right. There appears to be a clearer visual separation between the cancerous and healthy points in the right plot.

Despite the promise shown in Figure 38, upon closer visual inspection we can see that some of the dots corresponding to proposed peak intensities are not at the local maximum, but are part of the run-up or run-down of the peak. Thus, while such calibration may be helpful in aligning spectra, it will not eliminate the need for subsequent peak matching, if we are to compare the maximum intensities of each peak. Again, our hope was to align the spectra using all of the points, instead of using only the peaks to guide the alignment.

IV.2.2. Peak Matching

If we consider the spectral alignment using only the peaks, we begin with a list of locations and intensities of possible peaks for each spectrum. However, because of the accuracy of the mass spectrometer, peaks which correspond to the same peptide in different spectra may have slightly different m/z values. The most important input into a peak matching algorithm is the degree of offset, or mass accuracy, of a peak across spectra. One method for peak matching uses a one-dimensional clustering algorithm on the location axis (Tibshirani *et al.*, 2004). They observed that, after applying a supersmoother and log transformation to the m/z values and intensity values, the peak widths were approximately constant with width of .005 across each spectrum on the log m/z scale. Although these peak widths may be somewhat specific to the data, operator, or machine, there are some merits to performing alignment on the log m/z scale, which we will discuss later.

A general purpose idea is to cluster peaks together based on mass accuracy. This idea is used in constructing a peak super set (Fushiki *et al.*, 2006) as follows. Given

a peak at mass m, a peak from another spectrum is aligned to the given peak if its peak location lies within $[m/(1+a), m(1+a)]$, where m represents the m/z value of the given peak and a represents the mass accuracy. We can treat the mass accuracy as an adjustable input into a classification algorithm to see which value of the mass accuracy yields the best classification results.

We implement our peak matching algorithm similar to the continuous covariate case considered in the super set idea proposed by Fushiki. Initially, the spectra are sorted by the number of peaks detected after the peak detection stage. The spectrum with the smallest number of initial peaks initializes the clusters for a peak super set. Then the spectrum with the next smallest number of initial peaks is then aligned to this peak super set, based on the mass accuracy. We compare the peaks from this second spectrum with the peaks in the super set. If a peak from this second spectrum is within the pre-determined mass accuracy of one of the peaks in the super set, then each pair of aligned peaks is assumed to represent the same biological molecule. If a peak from this second spectrum is not aligned to the peak super set, it initializes a new peak cluster. This procedure is carried out until all of the spectra have been aligned with the peak super set.

Our peak detection method may not be perfect, so it is reasonable to believe that we may have missed some potentially important peaks or included some spurious peaks in our peak super set. To address the first issue, after the initial peak matching, we search for local maxima in the spectra with missing peaks in each location within the confines of the existing cluster limits on the m/z axis in the peak super set. This is analogous to the continuous covariate case described by Fushiki. To address the second issue, we eliminated clusters with mostly missing peaks. While analyzing the MALDI data, we eliminated peak locations with nine or fewer peaks. We chose nine since the use of 10-fold cross-validation to divide the training and test

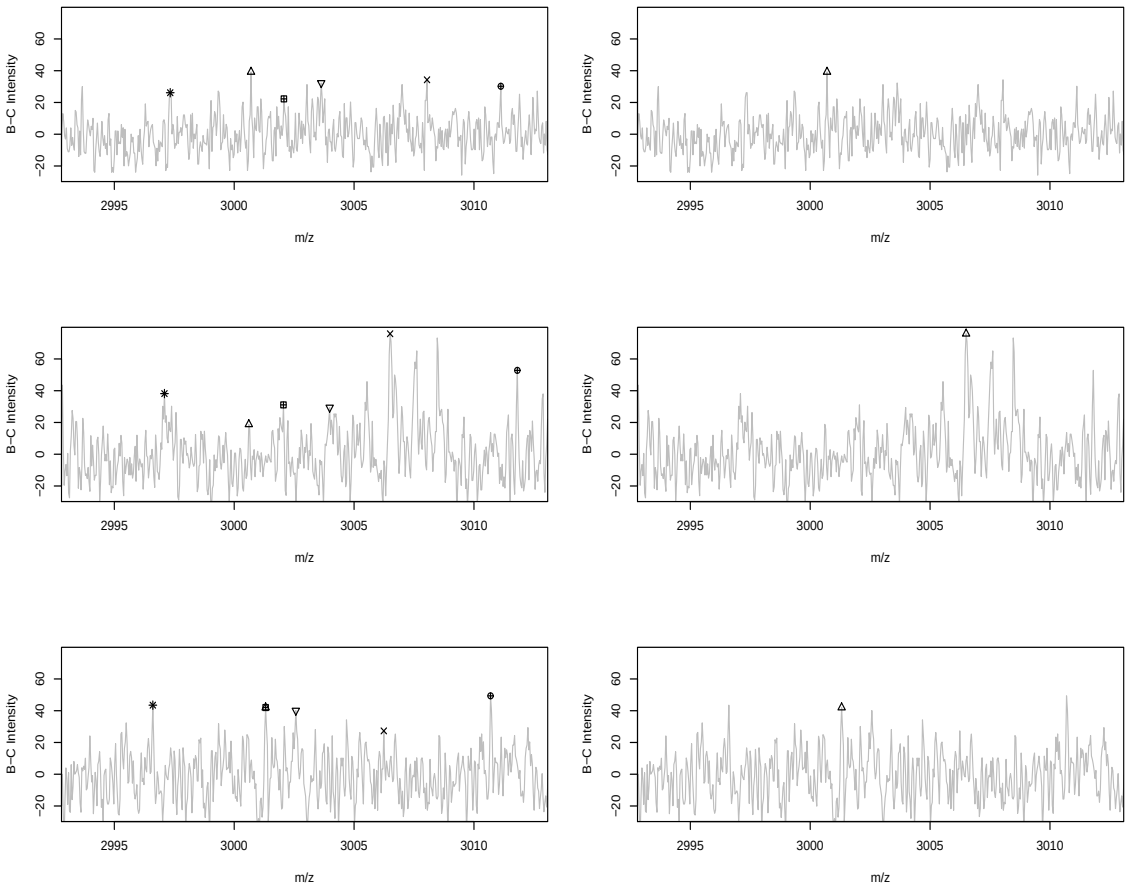

Fig. 39 Matched peaks across spectra. Peaks in different spectra with the same plot symbol are matched together in the same peak cluster. The left plots show peak detection and matching using a mass accuracy of .1%, while the right plots show peak detection and matching using a mass accuracy of .5%.

folds would guarantee that at least one peak from each peak cluster would always be retained in the training fold. Prior to classification, our peak clusters can be viewed as a rectangular matrix where each location has a quantitative measurement for each spectrum.

To show the difficulties in peak detection and peak matching, we show the peaks of three different spectra after peak detection and peak matching, in Figure 39. Peaks in different spectra with the same plot symbol are matched together in the same peak

Table 11. Number of clusters found with varied mass accuracy.

Mass Accuracy	.001	.002	.003	.004	.005
No. of Clusters	1149	653	443	352	281

cluster. The plots on the left initially find peaks that have the largest intensity in its mass accuracy range of .1%, and then are matched together if peaks in other spectra are within its mass accuracy range. The plots on the right show peaks which are detected and matched using a mass accuracy range of .5%. Note that in the bottom left spectra, a single peak at $m/z = 3001$ is matched to two different peaks in the spectra above it.

Clearly, the window width used in peak detection and peak matching routines will affect subsequent classification results as is evident from Table 11. Unless the mass accuracy of the mass spectrometer used is known exactly, this mass accuracy can be used as an adjustable parameter in classification procedures. In the next section, we make use of this fact.

IV.3. Classification

After performing peak detection and peak matching, peaks from all spectra will be matched together so that we have tens or hundreds of clusters of peaks. Each cluster of peaks represents peaks in various spectra with m/z values in some proximity of each other. This clustering facilitates a classification problem where each spectral observation has p covariates often consisting of some intensity measure at each of the p peak cluster locations. This representation enables the use of a wide range of classifiers, several of which were compared in a study by Wu *et al.* In this section, we compare the performance of some different classifiers for its own sake, as well as

a means to evaluate the merits of our work in previous chapters.

Using our MALDI data, we compared the performance of peak probability contrasts (PPC), locally adaptive discriminant analysis (LADA) (Wu and West, 2008) and adaptive boosting (AdaBoost) (Yasui *et al.*, 2003b) using the peak inputs from different pre-processing techniques including our model-based representation. As a benchmark for comparison, we had originally hoped to outperform an average of 23 misclassified spectra that was obtained using peak probability contrasts and ten-fold cross validation for this data. However, we were unable to duplicate those results using the supplementary code provided by the authors, so we used our own attempts to reproduce these results to provide two benchmarks, which are described later. We now present a brief summary of the PPC, LADA and AdaBoost methods.

IV.3.1. Peak Probability Contrasts

The peak probability contrasts method (Tibshirani *et al.*, 2004) begins by iteratively searching each peak cluster for an optimal split point which maximally discriminates between the cancerous and healthy spectral peak intensities at that location. The candidate split point is denoted as the α-quantile among the peak intensities at location i, $q(\alpha, i)$. For each candidate value of $q(\alpha, i)$, the proportion of cancerous spectra and healthy spectra at this location with peak intensity values above $q(\alpha, i)$ is computed as $p_{i1}(\alpha)$ and $p_{i2}(\alpha)$, respectively. The optimal split point at location i is denoted as the $q(\alpha, i)$ which maximizes the absolute difference in class proportions above the split point, $|p_{i1}(\alpha) - p_{i2}(\alpha)|$. If there are p locations, each disease class is represented by the vector of these maximally discriminated proportions, $(p_{11} \dots p_{p1})^T$ and $(p_{12} \dots p_{p2})^T$, respectively.

Classification for a new spectrum is done by defining a binary feature which has

a 1 at location i if its peak intensity at location i is at least as large as the optimal split point at location i and 0, otherwise. The new spectrum will be predicted to class 1 (say, cancerous) if the distance between its feature vector and the class 1 proportion vector, described in the previous paragraph, is smaller than the distance to the class 2 proportion vector in some metric. Additionally, feature selection is done by finding an optimal choice of a soft threshold parameter, δ, which shrinks the class proportions towards .5 and essentially removes locations where the discriminated class proportions are less than δ in absolute value. This approach provides a means for decreasing the importance of spurious or otherwise unimportant features.

IV.3.2. Locally Adaptive Discriminant Analysis

Locally adaptive discriminant analysis (Wu and West, 2008) is an ensemble classifier which initially fits several classifiers (say, c classifiers) to the peak intensities at each peak location. The LADA method does not require that every spectra have an intensity (peak or non-peak) at each location. At each location, the c classifiers are fit to the intensities in the training data and yield c predicted classifications, consisting of $C_j = -1$ for a healthy prediction or $C_j = +1$ for cancerous prediction, using classifier j. At each location, the error rate of classifier j is denoted by e_j, and the class prediction from the classifier with the minimizing e_j is retained. Thus, the equivalent feature vector for a single spectrum is a binary vector consisting of +1's and/or -1's corresponding to the class predictions at each peak location of the respective best classifiers. For each spectrum a score is then calculated from a weighted average of the elements of the binary classification vector, where the weight at location i, is computed as $w_i = 1 - e_i$. This representation places the largest weights on classifiers which perform best.

Through leave-one-out and ten-fold cross-validation, an optimal cutoff error rate, e_c, is determined so that $w_i = 0$ if $e_i > e_c$. For each spectra in the training set, s, a score is calculated as

$$score_s = \Sigma_i w_i C_i,$$

using the weights and best classifiers at locations where spectra s had a peak present. An optimal threshold score, t, is computed which maximally discriminates between the scores of the cancerous and healthy spectra. Prediction for a new spectra from a test set is done by first using the corresponding best classifiers where the test spectrum has detected peaks. The score for the new spectrum is calculated using these classification results and the weights computed from the training sample. If the score of the new spectrum is larger than the computed threshold, the new spectrum is classified as cancerous, otherwise, healthy.

Like the PPC method, LADA also uses a threshold parameter as variable selection criteria, where locations can be assigned a weight of zero if the corresponding classifier error rate is too large. This threshold can be similarly estimated using a ten-fold or leave-one-out cross-validation procedure. Another attractive property of the LADA method is that it can use one-dimensional as well as two-dimensional peak information from the spectra. In our LADA procedure, we fit classifiers using one-dimensional (intensity only) and two-dimensional (intensity and location) peak information with 1- and 3-nearest neighbor classification and linear and quadratic discriminant analysis. Using such information can dramatically increase predictive power if there is a systematic shift in mass-to-charge ratio at a particular location between disease classes. Also, the computational intensity of the LADA method can be adjusted by controlling the number of classifiers fit at each location.

IV.3.3. Adaptive Boosting

Adaptive boosting is similar to the LADA method in that it is ensemble classifier. The AdaBoost procedure uses the outputs of a sequence of M weak classifiers to form a much stronger classifier, by committee. These weak classifiers are formed iteratively using modified versions of the training data. Initially, each of the spectra are weighted equally with $w_i = 1/N$, where i indexes the training spectra and N is the number of spectra in the training set. For each iteration of the AdaBoost algorithm, a binary classifier, $G_m(x)$, is fit to the training data, where m indexes the classifier in sequence (initially, $m = 1$) and x indicates a vector of predictors. If y_i indicates the true disease status of the ith spectrum, the misclassification error of the mth classifier is computed as

$$err_m = \frac{\sum_{i=1}^{N} w_i I(y_i \neq G_m(x_i))}{\sum_{i=1}^{N} w_i}.$$

From this misclassification error rate, the measure $\alpha_m = (1-err_m)/err_m$ is computed, which indicates the importance of mth weak classifier in the final classification. Like LADA, the weak classifiers that perform better receive larger weight. The weights for each spectra are updated as follows: $w_{i+1} \leftarrow w_i \cdot \exp(\alpha_m \cdot I(y_i \neq G_m(x_i)))$. This iterative re-weighting increases the importance of spectra which were misclassified by the previous weak classifier, so the AdaBoost procedure increases weights for more accurate classifiers and rogue spectra that are difficult to predict.

After M weak classifiers are fit, we obtain the predicted class of disease status, $G_m(x), m = 1, \ldots, M$ and the importance measure of each classifier, $\alpha_m, m = 1, \ldots, M$. The final "committee" classifier is then

$$G(x) = \sum_{m=1}^{M} \alpha_m G_m(x).$$

Table 12. Average number (and standard error) of misclassified spectra after applying PPC to our model-based peaks, using different normalization techniques. Results in this table use all 89 spectra to identify peak cluster locations and split points.

Normalization	PPC	LADA	AdaBoost
Model-based σ_t	.272 (.004)	.216 (.005)	.258 (.011)
No normalization	.374 (.004)	.289 (.010)	.366 (.008)
Log and $P_{90} - P_{10}$	.297 (.002)	.198 (.007)	.275 (.015)
$P_{100} - P_0$	.285 (.002)	.266 (.003)	.391 (.011)
Mean Ion Current	.299 (.004)	.275 (.013)	.334 (.018)

In Table 12, we present results which compare misclassification results for different normalization procedures for three classifiers, PPC, LADA and AdaBoost. The model-based local variance normalization procedure is applied after fitting the mixture model and baseline in the spirit of Chapters II and III. After our pre-processing method is applied to all of the spectra, peaks are harvested as baseline-corrected intensities which are local maxima in a window of ± 100 points and whose largest responsibility is not from the normally distributed machine error component. After using the peak clustering algorithm from Tibshirani *et al.* (2004), the unnormalized baseline-corrected intensities are input into the three classifiers to produce the results in the "No normalization" row. The results in the "Model-based σ_t" row normalize the above baseline-corrected intensities *after* model fitting, by dividing these intensities by the square root of the local estimate of σ_t^2.

We use the estimate of σ_t as a normalization approach since spectra with relatively large peaks compared to other spectra, often possess larger noise estimates, as well. The mean ion intensity normalization approach stems from a similar idea, however, it is well known that the mean intensity in a mass spectrum will be highly affected a few very large peak intensities. The $P_{100} - P_0$ normalization technique

transforms the largest intensities in each spectrum to 1, so it can also be affected by one very large (or very small) intensity. Moreover, this transformation is especially problematic if the largest intensity in each spectrum is often due to the same biological feature, since this will result in a serious compromise of discriminatory power among the spectra at this location.

To compare the results of our approach, we carried out the pre-processing procedure for this data outlined in by Tibshirani *et al.* using code from the authors. Our attempt to reproduce the results using the authors' code yielded a mean misclassification rate of .343 with a standard error of .005, which is significantly larger than .258 (23/89) as claimed by the authors. Although the results use all 89 spectra to identify peak cluster locations and split points, there are several things to note in Table 12. Not surprisingly, our normalized peaks perform better than the unnormalized peaks for all three classifiers tested, which illustrates that it may be advantageous to use some model-based normalization techniques. Our model-based normalization uniformly outperforms the three other normalization techniques, except in one case where the tuned normalization from Tibshirani *et al.* performs better using the LADA classifier.

The last three rows in the table refer to normalization procedures applied *prior* to the fitting of our mixture model to the baseline-corrected intensities. Each of these pre-processing approaches uses a 'loess' baseline with a span of 1000/91360 after applying the normalization. The log and $P_{90} - P_{10}$ transformation is the normalization of choice for this data by Tibshirani *et al.*; after a log transformation was fit, a linear transformation mapped the 10th and 90th percentiles of the baseline-corrected intensities to 0 and 1, respectively. The $P_{100} - P_0$ transformation is a similar transformation which maps the minimum and maximum baseline-corrected intensities to 0 and 1, respectively, prior to model fitting. The mean ion current divides the inten-

sities by the average intensity in the spectrum prior to model fitting. After all of the spectra have been normalized and baseline-corrected, our mixture model is fit to the spectra to find the responsibility-based peaks.

The log and $P_{90} - P_{10}$ normalization circumvents some of the drawbacks that afflict the mean ion intensity and 0-1 normalization approaches. However, the $P_{90} - P - 10$ normalization is somewhat specific to this data set, which is convenient in that these percentiles correspond to points near the bounds of machine noise. However, this approach may not work in general, since not all spectra have 90% (or more) points in the machine noise component. We illustrate this fact with a low-resolution SELDI data in the next section, which has peaks which extend over much larger ranges of m/z than in this particular MALDI data set.

To obtain the results in the bottom three rows, the same initial baseline subtraction and normalization were applied and our mixture model was fit to the preprocessed data. Peaks were then harvested in the same manner as the results from unnormalized and post-fit normalized spectra in the two rows above. For all of the results in Table 12, we used the peaks from all spectra to identify the peak clusters, split points and train the classifiers, so the misclassification results are slightly optimistic, however, that should not affect the comparison of the results within Table 12.

Since the misclassification results in Table 12 possess some in-sample qualities and are slightly optimistic, we decided to compute the number of misclassified spectra using strictly out-of-sample results while comparing baseline estimation techniques. To obtain the results in Table 13, we first used a single baseline estimate from a 'loess' smooth with a span of 1000/91360 points prior to any normalization. For the results in the top two rows, we did not update the initial baseline estimate as we had in producing the results of Table 12. For the results in the bottom three rows, we normalized the baseline-corrected intensities prior to fitting our mixture model.

Table 13. Leave-one-out cross-validation of misclassified spectra after applying PPC, LADA and AdaBoost to our model-based peaks, using different normalization techniques and only an initial baseline estimate. Results in this table use only training spectra to identify peak cluster locations and train classifiers.

Normalization	PPC	LADA	AdaBoost
Model-based σ_t	.292	.393	.326
No normalization	.371	.404	.360
Log and $P_{90} - P_{10}$	.326	.360	.202
$P_{100} - P_0$	.382	.348	.303
Mean Ion Current	.371	.382	.225

Table 14. Leave-one-out cross-validation of misclassified spectra after applying PPC, LADA and AdaBoost to our model-based peaks, using different normalization techniques and an updated baseline estimate. Results in this table use only spectra in the training set to identify peak cluster locations and train classifiers.

Normalization	PPC	LADA	AdaBoost
Model-based σ_t	.348	.404	.303
No normalization	.393	.506	.337
Log and $P_{90} - P_{10}$	.438	.483	.371
$P_{100} - P_0$	.382	.348	.315
Mean Ion Current	.382	.360	.315

After our mixture models were fit, we harvested peaks using responsibilities from all spectra, regardless of normalization, and used a one-dimensional clustering to cluster the peaks. The peaks harvested from the top rows had their baseline-corrected intensities standardized by the local variance estimate, prior to entrance in the classifiers.

In Table 14, we used the same pre-processing approach as in Table 13, however we allowed the baseline to be updated regardless of normalization technique. In Table 12, we updated the baseline for the spectra for the results in the top two rows, but retained the original baseline throughout model fitting for the bottom three rows. Through the comparison of Tables 13 and 14, we can better ascertain where our superiority lies in Table 12; that is, how advantageous are our model-based baseline estimation and normalization approaches? In addition, we computed the number of misclassified spectra more honestly using out-of-sample classification performance.

After model fitting, peak detection and peak matching, we used leave-one-out cross-validation estimate the mean number of misclassified spectra. For each run, one of the spectra (and their associated peaks) are withheld, while the training of the classifier is performed using the remaining spectra (and their associated peaks). The cancer status of each withheld spectra is unbiasedly estimated using the trained classifier, and this process is repeated for the remaining nine folds in the fold choice. This is a computationally intensive procedure, so it is important to balance the computing time with the number of repetitions desired. We would also like to point out that we did not harvest local maxima from each spectrum after peak clustering where each spectrum had no detected peaks, as we had done previously.

The results in Table 13 are useful for comparison with Table 14, but are also useful for comparing our peak detection results with the peak detection algorithm in Tibshirani *et al.*. To adjust our earlier optimistic results, we used the same 10-fold cross-validation procedure to estimate the number of misclassified spectra. We briefly

describe our attempt to reproduce this result for matter of completeness. We carried out the pre-processing techniques where the spectral intensities underwent log and linear transformations, prior to applying a 'loess' baseline estimate and supersmoother to remove the isotopic envelop. Peaks were harvested as local maxima within a window of ± 100 points and then clustered together using a one-dimensional clustering on the $log(m/z)$ scale. After applying the PPC classifier for 25 folds, we found an average and standard error of .388 and .005 misclassified spectra, respectively. Again, this was higher than what was claimed by the authors, and our model-based normalization compares favorably with this result.

To summarize our findings from Tables 12, 13 and 14, it is clear that our baseline estimation procedure yielded *h*igher misclassification rates when compared with the initial baseline estimation. While this is somewhat disappointing, but it is not entirely surprising since the method of selecting λ from Chapter III was not exactly optimal. We consider this fact further in Chapter V, where we consider a grid search of the log-likelihood over many parameters, including the initial baseline smoothness. Our normalization idea worked well for the PPC classifier, but less so for the other classifiers. In the next section, we evaluate our normalization method with a SELDI dataset.

IV.4. Application to SELDI Data

Throughout this document, our examples have referred to a popular MALDI dataset. In this section, we show the application of our method to data that is different in many ways from the ovarian cancer data set used in the previous chapters. We used a set of mass spectra that are produced according to SELDI-TOF specifications and was previously analyzed in Petricoin *et al.* (2002c). The mass spectra were collected

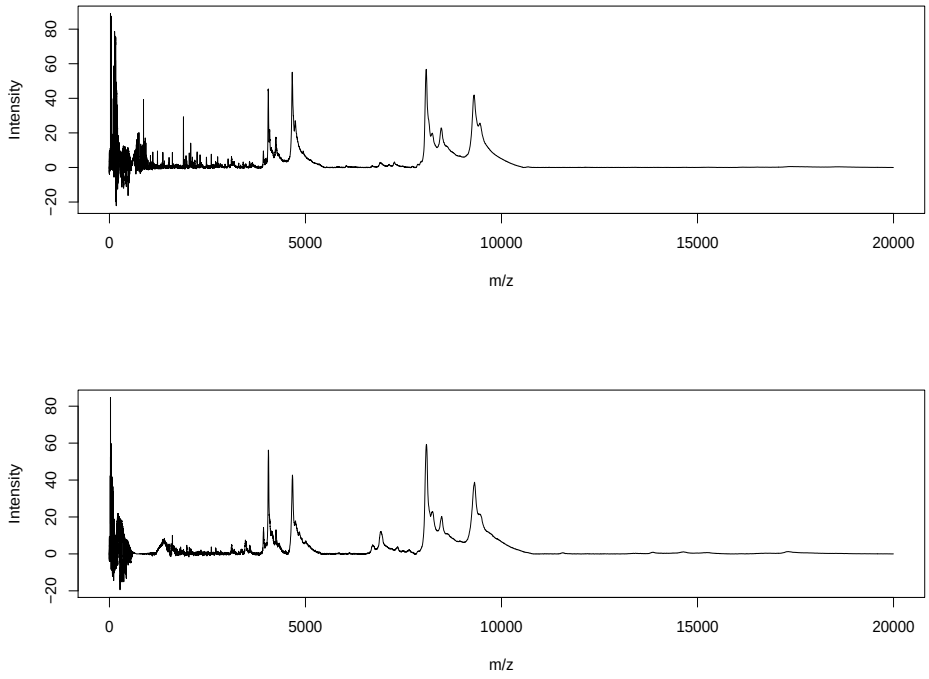

Fig. 40 SELDI mass spectra of a control subject with low PSA level and prostate cancer patient with a highly-elevated PSA level in the top and bottom plots, respectively.

from a study set of patients with a histopathologic diagnosis of prostate cancer or no evidence of prostate cancer. These low resolution SELDI mass spectra were measured at 15,191 locations ranging from 0 to 20,000 m/z with the baseline already subtracted. Mass spectra of two subjects are presented in Figure 40. The top spectra shows a subject with low PSA level and no evidence of cancer, while the bottom spectra represents a patient with highly-elevated PSA level. The spectra in this dataset had similar ranges of intensity; most spectral intensities were located between -20 and 80 for each spectra. The data were found at the Clinical Proteomics National Databank (http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp).

This data was selected for many reasons; we will state a few of them here. First, SELDI data do not possess the isotopic envelope that is often present in MALDI data, and SELDI data tend to be less "messy". Second, this SELDI data set has a much

larger range (0-20,000 m/z) than the MALDI data set (800-3,500 m/z) analyzed in previous chapters. Third, despite the much larger range, the resolution of the SELDI data is much lower; the number of points in this SELDI data set is about 1/6 of the number of points in the MALDI data set. Lastly, these data were already baseline-corrected, which facilitates a more honest comparison of normalization methods.

We first used our mixture model to fit the baseline-corrected intensities, and discarded all of the spectral information below $m/z = 800$ since the mass spectra are quite noisy and this region provides no biologically discriminating information. In an attempt to compare some of our results, we divided the data into training and test sets as in Petricoin *et al.* (2002c), where the training set comprised 25 patients with no evidence of disease and PSA levels ≤ 1 ng/mL and 31 patients with biopsy-proven prostate cancer and PSA levels ≥ 4 ng/mL. The test set contained 76 patients, with 38 patients in each of the aforementioned groups.

We divided the data into training and test sets as described in the previous paragraph 100 times. For each training set, the peaks are clustered together using the one-dimensional clustering technique described in Tibshirani *et al.* on the $\log(m/z)$ axis. There is an important point in clustering on the $\log(m/z)$ axis, as was briefly alluded to earlier in the chapter. Two peaks at $m/z = m_p$ and $m/z = m_p(1 + a)$ would aligned together which a relative mass accuracy of $100a\%$, however the absolute distance between these points depends on their mass, m_p. However, the distance between these points on the log scale is $log(m_p(1 + a)) - log(m_p) = log(1 + a)$, which is independent of the peak masses. Conveniently, $log(1 + a)$ is approximately equal to a for small and positive values of a, which is our range for the mass accuracy, a, as a decimal. Thus, peaks can be aligned according to relative mass accuracy on the log scale, regardless of mass.

We used four different values of the adjustable parameter `peak.gap`, to cluster

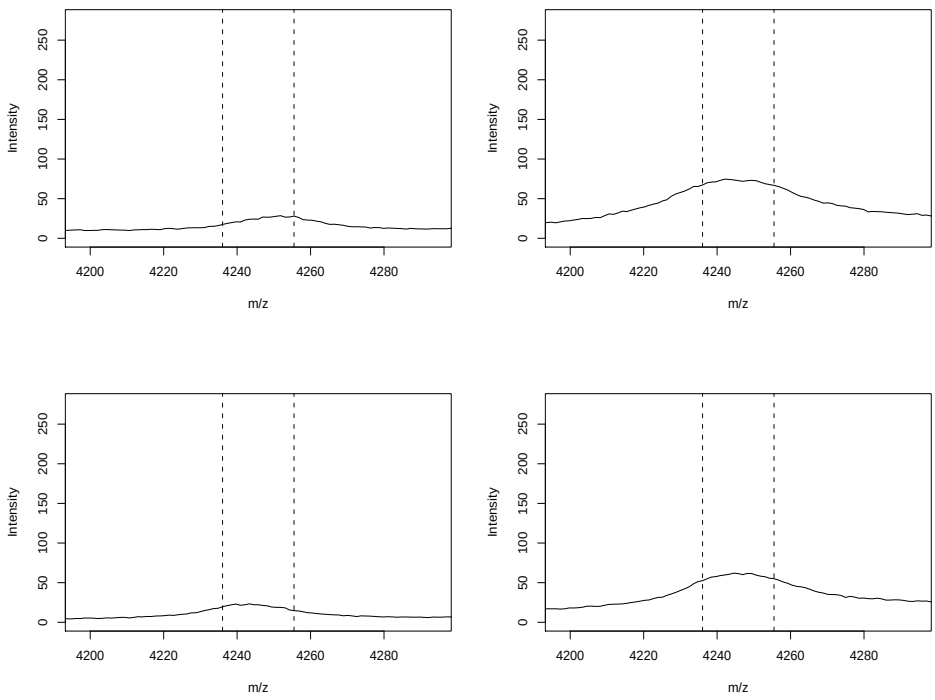

Fig. 41 SELDI mass spectra in the vicinity of a potentially discriminating location. The left graphs represent healthy spectra and the right graphs represent cancerous spectra.

peaks from the various training spectra. For comparative purposes, these values are equivalent to peak matching with mass accuracies of .125%, .250%, .374% and .499%, respectively. Thus, the equivalent mass accuracy, as a decimal, is roughly half of the peak gap.

After clustering the peaks together we then applied the PPC and LADA methods to the clusters of peaks. In Figure 41, we show one of the discriminating locations identified by PPC. The graphs on the left show the two largest peaks at this location from the healthy spectra in one of the training sets, while the graphs of the right show the two largest peaks from the cancerous spectra in the same training set. At this location, the peaks tend to be larger in the cancerous spectra. In Figure 42, we show a visual description of the PPC method applied at this location. The left graph shows the normalized peak intensities in the training data at this location, where the circles

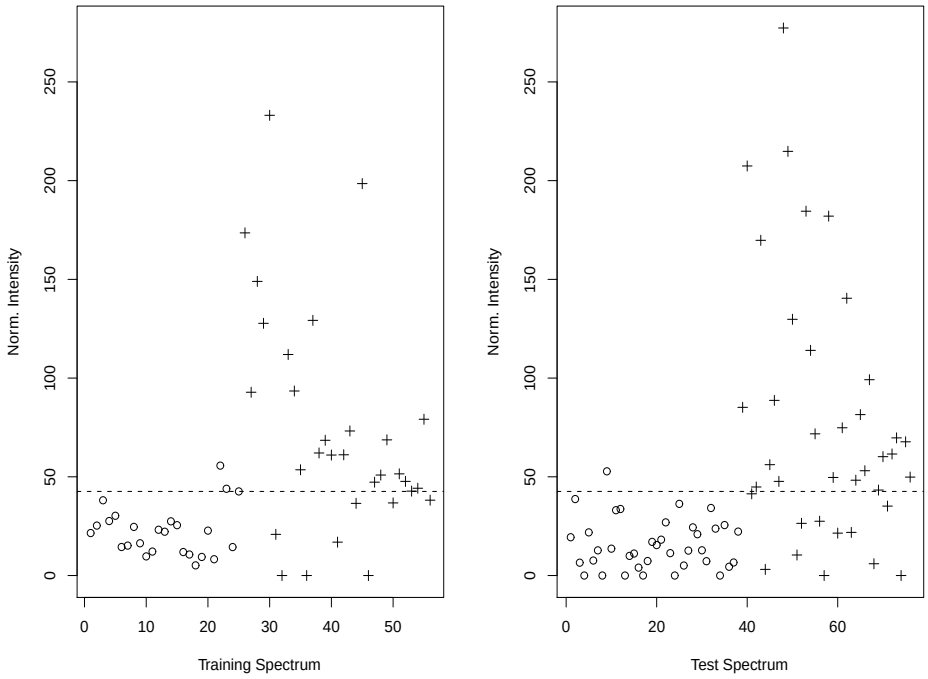

Fig. 42 Application of the PPC classifier to the peaks with centroid at $m/z = 4246.37$. The circles represent the normalized peak intensity for each of the healthy spectra, while the pluses represent the normalized peak intensity for each of the cancerous spectra at this location.

represent healthy spectra and the pluses represent cancerous spectra. The dotted line is the optimal split point estimated by PPC at this location for the training data. In the right graph, we show the normalized peak intensities in the test spectra with the same split point estimated from the training data. In this instance, spectra with normalized peak intensities larger than the split point will be classified as cancerous at this location and assigned a value of 1 in its feature vector. Spectra with normalized peak intensities below the split point will be classified as healthy, at this location, and assigned 0 at this location of its feature vector.

For each test set of 76 subjects, the sensitivity, specificity and misclassification proportion are computed, using our normalization procedure and competing procedures described previously. Note that since the spectra were already baseline-

Table 15. Average (and standard error) of misclassification rates after applying PPC, LADA and AdaBoost to our model-based peaks in the data from Petricoin *et al.* (2002c) using two different mass accuracies for peak matching. Results in this table use only training spectra to identify peak cluster locations and train classifiers.

Normalization	PPC, .005	LADA, .005	AdaB, .005
Model-based σ_t	.152 (.007)	.155 (.005)	.129 (.007)
No normalization	.194 (.007)	.197 (.006)	.127 (.010)
$P_{90} - P_{10}$	.146 (.007)	.163 (.005)	.110 (.007)
$P_{100} - P_0$	.163 (.008)	.179 (.005)	.142 (.011)
Mean Ion Current	.174 (.009)	.174 (.005)	.116 (.007)
Normalization	PPC, .01	LADA, .01	AdaB, .01
Model-based σ_t	.154 (.005)	.134 (.004)	.116 (.006)
No normalization	.168 (.006)	.179 (.005)	.126 (.007)
$P_{90} - P_{10}$	.161 (.008)	.154 (.006)	.101 (.008)
$P_{100} - P_0$	.172 (.007)	.191 (.005)	.113 (.008)
Mean Ion Current	.165 (.008)	.161 (.005)	.113 (.007)

corrected, the only the divisors of the normalization procedures were used and no log transformation was applied prior to the $P_{90} - P_{10}$ normalization. We want to point out that the comparison of our proposed normalization with unnormalized spectra is not entirely foolhardy since the range of intensities changes very little across spectra from this data set. From Figure 40, it may seem like little discriminatory information can be gained through normalization, however, the results in Table 15 would suggest otherwise, as the error rate is uniformly lower for the normalized spectra. In fact, the difference in error rates between our normalized and unnormalized peaks are highly significant (p-value $< .01$), with the exception of the largest peak gap.

Our proposed model-based normalization compares very favorably to competing normalization procedures. Our model-based approach had uniformly smaller misclassification rates than both the $P_{100} - P_0$ and the mean ion current normalizations, and performed better than the $P_{90} - P_{10}$ normalization in more than half of the comparisons. To show a more detailed view of where our improvement in the misclassification rates lies, we computed the sensitivities and specificities in Tables 16 and 17.

Table 16. Average (and standard error) of specificities after applying PPC and LADA to our model-based peaks in the data from Petricoin *et al.* (2002c). Results in this table use only spectra in the training set to identify peak cluster locations and train classifiers.

Normalization	PPC, .0025	PPC, .005	PPC, .0075	PPC, .01
Model-based σ_t	.789 (.0104)	.777 (.0129)	.759 (.0164)	.769 (.0163)
No normalization	.676 (.0147)	.682 (.0153)	.673 (.0150)	.732 (.0111)
$P_{90} - P_{10}$	.716 (.0178)	.744 (.0132)	.748 (.0142)	.721 (.0164)
$P_{100} - P_0$	.711 (.0171)	.716 (.0164)	.689 (.0156)	.707 (.0124)
Mean Ion Current	.692 (.0145)	.696 (.0187)	.695 (.0148)	.735 (.0155)
Normalization	LADA, .0025	LADA, .005	LADA, .0075	LADA, .01
Model-based σ_t	.711 (.0130)	.786 (.0117)	.796 (.0120)	.819 (.0100)
No normalization	.661 (.0110)	.720 (.0110)	.714 (.0108)	.753 (.0119)
$P_{90} - P_{10}$	.753 (.0099)	.761 (.0106)	.785 (.0108)	.785 (.0107)
$P_{100} - P_0$	.691 (.0114)	.754 (.0109)	.780 (.0096)	.777 (.0096)
Mean Ion Current	.658 (.0118)	.748 (.0107)	.762 (.0107)	.778 (.0010)

The high sensitivity for this data set was reported in Petricoin *et al.* (2002c), as they correctly classified 36 of the 38 patients with prostate cancer and high PSA levels in their test set. However, this estimate is believed to be computed from a single partition of the training and test set. We want to point out that the sensitivities for some of our test sets were exactly 1, and our mean sensitivity is not significantly different from the high sensitivity claimed by the authors. It is interesting to note that our sensitivities appear to be slightly lower using the PPC classifier, but with a larger increase in the specificities.

In this chapter, we have used our model-based pre-processing procedures as inputs into peak detection and several classification procedures. In Chapter I, we outlined several pre-processing steps which are typically performed in the analysis of mass spectral data. We have shown that our approach provides model-based procedures for baseline correction, normalization and peak detection. In this chapter, we have shown that there may be some merit to using such model-based procedures insofar as classification performance is performed, as we have obtained promising results using different data sets and classifiers. In the next chapter, we outline future

Table 17. Average (and standard error) of sensitivities after applying PPC to our model-based peaks in the data from Petricoin *et al.* (2002c). Results in this table use only spectra in the training set to identify peak cluster locations and build classifier.

Normalization Divisor	PPC, .0025	PPC, .005	PPC, .0075	PPC, .01
Model-based σ_t	.923 (.0055)	.920 (.0057)	.909 (.0070)	.923 (.0065)
No normalization	.945 (.0053)	.930 (.0058)	.923 (.0064)	.933 (.0059)
$P_{90} - P_{10}$	.957 (.0044)	.963 (.0039)	.961 (.0040)	.958 (.0045)
$P_{100} - P_0$	.963 (.0035)	.959 (.0039)	.958 (.0039)	.949 (.0049)
Mean Ion Current	.967 (.0032)	.956 (.0035)	.952 (.0043)	.935 (.0051)
Normalization Divisor	LADA, .0025	LADA, .005	LADA, .0075	LADA, .01
Model-based σ_t	.908 (.0074)	.904 (.0067)	.913 (.0065)	.914 (.0064)
No normalization	.896 (.0068)	.886 (.0078)	.884 (.0088)	.889 (.0082)
$P_{90} - P_{10}$	.913 (.0066)	.913 (.0071)	.919 (.0066)	.906 (.0063)
$P_{100} - P_0$	.901 (.0078)	.887 (.0089)	.865 (.0092)	.842 (.0097)
Mean Ion Current	.933 (.0062)	.904 (.0069)	.904 (.0075)	.900 (.0066)

improvements to our method.

CHAPTER V

CONCLUSION

Throughout the course of study that motivated this document, we have presented many important questions and often several attempts to provide suitable answers to these questions. For example, most of this document has assumed that mass spectra can be appropriately modeled with parametric mixture models. While our parametric mixture model is flexible, this structure may not be sufficiently accommodating as a general purpose algorithm for all mass spectra, regardless of ionization method, mass analyzer choice, etc. Thus, one improvement for our method would address a need for increased flexibility. Since the goal of pre-processing is to extract biologically relevant peaks from mass spectra, it may not be necessary to impose a parametric density on the intensities, if peak and non-peak intensities are sufficiently distinguishable using another choice of parametric components or some nonparametric assessment. Using a nonparametric approach raises a new set of questions, however, implementation of nonparametric densities does not prohibit the use of the EM Algorithm or model-based estimates of the baseline, peaks or normalization.

Before we delve into the specifics of the improvements of our existing method, it is important to exploit whatever advantages our method might hold over existing methods. The fitting of our model to the large number of intensities certainly makes computational implementation a concern, especially in cases where high-throughput mass spectral analysis is desirable. Along those lines, high-throughput procedures generally deem manually-tuned pre-processing techniques as impractical, thus we want our model-based procedures to be largely automatic with a minimum number

of user inputs. For the purposes of outlining future improvements for our method, we want to maintain a constant appreciation for automation in the name of high-throughput procedures.

An important matter of consideration is the optimal number of effective parameters, which is dependent on several inputs, including the smoothness of the baseline and the number of local noise variance estimates. In Chapter II, we addressed the optimal number of components using AIC and BIC, after fitting our model over a grid m. In Chapter III, we used generalized cross-validation and restricted maximum likelihood to determine an optimal smoothing parameter which maximized a penalized likelihood to obtain a suitable baseline. Thus, if we want to properly consider the optimal number of components using the AIC or BIC, a grid search for the number of fitted component densities should also include the number of local variance estimates and baseline estimate and its equivalent number of parameters. Clearly, such an exhaustive grid search increases the computation time to find this optimal number of components. Ideally, our model-based procedure would be able to automatically identify a reasonable number of parameters, while considering the variance and baseline estimates, as well, through a computationally feasible grid search.

In Table 18, we show the BIC values for a selection of parameters and inputs. The table shows two baseline estimates, $f_{9.138}$ and $f_{913.8}$, which are initial baseline estimates provided by the 'loess' function using the `enp.target` argument, which provides a baseline with an equivalent number of parameters equal to `enp.target`. These arbitrary values were obtained by dividing the number of points in the spectrum (91380) by 10,000 and 100, respectively. The value of m determines the number of mixture components; specifically, it includes 1 normal component and $m-1$ peak components. The values for $m.sec$ and $v.sec$ indicate the number of sections to divide the spectra on the m/z axis for fitting multiple components $\{\pi_j, \alpha_j, \theta_j\}$ and multiple

Table 18. BIC computed for 54 different combinations of model and parameter choice for a single spectrum.

		$f_{9.138}$			$f_{913.8}$		
		$m=2$	$m=3$	$m=4$	$m=2$	$m=3$	$m=4$
$m.sec=1$	$v.sec=1$	1077283	1076492	1075554	1131860	1131722	1131408
	$v.sec=3$	1058880	1057723	1057735	1089169	1088937	1088968
	$v.sec=5$	1051588	1050552	1050570	1068302	1068049	1068080
$m.sec=2$	$v.sec=1$	1077028	1075576	1075284	1131772	1131645	1131371
	$v.sec=3$	1058907	1057490	1057541	1089116	1088927	1089000
	$v.sec=5$	1051618	1050567	1050409	1068276	1068058	1068125
$m.sec=3$	$v.sec=1$	1076346	1074901	1074792	1131504	1131231	1131331
	$v.sec=3$	1057851	1056825	1056922	1088693	1088324	1088429
	$v.sec=5$	1050998	**1049945**	1050016	1067902	1067732	1067825

machine noise variances, σ_t^2. That is, when $m.sec = 2$, the spectrum is divided into two non-overlapping and exhaustive sections along the m/z and m components are estimated for each section.

From this small example, we can see that the BIC would select a smoother baseline with $m = 4$ mixture components, to be fit separately in each of $m.sec = 3$ sections. The BIC also prefers that the underlying noise estimate, σ_t^2, should be estimated over five sections of the spectrum, instead of one or three sections. The results from the AIC support these choices (table not shown). As presented in Chapter II and in Table 14, our approach selects the number of mixtures for each spectrum only after several models with different choices of m have been sufficiently fit according to some possibly stringent convergence assumptions. This is not computationally efficient, especially if the number of mixture components can be determined after a much smaller number of iterations or through some a priori determination. If determining the effective number of model parameters is indeed important to classification performance, finding a timely surrogate for the converged AIC or BIC criterion is important.

We have shown examples of both MALDI and SELDI spectra which have exhibited a noise variance which is nonconstant across the range of m/z. In Chapter IV,

we showed that there may be some promise in using the estimation of this machine noise to serve in a post-fit, model-based normalization procedure, so there may predictive merits to a well-estimated machine noise component. We obtained our local variance estimates by assuming that the change in machine noise variability was sufficiently gradual so that it could be viewed as constant over relatively small ranges of m/z. While this assumption provided reasonable estimates and classification results, a model that allows the variance to change more smoothly may be more appropriate.

In Chapter III, we investigated the premise of updating the baseline through a penalized likelihood form of our model. This approach can be somewhat prohibitive due to the large number of points that are typically found in mass spectral data. We have seen that we can obtain smooth baselines from very wiggly initial estimates and vice versa, but the computation time required to employ this idea cannot be overlooked. The use of a grid search as described earlier in the chapter may be able to identify an initial baseline that is similar in smoothness to an "optimal" baseline, which is likely to decrease the computation time. In addition, the number of points required to estimate a smoother baseline should be smaller than a wigglier baseline. Thus, if the notion of a "best baseline" is quite smooth, we can estimate it using a smaller number of points. We used this idea in Chapter III.

A natural solution to updating the baseline with a smaller number of points or knots may be the use of P-splines (Eilers and Marx, 1996). The use of P-splines is a combination of B-splines and difference penalties on coefficients of the B-splines so that the notion of roughness penalty developed in Chapter III is not moot. The advantages of using P-splines include the local fitting of polynomials at knots determined by the analyst, which makes the spacing of the raw intensities on the m/z axis irrelevant. Most importantly, the number of knots needed to produce a sufficiently flexible baseline may be far smaller (i.e., 50) than the number of intensity locations

in each mass spectrum. This makes for a much more practical computation time.

One part of the estimation procedure we have largely ignored is associated with the isotopic behavior of MALDI spectra. As we pointed out in Chapter I, some authors have attempted to model the periodic behavior that afflicts high resolution MALDI data. In Chapter III, our generalized cross-validation estimate of the smoothing parameter was affected by this property of the MALDI data set, however we determined that this baseline estimate was unsuitable due to its roughness and its penchant to disproportionately raze peaks of interest more than the isotopic noise envelop. It may be desirable to find a modeling solution for such behavior whether it be through an appropriately chosen value of the smoothing parameter or an extension of our model which models this somewhat periodic component, separately from the trend of the baseline.

In Chapter IV, we have used the responsibilities to indicate the presence of peaks in the mass spectra. However, due to our desire to handle the raw data without any prior normalization, the number of peaks varies in the mass spectra, even within disease classes. It would be nice to find a way to use the responsibilities across spectra in a way that keeps the number of peaks more uniform. This will help to disregard spurious peaks that are typically found in noisier spectra, where our model likelihood tends to be lower than for spectra that are better behaved.

REFERENCES

Adam, B.-L., Qu, Y., Davis, J., Ward, M., Clements, M., Cazares, L., Semmes, O., Schellhammer, P., Yasui, Y., Feng, Z. and Wright, Jr., G. (2002) Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, **62**, 3609–3614.

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

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

Breen, E., Hopwood, F., Williams, K. and Wilkins, M. (2000) Automated Poisson peak harvesting for high throughput protein identification. *Electrophoresis*, **21**, 2243–2251.

Carpenter, M., Melath, S., Zhang, S. and Grizzle, W. (2003) Statistical processing and analysis of proteomic and genomic data. *Proceedings of the Pharmaceutical SAS Users Group*, **21**, 545–548, Miami, FL.

Conrads, T., Fusaro, V., Ross, S., Johann, D., Rajapakse, V., Hitt, B., Steinberg, S., Kohn, E., Fishman, D. and Whitely, G. (2004) High-resolution serum proteomic features for ovarian cancer detection. *Endocrine-Related Cancer*, **11**, 163–178.

Coombes, K., Fritsche, H., Clarke, C., Chen, J., Baggerly, K., Morris, J., Xiao, L., Hung, M. and Kuerer, H. (2003) Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clinical Chemistry*, **49**, 1615–1623.

Coombes, K., Wang, J. and Baggerly, K. (2004) A statistical method for finding biomarkers from microarray data, with application to prostate cancer. *The University of Texas MD Anderson Cancer Center*, Technical Report UTMDABTR-001-04.

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

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

Dubitsky, W., Granzow, M. and Berrar, D. (2007) *Fundamentals of Data Mining in Genomics and Proteomics*. Boston: Manning Publications Co.

Eilers, P. (2004) Parametric time warping. *Analytical Chemistry*, **76**, 404–411.

Eilers, P. and Marx, B. (1996) Flexible smoothing using B-splines and penalties. *Statistical Science*, **2**, 89–121.

Furman, W. and Lindsay, B. (1994) Testing for the number of components in a mixture of normal distributions using moment estimators. *Computational Statistics and Data Analysis*, **17**, 473–492, 1994.

Fushiki, T., Fujisawa, H. and Eguchi, S. (2006) Identification of biomarkers from mass spectrometry data using a "common" peak approach. *BMC Bioinformatics*, **7**, 358–366.

Gras, M., Mueller, M., Gasteiger, E., Gay, S., Binz, P.-A., Bienvenut, W., Hoogland, C., Sanchez, J.-C., Bairoch, A., Hochstrasser, D. and Appel, R. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis*, **20**, 3535–3550.

Green, P. and Silverman, B. (1994) *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach.* London: Chapman & Hall.

Karlis, D. and Xekalaki, E. (1999) On testing for the number of components in finite poisson mixtures. *Annals of the Institute of Statistical Mathematics*, **51**, 149–162.

Karlis, D. and Xekalaki, E. (2003) Choosing initial values for the EM algorithm for finite mixtures. *Computational Statistics and Data Analysis*, **41**, 577–590.

Kozak, K., Amneus, M., Pusey, S., Su, F., Luong, M., Luong, S., Reddy, S. and Farias-Eisner, R. (2003) Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis. *Proceedings of the National Academy of Sciences*, **100**, 12343–12348.

Krivobokova, T. and Kauermann, G. (2007) A note on penalized spline smoothing with correlated errors. *Journal of the American Statistical Association*, **102**, 1328–1337.

Kuerer, H., Coombes, K., Chen, J., Xiao, L., Clarke, C., Fritsche, H., Krishnamurthy, S., Marcy, S., Hung, M. and Hunt, K. (2004) Association between ductal fluid proteomic expression profiles and the presence of lymph node metastases in women with breast cancer. *Surgery*, **136**, 1061–1069.

Li, J., Zhang, Z., Rosenzweig, J., Wang, Y. and Chan, D. (2002) Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, **48**, 1296–1304.

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

Pawitan, Y. (2001) *In All Likelihood: Statistical Modeling and Inference Using Likelihood.* New York: Oxford University Press.

Petricoin, E., Ardekani, A., Hitt, B., Levine, P., Fusaro, V., Steinberg, S., Mills, G., Simone, C., Fishman, D., Kohn, E. and Liotta, L. (2002a) Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, **359**, 572–577.

Petricoin, E., Ornstein, D., Paweletz, C., Arkedani, A., Hackett, P., Hitt, B., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C., Levine, P., Linehan, W., Emmert-Buck, M., Steinberg, S., Kohn, E. and Liotta, L. (2002b) Serum proteomic patterns for detection of prostate cancer. *Journal of the National Cancer Institute*, **94**, 1576–1578.

Petricoin, E., Zoon, K., Kohn, E., Barrett, J. and Liotta, L. (2002c) Translating benchside promise into bedside reality. *Drug Discovery*, **1**, 683–695.

Qu, Y., Adam, B.-L., Yasui, Y., Ward, M., Cazares, L., Schellhammer, P., Feng, Z., Semmes, O. and Wright, Jr., G. (2002) Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, **48**, 1835–1843.

Rai, A., Zhang, Z., Rosenzweig, J., Shih, I., Pham, T., Fung, E., Sokoll, L. and Chan, D. (2002) Proteomic approaches to tumor marker discovery. *Archives of Pathology and Laboratory Medicine*, **126**, 1518–1526.

Sorace, J. and Zhan, M. (2003) A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, **4**, 24–36.

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

Titterington, D., Smith, A. and Makov, U. (1985) *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley.

Torgrip, R., Aberg, M., Karlberg, B. and Jacobsson, S. (2003) Peak alignment using reduced set mapping. *Journal of Chemometrics*, **17**, 573–582.

Wagner, M., Naik, D. and Pothen, A. (2003) Protocols for disease classification from mass spectrometry data. *Proteomics*, **3**, 1692–1698.

Wu, B., Abbott, T., Fishman, D., McMurray, W., Mor, G., Stone, K., Ward, D., Williams, K. and Zhao, H. (2003) Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics*, **19**, 1636–1643.

Wu, Y. and West, W. (2008) Locally adaptive discriminant analysis of mass spectrometry data. Unpublished manuscript, Department of Mathematics, Cleveland State University.

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

Yasui, Y., Pepe, M., Thompson, M., Adam, B., Wright Jr., G., Qu, Y., Potter, J., Winget, M., Thornquist, M. and Feng, Z. (2003b) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, **4**, 449–463.

Yu, W., Wu, B., Lin, N., Stone, K., Williams, K. and Zhao, H. (2006) Detecting and aligning peaks in mass spectrometry data with applications to MALDI. *Computational Biology and Chemistry*, **30**, 27–38.

Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J. and Kovach, J. (2003) Detection of cancer-specific markers amid massive mass spectral data. *Proceedings of the National Academy Sciences*, **100**, 14666–14671.

Zhukov, T., Johnson, R., Cantor, A., Clark, R. and Tockman, M. (2003) Discovery of distinct protein profiles specific for lung tumors and pre-malignant lung lesions by SELDI mass spectrometry. *Lung Cancer*, **40**, 267–279.

VITA

John C. Wagaman

Department of Statistics

Texas A&M University

3143 TAMU

College Station, TX 77843-3143

EDUCATION

2009 Ph.D. Statistics, Texas A&M University

2003 M.S., Statistical Computing, University of Central Florida

2001 B.S., Mathematics, Millersville University

RESEARCH INTERESTS

Statistical Education, Applied Statistics