4

Analysis of Mass Spectrometry Data Using Bayesian Wavelet-Based Functional Mixed Models

Jeffrey S. Morris,
The University of Texas M.D. Anderson Cancer Center

Philip J. Brown,
University of Kent, Canterbury

Keith A. Baggerly,
The University of Texas M.D. Anderson Cancer Center

Kevin R. Coombes,
The University of Texas M.D. Anderson Cancer Center

Abstract

In this chapter, we demonstrate how to analyze MALDI-TOF/SELDI-TOF mass spectrometry data using the wavelet-based functional mixed model introduced by Morris and Carroll (2006), which generalizes the linear mixed models to the case of functional data. This approach models each spectrum as a function, and is very general, accommodating a broad class of experimental designs and allowing one to model nonparametric functional effects for various factors, which can be conditions of interest (e.g. cancer/normal) or experimental factors (blocking factors). Inference on these functional effects allows us to identify protein peaks related to various outcomes of interest, including dichotomous outcomes, categorical outcomes, continuous outcomes, and any interactions among factors. Functional random effects make it possible to account for correlation between spectra from the same individual or block in a flexible manner. After fitting this model using an MCMC, the output can be used to perform peak detection and identify the peaks that are related to factors of interest, while automatically adjusting for nonlinear block effects that are characteristic of these data. We apply this method to mass spectrometry data from an University of Texas MD Anderson Cancer Center experiment studying the serum proteome of mice injected with one of two cell lines in one of two organs. This methodology ap-

pears promising for the analysis of mass spectrometry proteomics data, and may have application for other types of proteomics data, as well.

4.1 Introduction

MALDI-TOF is a mass spectrometry based proteomics method that yields spiky functional data, with peaks corresponding to proteins present in the biological sample. SELDI-TOF is a type of MALDI-TOF instrument in which the surface of the chips is specially coated to bind only certain types of proteins. An introduction to these these technologies can be found in Chapter 1 of this book. In this chapter, we apply a new Bayesian method for modeling spiky functional data (Morris and Carroll, 2006) to analyze these data. This method yields posterior samples of fixed effect functions, which can be used to identify differentially expressed peaks while adjusting for potentially nonlinear block effects that are typical in these data.

In Section 4.2, we introduce our example data set and describing important preprocessing methods and standard analysis approaches in existing literature. In Section 4.3, we describe the functional mixed model upon which our method is based. Section 4.4 contains an introduction to wavelets and describes a Bayesian, wavelet-based method for fitting the functional mixed model. Section 4.5 describes how to apply this method to MALDI-TOF data to detect peaks, identify differentially expressed peaks, and adjust for block effects, and Section 4.6 contains some conclusions.

4.2 Overview of MALDI-TOF

4.2.1 Example

At Univesity of Texas MD Anderson Cancer Center, we conducted a SELDI-TOF experiment to study proteins in the serum of mice implanted with cancer tumors. The study included 16 nude mice. For each mouse, a tumor from one of two cancer cell lines was implanted into one of two organs (brain or lung). The cell lines were A375P, a human melanoma cancer cell line with low metastatic potential, and PC3MM2, a highly metastatic human prostate cancer cell line.

After a period of time, a blood sample was taken, from which the serum was extracted and then placed on two SELDI chips. One chip was run on the SELDI-TOF instrument using a low laser intensity and the

other using a high laser intensity, yielding two spectra per mouse. The low laser intensity spectrum tends to measure the low molecular weight proteins more efficiently, while the higher laser intensity yields more precise measurements for proteins with higher molecular weights. This resulted in a total of 32 spectra, two per mouse. Since the measurements for the very low mass regions are unreliable, for this analysis we kept only the part of the spectrum between 2000 and 14,000 Daltons, a range which contains roughly 24,000 observations per spectrum.

Our primary goals were to assess whether more proteins are differentially expressed by the host organ site or by the donor cell line type, and to identify any protein peaks differentially expressed by organ site, by cell line, and/or their interaction. Typically, spectra from different laser intensities are analyzed separately. This is inefficient since spectra from both laser intensities contain information on the same proteins. Thus, we wanted to perform these analyses combining information across the two laser intensities, which required us to adjust for the systematic laser intensity effect and account for correlation between spectra obtained from the same mouse.

4.2.2 Preprocessing MALDI-TOF/SELDI-TOF Data

There are a number of preprocessing steps that must be performed before modeling MALDI-TOF or SELDI-TOF data. It has been shown that inadequate or ineffective preprocessing can make it difficult to extract meaningful biological information from the data (Sorace and Zhan, 2003; Baggerly et al., 2003, 2004). These steps include baseline correction, normalization, and denoising. The baseline, which is frequently seen in spectra, is a smooth underlying function that is thought to be largely due to a large cloud of particles striking the detector in the early part of the experiment (Malyarenko, et al. 2005). This is an artifact that must be removed. Normalization refers to a constant multiplicative factor that is used to adjust for spectrum-specific variability, for example to adjust for different amounts of protein ionized and denoised from the sample. Denoising is done to remove white noise from the spectrum that is largely due to electronic noise from the detector. In recent years, various methods have been proposed to deal with these issues. In the analysis presented in this chapter, we used the methods described in Coombes, et al. (2005b). The first two columns of Figure 4.1 contain the raw spectrum and corresponding preprocessed spectrum for low and high laser intensity scans from one mouse, and demonstrate the effects of

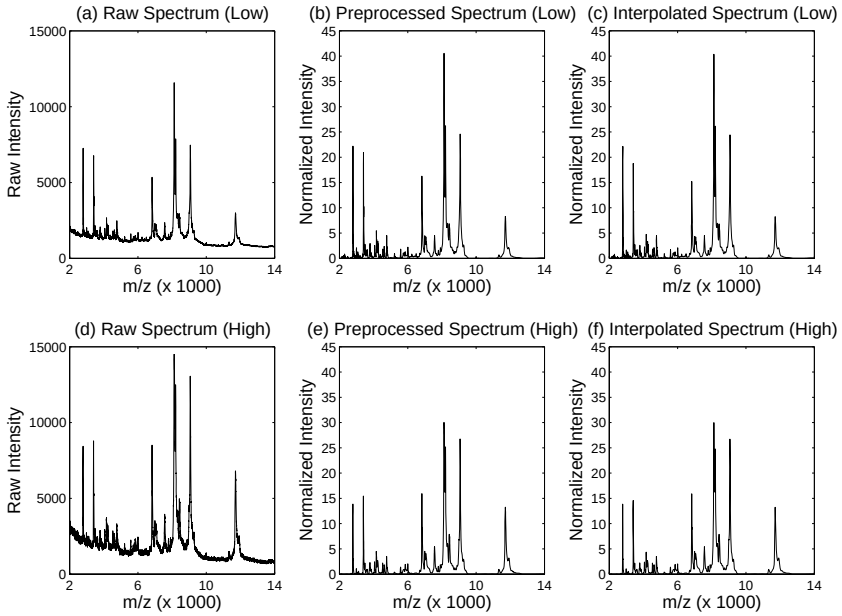

Fig. 4.1. *Sample Spectra.* Raw, preprocessed, and interpolated SELDI-TOF spectra for low and high laser intensity scans for one mouse. Note that these spectra are characterized by many peaks corresponding to proteins present in the sample.

preprocessing. A more thorough discussion of these issues can be found in Chapter 1 of this book.

In our example, we used linear interpolation to downsample the observations within each spectrum to a 2000-unit grid. The third column in Figure 4.1 contains the interpolated spectra corresponding to the preprocessed spectra in the second column. Visual inspection of the raw and interpolated spectra revealed virtually no differences. The interpolation was performed for computational convenience. Further optimization of our code for fitting the functional mixed model described in this paper will allow us to model the entire spectrum without downsampling. All analyses were performed on the interpolated, preprocessed spectra.

Some recent case studies (Baggerly et al. 2003, 2004, Sorace and Zhan 2004, Hu, et al. 2005, Coombes, et al. 2005a, Conrads and Veenstra 2005) have demonstrated that the MALDI-TOF instrument can be very sensitive to experimental conditions. Spectra can vary substantially for

samples collected at different locations or stored under different conditions. Also, spectra obtained on different days can appear different from one another. These differences can be manifest in changes both in the intensities and locations of the peaks (i.e. both the y and x axes, see Figure 4.7). They are sometimes so large in magnitude that they swamp the biological differences that are of primary interest (Coombes, et al. 2005a). Thus, it is important to take care in the experimental design phase to ensure that these factors are not counfounded with factors of interest, thus introducing systematic biases between the spectra from different treatment groups. Even when not confounded, it is still important to account for these block effects when modeling the spectra.

4.2.3 Peak Detection vs. Functional Modeling

It is common to use a two-step approach to analyze mass spectrometry data (Baggerly, et al. 2003, Yasui, et al. 2003, Coombes, et al. 2003, 2004, Morris, et al. 2005c). First, some type of feature detection algorithm is applied to identify peaks in the spectra, then a quantification for each peak is obtained for each spectrum, e.g. by taking the intensity at a local maximum or computing the area under the peak. Assuming there are p peaks and N spectra, this results in a $p \times N$ matrix of *protein expression levels* that is somewhat analogous to the matrix of mRNA expression levels obtained after preprocessing microarray data. Second, this matrix is analyzed using the similar methods to those used for microarrays identify peaks differentially expressed across experimental conditions.

This two-step approach is (i) intuitive since it focuses on the peaks, the most scientifically relevant features of the spectra, and (ii) convenient, it can borrow from a wide array of available methods developed for microarrays. However, it also has disadvantages. First, since group comparisons are only done after peak detection, this approach could miss out on important differences in low intensity peaks if the peak detection algorithm is not sensitive enough. Important information can be lost in the reduction from the full spectrum to the set of detected peaks. Second, in using this approach there may be no natural way to account for block effects that affect the spectra in both the x and y axes.

Mass spectra can also be viewed as functional data, since the observation $y_i(x)$ contains an intensity measurement from spectrum i associated with mass per unit charge value x. These functions are irregular, characterized by many peaks corresponding to proteins present in the tissue

sample. An alternative to the two-step approach described above is to model the spectra as functions, in the spirit of functional data analysis (Ramsay and Silverman 1997). This is the approach we take in this chapter. Specifically, we describe a Bayesian functional modeling approach based on the functional mixed model, a generalization of the linear mixed model equation to potentially irregular functional data. Since it involves modeling the entire spectrum, this method may detect significant group differences for very low abundance peaks that might be missed by peak detection algorithms. Further, by allowing a very flexible nonparametric representation of the fixed and random effects, this method can simultaneously model the functional effects of a number of factors, such as experimental factors of interest as well as nuisance factors related to the experimental design. These nonparametrically modelled effects can account for differences on both the x and y axes. In the subsequent sections, we introduce the functional mixed model, describe our method for fitting it, and demonstrate how to apply it to MALDI-TOF data.

4.3 Functional Mixed Models

Suppose we observe n functional profiles $Y_i(t), i = 1, \ldots, n$, all defined on the compact set $\mathcal{T} \in \Re^1$. A functional mixed model for these profiles is given by

$$Y_i(t) = \sum_{j=1}^{p} X_{ij} B_j(t) + \sum_{k=1}^{m} Z_{ik} U_k(t) + E_i(t), \tag{4.1}$$

where X_{ij} are covariates, $B_j(t)$ are functional fixed effects, Z_{ik} are elements of the design matrix for functional random effects $U_k(t)$, and $E_i(t)$ are residual error processes. Here, we assume that $U_k(t)$ are independent and identically distributed (iid) mean-zero Gaussian processes with covariance surface $Q(t_1, t_2)$, and $E_i(t)$ are iid mean-zero Gaussian processes with covariance surface $S(t_1, t_2)$, with $U_k(t)$ and $E_i(t)$ assumed to be independent. The matrix Q is the covariance across the random effect functions $k = 1, \ldots, m$, and S is the covariance across the residual error processes for the n curves, after conditioning on the fixed and random effects. This model is a special case of the one discussed in Morris and Carroll (2006), and is also equivalent to the functional mixed model discussed by Guo (2002).

Suppose all observed profiles are sampled on the same equally spaced grid $\mathbf{t}$ of length T. Let Y be the $n \times T$ matrix containing the observed

profiles on the grid, with each row containing one observed profile on the grid $\mathbf{t}$. A discrete, matrix-based version of this mixed model can be written as

$$Y = XB + ZU + E. \tag{4.2}$$

The matrix X is an $n \times p$ design matrix of covariates; B is a $p \times T$ matrix whose rows contain the corresponding *fixed effect functions* on the grid $\mathbf{t}$. B_{ij} denotes the effect of the covariate in column i of X on the response at time t_j. The matrix U is an $m \times T$ matrix whose rows contain *random effect functions* on the grid $\mathbf{t}$, and Z is the corresponding $n \times m$ design matrix. Each row of the $n \times T$ matrix E contains the residual error process for the corresponding observed profile. We assume that the rows of U are iid MVN$(\mathbf{0}, Q)$ and the rows of E are iid MVN$(\mathbf{0}, S)$, independent of U, with Q and S being $T \times T$ covariance matrices that are discrete approximations to the covariances surfaces in (4.1) on the grid.

This model is very flexible and can be used to represent a wide range of functional data. The fixed effect functions may be group mean functions, interaction functions, or functional linear effects for continuous covariates, depending on the structure of the design matrix. The random effect functions provide a convenient mechanism for modeling between-function correlation, for example when multiple profiles are obtained from the same individual. The model places no restrictions on the form of the fixed or random effect functions. Since the forms of the covariance matrices Q and S are also left unspecified, it is necessary to place some type of structure on these matrices before fitting this model.

Guo (2002) introduced frequentist methodology for fitting this model, whereby the functions were represented as smoothing splines and the matrices Q and S were assumed to follow a particular fixed covariance structure based on the reproducing kernel for the spline. By using smoothing splines, one implicitly makes certain assumptions about the smoothness of the underlying functions that are not appropriate for the irregular, spiky functions encountered in MALDI-MS. Also, the structure assumed on the Q and S matrices in that paper is not flexible enough to accommodate the complex types of curve-to-curve deviations encountered for irregular spiky functional data like MALDI-MS data. Morris and Carroll (2006) introduced a Bayesian wavelet-based method for fitting this model which uses wavelet shrinkage for regularization and allows more flexible structures for Q and S, and thus is better suited for the spiky functions encountered in MALDI-MS data.

4.4 Wavelet-Based Functional Mixed Models

We first give a brief overview of wavelets and wavelet regression, then describe the Bayesian wavelet-based approach for fitting the functional mixed model introduced in Morris and Carroll (2006), which extended the work introduced in Morris, et al. (2003). This method is described in detail in Morris and Carroll (2006), and applied to accelerometer data with an extension to partially missing functional data in Morris, et al. (2005a).

4.4.1 Wavelets and Wavelet Regression

Wavelets are families of basis functions that can be used to represent other functions, often very parsimoniously. A wavelet series approximation for an observed curve $y(t)$ is given by

$$y(t) \quad = \quad \sum_k c_{J,k}\phi_{J,k}(t) + \sum_{j=1}^{J}\sum_k d_{j,k}\psi_{j,k}(t), \tag{4.3}$$

where J is the number of scales, and k ranges from 1 to the K_j, the number of coefficients at scale j. The functions $\phi_{J,k}(t)$ and $\psi_{j,k}(t)$ are father and mother wavelet basis functions that are dilations and translations of a father and mother wavelet function, $\phi(t)$ and $\psi(t)$, respectively, with $\phi_{j,k}(t) = 2^{-j/2}\phi(2^{-j}t - k)$ and $\psi_{j,k}(t) = 2^{-j/2}\psi(2^{-j}t - k)$. These wavelet coefficients comprise a location-scale decomposition of the curve, with j indexing the scales and k indexing the locations within each scae. The coefficients $c_{J,k}, d_{J,k}, \ldots, d_{1,k}$ are the *wavelet coefficients*. The $c_{J,k}$ are called the *smooth* coefficients, and represent smooth behavior of the function at coarse scale J, and the $d_{j,k}$ are called the *detail* coefficients, representing deviations of the function at scale j, where smaller j correspond to finer scales. The wavelet coefficients at scale j essentially correspond to differences of averages of 2^{j-1} time units, spaced 2^j units apart. In addition, by examining the phase properties of the wavelet bases, we can associate each wavelet coefficient on each scale with a specific set of time points.

Theoretically, each coefficient can be computed by taking the inner product of the function and the corresponding wavelet basis function, although in practice more efficient approaches are used. If the function is sampled on an equally spaced grid of length T, then the coefficients may be computed using a pyramid-based algorithm called the discrete wavelet transform (DWT) in just O(T) operations. Applying the DWT

to a row vector of observations $\mathbf{y}$ produces a row vector of wavelet coefficients $\mathbf{d} = (c_{J,1}, \ldots, c_{J,K_J}, d_{J,1}, \ldots, d_{1,K_1})$. This transformation is a linear projection, so it may also be represented by matrix multiplication, $\mathbf{d} = \mathbf{y}W'$, with W' being the DWT projection matrix. Similarly, the inverse discrete wavelet transform (IDWT) may be used to project wavelet coefficients back into the data space, and can also be represented by matrix multiplication by the IDWT projection matrix W, the transpose of the DWT projection matrix. We use the method implemented in the Matlab Wavelet Toolbox (Misiti, et al. 2000) for computing the DWT, although other implementations could just as well have been be used.

Wavelets can be used to perform nonparametric regression using the following three-step procedure. First, noisy data $\mathbf{y}$ are projected into the wavelet domain using the DWT, yielding empirical wavelet coefficients $\mathbf{d}$. The coefficients are then thresholded by setting to zero any coefficients smaller in magnitude than a specified threshold, and/or nonlinearly shrunken towards zero using one of a number of possible frequentist or Bayesian approaches. These result in estimates of the true wavelet coefficients, which would be the wavelet coefficients for the true function if there was no noise. Finally, these estimates are projected back to the original data domain using the IDWT, yielding a denoised nonparametric estimate of the true function. Since most signals may be represented by a small number of wavelet coefficients, yet white noise is distributed equally among all wavelet coefficients, this procedure yields denoised function estimates that tend to retain dominant local features of the function. We refer to this property as *adaptive regularization*, since the function is regularized (i.e., denoised or smoothed) in a way that adapts to the characteristics of the function. This property makes the procedure useful for modeling functions with many local features like peaks. References on wavelet regression can be found in Chapters 6 and 8 of Vidakovic (1999), and in Donoho and Johnstone (1995), Chipman, Kolaczyk, and McCulloch (1997), Vidakovic (1998), Abramovich, Sapatinas, and Silverman (1998), Clyde, Parmigiani, and Vidakovic (1998), and Clyde and George (2000).

4.4.2 Wavelet-Based Modeling of Functional Mixed Model

Morris and Carroll (2006) used a similar three-step procedure to fit the functional mixed model discussed in Section 4.3. First, the DWT is used to compute the wavelet coefficients for the N observed functions,

effectively projecting these functions into the wavelet space. Second, a Markov Chain Monte Carlo is performed to obtain posterior samples of the model parameters in a wavelet-space version of the functional mixed model. Third, the IDWT is applied to the posterior samples, yielding posterior samples of the parameters in the data-space functional mixed model (4.2), which could be used to perform Bayesian inference. The wavelet space modeling allows parsimonious yet flexible modeling of the covariance matrices Q and S, leading to computationally efficient code, and providing a natural mechanism for adaptively regularizing the random and fixed effect functions.

The projection in the first step is accomplished by applying the discrete wavelet transform (DWT) to each row of Y, yielding a matrix of wavelet coefficients $D = YW'$, where W' is the DWT projection matrix. Row i of D contains the wavelet coefficients for profile i, with the columns corresponding to individual wavelet coefficients and double-indexed by scale j and location k. It is easy to show that the wavelet-space version of model (4.2) is

$$D = XB^* + ZU^* + E^*, \tag{4.4}$$

where each row of $B^* = BW'$ contains the wavelet coefficients corresponding to one of the fixed effect functions, each row of $U^* = UW'$ contains the wavelet coefficients for a random effect function, and $E^* = EW'$ contains the wavelet-space residuals. The rows of U^* and E^* remain independent mean-zero Gaussians, but with covariance matrices $Q^* = WQW'$ and $S^* = WSW'$.

Motivated by the whitening property of the wavelet transform, many wavelet regression methods in the single-function setting assume that the wavelet coefficients for a given function are mutually independent. In this context, this corresponds to making Q^* and S^* diagonal matrices. Allowing the variance components to differ across both wavelet scale j and location k yields $Q^* = \text{diag}(q_{jk})$ and $S^* = \text{diag}(s_{jk})$. This assumption reduces the dimensionality of Q and S from $T(T+1)/2$ to T, while still accommodating a reasonably wide range of nonstationary within-profile covariance structures for both the random effects and residual error processes. For example, it allows heteroscedasticity and differing degrees of smoothness at different regions of the curves, which are important characteristics of these matrices for MALDI-TOF spectra. Figure 1 of Morris and Carroll (2006) illustrates this point.

Next, a Markov Chain Monte Carlo scheme is used to generate posterior samples for quantities of model (4.4). We use vague proper pri-

ors for the variance components and independent mixture priors for the elements of B^*. Specifically, the prior for B^*_{ijk}, the wavelet coefficient at scale j and location k for fixed effect function i, was a spike-slab prior given by $B^*_{ijk} = \gamma_{ijk}\text{Normal}(0, \tau_{ij}) + (1 - \gamma_{ijk})\delta_0$, with $\gamma_{ijk} \sim \text{Bernoulli}(\pi_{ij})$ and δ_0 being a point mass at zero. This prior is commonly used in Bayesian implementations of wavelet regression, including Clyde, Parmigiani and Vidakovic (1998) and Abramovich, Sapatinas, and Silverman (1998). Use of this mixture prior causes the posterior mean estimates of the B^*_{ijk} to be nonlinearly shrunken towards zero, which results in adaptively regularized estimates of the fixed effect functions. The parameters τ_{ij} and π_{ij} are *regularization parameters* that determine the relative trade-off of variance and bias in the nonparametric estimation. They may either be prespecified or estimated from the data using an empirical Bayes method; see Morris and Carroll (2006) for details.

There are 3 major steps in the MCMC. Let Ω be the set of all covariance parameters indexing the matrices Q^* and S^*. The first step is a series of Gibbs steps to sample from the distribution of the fixed effect functions' wavelet coefficients conditional on the variance components and the data, $f(B^*|\Omega, D)$, which is a mixture of a point mass at zero and a Gaussian. See Morris and Carroll (2006) for an expression for the mixing parameters, means, and variances of these distributions. The second step is to sample from the distribution of the variance components conditional on the fixed effects and data, $f(\Omega|B^*, D)$. We accomplish this using a series of random walk Metropolis-Hastings steps, one for every combination of (j, k). We estimate each proposal variance from the data by multiplying an estimate of the variance of the MLE by 1.5, which was essential in order for our MCMC to be automated and thus computationally feasible to implement in this very high dimensional, highly parameterized setting. Note that we work with the marginalized likelihood with the random effects U^* integrated out when we update the fixed effects B^* and variance components Ω. This greatly improves the computational efficiency and convergence properties of the sampler over a simple Gibbs sampler that also conditions on the random effects. The stationary distribution for these first two steps is $f(B^*, \Omega|D)$. The third step is a series of Gibbs step to update the random effects' wavelet coefficients from their complete conditional distribution, $f(U^*|B^*, \Omega, D)$, which is a Gaussian. Note that this step is optional, and only necessary if one is specifically interested in estimating the random effect functions.

Posterior samples for each fixed effect function, $\{\mathbf{B}_i^{(g)}, g = 1, \ldots, G\}$,

on the grid $\mathbf{t}$ are then obtained by applying the IDWT to the posterior samples of the corresponding complete set of wavelet coefficients $\mathbf{B}_i^{*(g)} = [B_{i11}^{*(g)}, \ldots, B_{iJK_J}^{*(g)}]$, and similarly for the random effect functions $\mathbf{U}_i$. If desired, posterior samples for the covariance matrices Q and S may also be computed using matrix multiplication $Q^{(g)} = WQ^{*(g)}W'$ and $S^{(g)} = WS^{*(g)}W'$, respectively. Since $Q^{*(g)}$ and $S^{*(g)}$ are diagonal, this may be accomplished in an equivalent but more efficient manner by applying the 2-dimensional version of the IDWT (2d-IDWT) to $Q^{*(g)}$ and $S^{*(g)}$ (Vannucci and Corradi, 1999). These posterior samples of the quantities in model (4.2) may subsequently be used to perform any desired Bayesian inference. Code to fit the wavelet-based functional mixed model can be found on `http://biostatistics.mdanderson.org/Morris`.

4.5 Analyzing Mass Spectrometry Data Using Wavelet-Based Functional Mixed Models

In this section, we apply the Bayesian wavelet-based functional mixed model to analyze our example SELDI-TOF data set. Recall that this data set consisted of $N = 32$ spectra, $Y_i(t), i = 1, ..., 32$, one low laser intensity scan and one high laser intensity scan for each of 16 mice. Each mouse had one of two cancer cell lines (A375P or PC3MM2) injected in one of two organ sites (lung or brain). We were interested in identifying protein peaks differentially expressed between organs, between cell lines, and with significant interactions between any organ and cell line.

The functional mixed model we used to fit these spectra is given by

$$Y_i(t) = \sum_{j=0}^{4} X_{ij} B_j(t) + \sum_{k=1}^{16} Z_{ik} U_k(t) + E_i(t), \tag{4.5}$$

where $X_{i0} = 1$ and corresponds to the overall mean spectrum $\beta_0(t)$, $X_{i1} = 1$ if the mouse was injected with the A375P cell line, -1 if PC3MM2, and corresponds to the cell line main effect function $\beta_1(t)$. Also, $X_{i2} = 1$ if the injection site was the lung and -1 if the injection site was brain, and corresponds to the organ main effect function $\beta_2(t)$, while $X_{i3} = X_{i1} * X_{i2}$ and corresponds to the organ-by-cell line interaction function $\beta_3(t)$. Finally, we included a fixed effect function $\beta_4(t)$ to model the laser intensity effect, with corresponding covariate X_{i4}= 1 or -1 if the spectrum came from low or high intensity scans, respectively. We included random effects functions $U_i(t)$ for each mouse $i = 1, \ldots, 16$

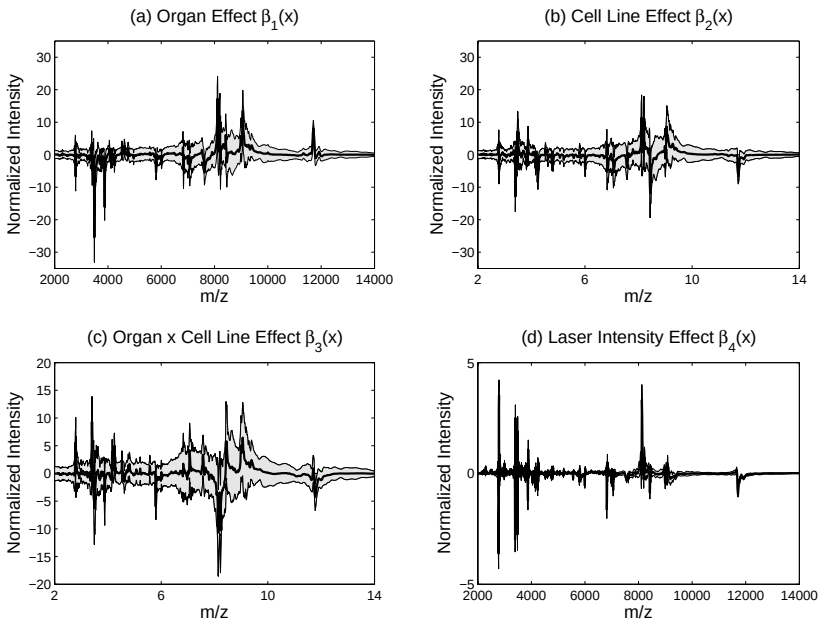

Fig. 4.2. *Fixed Effect Curves.* Posterior mean and 95% pointwise posterior credible bands for (a) organ main effect function, (b) cell line main effect function, (c) organ-by-cell line interaction function, and (d) laser intensity effect function.

to model the correlation between spectra obtained from the same mouse, so $Z_{ik} = 1$ if and only if spectrum i came from mouse k.

We modeled the spectra on the time scale t because they were equally spaced on that scale, but plotted our results on the mass per unit charge scale (m/z, x), since that scale is biologically meaningful. In our wavelet-space modeling, we chose the Daubechies wavelet with 4 vanishing moments and performed the DWT down to $J = 11$ levels. We used a modified empirical Bayes procedure to estimate the shrinkage hyperparameters π_{ij} and $\tau_{ij}, i = 1, \ldots, 5, j = 1, \ldots 10$, constraining $\tau \geq 10$ so there would be less bias in the estimation of peak heights, which we believed to be important in this context. We did almost no shrinkage ($\pi \approx 1, \tau = 1000$) for wavelet level 11 or the scaling coefficients. After a burn-in of 1000, we ran our MCMC for a total of 20,000 iterations, keeping every 10. The entire model fitting took 7 hours, 53 minutes in Matlab on a Windows 2000 Pentium IV 2.8GHz machine with 2GB

RAM. The Metropolis Hastings acceptance probabilities for the roughly 2000 sets of covariance parameters were all between 0.041 and 0.532, with median of 0.294 and 10th and 90th quantiles being 0.20 and 0.50.

Figure 4.2 contains the posterior means and 95% posterior credible bands for the the organ and cell line main effect functions, the interaction function, and the laser intensity effect function. The interpretation of the organ main effect function $\beta_1(x)$, for example, is the difference between the mean spectra for lung and brain-injected animals at m/z value x, after adjusting for the functional effects of cell-line, cell-line by organ interaction, and laser intensity. The spiky nature of these fixed effect functions indicate that differences in spectra between treatment groups are localized, and highlights the importance of using adaptive regularization methods with these data. Although difficult to see in these plots, there are a number of locations within the curves at which there is strong evidence of significant effects. These are evident in Figure 4.3, which contains the pointwise posterior probabilities of each fixed effect curve being greater than zero, $\text{Prob}(\beta_j(x) > 0|\mathbf{Y})$. These significant regions are also evident if one zooms in on certain regions of the plots, e.g. see Figures 4.5 and 4.6.

4.5.1 Peak Detection

While it is not necessary to perform peak detection when using this functional analysis approach, it still may be useful to perform a peak-level analysis since the peaks are the most biologically relevant features of the spectra, and by restricting to the peaks, we can reduce the multiplicity problems inherent to performing pointwise inference on these curves. Morris, et al. (2005c) demonstrated that for MALDI-TOF data, it was possible to obtain more sensitive and specific peak detection by performing the peak detection on the mean spectrum rather than on the individual spectra. In the present context, we can perform peak detection using the posterior mean estimate of the overall mean spectrum $\beta_0(t)$ and expect to see similar advantages. The adaptively regularization inherent to our estimation approach results in a natural denoising of this curve, reducing the number of spurious peaks detected.

In order to perform this peak detection, we first applied the first difference operator $\bigtriangledown$ to our regularized estimate of the mean spectrum $\Gamma_0(t) = \bigtriangledown\beta_0(t) = \beta_0(t+1) - \beta_0(t)$. We considered a location t to be a *peak* if its first difference and the first difference immediately preceding it were positive ($\Gamma_0(t-1) > 0$ and $\Gamma_0(t) > 0$), and the first differences for

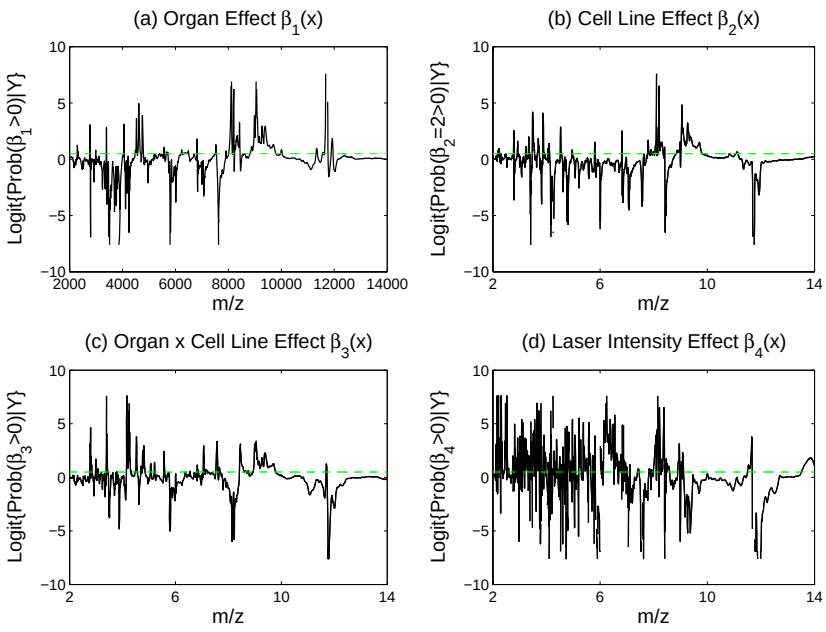

Fig. 4.3. *Posterior Probabilities.* Logit transformed pointwise posterior probabilities of being greater than zero for (a) organ main effect function, (b) cell line main effect function, (c) organ-by-cell line interaction function, and (d) laser intensity effect function.

the two locations immediately following it were negative ($\Gamma_0(t+1) < 0$ and $\Gamma_0(t+2) < 0$). This condition assured that this location was a local maximum, and the left and right slopes of the peak were monotone for at least two adjacent points.

Using this procedure, we found a total of 82 peaks out of the 2000 observations within the spectrum. Figure 4.4 contains the posterior mean overall mean curve with peaks locations indicated by the dots. Based on visual inspection, this procedure appears to have done a reasonable job of identifying the peaks.

4.5.2 Identifying Peaks of Interest

We further investigated each fixed effect function at m/z values corresponding to detected peaks. For each of the $j = 1, \ldots, 82$ peaks with locations t_j, and $i = 1, \ldots, 3$ comparisons of interest (organ main effect,

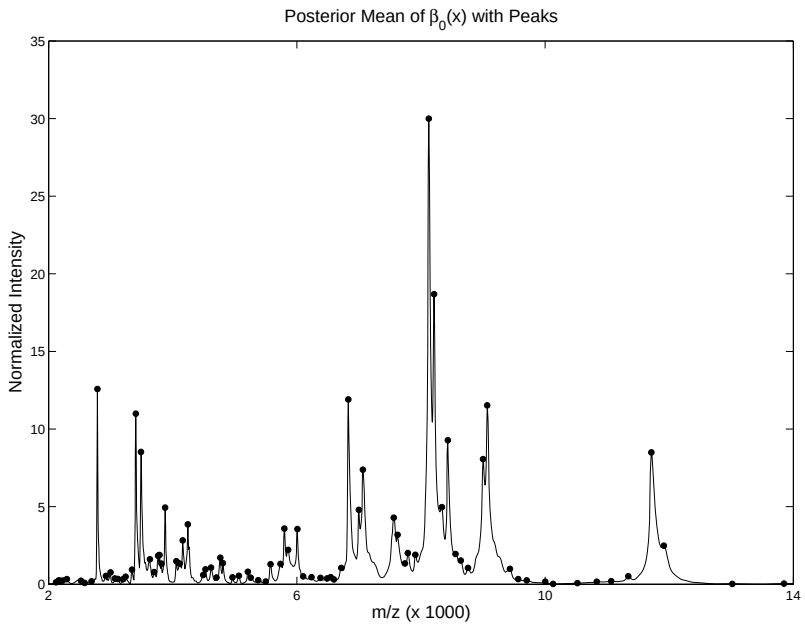

Fig. 4.4. *Peak Detection.* Posterior mean for the overall mean spectrum, $\beta_0(t)$, with detected peaks indicated by the dots. Peaks are defined to be locations for which its first difference and the first difference immediately preceding it are positive, and the first differences for the two locations immediately following it are negative.

cell line main effect, organ-by-cell line interaction), we computed the minimum posterior probability for each fixed effect function to be greater than or less than zero, i.e. $p_{ij} = \min[\Pr\{\beta_i(t_j) > 0|D\}, \Pr\{\beta_i(t_j) < 0|D\}]$. We determined a threshold ϕ below which a comparison was considered interesting and worthy of further investigation. To obtain ϕ, we first specified a small positive number α and sorted the p_{ij} from smallest to largest, $p_{(1)}, \ldots, p_{(246)}$. We defined ϕ to be $p_{(\delta)}$, where δ was the largest integer for which $\sum_{k=1}^{\delta}\{2p_{(k)} + (2G)^{-1}\} < \alpha$. Recall G is the number of MCMC samples used to compute the posterior probabilities. The factor of two is included to adjust for the two-sided nature of the analysis, and the factor involving G adjusts for the limitation in precision for estimating p_{ij} that is due to the number of MCMC samples run.

We applied this procedure to our data using $\alpha = 0.01$, and found that

Table 4.1. *Flagged peaks from proteomics example. Location of peak (in Daltons per coulomb) is given, along with which effect was deemed significant, the associated posterior probability p, and a description of the effect.*

Peak location	Effect type	p	Comment
3412.6	interaction	<0.0005	PC3MM2>A375P for brain-injected only
3496.6	organ	<0.0005	Only expressed in brain-injected mice
3886.3	organ	<0.0005	Only expressed in brain-injected mice
4168.2	interaction	0.0005	PC3MM2>A375P in brain-injected only
4252.1	interaction	<0.0005	PC3MM2>A375P in brain-injected only
4814.2	cell line	0.0030	PC3MM2>A375P
5805.3	interaction	<0.0005	brain>lung only for mice with A375P cell line
6015.2	cell line	<0.0005	PC3MM2>A375P
7628.1	organ	0.0015	Only expressed in brain-injected mice
8438.1	cell line	0.0015	PC3MM2>A375P
9074	organ	0.0020	lung>brain
11721.0	organ	<0.0005	lung>brain
11721.0	cell line	<0.0005	PC3MM2>A375P

$\phi = 0.0033$, and $\delta = 18$ of the p_{ij} were flagged as interesting. These 18 p_{ij} were from a total of 12 peaks. Table 4.1 lists the m/z values for these peaks, along with p_{ij} and a description of the interesting effect. Whenever an interaction effect was found to be interesting, the main effects for that peak were not considered. Out of these 12 peaks, we found 4 with organ main effects, 3 with cell-line main effects, 1 with both organ and cell-line main effects, and 4 associated with the organ-by-cell line interaction effects.

We attempted to find information about the possible identity of the flagged peaks by running the estimated m/z values of the corresponding peaks through TagIdent, a searchable database (available at `http://us.expasy.org/tools/tagident.html`) that contains the molecular masses and pH for proteins observed in various species. We searched for proteins eminating from both the source (human) and the host (mouse) whose molecular mass was within the estimated mass accuracy (0.3%) of the SELDI instrument. This only gives an educated guess at what the protein identity of the peak could be; it would be necessary to perform an additional MS/MS experiment in order to definitively identify the peak.

For illustration, we plotted posterior means and posterior pointwise

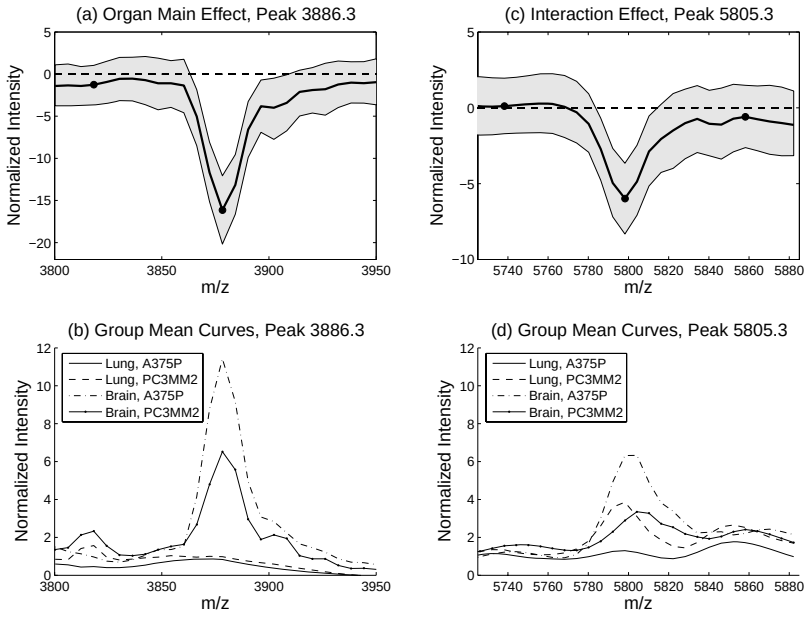

Fig. 4.5. *Peaks of Interest.* (a) Posterior mean for organ main effect function and 95% pointwise posterior credible bands for peak near 3886.3. (b) Posterior mean for mean functions for each organ × cell line group near peak at 3886.3. (c) Posterior mean organ-by-cell line interaction effect function and 95% pointwise posterior credible bands near peak at 5805.3. (e) Posterior mean for mean functions for each organ × cell line group near peak at 5805.3.

credible intervals for the interesting effect functions in the neighborhood of the peaks at 3886.3, 5805.3, 7628.1, and 11721.0 (Figures 4.5 and 4.6). In Figure 4.5(a), we see that the peak at 3886.3 is expressed more highly in brain-injected mice than in lung-injected mice. In fact, our observation of the posterior mean curves for the 4 group organ-by-cell line group mean curves $(\beta_0(x) + \beta_1(x) + \beta_2(x) + \beta_3(x), \beta_0(x) + \beta_1(x) - \beta_2(x) - \beta_3(x), \beta_0(x) - \beta_1(x) + \beta_2(x) - \beta_3(x)$, and $\beta_0(x) - \beta_1(x) - \beta_2(x) + \beta_3(x))$, indicates that this peak does not even appear to be present in the serum proteomic profile of lung-injected mice, only brain-injected mice. Using TagIdent, this peak closely matched calcitonin gene-related peptide II precursor (CGRP-II, 3882.34 Daltons, 5.41 pH). This peptide is in the mouse proteome, dilates blood vessels in the brain, and has been

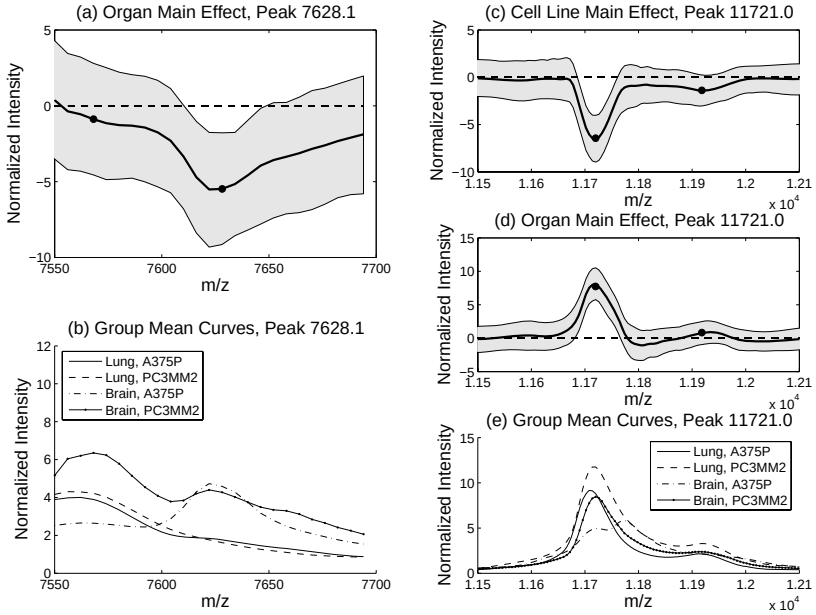

Fig. 4.6. *Peaks of Interest.* (a) Posterior mean for organ main effect function and 95% pointwise posterior credible bands for peak near 7628.1. (b) Posterior mean for mean functions for each organ × cell line group near peak at 7628.1. (c) and (d) Posterior mean cell line and organ main effect functions, respectively, and 95% pointwise posterior credible bands near peak at 11721.0. (e) Posterior mean for mean functions for each organ × cell line group near peak at 11721.0.

observed to be abundant in the central nervous system. This result may represent an important host response to the implanted tumor.

Figure 4.5(c) and (d) contain the posterior mean interaction main effect curve $\beta_3(x)$ and group mean functions, respectively, in the neighborhood of the peak at 5805.3 Daltons. This protein is higher in brain-injected mice than lung-injected mice only for those mice given cell line A375P. There is a protein in the human proteome KiSS-16 with molecular weight of 5794.7 Daltons that is known to be highly expressed in metastatis-suppressed chromosome 6 melanoma hybrids.

Figure 4.6(a) and (b) contain the posterior mean organ main effect curve $\beta_1(x)$ and group mean functions, respectively, in the neighborhood of the peak at 7628.1 Daltons. This protein is only present in spectra

from brain-injected mice. There is a protein Neurogranin in the human proteome with a molecular weight of 7618.47 Daltons that is active in synaptic development and remodeling in the brain.

Figures 4.6(c) and 4.6(d) contain the posterior mean and 95% posterior pointwise credible intervals for the organ and cell line main effects curves $\beta_1(x)$ and $\beta_2(x)$ in a neighborhood around the peak at 11721.0 Daltons, and Figure 4.6(e) contains the corresponding group mean curves. This protein has higher expression in the metastatic cell line PC3MM2 than the non-metastatic cell line A375P, and in lung-injected mice than brain-injected mice. There is a protein MTS1 in the mouse proteome with molecular mass 11721.4 Daltons that is known to be specifically expressed in different metastatic cells (Tulchinsky, et al. 1990). A similar protein with molecular mass of 11728.5 Daltons is present in the human proteome.

4.5.3 Nonparametric Modeling of Block Effects

The analysis described above identified a number of interesting peaks. Our ability to detect these differences was aided by the fact that we were able to combine information from both the low and high laser intensity spectra to perform our analysis, giving us greater power to detect differences. Recall that it is typical to analyze spectra with different laser intensities separately because there are systematic differences between the spectra, but this is inefficient since spectra from both laser intensities contain information about the same protein peaks.

Our inclusion of a nonparametric functional laser intensity effect $\beta_4(x)$ in our modeling allowed us to combine these in a common model. The interpretation of this effect is the difference between the mean spectrum from the two laser intensities, after adjusting for the other functional effects in the model. The flexibility of the nonparametric modeling allows this factor to adjust for systematic differences in both the x and y axes. Figure 4.7 illustrates this point. Figure 4.7(a) contains the posterior mean laser effect function $\beta_4(x)$ and 95% posterior bounds in the region of two peaks at 3412.6 and 3496.6, while Figure 4.7(b) contains the posterior mean for the overall mean spectrum $\beta_0(x)$ in the same region. The pulse-like characteristics in the laser effect curve near the peak location demonstrates that the inclusion this effect in the model adjusts for a slight misalignment in the peaks across the different laser intensity blocks, i.e. differences in the x axis. The mean curves for low laser intensity $\beta_0(x) + \beta_4(x)$ and high laser intensity $\beta_0(x) - \beta_4(x)$

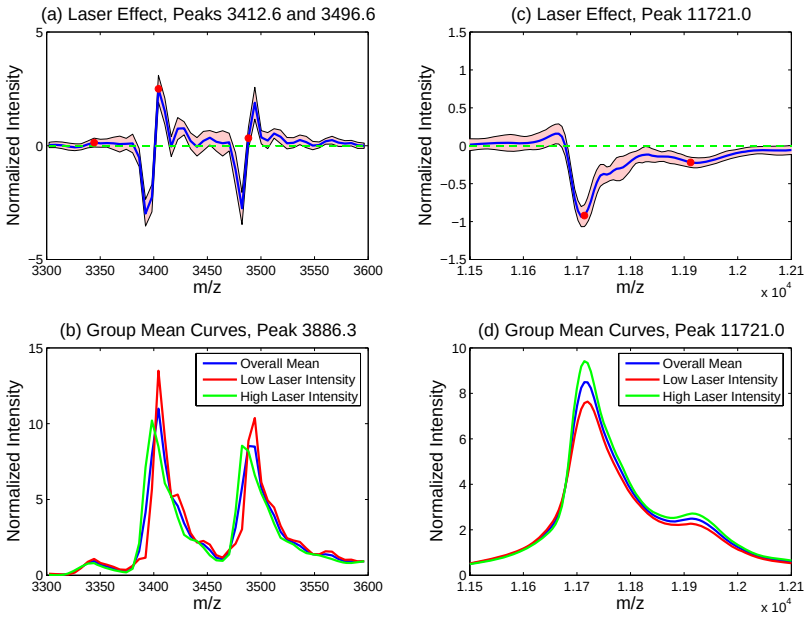

Fig. 4.7. *Laser Intensity Effect*. Posterior mean laser intensity effect near peaks of interest at (a) 3412.6 and 3496.6 and (c) 11721, along with 95% pointwise posterior credible bands. The red dots indicate the locations of peaks detected in the fitted mean spectrum. Panels (b) and (d) contain the corresponding fitted posterior mean curves for the overall mean and the laser intensity-specific mean spectra in the same two regions. Note that the nonparametrically estimated laser intensity effects are able to adjust for both shifts in location (x axis, see (a) and (b)), and shifts in intensity (y axis, see (c) and (d)).

(Figure 4.7(b)) demonstrate that the peak at 3412.6 in the overall mean curve is at a slightly higher m/z value for the low intensity spectra and slightly lower m/z value for the medium intensity spectra. Figures 4.7(c) and 4.7(d) contain $\beta_4(x)$ and $\beta_0(x)$ in the neighborhood of the peak at 11721.0, and demonstrate that the nonparametric laser effect can also adjust for an additive offset in the mean peak intensity across blocks, i.e. the y axis.

This same strategy can be used in other MALDI-TOF data sets to adjust for systematic batch effects when spectra are run in batches, laboratory effects when they are run at different laboratories, and sample effects when samples are obtained from different locations, as long as

these factors are not completely confounded with one of the other covariates of interest. This is important, since the MALDI-TOF instrument is known to be quite sensitive to these types of factors, and it is necessary to deal successfully with these factors in order to obtain reproducible results.

4.6 Conclusion

In this chapter, we have demonstrated how to use the newly developed Bayesian wavelet-based functional mixed model to model MALDI-TOF proteomics data. This method appears well suited to this context, for several reasons. The functional mixed model is very flexible, able to simultaneously model nonparametric functional effects of many covariates simultaneousely, both factors of interest and nuisance factors such as block effects, plus the random effect functions can be used to model correlation structure among spectra that might be induced by the experimental design. The wavelet-based modeling approach works well for modeling functional data with many local features like MALDI-TOF peaks since it results in adaptive regularization of the fixed effect functions, avoiding attentuation of the effects at the peaks, and is reasonably flexible in modeling the between-curve covariance structure, accomodating autocovariance structures induced by peaks and heteroscedasticity to allow different between-spectrum variances for different peaks. Given the posterior samples produced by this method, we were able to perform peak detection and flag a number of peaks as interesting and worthy of future investigation. The efficiency of our analysis was increased because our use of random effect functions and a nonparametric laser intensity effect function allowed us to combine information across spectra obtained from different laser intensities. This strategy has more general application for calibrating spectra so data can be combined across different laboratories or batches. Our approach may also be useful for analyzing data from other proteomic platforms that generate functional data, and may be extended to model functional data on two-dimensional domains, including data from 2d gel electrophoresis and liquid chromatography mass spectrometry.

Acknowledgements

This work was partially supported by a grant from the National Cancer Institute (CA-107304).

References

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998) Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society, Series B*, 60, 725–749.

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

Baggerly, K. A., Morris J. S., and Coombes K. R. (2004). Reproducibility of SELDI Mass Spectrometry Patterns in Serum: Comparing Proteomic Data Sets from Different Experiments. *Bioinformatics*, 20(5): 777-785

Chipman, H. A., Kolaczyk, E. D. and McCulloch, R. E. (1997) Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association*, 92, 1413–1421.

Clyde, M. and George, E. I. (2000) Flexible empirical Bayes estimation for wavelets. *Journal of the Royal Statistical Society, Series B*, 60, 681–698.

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

Conrads TP and Veenstra TD (2005). What Have We Learned from Proteomic Studies of Serum? *Expert Review of Proteomics*, 2(3): 279-281.

Coombes KR, Fritsche HA Jr., Clarke C, Cheng JN, Baggerly KA, Morris JS, Xiao LC, Hung MC, and Kuerer HM (2003). Quality Control and Peak Finding for Proteomics Data Collected from Nipple Aspirate Fluid Using Surface Enhanced Laser Desorption and Ionization. *Clinical Chemistry*. 49(10): 1615-1623.

Coombes KR, Morris JS, Hu J, Edmondson SR, and Baggerly KA (2005a). Serum Proteomics Profiling: A Young Technology Begins to Mature. *Nature Biotechnology*, 23(3): 291-292.

Coombes KR, Tsavachidis S, Morris JS, Baggerly KA, and Kobayashi R (2005b). Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorp-

tion and Ionization by Denoising Spectra using the Undecimated Discrete Wavelet Transform. *Proteomics*, In Press.

Guo, W. (2002) Functional mixed effects models. *Biometrics*, 58, 121–128.

Hu J, Coombes KR, Morris JS, and Baggerly KA (2005). The Importance of Experimental Design in Proteomic Mass Spectrometry Experiments: Some Cautionary Tales. *Briefings in Genomics and Proteomics*, 3(4), 322-331.

Malyarenko DI, Cooke WE, Adam BL, Gunjan M, Chen H, Tracy ER, Trosset MW, Sasinowski M, Semmes OJ, and Manos DM (2004) Enhancement of sensitivity and resolution of SELDI TOF-MS records for serum peptides using time series analysis techniques. *Clinical Chemistry*, 51(1): 65-74.

Misiti M., Misiti Y., Oppenheim G., and Poggi J. M. (2000), *Wavelet Toolbox For Use with Matlab: User's Guide*. Natick, MA: Mathworks, Inc.

Morris JS, Arroyo C, Coull B, Ryan LM, Herrick R, and Gortmaker SL (2005a). Using Wavelet-Based Functional Mixed Models to characterize Population Heterogeneity in Accelerometer Profiles: A Case Study. *Journal of the American Statistical Association*, Under revision.

Morris JS and Carroll RJ (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*. In Press.

Morris JS, Coombes KR, Kooman J, Baggerly KA, and Kobayashi R (2005c). Feature Extraction and Quantification for Mass Spectrometry Data in Biomedical Applications Using the Mean Spectrum. *Bioinformatics*, 21(9): 1764-1775.

Morris JS, Vannucci M, Brown PJ, and Carroll RJ (2003). Wavelet-Based Nonparametric Modeling of Hierarchical Functions in Colon Carcinogenesis. *Journal of the American Statistical Association*, 98: 573-583.

Ramsay, J. O. and Silverman, B. W. (1997) *Functional Data Analysis*. Springer, New York.

Sorace JM and Zhan M (2003). A data review and re-assessment of

ovarian cancer serum proteomic profiling. *BMC Bioinformatics*, 2003 Jun 9, 4-24.

Tulchinsky, E. M., Grigorian, M. S., Ebralidze, A. K., Milshina, N. I. and Lukanidin, E. M. (1990) Structure of gene MTS1, transcribed in metastatic mouse tumor cells. *Gene*, 87(2): 219–223.

Vannucci, M. and Corradi, F. (1999) Covariance structure of wavelet coefficients: Theory and models in a Bayesian perspective. *Journal of the Royal Statistical Society, Series B*, 61, 971–986.

Vestal, M. and Juhasz, P. (1998) Resolution and Mass Accuracy in Matrix-Assisted Laser Desorption Ionization-Time-of-Flight *Journal fo the American Society of Mass Spectrometry*. 9, 892-909.

Vidakovic, B. (1998) Nonlinear wavelet shrinkage with Bayes rules and Bayes factors. *Journal of the American Statistical Association*, 93, 173–179.

Vidakovic, B. (1999) *Statistical Modeling by Wavelets*. Wiley, New York.

Yasui T., Pepe M., Thompson M. L., Adam B. L., Wright G. L. Jr., Qu Y., Potter J. D., Winget M., Thornquist M. and Feng Z. (2003) A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics*, 4(3): 449–463.