Contents

0-8493-0052-5/00/$0.00+$.50

Chapter 1

A Bayesian Mixture Model for Protein Biomarker Discovery

Peter Muller[1], **Keith Baggerly**[1], **Kim Anh Do**[1] **and Raj Bandyopadhyay**[2]

[1] *U.T. M.D. Anderson Cancer Center and* [2] *Rice University*

Abstract

Early detection is critical in disease control and prevention. Molecular biomarkers provide valuable information about the status of a cell at any given time point. Biomarker research has benefited from recent advances in technologies such as gene expression microarrays, and more recently, proteomics. Motivated by specific problems involving proteomic profiles generated using Matrix-Assisted Laser Desorption and Ionization (MALDI-TOF) mass spectrometry, we propose model-based inference with mixtures of beta distributions for real-time discrimination in the context of protein biomarker discovery. Most biomarker discovery projects aim at identifying features in the biological proteomic profiles that distinguish cancers from normals, between different stages of disease development, or between experimental conditions (such as different treatment arms). The key to our approach is the use of a fully model-based approach, with coherent joint inference across most steps of the analysis. The end product of the proposed approach is a probability model over a list of protein masses corresponding to peaks in the observed spectra, and a probability model on indicators of differential expression for these proteins. The probability model provides a single coherent summary of the uncertainties in multiple steps of the data analysis, including baseline subtraction, smoothing and peak identification. Some ad-hoc choices remain, including some pre-processing and the solution of the label switching problem when summarizing the simulation output.

KEY WORDS: Density Estimation; Mass spectrometry; Mixture Models;

0-8493-0052-5/00/$0.00+$.50

Nonparametric Bayes; Proteomics; Spectra.

1.1 Introduction

We propose a model-based approach to analyze data from Matrix-Assisted Laser Desorption and Ionization – Time of Flight (MALDI-TOF) experiments. We construct a mixture of Beta model to represent protein peaks in the spectra. An important feature of the proposed model is a hierarchical prior with indicators for differential expression for each protein peak. The posterior distribution on the number of peaks, the locations of the peaks and the indicators for differential expression summarizes (almost) all relevant uncertainty related to the experiment. This is made possible by using one coherent model to implement joint inference for multiple steps in the inference, including baseline subtraction, noise removal, peak detection and comparison across biologic conditions.

Molecular biologists and geneticists have been guided by the central dogma that DNA produces RNA, which makes proteins, the actual agents that perform the cellular biologic functions (Alberts et al., 1994). Researchers are interested in seeking both genetic and protein biomarkers of disease. Advances in genomics have given scientists the ability to assess the simultaneous expression of thousands of genes commonly collected from cDNA microarrays and oligonucleotide gene chips. More recent advances in mass spectrometry have generated new data analytic challenges in proteomics, similar to those created by gene expression array technologies. Proteomics is valuable in the discovery of biomarkers because the proteome reflects both the intrinsic genetic program of the cell and the impact of its immediate environment (Srinivas et al., 2001). Specifically, in the context of medical and cancer research, the clinician's ultimate aim of finding new targets for diagnosis and therapeutic strategies can benefit from a better understanding of the molecular circuitry. Valuable information can be obtained by investigating protein profiles over a wide range of molecular weights in small biological specimens collected from different pathological groups, such as different disease stages, different treatment arms, or normal versus cancer subjects. See, for example, Baggerly et al. (2006) for a review of the experimental setup. In this paper, motivated by specific cancer research challenges involving Matrix-Assisted Laser Desorption and Ionization (MALDI) proteomic spectra, we suggest techniques based on density estimation for purposes of discrimination, classification, and prediction.

1.1.1 Background

Proteomics was originally defined to represent the analysis of the entire protein component of a cell or tissue. Proteomics now encompasses the study of expressed proteins; specifically, this refers to a group of technologies that attempt to separate, identify and characterize a global set of proteins, as discussed in Arthur (2003). Proteomic methods may be used to simultaneously quantify the expression levels of many proteins, thus providing information on biological cell activity such as the structure-function relationship under healthy conditions and disease conditions, for example cancer. Proteomic technologies can be employed to identify markers for cancer diagnosis, to monitor disease progression, and to identify therapeutic targets. These methods may be applied to easily obtained samples from the body (blood serum, saliva, urine, breast nipple aspirate fluid) in order to measure the distribution of proteins in that sample. Among the most important proteomic methods are 2D gel electrophoresis, and mass spectrometry. Mass spectrometry (MS) measures the relative amounts of individual molecules in a sample, converted to ions. The quantity that is actually measured is the mass-to-charge, or the m/z ratio. Masses of atoms and molecules are usually measured in a unit called the *Dalton*, which is defined as 1/12 the mass of ^{12}C. The unit of charge z is that present on an electron or a proton. The m/z ratio therefore refers to the number of Daltons per unit charge. In the case of singly charged ions, such as (most of) those generated in MALDI, this ratio is numerically equal to the ionic mass of the ions (in Daltons), plus one, due to the added mass of the unbalanced proton. In general, MS works by converting molecules to ions and measuring the proportions of the resulting ions in the mixture, after sorting them by the m/z ratio. This results in a histogram of m/z values, usually termed as a *mass spectrum*. Two common MS technologies are the Surface Enhanced Laser Desorption and Ionization (SELDI) technique and the MALDI method. MS, especially using the MALDI technique, has been very successfully applied to determine the molecular weight of both peptides and proteins, as well as to obtain structural information. This procedure has the advantages of rapid setup, high sensitivity, and tolerance for heterogeneous samples. Details of the experimental setup are described, for example, in Baggerly et al. (2003) or Baggerly et al. (2006). Briefly, the biological sample for which we wish to determine protein abundance is fixed in a matrix. A laser beam is used to break free and ionize individual protein molecules. The experiment is arranged such that ionized proteins are exposed to an electric field that accelerates molecules along a flight tube. On the other end of the flight tube molecules hit a detector that records a histogram of the number of molecules that hit over time. Assumig that all ionized molecules carry a unit charge, the time of flight is deterministically related to the molecule mass. The histogram of detector events over time can therefore be be changed to a histogram of detector events over protein masses. Allowing for multiple charges, the mass scale is replaced by a scale of mass/charge ratios. The

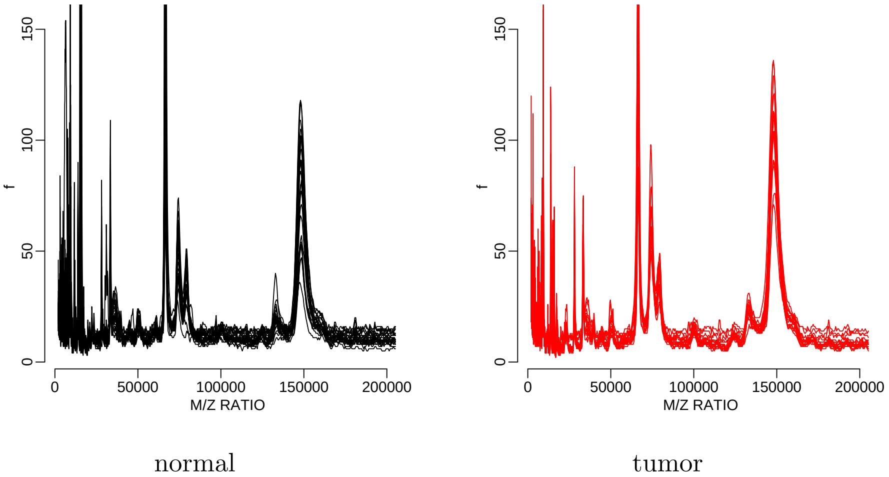


FIGURE 1.1: Raw data. The left panel shows the recorded spectra for normal samples. The right panel shows the spectra for tumor samples. The X axis are the m/z values.

histogram of detector events is known as mass/charge spectrum. More technical details are available, for example, in Siuzdak (2003) or de Hoffman and Stroobant (2002). Our motivating data set was generated using the MALDI technique.

Figure 1.1 shows typical mass spectra, plotted against mass/charge ratios. Ideally, one would expect a mass spectrum to be mostly flat, with spikes at the masses of proteins in the mixture. However, that is not the case. Firstly, each charged particle, on being energized by the laser, has a random initial velocity, before being accelerated by the electromagnetic field. These phenomena lead to the existence of a fairly smooth peak around the mass of the protein concerned. For lower m/z molecules, the peak is very narrow, and it gets broader with increasing m/z. Moreover, the ions generated may have slightly different masses, due to addition or removal of charged particles. In order to improve the resolution of the mass spectrometer, the sample is usually fractionated. This is achieved, for example, by running the sample through a gel, which separates out different molecules according to their pH values. The gel is then divided into pieces or fractions. Ideally, each fraction should contain a definite and distinct subset of proteins in the sample. The portion of the sample from each fraction is isolated and analyzed by the mass spectrometer. In practice, however, we find that the fractions do not separate as cleanly as expected. In some cases, the same protein may be found in multiple fractions, and the fractions may not be consistent across samples. This is the case, for example, for the motivating data set described later. Various other sources of noise exist in mass spectrometry. In MALDI, noise may arise from ionized molecules from the underlying matrix. This usually occurs at the lower end of the m/z spectrum. Also, electrical noise may interfere, particularly if the sensitivity is low or extremely small proportions of samples are used. This noise may give rise to spurious peaks. At very low m/z values, large numbers of particles can saturate the detector and introduce additional artifacts.

1.1.2 Statistical methods for classification of mass spectrometry data

Recent literature in the cancer classification arena that used MS generated data focused mainly on identifying biomarkers in serum to discriminate cancer from normal samples. Often the process involves a split of the data into a training set and a validation set. The training set is used to identify a subset of interesting biomarkers; the validation set is used to assess the selected biomarkers either individually or simultaneously by their ability to classify samples accurately in the separate test set. A number of statistical methods have been discussed for biomarker selection, including $t-$statistics by Chen et al. (2002), tree-based classification methods by Adam et al. (2002), genetic algorithms and self-organizing maps by Petricoin et al. (2002) and Baggerly et al. (2003), and artificial neural networks by Ball et al. (2002). The common classification methods used include the classical approaches such as linear discriminant analysis, quadratic discriminant analysis and k-nearest neighbors. More recent publications have discussed the use of bagging and boosting to the construction of a classifier, see Yasui et al. (2003) and Wu et al. (2003). Moreover, Wu et al. (2003) also compared the performance of random forest to other methods of classification and variable selection.

1.2 The Data

We consider the data set used as part of the First Annual Conference on Proteomics and Data Mining at Duke University. The ultimate goal is to identify protein biomarkers that distinguish between lung cancer and normal subjects. Assuming that the up- (or down-) regulation of certain proteins is the consequence of a transformed cancerous cell and its clonal expansion, an early detection research project may focus on the identification of such early molecular signs of lung cancer via the assessment of protein profiles from specific biological specimens. Researchers can thus analyze the collected protein profiles and identify signature *fingerprints* for the classification between lung cancer and normal samples. Based on the identified signature profiles, researchers can ultimately study the biological significance of those specific proteins or peptides. Such advances can potentially lead to clinical detection tools. Different research groups have attempted to develop techniques to classify or cluster this data set. Our analysis initially employs the preprocessing steps described in Baggerly et al. (2003). This motivating data set consists of MALDI-MS spectra of serum for 24 individuals with lung cancer and 17 normal individuals (without cancer). For each subject (sample), the raw data contained recordings of 20 fractions. Each such spectrum had readings for 60831 m/z values.

Traditional inference for mass/charge spectra proceeds in a step-wise fashion, often ignoring uncertainties involved in each step of the process. For example, Baggerly et al. (2003) analyzed this data set using the following steps.

First, a baseline, computed using a windowed local minimum technique, is subtracted from the data. This baseline correction has to be performed separately for each spectrum in each sample. It is a crucial step in the preprocessing, as we cannot combine the spectra otherwise in a meaningful manner. Next, using a Fourier transform, periodic noise most likely associated with electrical activity is removed. Next, the spectra are scaled by dividing by the total current over all the readings. Next, exploratory data analysis showed that several peaks were scattered across fractions, appearing in different fractions for different samples. Therefore, the normalized fractions were combined to generate one spectrum per sample. In a final pre-processing step, the dimensionality of the data was drastically reduced by carrying out a windowed peak identification. Taking the maximum intensity in each window (of 200 readings) and taking windows in which at least 8 of the samples contained a peak, using an ad-hoc definition of peaks, the dimensionality was reduced from 60381 to 506. The combined result of these pre-processing steps is a 506×41 peak matrix.

Baggerly et al. (2003) combined a genetic algorithm and an exhaustive search to extract a small subset of peaks which were good discriminators between cancers and normals. Perfect classification was achieved with a set of 5 peaks with intensities at the following M/Z values: 3077, 12886, 74263, 2476, and 6966 Dalton.

1.3 Likelihood based inference for proteomic spectra

A common feature of currently used methods is the use of some form of smoothing to separate the observed mass/charge spectrum into noise and signal, by directly smoothing the raw spectrum, by considering principal components, or by using reasonable exploratory data analysis tools like the windowing of the raw data described before. Also, most methods involve multiple steps, including separate steps related to noise removal, baseline subtraction, peak identification, and finally identification of differentially expressed peaks. Such methods are highly appropriate when the focus of inference is the search for peaks corresponding to specific proteins and the identification of peaks that are correlated with the biologic condition of interest. A critical limitation, however, is the lack of a joint probability model that characterizes the combined uncertainty across all steps, and can be the basis for probability statements related to the desired inference. As an alternative we propose a likelihood-based approach that allow us to implement joint inference across all steps. The estimated peaks differ little from what is obtained with other methods. The main difference is in the full probabilistic description of all relevant uncertainties. Instead of a point estimate we report a posterior distribution on the unknown true spectrum of mass/charge ratios. All uncertainties related to denoising and baseline substraction are appropriately propagated and accounted for. The proposed approach proceeds as follows. We treat the spectrum as a histogram of observed detector events. We assume that the original time-of-flight scale has been transformed to counts on a mass/charge scale. Let $p_k(m)$ denote the frequency of mass/charge ratios m in the k-th sample. We treat the recorded data as an i.i.d. sample of draws from p_k. This naturally turns the inference problem into a density estimation problem of inference on the unknown distribution p_k, conditional on a random sample summarized by the observed spectrum. The problem differs from traditional density estimation problems due to the nature of the prior information. The spectrum p_k is known to be multimodal with sharp peaks corresponding to different proteins, plus a smooth baseline. We are thus lead to consider models for random distributions p_k on a compact interval, allowing for relatively narrow peaks corresponding to different proteins. Without loss of generality we assume that the range of mass/charge ratios is rescaled to the interval $[0, 1]$. A convenient model for random distributions on the unit interval that allows for the desired features is a mixture of Beta distributions. Restricting the Beta distributions in the mixture to integer parameters over a certain range leads to Bernstein priors discussed in Petrone (1999a) We follow Robert and Rousseau (2003) who argue for the use of Beta mixtures with unconstrained parameters to achieve more parsimonious models. We introduce an additional level of mixture to deconvolve the random distribution into one part, f_k, corresponding to the peaks arising from specific proteins and one part, B_k, corresponding to a non-zero baseline arising from background noise in the detector, the matrix used to fix the probe on the sample plate and other unspecified sources unrelated to the biologic effects of interest. A hierarchical prior distribution completes the model. In words, the hierarchical prior probability model is described as follows. We start with a distribution for a random numbers J of peaks, continue with a prior for the location and scale of the J peaks, and weights for each peak in each of the biologic samples. Samples collected under different biologic conditions, for example, tumor and normal, might require different weights, corresponding to different abundance of the respective protein in the samples. In addition to the peaks related to specific proteins in the probes the mixture also includes terms to represent a smooth baseline. Inference about the baseline is usually not of interest in itself. It is a nuisance parameter. We use J_k to denote the number of Beta terms that constitute the baseline, allowing for a different size mixture for each sample k. Details of the prior are described below. In the context of this hierarchical Beta mixture the desired inference about relative abundance of proteins in the probes reduces to inference about the weights in the Beta mixtures. We implement inference with a reasonably straightforward Markov chain Monte Carlo (MCMC) posterior simulation. The random number J and J_k of terms in the mixtures complicates inference by introducing a variable dimension parameter space. We use a reversible jump MCMC implementation to achieve the desired posterior simulation.

1.4 A hierarchical beta mixture model

We assume that the mass/charge spectrum is recorded on a grid m_i, $i = 1, \ldots, I$. For convenience we rescale to $m_i \in [0, 1]$. Let $y_k(m_i)$, $k = 1, \ldots, K$ denote the observed count in sample k for mass/charge ratio m_i. Due to the nonlinear nature of the transformation from time-of-flight to mass/charge, the grid on the mass/charge scale is not equally spaced. Sample k is observed under biologic condition x_k. For example, $x_k \in \{0, 1\}$ for a comparison of normal versus tumor tissue samples. We write $y_k = (y_k(m_1), \ldots, y_k(m_I))$ for the data from the k-th sample, $y = (y_1, \ldots, y_K)$ for the entire data set, and θ to generically indicate all unknown parameters in the model. We use an unconventional parametrization of Beta distributions, letting $\mathrm{Be}(m, s)$ denote a Beta distribution with mean m and standard deviation s (with an appropriate constraint on the variance). This notation simplifies the description of reversible jump moves and other technical details below. Finally, we generically use $p(a)$ and $p(a \mid b)$ to denote the distribution of a random variable a and the conditional distribution of a given b. We use notation like $\mathrm{Be}(x;\ m, s)$ to denote a Beta distribution with parameters (m, s) for the random variable x. We assume the following sampling model:

$$p(y_k \mid \theta) = \prod_{i=1}^{I} p_k(m_i)^{y_k(m_i)}, \quad p_k(m) = p_{0k}\, B_k(m) + (1 - p_{0k})\, f_k(m). \tag{1.1}$$

In words, we assume i.i.d. sampling from an unknown distribution p_k. The distribution p_k is assumed to arise as a convolution of a baseline $B_k(m)$ and a spectrum $f_k(m)$. We refer to the data y_k as the *empirical spectrum*, f_k as the unknown true *spectrum*, B_k as the baseline, and p_k as baseline plus spectrum, or the unknown distribution of mass/charge ratios. Both baseline and spectrum, are represented as mixtures of beta distributions. The means of the Beta kernels in the mixture for f_k are the mass/charge ratios of the detected proteins. Specifically we define

$$f_k(m) = \sum_{j=1}^{J} w_{xj} \text{ Be}(m;\ \epsilon_j, \alpha_j), \tag{1.2}$$

where $x = x_k$ denotes the biologic condition of sample k. We assume that $\theta_0 = (J, \epsilon_j, \alpha_j,\ j = 1, \ldots, J)$, the number and location and scale of the Beta terms, is common across all samples. The weights w_{xj} are specific to each biologic condition. This is consistent with the interpretation of the peaks in the spectrum as arising from proteins in the sample. The weights are the abundance of protein j in the k-th sample. For the baseline we use a similar mixture of Beta model:

$$B_k(m) = \sum_{j=1}^{J_k} v_{kj} \text{ Be}(m;\ \eta_{kj}, \beta_{kj}). \tag{1.3}$$

Here J_k is the size of the mixture, v_{kj} are the weights, and (η_{kj}, β_{kj}) are the parameters of the j-th term in the mixture for the baseline of sample k. The parameters $\theta_k = (J_k, v_{kj}, \eta_{kj}, \beta_{kj},\ j = 1, \ldots, J_k)$ describe the Beta mixture model for the baseline in the k-th sample. The choice of the mixture of Beta representation for the baseline is for convenience of the implementation. Any alternative non-linear regression model, such as regression splines, could be used with little change in the remaining discussion.

The likelihood (1.1) describes the assumed sampling model, conditional on the unknown parameters $\theta = (\theta_0, \theta_k, w_k,\ k = 1, \ldots, K)$, with (1.2) and (1.3) defining how the paramters determine the sampling model. The model is completed with a hierarchical prior. Let $\text{Poi}^+(\lambda)$ denote a Poisson distribution with parameter λ, constrained to non-zero values, let $\text{Ga}(a, b)$ denote a Gamma distribution with expectation a/b, and let $U(a, b)$ denote a uniform distribution on $[a, b]$. For the baseline mixture we assume

$$J_k \sim \text{Poi}^+(R_0), \eta_{kj} \sim U(0,1),\ \beta_{kj} \sim U(\underline{\beta}, \overline{\beta}),\ \text{and } v_k \sim Dir(1, \ldots, 1). \tag{1.4}$$

Here $R_0, \underline{\beta}$ and $\overline{\beta}$ are fixed hyperparameters. For the peaks in the spectra we assume

$$J \sim \text{Poi}^+(R_1), \epsilon_j \sim U(0,1),\ \alpha_j \sim U(\underline{\alpha}, \overline{\alpha}), \tag{1.5}$$

with fixed hyperparameters $R_1, \underline{\alpha}$ and $\overline{\alpha}$. A constraint $\overline{\alpha} < \underline{\beta}$ ensures identifiability by uniquely identifying any given Beta kernel as a term in either the baseline mixture (1.4) or a peak in (1.5).

Finally, for the weights w_{xj} we assume common values for all samples under the same biologic condition. Thus the weights are indexed by biologic condition x and peak j. The prior model includes positive prior probability for ties $w_{0j} = w_{1j}$. Let $\lambda_j = I(w_{0j} = w_{1j})$ be an indicator for a tie and let $\Gamma = \{j : \lambda_j = 1\}$ and $L = \sum \lambda_j$ denote the set of indices and the number of peaks with $\lambda_j = 1$ and $W_1 = \sum_{j \in \Gamma} w_{0j}$. Let $w^{\star 1}$ and $w_x^{\star 0}$, $x = 0, 1$, denote the (standardized) subvectors of the weights w_{xj} defined by λ_j as follows:

$$w^{\star 1} = (w_1^{\star 1}, \ldots, w_L^{\star 1}) \equiv \frac{1}{W_1}(w_{0j};\ j \in \Gamma) \text{ and}$$
$$w_x^{\star 0} = (w_{x1}^{\star 0}, \ldots, w_{x,J-L}^{\star 0}) \equiv \frac{1}{1 - W_1}(w_{xj};\ j \notin \Gamma).$$

We assume

$$\begin{gathered} Pr(\lambda_j = 1) = \pi,\ j = 1, \ldots, J \\ w^{\star 1} \sim Dir(C_w, \ldots, C_w), \text{ and } w_x^{\star 0} \sim Dir(C_w, \ldots, C_w),\ x = 0, 1, \end{gathered} \tag{1.6}$$

and $p(W_1 \mid \lambda) = Be(L\, C_w, (J - L)\, C_w)$. In words, we assume that the weights $w_0 = (w_{01}, \ldots, w_{0J})$ and $w_1 = (w_{11}, \ldots, w_{1J})$ are generated as product of independent rescaled Dirichlet distributions on the subsets of differentially and non-differentially expressed peaks, with positive prior probability π of any of the peaks $j = 1, \ldots, J$ being identical across $x = 0, 1$. The model is completed with a hyper prior $\pi \sim \text{Be}(A_\pi, B_\pi)$ (using the conventional parametrization of a Beta distribution).

We recognize that the model specification includes several simplifying assumptions. For example, we assume equal weights w_{xj} across all samples under the same biologic condition. A more realistic prior would require a hierarchical extension with sample specific weights, centered at distinct means under each biologic condition. Instead, we chose to use the simplified model and subsume the model misspecification error in the multinomial sampling model. Another simplification is the uniform prior on the peak locations ϵ_j and peak widths α_j. A more informative prior might formalize the fact that peaks at higher masses tend to be wider, for reasons related to physics of the experimental arrangment. Such dependent priors could easily be substituted for (1.5).

1.5 Posterior inference

Inference in the proposed mixture of Beta model is implemented by Markov chain Monte Carlo (MCMC) simulation. Most details of the implementation are straightforward applications of standard MCMC algorithms. See, for example Tierney (1994). We describe the outline of the algorithm by indicating for each step the random variables that are updated, and the random quantities that are conditioned upon their currently imputed values. We use notation like $[x \mid y, z]$ to indicate that x is being updated, conditional on the known or currently imputed values of y and z. We generically use θ^- to indicate all parameters, except the parameter on the left side of the conditioning bar. Each iteration of the MCMC simulation includes the following steps: $[v_k \mid \theta^-, y^*]$, $[\lambda_j \mid \theta^-, y^*]$, $[W_1 \mid \theta^-, y^*]$, $[w^{\star 1} \mid \theta^-, y^*]$, $[w_x^{\star 0} \mid \theta^-, y^*]$, $[\pi \mid \theta^-, y^*]$, $[p_{0k} \mid y^*]$, $[J_k \mid \theta^-, y]$, $[J \mid \theta^-, y]$, $[\beta_{kj} \mid \theta^-, y]$, $[\eta_{kj} \mid \theta^-, y]$, $[\alpha_j \mid \theta^-, y]$, and $[\epsilon_j \mid \theta^-, y]$.

All steps except for the steps that update J and J_k are carried out with Metropolis-Hastings transition probabilities. We considered the use of imputed latent variables to replace the mixtures with conditionally conjugate hierarchical models, but found this to be computationally inferior. The transition probabilities used to update J and J_k require more attention. Changing the size J and J_k, respectively, of the mixtures implies a change in dimension of the parameter vector. We implement this by an application of reversible jump MCMC (RJ) transitions (Green, 1995).

The reversible jump moves for changing J and J_k use split/merge and birth/death proposals. The construction of the moves follows Richardson and Green (1997) who define RJ for mixture of normal models. However, the nature of the spectra with multipe highly peaked local modes requires a careful implementation. Below we explain the moves to update J. The moves for J_k are similar. The nature of B_k as relatively smooth baseline makes the transition probabilities to change J_k computationally easier to carry out.

We introduce a matching pair of *split* and *merge* moves to propose increments and decrements in J. The split move implements a proposal to replace a currently imputed peak j in f_k by two daughter peaks j_1 and j_2, maintaining the first two moments, and restricting the move such that the two new peaks are each others' nearest neighbors. The latter restriction simplifies the matching merge move. To select the peak to be split we pick with equal probabilty any of the current peaks. Without loss of generality assume $j = J$, $j_1 = J$ and $j_2 = J + 1$, and we assume $\lambda_J = 0$, i.e., the selected peak is imputed to be differentially expressed across biologic conditions $x = 0, 1$. The modifications for $\lambda_J = 1$ are straightforward. Also, in the following description we mark parameter values for the proposal with tilde, as in $\tilde{\theta}$. To propose new weights we generate two auxiliary variable $u_{1x} \sim Be(2, 2)$, $x = 0, 1$ and define $\tilde{w}_{xJ} = u_{1x}\, w_{xJ}$ and $\tilde{w}_{x,J+1} = (1-u_{1x})\, w_{xJ}$. To propose location and scale for the two new daughter peaks we generate two auxiliary variables $u_2 \sim \mathrm{Be}(2, 2)$ and $u_3 \sim \mathrm{Ga}(5, 5)$ and define $\tilde{\epsilon}_J = \epsilon_J + u_2\sqrt{\tilde{w}_{x,J+1}/\tilde{w}_{xJ}}$ and $\tilde{\epsilon}_{J+1} = \epsilon_J - u_2\sqrt{\tilde{w}_{xJ}/\tilde{w}_{x,J+1}}$ for the locations and $\tilde{\alpha}_J = \sqrt{(1-u_3)(1-u_2^2)\alpha_J^2 w_{xJ}/\tilde{w}_{xJ}}$ and $\tilde{\alpha}_{J+1} = \sqrt{u_3(1-u_2^2)\alpha_J^2 w_{xJ}/\tilde{w}_{xJ}}$. With the appropriate RJMCMC acceptance probabilty we accept the proposal as the new state of the MCMC simulation. Otherwise we discard the proposal. The matching *merge* proposal starts by selecting a pair of adjacent peaks, (j_1, j_2) and uses the inverse transformation to propose a merge.

Another pair of transition probabilities to change J are *birth* and *death* moves. To prepare a *birth* proposal it is important to slightly modify traditional *birth/death* proposals as used, for example, in Richardson and Green (1997). Because of the extremely narrow nature of the peaks in the spectrum, a randomly proposed new peak would have practically zero chance of of being accepted when evaluating the ratio of likelihood values in the RJ acceptance probability. Instead we exploit the availability of reasonable ad-hoc solutions. Let $A = \{\alpha_h^o, \epsilon_h^o, w_{xh}^o, x = 0, 1, \pi_h^o;\ h = 1, \ldots, H\}$ denote a list of peaks identified by a preliminary data analysis step, using for example the approach outlined in Section 1.2. We refer to A as the reference solution. In words, the *birth* move proceeds by proposing a new peak as a slightly jittered copy of one of the reference peaks. The new peak is restricted to be, among all currently imputed peaks, the closest neighbor to the identified reference peak. The additional randomness and the restriction to nearest neighbors is important to facilitate the matching *death* move. Specifically, we start by randomly selecting an index $h \in \{1, \ldots, H\}$ to identify a peak in the reference solution. Then evaluate $\Delta = \min\left\{|\epsilon_h^o - \epsilon_j|, j = 1 \ldots J,\ \frac{1}{2}|\epsilon_h^o - \epsilon_g^o|, g \neq h\right\}$ and $\sigma = \min\{\Delta, \alpha_h^o\}$ and generate auxiliary variables u, s, v_0, v_1, r

$$u \sim N(0, \sigma^2)\, I(|u| \leq \Delta),\ s \sim \mathrm{Ga}(c_b, c_b),\ Pr(r = 1) = \pi_h^o, v_x \sim \mathrm{Be}(c_b, c_b).$$

Again we use $\tilde{\theta}$, etc., to mark proposed parameter values. We propose $\tilde{\epsilon}_{J+1} = \epsilon_h^o + u$, $\tilde{\alpha}_{J+1} = \alpha_h^o\, s$ and $\tilde{\lambda}_{J+1} = r$. We define weights for the newly proposed peak as $\tilde{w}_{x,J+1} = w_{xh}^o\, v_x$, with $x = 0, 1$ if $\tilde{\lambda}_{J+1} = 0$ and $x = 0$ if $\tilde{\lambda}_{J+1} = 1$. Finally, we re-standardize the weights $\tilde{w}_{xj}$ to sum to W_1 over $\{j :\ \lambda_j = 1\}$ and to $(1 - W_1)$ over $\{j :\ \lambda_j = 0\}$.

The matching *death* move proceeds by identifying one one the reference peaks, again by $h \sim \mathrm{U}\{1, \ldots, H\}$, and finding the currently imputed peak j that is closest to ϵ_h^o. When evaluating the appropriate RJ acceptance probability we keep in mind that ϵ_j might be nearest neighbor to more than one reference peak.

1.6 Results

We implemented the proposed algorithm to analyze the lung cancer data set described in Section 1.2. We exclude the very low part of the m/z scale to avoid saturation artifacts. Figures 1.2 through 1.8 summarize the inference. We use Y to generically indicate the observed data. Figure 1.2 shows the estimated size of the beta mixtures. The baseline is adequately modeled with mixtures of between 1 and 4 beta kernels. Although the baseline is not of interest by itself, appropriate modeling of the baseline is critical for meaningful inference on the spectrum f_k. The number of distinct proteins seen in the spectrum is *a posteriori* estimated between 15 and 29 (Figure 1.2). The raw spectra include many more local peaks. But only a subset of these are a posteriori identified as protein peaks. Others are attributed to noise, or included in the estimated baseline. Figure 1.3 shows the estimated mean spectra. The figure plots $E[f_k(m) \mid Y, x_k = x]$ for a future, $k = (K+1)$-st sample of normal ($x = 0$) and cancer ($x = 1$) tissue, respectively. The posterior mean is evaluated as

$$E[f_k(m) \mid Y, x_k] = E[\sum_{j=1}^{J} w_{xj} Be(m;\ \epsilon_j, \alpha_j) \mid Y].$$

The posterior expectation is approximated as the ergodic average over the iterations of the MCMC simulation. The model includes the unknown distributions f_k as random quantities, parametrized by θ_0. Thus, in addition to point estimates, posterior inference provides a full probabilistic description of uncertainties. Figure 1.4 illustrates the posterior distribution $p(f_k \mid Y)$ by showing 10 random draws for the random distributions.

Posterior probabilities for any event of interest related to f_k can be reported. In particular, we find posterior probabilities for differential expression across the two biologic conditions. This is shown in Figure 1.5 which summarizes posterior inference on the indicators for differential expression, $1-\lambda_j$. (Recall that λ_j is defined as indicator for $w_{0j} = w_{1j}$, i.e., non-differential expression). The figure shows estimated marginal probabilities of differential expression for all distinct peaks. A minor complication arises in reporting and summarizing posterior inference about distinct proteins. The mixture f_k only includes exchangeable indices j, leading to the complication that the Beta kernel corresponding to a given protein might have different indices at different iterations of the posterior MCMC simulation. In other words, the protein identity is not part of the probability model. To report posterior inference on specific proteins requires additional post-processing to match Beta kernels that correspond to the same protein across iterations. We use an ad-hoc rule. Assume two peaks, j and h are recorded in the MCMC simulations and assume that the j-th peak occured first in the MCMC output. The two peaks j and h are counted as arising from the same protein if the difference in masses is below a certain threshold. Specifically, we use $|\epsilon_j - \epsilon_h| < 0.5\alpha_j$ to match peaks. Alternatively to this ad-hoc choice one could use a threshold related to the nominal mass accuracy of the instrument. The problem of reporting inference related to the terms in a mixture is known as the label switching problem. Figure 1.5 shows unique peaks using this rule. Also, only proteins that occur in at least 5% of the MCMC simulations are reported. Different sets of peaks appear in different iterations of the MCMC, i.e., the number of peaks shown in Figure 1.5 does not match J. All proteins with $Pr(\lambda_j = 0 \mid Y) > 50\%$ are considered differentially expressed and are are marked as solid dots. Figure 1.6 shows the estimated relative abundance $E(w_{xj} \mid Y)$ for all detected proteins. For peaks corresponding to differentially expressed proteins we plot the estimated abundance for $x = 0$ and $x = 1$, connected by a short line segment. Table 1.1 gives a brief description of known proteins with mass close to the identified peaks.

Inference on λ allows evaluation of posterior expected false discovery rates (FDR). The notion of FDRs is a useful generalization of frequentist type-I error rates to multiple hypothesis testing, introduced in Benjamini and Hochberg (1995). Let δ_j denote an indicator for rejecting the j-th comparison in a multiple comparison problem, i.e., deciding $\lambda_j = 0$ (no tie), in our example. FDR is defined as $\text{FDR} = \sum \lambda_j \, \delta_j / \sum \delta_j$, the fraction of false rejections, relative to the total number of rejections. Applications of FDR to high throughput gene expression data are discussed, among others, by Storey and Tibshirani (2003). Posterior expectated FDR is easily evaluated as $\overline{\text{FDR}} = E(\text{FDR} \mid Y) = [\sum E(\lambda_j \mid Y) \, \delta_j] / \sum \delta_j$. Let $\overline{\lambda}_j = E(\lambda_j \mid Y)$ denote the marginal posterior probability of non-differential expression for peak j. Consider now decision rules that classify a protein as differentially expressed if $1 - \overline{\lambda}_j > \gamma^*$. In analogy to classical hypothesis testing, we fix γ^* as the minimum value that achieves a certain pre-set false discovery rate, $\overline{\text{FDR}} \leq \alpha$. It can be shown (Müller et al., 2004) that under several loss functions that combine false negative and false discovery counts and/or rates the optimal decision rule is of this form. Newton et al. (2004) comment on the dual role of $\overline{\lambda}_j$ in decision rules like $\delta_j = I(1 - \overline{\lambda}_j > \gamma^*)$. It determines the decision, and at the same time already reports the probability of a false discovery as $\overline{\lambda}_j$ for $\delta_j = 1$ and the probability of a false negative as $1 - \overline{\lambda}_j$ for $\delta_j = 0$.

An important feature of the proposed model is the multilevel nature of the hierarchical probability model defined in (1.1) through (1.6). The posterior on the indicators λ_j contains all relevant information about differential levels of protein expression. Thus any inference about patterns of protein expression across different biologic conditions can be derived from $p(\lambda \mid Y)$ only. The discussed inference for the multiple comparison is one example. Another important decision problem related to the λ_j indicators is the identification of a minimal set of of proteins to classify samples according to biologic condition. In contrast to the multiple comparison decision the goal now is to select a small number of differentially expressed proteins. A convenient formalization of this question is to find locations m_i with maximum posterior probability $P(\lambda_j = 1 \mid Y)$ for peaks located at m_i. Figure 1.7 shows the optimal sets of peak locations as a function of the desired size of the set.

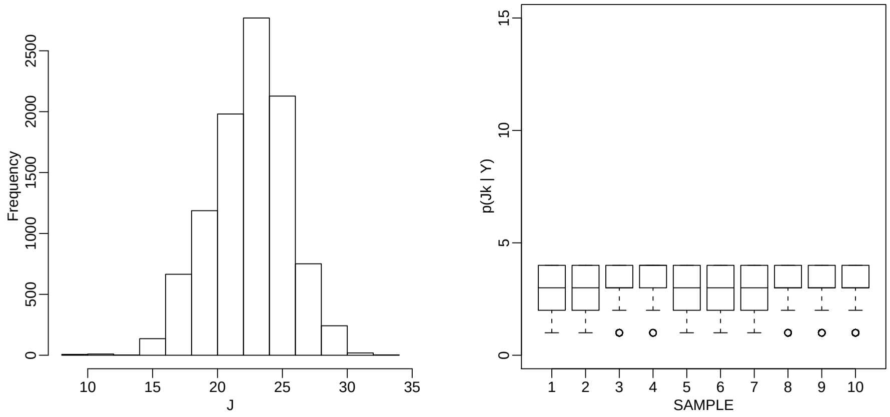
Finally, Figure 1.8 shows some aspects of the posterior MCMC, plotting the trajectory of imputed values for J, and for the unique peak locations ϵ_j against iterations.

1.7 Discussion

We have proposed a likelihood-based approach to inference about differential expression of proteins in mass/charge spectra from SELDI or MALDI

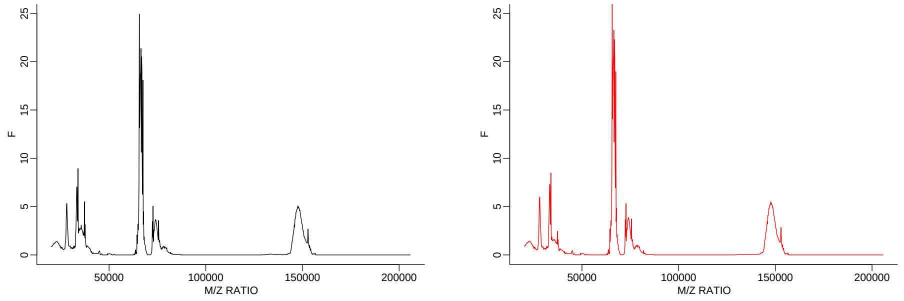
FIGURE 1.1: Detected proteins. The table reports all human proteins within 0.1% of the reported masses ϵ_j *and* with `Swissprot` entries reporting terms ”tumor, cancer, lung,” or “carcinoma.”

Mass	name	description
11662	S10AE	Expressed at moderate level in lung
13016	MAGA5	Expressed in many tumors of several types, such as melanoma, head and neck squamous cell carcinoma, lung carcinoma and breast carcinoma, but not in normal tissues except for testes
13018	MAGB5	Expressed in testis. Not expressed in other normal tissues, but is expressed in tumors of different histological origins
15126	HBA	Involved in oxygen transport from the lung; Defects cause thalassemia
15864	ERG28	Ubiquitous; strongly expressed in testis and some cancer cell lines
15867	HBB	Involved in oxygen transport from the lung; Defects cause sickle-cell anemia
15989	PA2GE	Present in lung
29383	FA57A	Not detected in normal lung
29937	LAPM5	High levels in lymphoid and myeloid tissues. Highly expressed in peripheral blood leokocytes, thymus, spleen and lung
35010	GPR3	Expressed in lung at low level
35049	PLS1	Expressed in lung
35055	MAGA2	Expressed in many tumors of several types, such as melanoma, head and neck squamous cell carcinoma, lung carcinoma and breast carcinoma, but not in normal tissues except for testes
35844	SPON2	Expressed in normal lung tissues but not in lung carcinoma cell lines
65369	SEPT9	Chromosomal aberration involving SEPT9/MSF is found in therapy-related acute myeloid leukemia
65418	IL1AP	Detected in lung



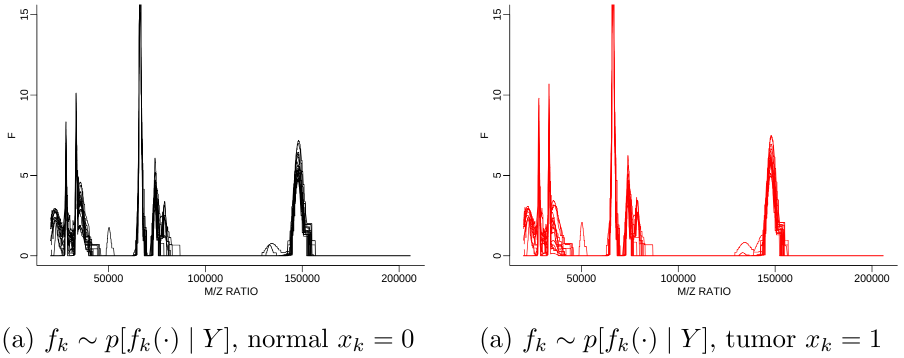
(a) $p(J \mid Y)$ (b) $p(J_k \mid Y)$, $k = 1, \ldots, 10$

FIGURE 1.2: $p(J \mid Y)$ and $p(J_k \mid Y)$. The histogram in the left panel shows the posterior probabilities for the number of distinct peaks. We find the posterior mode around 18 peaks. The boxplots (right panel) summarize the posterior distributions $p(J_k \mid Y)$. The number of terms in the baseline mixtures were constrained at $J_k \leq 5$.



(a) $E[f_k(\cdot) \mid Y]$, normal $x_k = 0$ (a) $E[f_k(\cdot) \mid Y]$, tumor $x_k = 1$

FIGURE 1.3: $E[f_k(m) \mid Y, x_k = x]$. Estimated spectrum for normal and tumor samples.



(a) $f_k \sim p[f_k(\cdot) \mid Y]$, normal $x_k = 0$ (a) $f_k \sim p[f_k(\cdot) \mid Y]$, tumor $x_k = 1$

FIGURE 1.4: Posterior inference defines a probability model on the unknown spectra. The two panels show ten draws from $p(f_k \mid Y)$ for $x_k = 0$ (left) and $x_k = 1$ (right).

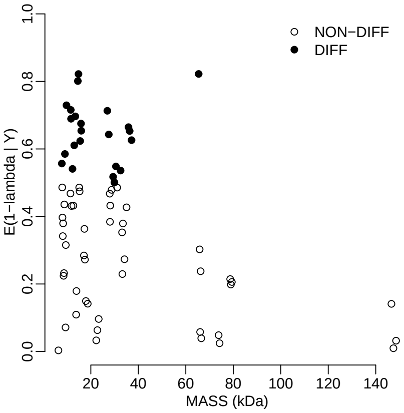
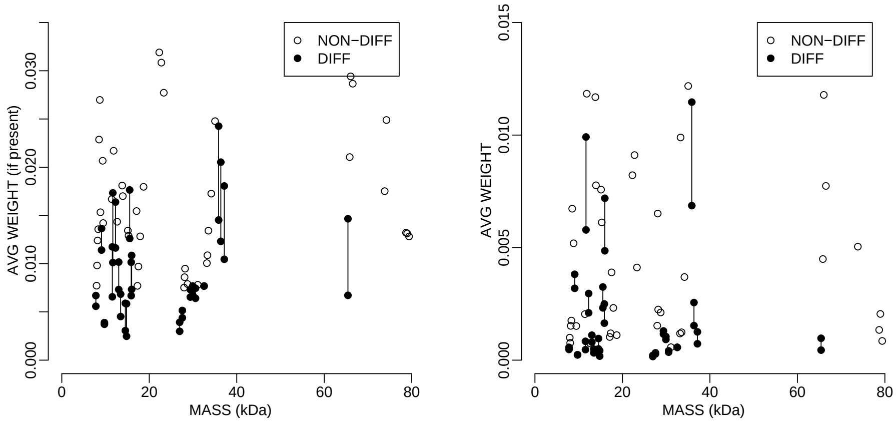


FIGURE 1.5: $E(1 - \lambda_j \mid Y)$. Posterior expected probability of differential expression for the reported peaks.



(a) conditional on presence (b) averaging over presence

FIGURE 1.6: Detected proteins. The figure plots the posterior estimated weights $E(w_{xj} \mid Y)$ against mass/charge ratios $E(\epsilon_j \mid Y)$. Panel (a) plots the average weight $E(w_{xj} \mid Y)$, averaging over all iterations that reported a protein peak at ϵ_j. Panel (b) scales the average weight by the probability of a peak being detected at ϵ_j (i.e., no detection is counted as $w = 0$). For genes with $E(\lambda_j \mid Y) > 0.5$ we plot the posterior expectation of w_{0j}. For differentially expressed genes we plot posterior expectations for w_{xj} for both conditions, $x = 0, 1$, linked with a line segment. Only proteins with posterior expected weight greater equal 0.05 are shown.

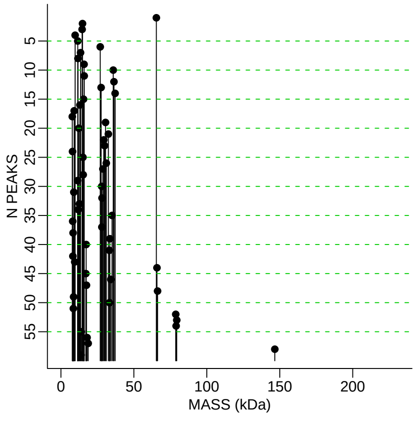
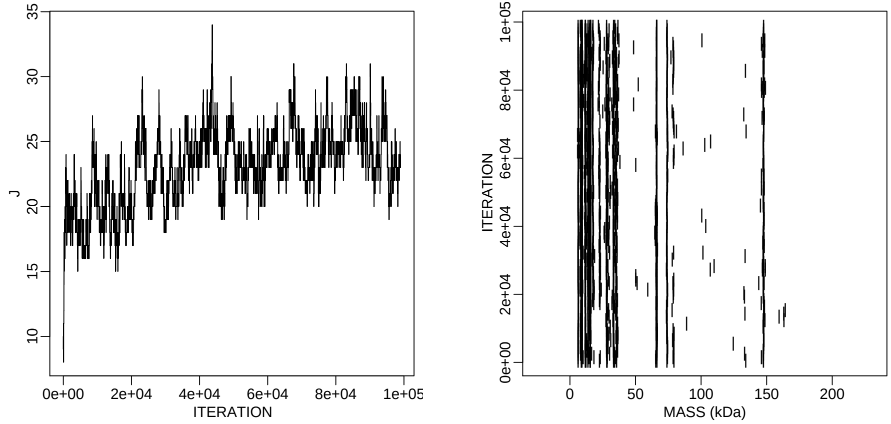


FIGURE 1.7: Proteins included in a panel of biomarkers to classify samples according to biologic condition x. Let $E_R = \{\epsilon_1, \ldots, \epsilon_R\}$ denote the locations of peaks, i.e., proteins, included in a subset of size R. The figure plots for each location ϵ_r the minimum size R of subsets that include ϵ_r.



(a) J against iteration m (b) ϵ_j against iteration

FIGURE 1.8: Some aspects of the MCMC simulation. The left panel plots the imputed number of proteins J, against iteration. The right panel plots the locations ϵ_j of the imputed beta kernels (x-axis) over iterations (y-axis).

experiments. We argued that the appropriate likelihood is based on random sampling. The usual approach of smoothing the raw spectrum is reasonable and leads to almost identical point estimates.

An important advantage of the proposed model is the easy generalization to more complicated experimental setups. Conditional on the λ indicators the rest of the model is independent of the biologic conditions for each sample. Consider, for example, an experiment with more than two biologic conditions (more than two tumor types, etc.). For more than two conditions it is convenient to describe ties by configuration indicators $s_{jx} \in \{1, \ldots, S_j\}$. For example, for four conditions a configuration of $s_j = (1, 1, 1, 2)$ would indicate that the first three conditions share the same peak, whereas expression is different under the fourth condition.

Our approach allows added flexibility in addressing a number of statistical challenges presented by MALDI-TOF mass spectra, including the interpretation of multiple tests, modelling and overfitting, and inadequate covariance in estimation, as well as substantial autocorrelation within a spectrum. In particular, a typical characteristic of MS data is that variance (and higher moments) appear to be related to mean intensity, resulting in measured intensities at the highest protein "peaks" exhibiting greater variability than do less abundant species. This property makes the magnitude of high intensity peaks less reliable for ensuing in ference.

References

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

Alberts, B., Bray, D., Lewis, J., Ra, M., Roberts, K., and Watson, J. D. (1994), *Molecular biology of the cell (3rd ed.)*, New York, NY: Garland.

Arthur, J. M. (2003), “Proteomics,” *Current opinion in nephrology and hypertension*, 12, 423–430.

Baggerly, K. A., Coombes, K. R., and Morris, J. S. (2006), “Bayesian Inference for Gene Expression and Proteomics,” Cambridge University Press, chap. An Introduction to High-Throughput Bioinformatics Data, pp. xxx–xxx.

Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L. C., and Coombes, K. R. (2003), “A comprehensive approach to analysis of MALDI-TOF proteomics spectra from serum samples,” *Proteomics*, 9, 1667–1672.

Ball, G. S., Mian, F., Holding, F., Allibone, R. O., Lowe, J., Ali, S., G., L., McCardle, S., Ellis, I. O., Creaser, C., and Rees, R. C. (2002), “An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumors and rapid identification of potential biomarkers,” *Bioinformatics*, 18, 395–404.

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

Chen, G., Gharib, T. G., Huang, C.-C., Thomas, D. G., Shedden, K. A., Taylor, J. M. G., Kardia, S. L. R., Misek, D. E., Giordano, T. J., Iannettoni, M. D., Orringer, M. B., Hanash, S. M., and Beer, D. G. (2002), “Proteomic analysis of lung adenocarcinoma: identification of a highly expressed set of proteins in tumors,” *Clinical Cancer Research*, 8, 2298–2305.

de Hoffman, E. and Stroobant, V. (2002), *Mass Spectrometry: Principles and Applications*, John Wiley.

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

0-8493-0052-5/00/$0.00+$.50

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004), "Optimal Sample Size for Multiple Testing: the Case of Gene Expression Microarrays," *Journal of the American Statistical Association*, 99.

Newton, M., Noueriry, A., Sarkar, D., and Ahlquist, P. (2004), "Detecting differential gene expression with a semiparametric heirarchical mixture model," *Biostatistics*, 5, 155–176.

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

Petrone, S. (1999a), "Bayesian density estimation using Bernstein polynomials," *Canadian Journal of Statistics*, 27, 105–126.

Richardson, S. and Green, P. J. (1997), "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society B*, 59, 731–792.

Robert, C. and Rousseau, J. (2003), "A mixture approach to Bayesian goodness of fit," Tech. rep., CREST/INSEE, Paris.

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

Srinivas, P. R., Srivastava, S., Hanash, S., and Wright, Jr., G. L. (2001), "Proteomics in early detection of cancer," *Clinical Chemistry*, 47, 1901–1911.

Storey, J. S. and Tibshirani, R. (2003), "SAM Thresholding and False Discovery Rates for Detecting Differential Gene Expression in DNA Microarrays," in *The analysis of gene expression data: methods and software*, eds. Parmigiani, G., Garrett, E. S., Irizarry, R. A., and Zeger, S. L., New York: Springer.

Tierney, L. (1994), "Markov chains for exploring posterior distributions," *The Annals of Statistics*, 22, 1701–1762.

Wu, B., Abbott, T., Fishman, D., McMurray, W., More, 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*, 13, September.

Yasui, Y., Pepe, M., Thompson, M. L., Adam, B. L., Wright Jr., G. L., 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, 449–463.