

Nonparametric Models for Peak Identification and Quantification in MALDI-TOF Mass Spectroscopy

Leanna L. House Merlise A. Clyde

Robert L. Wolpert

Institute of Statistics and Decision Sciences

Department of Statistical Science

Duke University, Durham, NC 27708-0251

January 12, 2007

Abstract

We present a novel nonparametric Bayesian model using Lévy random field priors for identifying the presence and abundance of proteins from mass spectrometry data. Informed prior distributions, based on expert opinion and on preliminary laboratory experiments, help distinguish true peaks from background noise and help resolve uncertainty about peak multiplicity.

Key words: Bayes; Kernel Regression; Lévy random fields

1 Introduction

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

Identifying peak locations (which represent proteins) and quantifying protein abundance is often preceded by a two or more stage analysis, involving calibration, normalization, baseline subtraction and filtering of noise [Morris *et al.*, 2005; Tibshirani *et al.*, 2004; Yasui *et al.*, 2003; Carpenter *et al.*, 2003]. A problem with such multistage analyses is that each individual step potentially introduces errors or biases that may subsequently create challenges for later stages such as classification of subjects or identification of biomarkers; methods that simultaneously model background, noise and features may lead to improved classification or inferences [Coombes *et al.*, 2005b]. Nonparametric models such as wavelets have proved successful in simultaneously modeling background and denoising, allowing one to extract features or regions of spectra that differentiate groups [Yasui *et al.*, 2003; Coombes *et al.*, 2005b]. While wavelets are well suited for modeling local features like spectral peaks, the coefficients and basis functions used in the representation of expected intensity have no inherent biological interpretation. In this paper, we propose a novel nonparametric method employing Lévy Adaptive Kernel Regression Models (LARKs) [Tu *et al.*, 2006; House, 2006; Clyde and Wolpert, 2006] that provides the adaptivity and flexibility that make wavelet methods advantageous, but more importantly uses a model parameterization for features with direct biological interpretations.

We begin in Section 2 with a brief overview of MALDI-TOF mass spectrometry. In

Section 3 we develop a statistical model for protein abundance as a function of time-of-flight using a novel nonparametric Bayesian approach. The model encompasses both signal (the protein abundance) and noise (due to artifacts of the MALDI-TOF technology), including run-to-run variability. Based on physical models for mass spectroscopy [Coombes *et al.*, 2005a], the distribution of the time of flight of a given protein may be represented by a kernel density, such as a Gaussian or Cauchy density with location parameter representing the expected time of flight and width parameter governed by both the mass of the protein and resolution of the machine (and its settings) used for MS. The unknown protein signal is then represented as a convolution of these kernels with a distribution that characterizes protein abundance at expected times of flight. Solving this deconvolution problem provides estimates of the number of proteins, their times of flight, and abundances. As deconvolution problems typically have no unique solution, we utilize a Bayesian approach that incorporates prior knowledge about the process which facilitates resolving the number of peaks (proteins). Prior distributions for the Bayesian model are developed in Section 4 from expert knowledge about the MALDI-TOF procedure and from exploratory analysis of MALDI-TOF data from related experiments. Inference about parameters of clinical interest, based on posterior distributions, are described in Section 5. In Section 6 we validate our method and compare it to the conventional peak-finding algorithm `PROcess` [Li, 2005] using simulated data. We illustrate our methodology in Section 7 using data from a recent lung cancer study conducted at Duke University. We conclude with a discussion and suggestions for future work in Section 8.

2 MALDI-TOF Data

In Matrix Assisted Laser Desorption Time-of-Flight Mass Spectrometry, or MALDI-TOF MS, inference about the molecular composition of a compound is based on indirect measure-

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

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

In actual MALDI-TOF spectra (see Figure 1.) we observe irregular peaks rather than one-dimensional spikes because molecules of equal size and charge do not all reach the detector at the same time. The most important of the many causes of TOF dispersion is variability in the amount of ionizing laser energy received by molecules of varying location within the matrix; those further from the matrix surface or from the center of the laser pulse may receive less kinetic energy and thus have lower initial velocities than similarly-sized molecules located closer to the center, delaying their arrival at the detector. Molecules may exchange energy in collisions, and may lose or gain mass through fragmentation and agglomeration, respectively. All these lead to TOF variation for each molecular species [Coombes *et al.*, 2005a; Zhigilei and Garrison, 1998; Franzen, 1997].

The interpretation and analysis of MALDI-TOF data are complicated by several other

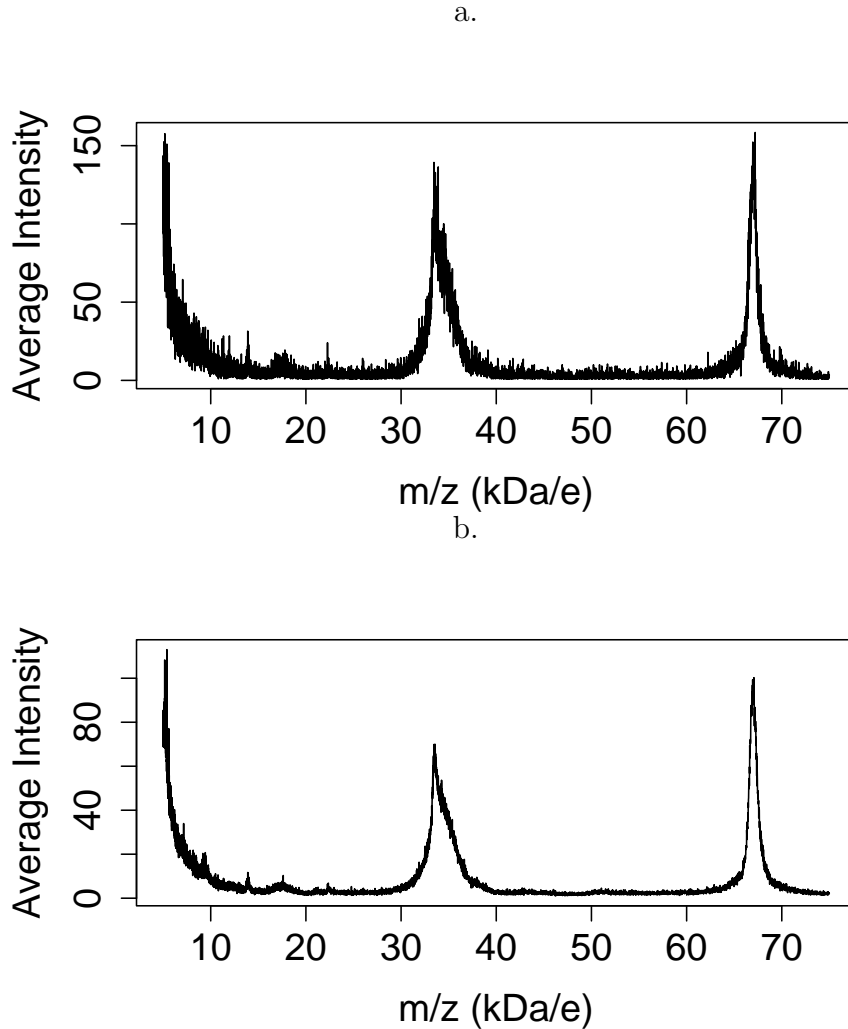


Figure 1: Single spectrum (a.) and mean of ten spectra (b.) from a single fraction, single subject. Note noise reduction and peak broadening in (b.)

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

One way to accommodate these sources of variability is to construct models for peak identification and quantification that incorporate all these recognized sources of variability, as in the wavelet approach of [Morris *et al.* \[2005\]](#). Our approach, described in Section 3, has the advantage that each of the model parameters has a direct physical interpretation.

3 A Model for MALDI-TOF

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

$$Y_t = \frac{Y_t^{\text{ob}} - \min(\mathbf{Y}^{\text{ob}})}{l} \quad (1)$$

based on a raw spectrum $\mathbf{Y}^{\text{ob}} = \{Y_t^{\text{ob}}\}_{T_0 \leq t \leq T_1}$ with l laser shots. [Dass \[2001, p. 75\]](#) suggests that the initial molecular velocities will be approximately Gaussian in distribution. This and the physical modeling of the MALDI-TOF process by [Coombes *et al.* \[2005a\]](#) suggest that TOFs for a single isotopic peak will also have symmetric bell-shaped distributions in the time domain, leading us (and others— see [[Morris *et al.*, 2005](#); [Kempka *et al.*, 2004](#); [Malyarenko *et al.*, 2005](#)]) to prefer TOF (in μs) rather than m/z (in Da/e) for spectral modeling.

3.1 Peak Shape

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

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

and Cauchy (sometimes called Lorentzian in the MS literature)

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

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

$$f(t) = \sum_{j=1}^J k(t; \tau_j, \omega_j) \eta_j, \quad (3)$$

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

3.2 Peak Width and Resolution

Protein peaks tend to be broader for late-arriving molecules than for earlier ones, with width nearly proportional to arrival time [Siuzdak, 2003, p. 44]; for this reason it is conventional in mass spectrometry to quantify the precision (narrowness) of a kernel $k(\cdot; \tau, \omega)$ not by the width ω , but by the *resolution*

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

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

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

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

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

Prior knowledge about precision can be used to resolve the ambiguity illustrated in Figure 2., where the observed spectrum may arise from either a single wide peak or a pair of near-by narrower peaks.

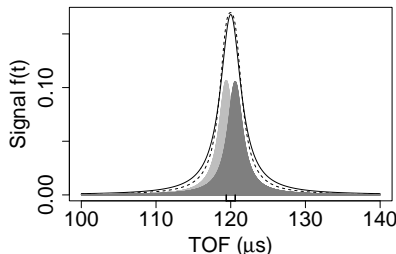


Figure 2: The (nearly indistinguishable) solid and dotted lines represent simulated protein signals from a sample mixture with either one wide or two narrow peaks.

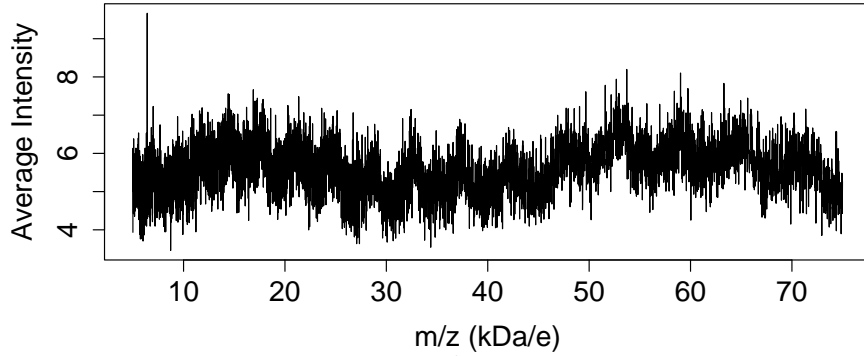
3.3 Background Noise Sources

Even in the absence of any protein molecules (*i.e.*, with $f(t) \equiv 0$) the MALDI-TOF spectrum does not vanish. Figure 3 a. shows the nearly-constant level of thermal noise from a run with an empty plate, while Figure 3 b. shows the rapidly-decreasing signal with only the sinapinic acid matrix, showing the arrival at the detector of ionized matrix molecules (far lighter than the proteins under study, hence near the left of the spectrum). Together these contribute a background that falls off nearly exponentially to a non-zero asymptote. Exploratory analysis suggests that the matrix molecular signal $\beta_0(t)$ can be modeled adequately as an exponential function,

$$\beta_0(t) = k_0(t; \omega_0) \eta_0 = \frac{\eta_0}{\omega_0} \exp\{-t/\omega_0\} \mathbf{1}_{(t>0)}, \quad (5)$$

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

a.



b.

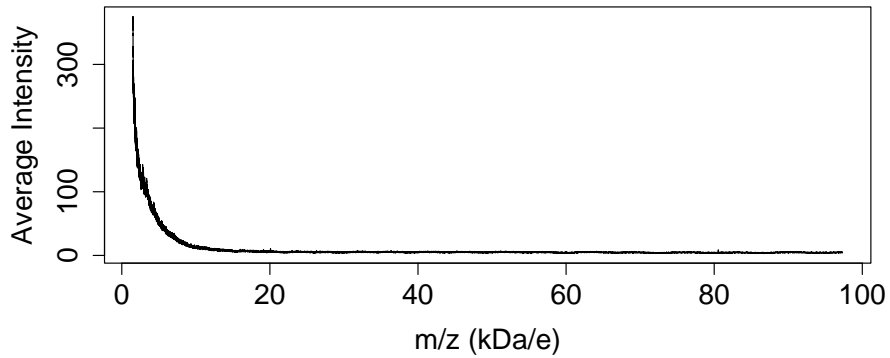


Figure 3: Figure 3a. shows the near-uniform thermal noise spectrum (or “ringing”) from an empty plate, while Figure 3b. shows the rapidly-decreasing spectrum from the sinapinic acid matrix without any protein sample. Note the vertical axes are scaled differently.

3.4 Mean Spectrum

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

$$\mu(t) = \zeta \left\{ (1 - S) + S[f(t) + \beta_0(t)] \right\} \quad (6)$$

for an overall scale ζ , a dimensionless number $S \in [0, 1]$, the protein signal $f(t)$ from Equation (3), and the matrix molecular signature $\beta_0(t)$ from Equation (5). The term S represents the proportion of observed intensity produced by molecular signal (both matrix and protein),

rather than by the ringing and thermal noise of Figure 3 a..

3.5 Likelihood

Both gamma and log-normal distributions are commonly used to model positive data like Y_t . We based our choice on the observation that the variance is proportional to the mean for gamma distributions and to the square of the mean for log-normals. Exploratory data

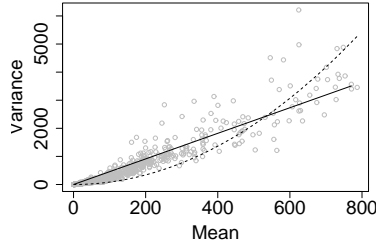


Figure 4: Linear and quadratic fits of mean intensity *vs.* variance of intensity for 200 μ s blocks of observations from a single spectrum.

analysis (from both a Box-Cox approach, and a regression comparison illustrated in Figure 4.) suggests that the variance of standardized MS data Y_t , given the mean, is nearly proportional to the first power of the mean, supporting the gamma model

$$Y_t \mid \mu(\cdot), \varphi \stackrel{\text{ind}}{\sim} \text{Ga}(\varphi\mu(t), \varphi), \quad (7)$$

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

$$\mathcal{L}(\boldsymbol{\theta}; \mathbf{Y}) = \prod_{i=1}^n \text{Ga}(Y_{t_i}; \varphi\mu(t_i), \varphi) \quad (8)$$

for the parameter vector $\boldsymbol{\theta}$ comprising the conditional mean function $\mu(\cdot)$ (or, equivalently from Equation (6), all of ζ , J , $\{\tau_j, \omega_j, \eta_j\}_{1 \leq j \leq J}$, S , ω_0 , and η_0) and φ . Here $\mathbf{Y} = \{Y(t_i)\}_{1 \leq i \leq n}$ represents the vector of standardized intensities, and $\text{Ga}(y; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\beta y} \mathbf{1}_{(y>0)}$ is the probability density function at $y \in \mathbb{R}$ for the gamma $\text{Ga}(\alpha, \beta)$ distribution.

Typically the likelihood function of Equation (8) has many modes because it is difficult to distinguish wide peaks from clusters of narrow ones, or small peaks from noise, from the data alone. Estimating θ (and in particular J , the number of protein peaks) by direct maximization of the likelihood leads to over-fitting the data and to over-estimating J . This can be overcome by regularization [Tikhonov, 1963] or by a Bayesian approach like ours, in which prior distributions penalize overly complex models.

4 Prior Distributions for MALDI-TOF

We now address the problem of constructing a joint prior distribution for all the unknown parameters of the model of Section 3,

$$\begin{aligned}
 Y_t \mid \mu(\cdot), \varphi &\stackrel{\text{ind}}{\sim} \text{Ga}(\varphi\mu(t), \varphi) & (9) \\
 \mu(t) &= \zeta \left\{ (1 - S) + S[f(t) + \beta_0(t)] \right\} \\
 f(t) &= \sum_{j=1}^J k(t; \tau_j, \omega_j) \eta_j, \quad \beta_0(t) = \frac{\eta_0}{\omega_0} \exp\{-(t - T_0)/\omega_0\} \mathbf{1}_{(t > T_0)}
 \end{aligned}$$

4.1 Measurement Error φ and Overall Level ζ

The exploratory data analysis of Section 3.5 suggests that the sample mean of the $\{Y_t\}$ is nearly proportional to the variance, with a ratio of approximately $\varphi \approx 0.223$. We use a gamma prior distribution $\varphi \sim \text{Ga}(a_\varphi = 0.5, b_\varphi = 1)$ to place about 90% of the prior mass in the interval $[0.002, 2]$.

The parameter ζ may be interpreted as the mean level or scale for Y_t , since $\text{E}[f(t)] \approx 1$ (see Section 4.2). Since experimental levels depend on a wide range of exogenous variables and vary widely among trials, we use a rather tight data-dependent prior distribution for ζ

centered at the empirical mean \bar{Y} ,

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

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

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

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

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

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

Conditional on J , we take the triplets $\{(\tau_j, \omega_j, \eta_j)\}_{1 \leq j \leq J}$ to be independent and identically-distributed. For peak abundances $\{\eta_j\}$ we use the truncated gamma distribution $\text{Ga}(0, \lambda, \epsilon)$ with parameters λ and ϵ chosen below in Section 4.4. Here $\text{Ga}(\alpha, \lambda, \epsilon)$ denotes the truncated gamma distribution with density function

$$\text{Ga}(\eta; \alpha, \lambda, \epsilon) \equiv \frac{\lambda^\alpha}{\Gamma(\alpha, \lambda\epsilon)} \eta^{\alpha-1} e^{-\lambda\eta} \mathbf{1}_{(\eta > \epsilon)}, \quad (10)$$

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

There is little reason to give higher prior probability to one range of TOFs than another without prior knowledge of the collection of proteins present in the samples. Thus we take

$\{\tau_j\}_{1 \leq j \leq J} \stackrel{\text{iid}}{\sim} \text{Un}(T_0, T_1)$ (independently of J and $\{\lambda_j\}_{1 \leq j \leq J}$), for some interval large enough to exceed the TOF for all molecules of interest. To eliminate saturation by matrix molecules at the low end, and include as wide a range as possible of the biologically relevant molecules, we chose the range $[5 \text{ kDa/e} \leq m/z \leq 75 \text{ kDa/e}]$, leading to TOF range $[T_0, T_1] = [32 \mu\text{s}, 278 \mu\text{s}]$ of length $T = T_1 - T_0 = 246 \mu\text{s}$.

To construct a prior distribution on the widths $\{\omega_j\}_{1 \leq j \leq J}$ we first use expert opinion to construct an informed prior distribution on the resolutions $\{\rho_j\}_{1 \leq j \leq J}$ (see Section 3.2). Siuzdak [2003, p. 44] suggests that individual peak resolutions ρ_j *should* be nearly constant across the entire TOF range, but in practice they are observed to vary [Coombes *et al.* 2005a; Applied Biosystems 2001, p. 6-32]. To reflect this variation we construct a hierarchical prior probability distribution for the resolution parameters $\{\rho_j\}_{1 \leq j \leq J}$. Independently of J and $\{(\lambda_j, \tau_j)\}_{1 \leq j \leq J}$, we take

$$\begin{aligned} \varrho &\sim \text{LN}(\log(\varrho\mu), \varrho\sigma^2) \\ \rho_j \mid \varrho &\stackrel{\text{iid}}{\sim} \text{LN}(\log(\varrho), \rho\sigma^2), \end{aligned}$$

centered around a hyperparameter $\varrho\mu$. We take $\varrho\mu = 200$ and set $\varrho\sigma = 0.7$ so that the LogNormal prior distribution of ϱ covers the range 50-800 with 95% probability, based on reported ranges of resolution from the literature [Applied Biosystems, 2001, Table 6-2 and Table H-6] The standard deviation for the individual resolutions was set to $\rho\sigma = 0.35$, one half of the population standard deviation for resolution. For a population resolution of $\varrho\mu = 50$ this leads to a prior 95% interval for individual resolutions of (25, 100), while at the upper extreme with a population resolution of $\varrho\mu = 800$, the prior 95% interval covers (403, 1589). The relationship between width, TOF, and resolution given by Equation (4)

now induces a prior distribution on the width parameters, *e.g.*,

$$\{\omega_j\}_{1 \leq j \leq J} \mid \tau, \varrho \stackrel{\text{ind}}{\sim} \text{LN}(\log(\tau_j/2\varrho), 0.05^2)$$

for the Cauchy kernel.

A random variable $\eta \sim \text{Ga}(\alpha, \lambda, \epsilon)$ with the truncated gamma distribution of Equation (10) has mean $\mathbb{E}[\eta] = (\alpha/\lambda) + \epsilon^\alpha \lambda^{\alpha-1} e^{-\lambda\epsilon} / \Gamma(\alpha, \lambda\epsilon)$ in general or, in our case with $\alpha = 0$,

$$\mathbb{E}[\eta] = \frac{1}{\lambda e^{\lambda\epsilon} \mathbb{E}_1(\lambda\epsilon)}.$$

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

$$\mathbb{E}[f(t)] \approx \frac{\mu_J}{T \lambda e^{\lambda\epsilon} \mathbb{E}_1(\lambda\epsilon)} \quad T_0 \ll t \ll T_1 \quad (11)$$

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

4.3 Prior Distribution for Matrix Background

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

The exponential fall-off rate $\frac{1}{\omega_0}$ of $\beta_0(t)$ (see Equation (5)) can be estimated by logarithmic regression of an initial segment of the blank spectrum from Figure 3b. From the estimate $\frac{1}{\omega_0} \approx 0.0492 \pm 0.0001367$ (mean \pm one standard error, in units of μs^{-1}) we infer

(using the delta method) that $\log(\omega_0) \approx 11.5 \pm 0.00278$. To accommodate possible variation between the blank spectrum experiment and the protein analysis, we use a much broader prior distribution,

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

We use a truncated gamma model for the abundance η_0 ,

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

with λ_0 chosen to match the mean $\text{E}[\eta_0] = \{\lambda_0 e^{\epsilon\lambda_0} \text{E}_1(\epsilon\lambda_0)\}^{-1} \approx \hat{\eta}_0$ with the estimate $\hat{\eta}_0$ from nonlinear regression on an initial segment of the protein data set short enough that it does not appear to include any peaks associated with proteins (we used $0 < t < 40 \mu\text{s}$, corresponding to the range $0 < m/z < 6 \text{ kDa/e}$). The solution is $\lambda_0 = x/\epsilon$ for the solution x to the equation

$$x e^x \text{E}_1(x) = \frac{\epsilon}{\hat{\eta}_0}$$

(easily found using MathematicaTM [2005] or MapleTM [2005], for example, or the approximations in Abramowitz and Stegun [1964, §5.1.53–56]).

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

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

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

4.4 Random Field Formulation

The distribution constructed in Section 4.2 for J and $\{(\tau_j, \omega_j, \eta_j)\}_{1 \leq j \leq J}$ induces one for $f(\cdot) = \sum_{j=1}^J k(t; \tau_j, \omega_j) \eta_j$, which can be written in integral form

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

with a random measure

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

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

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

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

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

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

with finite Lévy measure

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

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

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

5 Posterior Analysis

To support inference about protein peak locations and abundance, and about other model parameters, we construct an ergodic Markov chain in the space Θ of possible parameter vectors $\boldsymbol{\theta} = \{\zeta, J, \{\tau_j, \omega_j, \eta_j\}_{1 \leq j \leq J}, S, (\omega_0, \eta_0), \lambda_J, \rho\}$ with the posterior distribution as its stationary distribution [Besag *et al.*, 1995; Tierney, 1994; Gelfand and Smith, 1990]. At each Markov chain step we select one of the components of $\boldsymbol{\theta}$ and either replace it with a draw from its complete conditional posterior distribution, given the other components (a Gibbs step) or, if this is impractical, propose a small change in that component which is then accepted or rejected according to the Hastings probabilities (a random-walk Metropolis-Hastings or M-H step). Note that each proposed change in J (which we always take to be an M-H step of size one) changes the *dimension* of $\boldsymbol{\theta}$ (by three), requiring some delicacy in computing the Hastings ratios; such schemes, called “reversible jump MCMC algorithms,” were introduced by Green [1995]. Our approach is modeled after that of Wolpert and Ickstadt [2004], who introduced a general RJMCMC procedure for Lévy random field models. For updating the varying dimensional parameters $\{\tau_j, \omega_j, \eta_j\}_{1 \leq j \leq J}$ we consider five possible moves: peak birth (incrementing J by one and introducing a new triplet $(\tau_*, \omega_*, \eta_*)$), peak death (decrementing J by one and removing a randomly-chosen triplet $(\tau_j, \omega_j, \eta_j)$), and peak update (moving a randomly-chosen triplet $(\tau_j, \omega_j, \eta_j)$ within \mathbb{R}^3). Peak splitting, in which a single large peak is replaced by a pair of smaller ones, and peak merging, in which two nearby peaks are replaced

with a single larger one lead to a vast improvement in algorithmic efficiency over RJMCMC algorithms using only birth/death and update steps.

For sufficiently large spectra or complex protein mixtures, convergence to the posterior distribution from random starting values may require millions of MCMC iterations. To reduce computation time we begin the Markov chain close to its mode, located by applying the EM algorithm [Dempster *et al.*, 1977] to a simple Gaussian approximation to our LARK model; see House [2006] for details.

5.1 Peak Identification

At each iteration of the RJ-MCMC sampler, locations of the expected TOF (τ_j) are updated, with the number of peaks potentially changing. A technical issue with using RJ-MCMC algorithms for peak identification involves summarizing a high dimensional parameter vector of varying dimension. There are two approaches that we have implemented to identify peaks in the spectrum. The first, denoted as HP, uses the locations τ_j^{HP} corresponding to the iteration in the Markov chain that has highest posterior density, as an approximation to the posterior modal estimate of peak location. The second approach uses the posterior mean of the expected intensity and is based on estimating locations of local modes of the expected intensity, by solving for the set of TOFs such that

$$\frac{d}{dt}\mathbb{E}[\mu(t)] = 0 \tag{12}$$

where $\mu(t)$ is given in (9) and the expectation is taken with respect to the posterior distribution of parameters in $\mu(t)$. Interchanging expectation and taking derivatives, the posterior mean of the derivative may be estimated by the ergodic average of $\mu'(t)$ over the draws of θ from the Markov chain. Peak identification is carried out by finding where the posterior mean of the derivative process crosses zero. This typically results in fewer peaks than the

the HP draw, but identifies major peaks.

6 Simulation Study

In this section we describe a simulation study intended to explore how well our approach succeeds in locating true peaks within spectra of varying signal-to-noise ratios. We fit the model to simulated datasets for five values of the signal proportion: $S \in [0.10, 0.40, 0.70, 0.85, 0.95]$ (*i.e.*, signal-to-noise ratios of [0.1, 0.7, 2.3, 5.7, 19]). Twenty-five datasets were generated for each of these values of S , with fixed peak locations and with the remaining model parameters set to nominal values ($J = 35$, $\zeta = 130$, $\varphi = 0.50$, $(\alpha_J, \beta_J) = (1, 0.02)$, $\varrho = 56$, $\eta_j \equiv 3.9 \mu\text{s}$, $\eta_0 = 35.6 \mu\text{s}$, $\omega_0 = 46.2 \mu\text{s}$) chosen so that the simulated spectra appeared similar to observed spectra. With a few exceptions the hyperparameter and RJMCMC specifications remained the same as those described in Section 4. Departures include reducing μ_J (by half) to 50 and fixing the overall spectrum resolution at $\varrho_\mu = 50$. As starting values for the peak locations $\{\tau_j\}$ we took 50 equidistant peaks in the range $[32, 278] \mu\text{s}$, corresponding to the range $[5\,000, 67\,000] \text{m/z}$. We used 500 000 RJMCMC iterations, and retained the last 5 000 for further analysis. We compare our approach to a conventional peak finding algorithm using the R package `PROcess` [Li, 2005], which was originally designed for analysis of SELDI-TOF data, but is also applicable to MALDI-TOF. Li [2005] suggests removing background via subtracting a smoothed estimate of the spectrum local minima. Peaks are then defined as local maxima, with a signal greater than a user specified threshold and have a signal to noise ratio greater than a user specified value. Further more, an observation initially flagged as a peak may become disqualified if its area divided by the maximum peak area is less than a specified ratio. For both approaches, a discovered peak is regarded as a true peak if the TOF falls within $\pm 0.2\%$ of that of a true peak.

Figure 5 a,b. illustrates the fractions of true peaks found (TDR) and of spurious peak

discoveries (FDR), respectively, by our approach and by `PROcess` at several signal fractions S . Both the single highest-probability simulation outcome (thin solid lines) and the posterior mean (thick solid lines) are shown, along with two estimates from `PROcess` (dashed lines showing default program settings and dotted lines representing carefully tuned settings). Figure 5. shows that we find an average of 75% of the true peaks, with a FDR in the range of 13–56%, across the range of signal fractions for both model output summaries; the FDR using the posterior mean remains below 16% and the mean TDR above 83% for all signal-to-noise ratios above one. The simulation shows that the kernel based approach is superior in performance to `PROcess` over a range of scenarios.

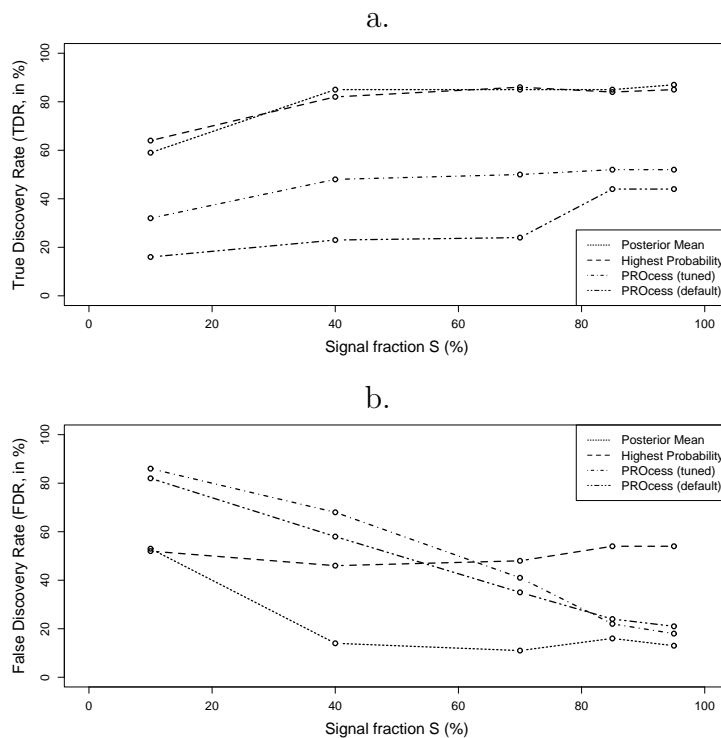


Figure 5: True and False Discovery Rates from simulation study. Note superiority of Posterior Mean (thick solid line) to Posterior Mode (thin solid line) and both `PROcess` approaches (dotted and dashed lines).

7 Examples

In this section, we apply our approach to three datasets (a blank spectrum, a matrix spectrum, and sample protein spectra) and compare the results to those produced by the peak-finding algorithm `PR0cess` [Li, 2005]. All data were generated using by a linear MALDI-TOF Delayed Extraction Mass Spectrometer, Applied Biosystems Voyager DE. The same prior distributions were used for the three datasets as in the simulation study (see Section 4), with the following exceptions. For the blank and matrix spectra, we took $\mu_J = 20$ (or $\beta_J = 0.05$), in anticipation of a small number of peaks. While for the protein samples, we set $\mu_J = 100$ (*i.e.*, $\beta_J = 0.01$). We restricted attention to the common range of 5–75 kDa/e for each dataset (TOF 30–280 μ s) and standardized intensities as in Equation (1).

7.1 Blank Spectrum

Figure 6. shows the recorded spectrum from the average of ten blank plate spectra based on 32 laser shots. Since no proteins are present there can be no protein signature, but nevertheless the spectrum shows low-resolution peaks that reflect fluctuations in the laser or resonances in the detector. The tick-marks on the horizontal axis represent peak locations identified as local modes of the posterior mean, shown as a solid curve.

The highest-posterior realization included $J = 27$ peaks, while the posterior mean (and standard deviation) was $\mathbf{E}[J] = 38.09(4.07)$. In this example `PR0cess` fits only a single peak, at 60 kDa/e. The posterior expected resolution was $\mathbf{E}[\varrho] = 20.06$, lower than the prior mean and (as we will see below) lower than the typical resolutions for protein peaks.

Sinusoidal patterns appear at two frequencies, found by Fourier techniques to be 10.66 kHz and 91 kHz (corresponding to periods of 94 μ s and 11 μ s). Periods and intensities may vary unpredictably across spectroscopic samples, so we chose not to use a sinusoidal filter for the background components but rather to allow the adaptive LARK model to identify

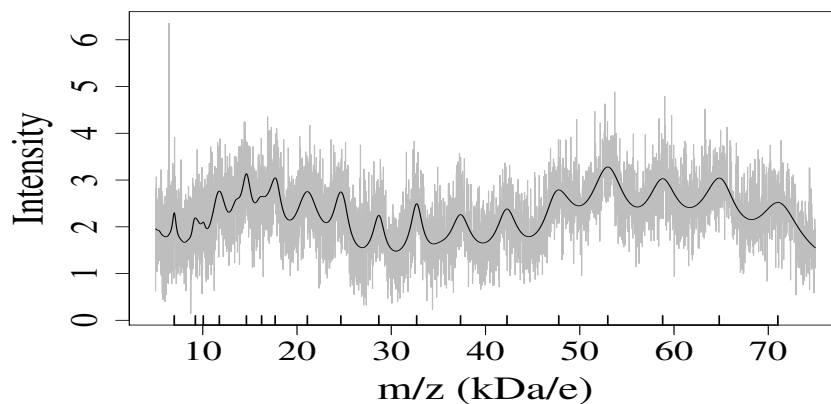


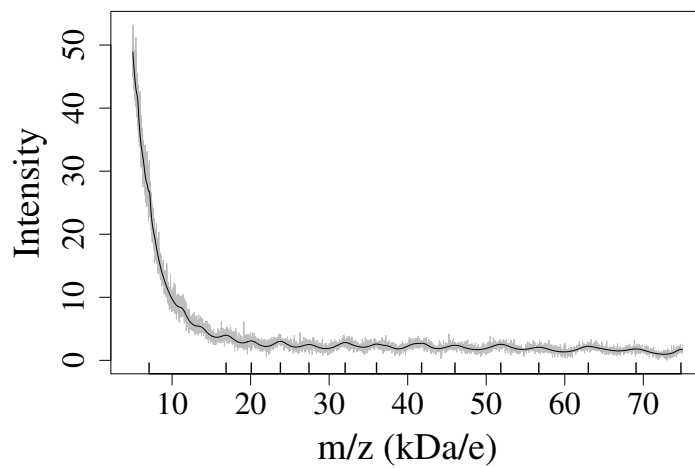
Figure 6: Blank spectrum from ten replicates, with posterior mean shown as a dark curve and identified peaks as tick-marks on the horizontal axis.

and discount these low-resolution peaks. While perhaps less parsimonious than a filtering approach, the LARK model offers more flexibility to accommodate the wide range of noises and distortions in background spectra.

7.2 Matrix Spectrum

Figure 7. shows the average of ten spectrum based on 32 laser shots each from a sinapaic matrix solution with no protein serum sample. Estimated peak locations identified by LARK (a.) and by PROCess (b.) are shown as tick marks on the horizontal axes; (a) also shows the posterior mean curve for LARK. The posterior mean under the LARK model shows the characteristic near-exponential spectral fall-off arising from the very low-molecular-weight sinapaic acid matrix ions (initial peak and a secondary peak at 7.1 kDa), as well as several low-resolution peaks from the detector seen before in the blank spectrum (see Section 7.1).

a.



b.

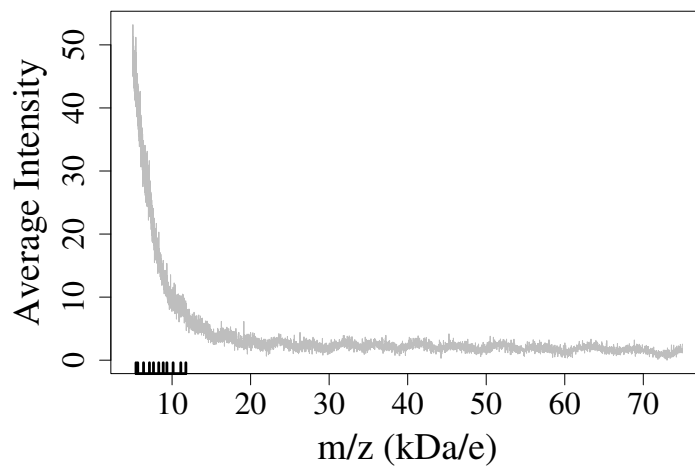


Figure 7: LARK (a.) and PROcess (b.) peak reconstructions for matrix spectra.

7.3 Protein Spectrum

The protein spectrum is based on data from an observational study at the Duke Medical Center Radiology Department [Campa *et al.*, 2003], intended to assess proteomic differences between cancerous and healthy patients. Serum from 30 diseased and control Caucasian males was collected and analyzed by a linear MALDI-TOF Delayed Extraction Mass Spectrometer, Applied Biosystems Voyager DE. Each sample was fractionated to 20 pH levels prior to the MALDI-TOF analysis to promote the ionization of a range of proteins. Ten replicate spectra were stored for each fraction, each from ten laser shots. For the present analysis we selected arbitrarily one fraction from one subject, from which we generated two datasets (see Figure 1.): a single spectrum from the chosen subject-fraction, and the mean spectrum of all ten replicates. The higher signal-to-noise ratio of the mean spectrum should make it better for supporting inference [Morris *et al.*, 2005], but the single spectrum offers an opportunity to show how the LARK model accommodates noisy data.

Table 1. Table 2. and Figure 8. display posterior parameter estimates and model fits, respectively, from the last 1 200 000 draws from 4.2 million RJMCMC iterations for the single and mean spectra. Three model estimates are given in Table 2. for the number of proteins found: the posterior mean J^{PM} , the single highest-probability value found in the simulation J^{HP} , and the number J^{∇} of local maxima found in the posterior mean $\mathbb{E}[\mu(t)]$ (see Equation (12)). The difference $J^{\text{PM}} - J^{\nabla}$ increases with decreasing resolution as multiple nearby peaks merge [Dass, 2001, p. 119]. Such peak merging would be appropriate for proteins with multiple isotopes, but it will lead to distortion when two or more similar-sized but distinct proteins appear as one. Figure 9. shows the difference between the two model fits for the mean spectrum at different TOF ranges.

Overall, we feel that our approach is far more satisfactory than algorithmic feature extraction methods such as that in PRO. When strong peaks are present, PROcess mistakes smaller peaks as noise, as is the case in Figure 8. where PROcess is only able to identify

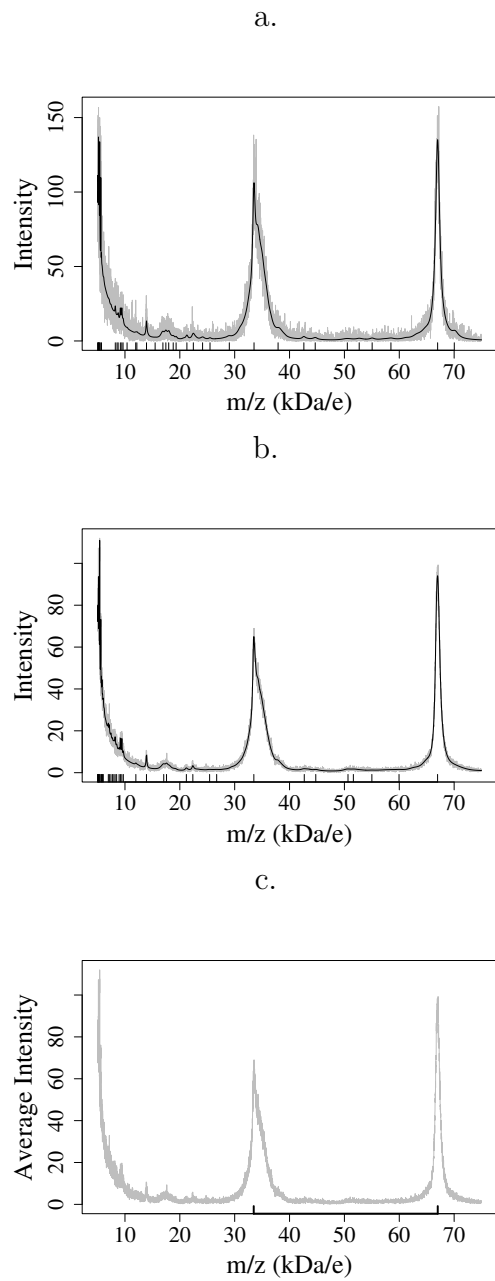


Figure 8: Peak reconstructions for the single spectrum using LARK (a) and the mean-spectra using LARK (b) and PR0cess(c).

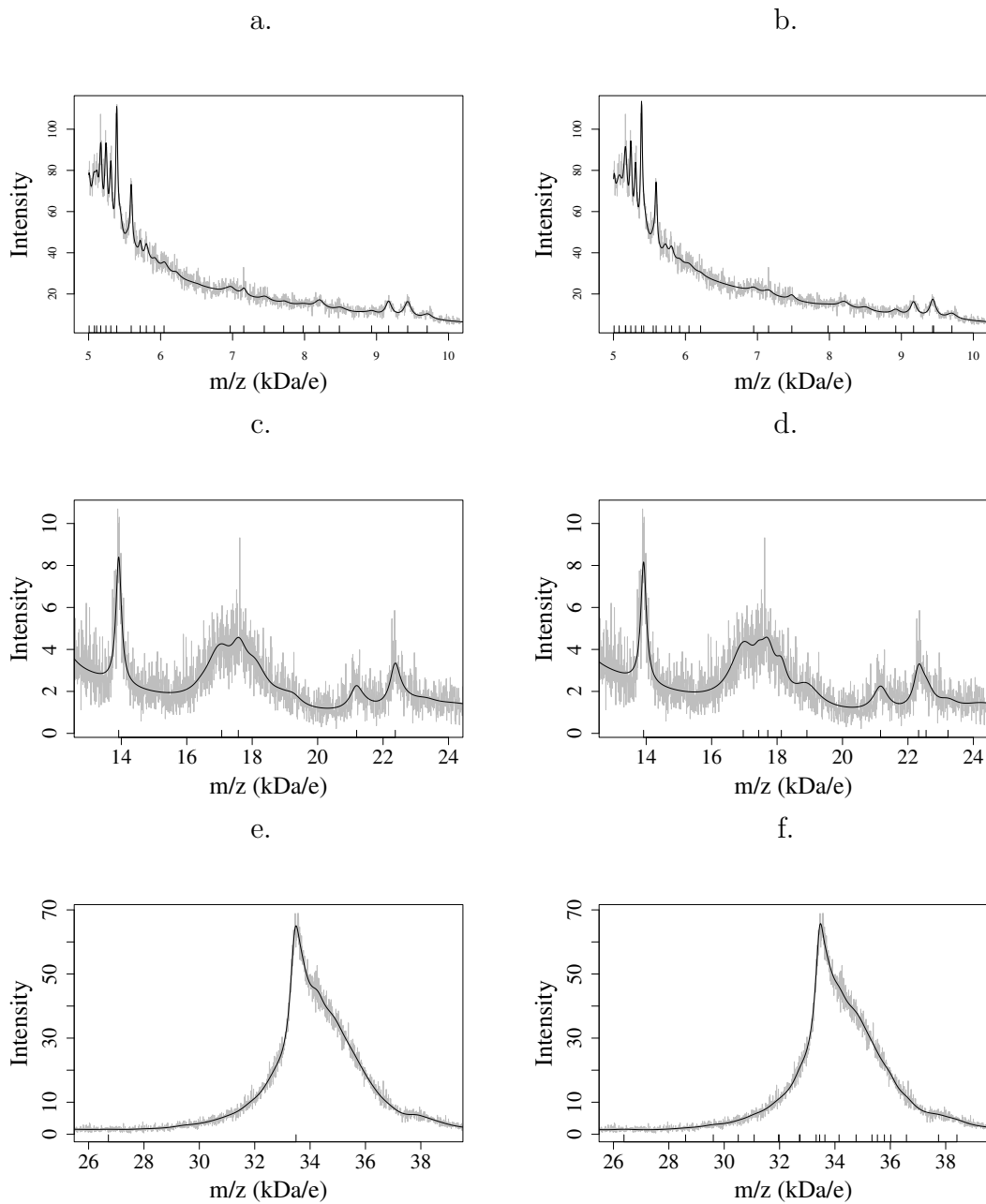


Figure 9: Local maxima from posterior mean (left) and highest probability model (right) estimates for various regions in mean spectra.

Table 1: Posterior Model Parameter Estimates (and Standard Deviations) for the Blank, Matrix, Single Spectrum and Mean Spectrum examples.

Dataset	S	φ	ϱ	η_0	ω_0	J^{PM}
Blank	0.397 (0.074)	7.13 (0.13)	20.06 (2.21)	5.12 (2.11)	30.98 (12.14)	38.09 (4.07)
Matrix	0.026 (0.014)	8.84 (0.17)	18.42 (1.45)	33.13 (0.45)	17.52 (0.18)	27.51 (0.98)
Single	0.005 (0.003)	0.33 (0.01)	62.75 (4.64)	92.77 (2.08)	18.40 (0.51)	69.60 (6.18)
Mean	0.036 (0.008)	4.96 (0.08)	56.47 (3.24)	96.10 (1.96)	17.88 (0.38)	86.96 (4.64)

Table 2: Number of peaks extracted by LARK and `Process`. Three model summaries are provided: the posterior mean J^{PM} , the single highest-probability value found in the simulation J^{HP} , and the number J^{∇} of local maxima.

Dataset	J^{PM}	J^{HP}	J^{∇}	J^{PRO}
Blank	38.09	27	18	1
Matrix	27.51	25	14	11
Single	69.60	55	39	2
Mean	86.96	76	39	2

the two large peaks in the mean spectrum (shown) and the single spectrum (not shown). Without knowledge of the true protein distribution in the serum samples, we cannot know if we are over- or under- estimating the number of proteins. We do see that the method finds more peaks in the mean spectrum than in the single spectrum (Table 2). The difference between the mean and single spectra is not surprising since the posterior estimates for φ are different (Table 1). The reduced level of noise that results from averaging several spectra enables the detection of small peaks [Morris *et al.*, 2005].

8 Discussion

Our novel LARK approach provides estimates of quantities of interest to experimenters, such as protein masses and abundances. From posterior samples we have constructed point estimates for the number of proteins, protein mass, and protein abundance, using posterior means or the sample with the highest posterior density (an approximation to the posterior mode). The Bayesian model-based approach can also be used to provide measures of uncertainty for any of these quantities. We have demonstrated through simulation studies that the procedure has desirable true and false discovery rates. Incorporating information about resolution and peak shape leads to improved peak detection.

While one might expect background levels to be constant over time, in fact our analyses have revealed a systematic variation: a nearly periodic series of small peaks, possibly due to thermal ringing in the detector. While it is potentially difficult to distinguish these peaks from proteins, in fact their lower resolution allows the user to differentiate artefactual background peaks from true protein peaks. The LARK approach makes it simple for a user to accommodate such features.

The methods developed in this work for a single spectrum may be extended to hierarchical models for multiple spectra from different subjects or multiple subjects within groups. The analysis of multiple spectra is complicated by misalignment, due to the variability of TOFs across shots within the same subject or experimental conditions. Averaging spectra across shots for the same subject may lead to additional difficulties, such as the broadening of peaks or possible loss of small peaks. Automatic calibration and alignment of spectra may be achieved by a hierarchical model that allows the TOF parameters to vary from shot to shot, but remain centered at subject-specific expected TOFs. The hierarchical version of our single spectrum model can be extended to identify peaks with differential abundance or presence/absence across experimental conditions, as in the functional data analysis approach

of [Morris et al. \[2006\]](#).

Acknowledgments

The authors would like to thank Michael J. Campa, Michael C. Fitzgerald, Edward Patz, Jr., and Petra L. Roulhac for providing the data used in our examples and for many helpful conversations. This work was supported by the National Science Foundation under Grant Number DMS-0342172, DMS-0422400 and DMS-0406115. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

- C. Dass. *Principles and Practice of Biological Mass Spectrometry*. John Wiley & Sons, 2001.
- K. R. Coombes, J. M. Koomen, K. A. Baggerly, J. S. Morris, and R. Kobayashi. “Understanding the Characteristics of Mass Spectrometry Data Through the Use of Simulation.” *Cancer Informatics*, vol. 1(1), pp. 41–52, 2005a.
- K. A. Baggerly, J. S. Morris, and K. R. Coombes. “Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments.” *Bioinformatics*, vol. 20(5), pp. 777–785, 2004.
- K.-A. Do, P. Müller, and M. Vannucci, eds. *Bayesian Inference for Gene Expression and Proteomics*. Cambridge University Press, Cambridge, UK, 2006.
- J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi. “Feature Extraction and Quantification for Mass Spectrometry in Biomedical Applications Using Mean Spectrum.” *Bioinformatics*, vol. 21(9), pp. 1764–1775, 2005.

- R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q.-T. Le. “Sample Classification from Protein Mass spectrometry, by ‘peak probability contrasts’.” *Bioinformatics*, vol. 20(17), pp. 3034–3044, 2004.
- Y. Yasui, D. McLerran, B.-L. Adam, M. Winget, M. Thornquist, and Z. Feng. “An Automated Peak Identification/Calibration Procedure for High Dimensional Protein Measures From Mass Spectrometers.” *Journal of Biomedicine and Biotechnology*, vol. 4, pp. 242–248, 2003.
- M. Carpenter, S. Melath, S. Zhang, and W. E. Grizzle. “Statistical Process and Analysis of Proteomic and Genomic Data.” In “Proceedings of the Pharmaceutical SAS Users Group, Miami, FL,” pp. 545–548. 2003.
- K. R. Coombes, S. Tsavachidis, J. S. Morris, K. A. Baggerly, M.-C. Hung, and H. M. Kuerer. “Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform.” *Proteomics*, vol. 5(16), pp. 4107–4117, 2005b.
- C. Tu, M. A. Clyde, and R. L. Wolpert. “Lévy Adaptive Regression Kernels.” Discussion Paper 2006-08, Duke University ISDS, 2006.
- L. L. House. “Nonparametric Bayesian Models in Expression Proteomic Applications.” Ph.D. dissertation, Duke University, Durham, NC, 2006. URL <http://stat.duke.edu/people/theses/leanna.pdf>.
- M. A. Clyde and R. L. Wolpert. “Nonparametric Function Estimation using Overcomplete Dictionaries.” In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., “Bayesian Statistics 8,” p. In Press. Oxford University Press, Oxford, UK, 2006.

- X. Li. *PROcess*, 2005. Version 2.10 R package, <http://cran.r-project.org>.
- L. V. Zhigilei and B. J. Garrison. “Velocity Distributions of Analyte Molecules in Matrix Assisted Laser Desorption from Computer Simulations.” *Rapid Communications in Mass Spectrometry*, vol. 12, pp. 1273–1277, 1998.
- J. Franzen. “Improved Resolution for MALDI-TOF Mass Spectrometers: A Mathematical Study.” *International Journal of Mass Spectrometry and Ion Processes*, vol. 164(1), pp. 19–34, 1997.
- M. Kempka, J. Södahl, A. Björk, and J. Roeraade. “Improved Method for Peak Picking in Matrix-Assisted Laser Desorption/Ionization Time-of-Flight Mass Spectrometry.” *Rapid Communications in Mass Spectrometry*, vol. 18(11), pp. 1208–1212, 2004.
- D. I. Malyarenko, W. E. Cooke, B.-L. Adam, G. Malik, H. Chen, E. R. Tracy, M. W. Trosset, M. Sasinowski, O. J. Semmes, and D. M. Manos. “Enhancement of Sensitivity and Resolution of Surface-Enhanced Laser Desorption/Ionization Time-of-Flight Mass Spectrometric records for serum peptides using time-series analysis techniques.” *Clinical Chemistry*, vol. 51(1), pp. 65–74, 2005.
- Applied Biosystems. *Voyager Biospectrometry Workstation with Delayed Extraction Technology User Guide Version 5.1*. Applied Biosystems, Foster City, CA, 2001. URL <http://docs.appliedbiosystems.com/pebiiodocs/04317707.pdf>.
- G. Siuzdak. *The Expanding Role of Mass Spectrometry in Biotechnology*. MCC Press, San Diego, CA, 2003.
- A. N. Tikhonov. “Solution of incorrectly formulated problems and the regularization method.” *Soviet Mathematics Doklady*, vol. 4, pp. 1035–1038, 1963. English translation of *Doklady Akademii Nauk SSSR*, 151(3), 501–504.

- M. J. Campa, M. Z. Wang, B. A. Howard, M. C. Fitzgerald, and E. F. Patz, Jr. “Protein Expression Profiling Identifies MIF and Cyclophilin A as Potential Molecular Targets in Non-Small Cell Lung Cancer.” *Cancer Research*, vol. 63(7), pp. 1652–1656, 2003.
- M. Abramowitz and I. A. Stegun, eds. *Handbook of Mathematical Functions With Formulas, Graphs, and Mathematical Tables*, vol. 55 of *Applied Mathematics Series*. National Bureau of Standards, Washington, D.C., 1964.
- Wolfram Research, Inc. *Mathematica*. Wolfram Research, Inc., Champaign, IL, 5th edn., 2005.
- M. B. Monagan, K. O. Geddes, K. M. Heal, G. Labahn, S. M. Vorkoetter, J. McCarron, and P. DeMarco. *Maple 10 Programming Guide*. Maplesoft, Waterloo ON, Canada, 2005.
- A. Y. Khinchine and P. Lévy. “Sur les lois stables.” *Comptes rendus hebdomadaires des seances de l’Académie des sciences. Académie des science (France). Serie A. Paris*, vol. 202, pp. 374–376, 1936.
- L. C. G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales*, vol. 1. John Wiley & Sons, New York, NY, 2nd edn., 1994.
- J. Besag, P. J. Green, D. Higdon, and K. Mengersen. “Bayesian computation and stochastic systems (with discussion).” *Statistical Science*, vol. 10(1), pp. 3–66, 1995.
- L. Tierney. “Markov chains for exploring posterior distributions (with discussion).” *Annals of Statistics*, vol. 22(4), pp. 1701–1762, 1994.
- A. E. Gelfand and A. F. M. Smith. “Sampling-based approaches to calculating marginal densities.” *Journal of the American Statistical Association*, vol. 85(410), pp. 398–409, 1990.

- P. J. Green. “Reversible jump Markov chain Monte Carlo computation and Bayesian model determination.” *Biometrika*, vol. 82(4), pp. 711–732, 1995.
- R. L. Wolpert and K. Ickstadt. “Reflecting Uncertainty in Inverse Problems: A Bayesian Solution using Lévy Processes.” *Inverse Problems*, vol. 20(6), pp. 1759–1771, 2004.
- A. P. Dempster, N. M. Laird, and D. B. Rubin. “Maximum likelihood from incomplete data via the EM algorithm (with discussion).” *Journal of the Royal Statistical Society, Ser. B: Statistical Methodology*, vol. 39(1), pp. 1–38, 1977.
- J. S. Morris, P. J. Brown, K. A. Baggerly, and K. R. Coombes. “Analysis of Mass Spectrometry Data Using Bayesian Wavelet-Based Functional Mixed Models.” In [Do et al. \[2006\]](#), chap. 14, pp. 269–292.