Statistics and Its Interface Volume 2 (2009) 329–340

Alignment of protein mass spectrometry data by integrated Markov chain shifting method

Yang Feng, Weiping Ma, Zhanfeng Wang, Zhiliang Ying*, and Yaning Yang

Mass spectrometers such as SELDI-TOF (surface enhanced laser desorption/ionization time-of-flight) and MALDI-TOF (matrix assisted laser desorption and ionization time-of-flight) measure the relative abundance of different protein ions or protein fragments (peptides) indexed by the mass-to-charge ratio (m/z). A special characteristic of the MS spectra is its variabilities in both m/z values and intensity magnitudes. We propose modelling the log-intensities by a semiparametric model and the m/z by the integrated Markov chain shifting (IMS) model, for which the second-order differences of the random effects are assumed to follow a second-order Markov chain. Alignment of spectra is done through averaging over the random shifts conditional on the observed intensity information. The unknown parameters are estimated by an iterative nonparametric maximum profile likelihood method and a Gaussian kernel approximation. The bandwidths in kernel approximation are taken to be 0.04%–0.08% of the m/z values. Simulation results show that the proposed approach can achieve satisfactory alignment by reducing the intensity variations of the misalignment spectra by a factor of around 75%. Most alignment algorithms align spectra by clustering neighboring peaks and do not incorporate peak height information. Our semiparametric random shifting method builds a model taking into consideration of both the random shift effects of neighboring m/z values and similarity of the intensity magnitudes of common peaks within the ranges of about 50% of the intensity values.

AMS 2000 subject classifications: Primary 62P10, 62G99; secondary 92F05.
Keywords and phrases: MS spectra, semiparametric model, markov chain, integrated markov chain shifting, profile likelihood.

1. BACKGROUND

The field of proteomics is an evolving area, which may shed light on the proteins associated with diseases and tumors. A powerful tool for proteomic profiling is the mass spectrometer (MS). It produces high throughput relative abundance of protein ions or peptides (fragment of proteins) of the studied sample, in the form of spectrum. Screening multiple MS spectra can provide insight into the protein ion composition of different samples under different conditions.

Mass spectrometers such as SELDI-TOF (surface enhanced laser desorption/ionization time-of-flight) and MALDI-TOF (matrix assisted laser desorption and ionization time-of-flight) measure the relative abundance of different protein ions or peptides indexed by the mass-to-charge ratio (m/z). MS spectra can reveal proteomic patterns or features (e.g. biomarker ions) which may be related to specific characteristics of biological samples and can be used for diagnostic purposes. They can also be used for prognosis and for monitoring disease progression and evaluating treatment or intervention. Other applications of MS spectrum include pharmaceutical analysis, biomolecule characterization, environmental assessment and forensic analysis.

High throughput MS spectra data have raised challenging issues such as data preprocessing, baseline correction and peak alignment, among others (Baggerly et al., 2003, 2004; Diamandis, 2004). An outstanding feature of a typical MS spectrum is that both the x- and y-axes have measurement variabilities (Fig. 1). The y-axis of a spectrum is the intensity (relative abundance) of protein/peptide, the x-axis the mass-to-charge ratio. It is known that the SELDI intensity measures have errors up to 50% and that the m/z may shift its value by up to 0.1%–0.2% (Yasui et al., 2003). The variability of the intensity is largely due to measurement errors, which can be modelled by adding an additive noise term. On the other hand, the shift of m/z may cause migration of important features, resulting in, among other things, difficulties in data analysis. To characterize the MS spectrum of a sample from a population or to compare spectral patterns under different conditions, it is essential to identify the same informative peaks or signals by shifting or aligning the m/z values properly. Alignment by internal and/or external calibration is usually a first step for aligning MS spectra. This is done by aligning one or several reference peptides, known as the landmark registration and implemented in standard MS softwares (Guilhaus, 1995). However, more delicate alignment is needed after the first round calibration due to at least the following three reasons. First, reference signal(s) of known peptides may not be accurately spotted; second, alignment of the reference peptides or protein ions does not

*Corresponding author.

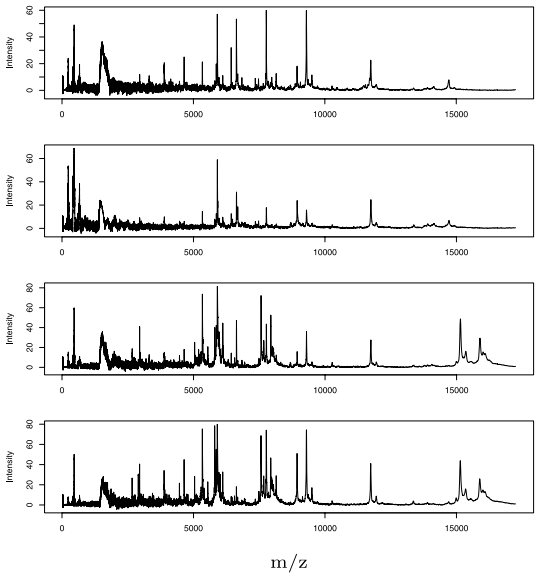

Figure 1. Illustration of MS spectra.

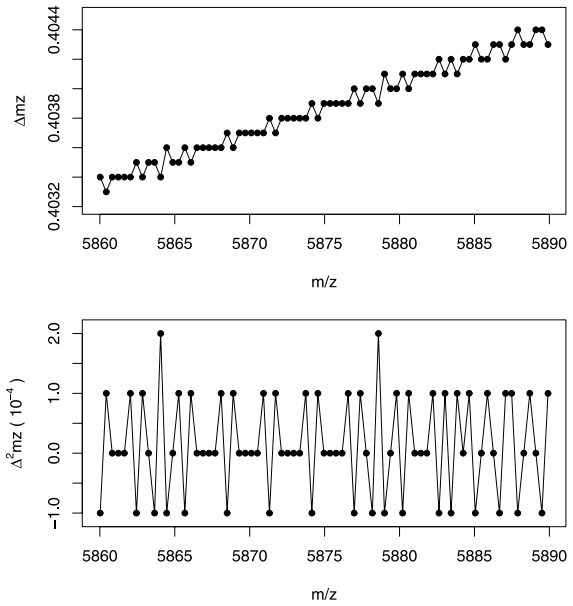

Figure 2. First and second order differences of 75 consecutive m/z values.

necessarily imply that other signals are properly aligned; third, the shifting of m/z is random and cannot be decided in a deterministic way. In this connection, several alignment algorithms have been proposed in the literature. They rely heavily on locating peaks in a given spectrum and align the spectra accordingly (Yasui et al., 2003; Tibshirani et al., 2004; Jeffries, 2005; Wolski et al., 2005).

In functional data analysis in areas such as image processing and speech recognition, alignment of several curves with different time scales via suitable transformation is called curve registration (Ramsay and Li, 1998; Ramsay and Silverman, 2005). The transformation function for time or phase is called a warping function. There are many warping methods for functional data registration. Landmark registration (Kneip and Gasser, 1992) uses internal/external references (landmarks) as the guide for warping. The self-modelling method is an automatic alignment procedure without using such landmark registration (Lawton et al., 1972; Ramsay and Silverman, 2005; Ronn, 2001; Gervini and Gasser, 2004; Ramsay and Li, 1998; Wang and Gasser, 1999). Other methods include shape invariant modelling (Kneip and Engel, 1995), local regression (Kneip et al., 2000), among others. All the above mentioned methods assume that the shifting effect is deterministic.

TOF MS spectra demonstrate some special characteristics. Ideally the m/z value of a TOF MS spectrum is a quadratic function of the time of flight (Jeffries, 2005; Guilhaus, 1995). However, we found that the observed m/z values deviate from these theoretical values randomly, which can not be properly captured by a few deterministic parameters (fixed-effects), and on average the observed m/z values tend to shift to the right. In this paper, we propose modelling the TOF MS spectra data by a random shift model to incorporate the randomness of the shift effect. It contains a random effect model on the m/z shifts, the second-order difference of random effect is a second-order Markov chain, for which the current state depends on previous states only through the last two states. It also contains a nonparametric part for the log-intensity curves, representing the true spectrum. An important advantage of the random shift model is its flexibility and efficiency. It models the errors in m/z values by many small random shift effects. These random variables are averaged out through marginalization, resulting in only a few parameters for m/z shifts. Therefore the complexity and delicate structure can be captured appropriately by these random effects without losing efficiency. Random shifting method using normal random effects in functional data analysis can be found in Ronn (2001), Brumback and Lindstrom (2004).

2. METHODS

The second-order Markov chain property for the second-order m/z differences is a key to our method. Exploratory analysis of several SELDI-TOF MS data sets produced from Ciphergen instrument (Biosystems, Inc.) shows that the m/z spacings (difference of two consecutive m/z values) deviate randomly from linearity discretely, jumping up and down with only a few states. Figure 2 illustrates the first-order and second-order differences of m/z from a typical spectrum. The first-order differences have an overall linear trend, compatible with the quadratic property of theoretical m/z values as a function of TOF (Guilhaus, 1995; Jeffries, 2005). The slight oscillations can be seen clearly from the plot of second-order differences. Table 1 illustrates the frequencies

Table 1. Distribution of random jumps of $\Delta^2(m/z)$

Jump ($\times 10^{-4}$Da)	-1	0	1	2
Frequency	0.186	0.494	0.315	0.005

of the four states of jumps in one such case we examined (m/z values between 2,000–10,000Da). This distribution differs slightly for different spectra.

Further examination of the second-order differences shows that transition from one state to another presents a second-order Markovian property, i.e, the current jump depends on the previous two. For example, at the scale of 10^{-4}Da, if the previous two jumps are 0, then the next one could be -1, 0, 1 with probabilities 0.003, 0.642 and 0.355; if the previous two jumps are 0 and $+1$, then the next one could be -1 and 0 with probability 0.744 and 0.256 respectively. Patterns of (0,1,1) or (0,1,2) never occurs for three consecutive jumps. In summary, second-order difference of m/z is a constant if there are no errors, the observed m/z deviate from this constant by a random discrete jump, which takes only a few distinct values and all the jumps form a second-order Markov chain. The cumulative effect of these small random errors of the second-order differences results in random shifting of m/z measurements.

Based on these findings, we propose to align multiple SELDI MS spectra data by a semiparametric random effect model, and the random effects are the integrated Markov chain of random jumps on the second-order differences. The likelihood for the spectra is the marginal likelihood of observed intensities and m/z values with the random effects being integrated out. We develop an iterative nonparametric maximum profile likelihood method for solving the interested parameters and a Monte Carlo simulation method for calculating the conditional expectations. The nonparametric maximum likelihood estimates (NPMLE) of the unknown parameters turn out to be related to expectations with respect to conditional distribution of the random shift effects given the observed intensities. For simplicity, we use one-sample and multi-sample cases as illustrations to our method.

2.1 The one-sample integrated Markov chain shifting model

As demonstrated in the previous section, there exist random shifts in MS spectra. Consequently, cross-sectional means of multiple spectra can be very misleading and alignment of spectral curves is crucial for analysis of protein mass spectrometry data. In the sequel, we assume preprocessing such as normalization, baseline subtraction and external calibration had been performed with the MS data.

Let t denote the TOF, the time before the particle hits the detector, and let $x_i(t)$ and $y_i(t)$ be respectively the m/z and log-intensity function for the i-th spectrum. Let $z_i(t) = \frac{1}{2}a_i t^2 + b_i t + c_i$, $t \geq t_0$, be the true m/z with second-order difference $\Delta^2 z_i(t) = a_i$, a constant that is not dependent on time t. Note that we allow different spectra to have different mean values of second-order differences a_i's, and therefore different b_i or c_i's. The discrepancy of these coefficients causes different systematic drifts in true m/z values. Our algorithm below can automatically account for this kind of machine drifts besides the random shifts.

Recall that the second-order differences of observed m/z values deviate from the mean value by a discrete error term, denoted by $\delta_i(t)$, with second-order Markovian property. We therefore assume the following model for the observed m/z (the x-axis of a spectrum)

$$\Delta^2 x_i(t) = a_i + \delta_i(t), \quad E(\delta_i(t)) = 0. \tag{1}$$

Equivalently we assume the following random shift model

$$x_i(t) = z_i(t) + s_i(t), \tag{2}$$

where the random shift term has the representation

$$s_i(t) = \int_0^t \int_0^v \delta_i(u)\,du\,dv, \tag{3}$$

with $\delta_i(t)$ being a mean 0 discrete error term satisfying second-order Markovian property, $P(\delta_i(t_k) = x|\delta_i(t_j), t_j < t_k) = P(\delta_i(t_k) = x|\delta_i(t_{k-1}), \delta_i(t_{k-2}))$. We call the random shift effect $s_i(t)$ an integrated Markov chain shifting (IMS) process.

Suppose there are n individual MS curves from a sample or group that share the same shape function $m(\cdot)$. For the log-intensities, we assume the following semiparametric model

$$y_i(t) = \alpha_i + \beta_i m(z_i(t)) + \epsilon_i(t), \quad i = 1, \ldots, n, \tag{4}$$

where $m(\cdot)$ is the true spectrum, $\epsilon_i(t)$ is the random error of log-intensity and we assume $\epsilon_i(t)$'s are independent and follow normal distribution $N(0, \sigma^2)$. Parameters α_i and β_i are respectively the individual magnitude and scale effect for spectrum i. This model describes a functional relationship between two processed y_i and x_i.

Combining equations (2) and (4), we have the following random shift model

$$y_i(t) = \alpha_i + \beta_i m(x_i(t) - s_i(t)) + \epsilon_i(t). \tag{5}$$

Notice here that the β_i and α_i are not fully identifiable for all i. So we must designate a baseline for them, such as $\beta_1 = 1$, $\alpha_1 = 0$.

In deriving the estimates of the unknown quantities in this model, we use a nonparametric profile maximum likelihood method, in which we use a Gaussian kernel approximation method to estimate the shape function $m(\cdot)$. A bandwidth parameter, h, in the kernel approximation determines the smoothness of the curve fitting. To incorporate the m/z

variabilities, we recommend choices in the range of 0.04%–0.08% of the m/z values for bandwidths. This choice can effectively keep the peak signals of the curves (see Section 2.3 for details).

2.2 The multi-sample Integrated Markov chain shifting model

A more important issue in proteomic study is to compare proteomic structures under different conditions. The one-sample semiparametric model can be extended to incorporate multiple samples (groups) by introducing different shape functions for different samples. Suppose there are G different samples, and there are n_g individual MS curves in sample g. Let the shape function for sample g be $m_g(\cdot)$ and assume the following semiparametric model for observed curves

$$y_{gi}(t) = \alpha_{gi} + \beta_{gi} m_g(x_{gi}(t) - s_{gi}(t)) + \epsilon_{gi} \tag{6}$$

where $\epsilon_{gi} \sim N(0, \sigma^2)$ are independent normal errors, $x_{gi}(t)$ the observed m/z values and $s_{gi}(t)$ the unobserved random shift effects at time t, where $g = 1, \ldots, G$ and $i = 1, 2, \ldots, n_g$. The random shift is the double integral of a second-order Markov chain as described above. In order for the model to be identifiable, assume $\alpha_{g1} = 0$ and $\beta_{g1} = 1$. Algorithm for fitting this model is similar to the one-sample case (see Methods).

2.3 Algorithm

We describe our algorithm for the one-sample model (5). The algorithm for multi-sample model (6) remains almost the same (see the Remarks at the end of this section). For the one-sample model (5), we develop a nonparametric profile likelihood method (Ronn, 2001) to estimate parameters θ and unknown shape function $m(t)$. It has three important ingredients: (i) an iterative algorithm, alternating between estimation of the parametric component θ and the nonparametric component $m(\cdot)$; (ii) kernel smoothing for dimension reduction for the (nonparametric) shape function $m(\cdot)$; and (iii) integration via Monte Carlo importance sampling. For given $m(\cdot)$, θ is estimated using Monte Carlo method. For given θ and posterior probability function of the shift variables given observations, the shape function is estimated by nonparametric maximum likelihood method, resulting in a kernel-like estimate, with the "kernel" estimated from data. Let $\mathbf{y}_i = (y_i(u), u = 1, \ldots, N)$, $\mathbf{x}_i = (x_i(u), u = 1, \ldots, N)$, $m(\mathbf{x}_i - \mathbf{s}_i) = (m(x_i(u) - s_i(u)), u = 1, \ldots, N)$, and $\bar{y}_i = \sum_{u=1}^N y_i(u)/N$, $\bar{m}_i = E_{\mathbf{s}_i|\mathbf{y}_i}[\sum_{u=1}^N m(x_i(u) - s_i(u))/N]$. Let $f_{\mathbf{s}_i|\mathbf{y}_i}(\mathbf{s_i})$ be the conditional probability function of $\mathbf{s}_i$ given $\mathbf{y}_i$. Using the nonparametric maximum profile likelihood method and the kernel approximation (Supplementary Notes), we have the following algorithm:

(a) Initialize estimate $m^{(0)}$ by the cross-sectional mean.

(b) Given $m^{(k-1)}$, $k = 1, 2, \ldots$, calculate

$$\hat{\beta}_i^{(k)} = \frac{\sum_{u=1}^N E_{s_i(u)|\mathbf{y}_i}[m^{(k-1)}(x_i(u) - s_i(u)) - \bar{m}_i^{(k-1)}]y_i(u)}{\sum_{u=1}^N E_{s_i(u)|\mathbf{y}_i}[m^{(k-1)}(x_i(u) - s_i(u)) - \bar{m}_i^{(k-1)}]^2},$$

$$\hat{\alpha}_i^{(k)} = \bar{y}_i - \bar{m}_i^{(k-1)} \hat{\beta}_i^{(k)},$$

$$\hat{\beta}_i^{(k)} = \hat{\beta}_i^{(k)} / \hat{\beta}_1^{(k)},$$

$$\hat{\alpha}_i^{(k)} = \hat{\alpha}_i^{(k)} - \hat{\alpha}_1^{(k)},$$

$$\hat{\sigma}^{(k)2} = \frac{\sum_{i,u} E_{s_i(u)|\mathbf{y}_i}(y_i(u) - \alpha_i^{(k)} - \beta_i^{(k)} m^{(k-1)}(x_i(u) - s_i(u)))^2}{nN}.$$

(c) For a pre-specified kernel function $K(\cdot)$, bandwidth h, estimated values of $\beta_i^{(k)}$, $\alpha_i^{(k)}$, and $\sigma^{2(k)}$, estimate $m(\cdot)$ by

$$\hat{m}^{(k)}(t) = \frac{\sum_{i=1}^n \sum_{u=1}^N \beta_i^{(k)}(y_i(u) - \alpha_i^{(k)}) E_{\mathbf{s}_i|\mathbf{y}_i} K\left(\frac{s_i(u) - (x_i(u) - t)}{h}\right)}{\sum_{i=1}^n \sum_{u=1}^N \beta_i^{(k)2} E_{\mathbf{s}_i|\mathbf{y}_i} K\left(\frac{s_i(u) - (x_i(u) - t)}{h}\right)} + \frac{\sum_{i=1}^n \sum_{u=1}^N \beta_i^{(k)2} E_{\mathbf{s}_i|\mathbf{y}_i}\left\{\Delta_i^{(k-1)}(t,u) K\left(\frac{s_i(u) - (x_i(u) - t)}{h}\right)\right\}}{\sum_{i=1}^n \sum_{u=1}^N \beta_i^{(k)2} E_{\mathbf{s}_i|\mathbf{y}_i} K\left(\frac{s_i(u) - (x_i(u) - t)}{h}\right)},$$

where $\Delta_i^{(k-1)}(t,u) = \hat{m}^{(k-1)}(t) - \hat{m}^{(k-1)}(x_i(u) - s_i(u))$.

(d) Repeat steps (b) and (c) until $\hat{\beta}_i^{(k)}$, $\hat{\alpha}_i^{(k)}$, $\hat{\sigma}^{(k)}$ and $\hat{m}^{(k)}$ converge.

Remarks: The NPMLE algorithm for multi-sample model (6) is basically the same as that for the one-sample model. The only differences are that the updating formula for m_g in the original algorithm should take summation over observations $i = 1, 2, \ldots, n_g$, and in each iteration α_{g1}, β_{g1}, $g = 1, 2, \ldots, G$, should be normalized.

In steps (b) and (c) of the algorithm, for a specific function ψ, one can approximate the conditional expectation

$$E_{\mathbf{s}_i|\mathbf{y}_i}\psi(\mathbf{s}_i) = \int \left(\frac{\psi(\mathbf{s}_i) f_{\mathbf{y}_i|\mathbf{s}_i}(\mathbf{y}_i)}{f_{\mathbf{y}_i}(\mathbf{y}_i)}\right) f_{\mathbf{s}_i}(\mathbf{s}_i)$$

by Monte Carlo importance sampling method. Specifically, one can approximate this integral by generating random numbers repeatedly from $\mathbf{s}_i$ instead of directly from $f_{\mathbf{s}_i|\mathbf{y}_i}$ and average over the quantities in the bracket.

We found that the importance sampling method is advantageous in reducing the amount of computational resources. We use the marginal distribution of the random shift as the proposal distribution instead of drawing random samples

from target distribution $f_{\mathbf{s}_i|\mathbf{y}_i}$, for which a Gibbs sampler is preferred in updating various parameters and samples repeatedly. The advantage of using this proposal is that the distributions of random effects do not depend on the unknown parameters. The resulting Monte Carlo simulation of the conditional means can be relatively stable than updating the proposed distribution iteratively.

The bandwidth specifies the number of neighboring m/z values around a point that are used for local smoothing. A suitable choice for the bandwidth should take into consideration of the trade-off between smoothness and accuracy. For example, if h is large, the fitted spectrum $m(\cdot)$ will be smooth and the accuracy will be low; on the other hand if it is small, then the fitted curve lacks smoothness and has high accuracy of retaining the original signals. Since it is desirable for an alignment algorithm to keep the intensity signals, we suggest using relatively small bandwidths that are comparable to the common choice of 0.1%–0.2% sliding window method. When a Gaussian kernel function is used, the 99% confidence interval is treated as a window which has width of $2 \times 2.576h$. Therefore a suitable choice of bandwidth analogous to the 0.1%–0.2% sliding window method is $h = 0.001/2.576 \approx 0.04\%$ to $0.002/2.576 \approx 0.08\%$ of the m/z values.

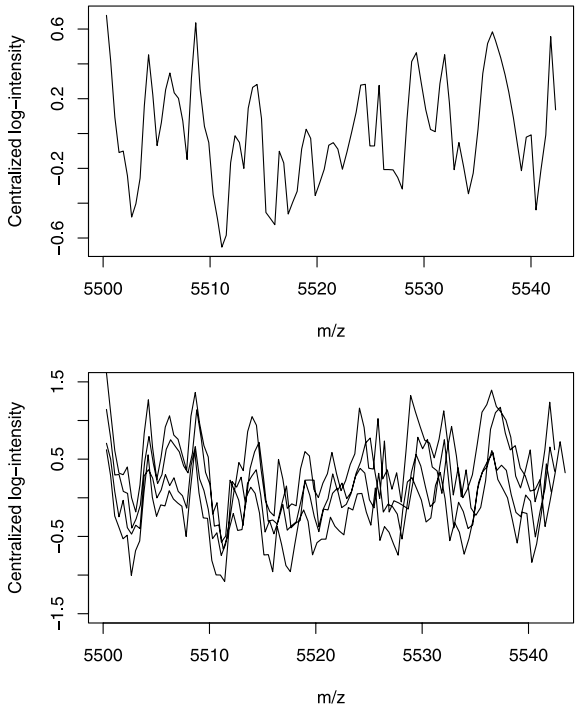

Figure 3. True curve (top row) and simulated curves (bottom row).

3. RESULTS

3.1 One-sample case

Extensive simulations are conducted to check performance of the proposed method. Special attention is being paid to the unbiasedness and accuracy of the estimates of parameters α, β and function $m(\cdot)$. We first generate a reference spectrum, mimicking the underlying true MS spectrum. We then generate n random copies from model (5). Estimates of α, β and $m(\cdot)$ are obtained using the proposed algorithm in the last section.

To run a large number of simulation replicates, the computational burden becomes prohibitively heavy when the number of m/z points is also large. To ease the computational intensity, we choose $N = 100$ m/z points. The reference spectrum (true spectrum) contains many spikes and is shown in Fig. 3. From this, all simulated spectrum curves $y_i(t)$ are generated by including random shifting effects in m/z (x-axis of a spectrum) and random errors for the intensities (y-axis). Since smaller bandwidth in the kernel function is usually preferred in order to keep the original signals, we use 0.04%–0.08% bandwidths in our simulations.

We take the sample size (number of spectral curves) $n = 20$. The true values of the parameters α_i, β_i, $i = 1, 2, \ldots, n$ are chosen as in columns 2–3 of Table 2. The true standard deviation of the random errors is set at $\sigma = 0.1$. For the random shifting effects, the transition probability matrices for the second-order Markov chains are taken to be the estimated frequencies from a sample of 20 sets of m/z values estimated from a real data set, all of which have four states $0, \pm 1, 2$ (we use the scale of 0.005Da instead of the true scale 10^{-4}Da in our simulation in order for the random shifts to be discernable). Analogous to the true mass spectrometry data, here we use the log-intensity for the analysis because it has peaks of the similar width and has stable variances. In each replicate of the simulations, the log-intensities are generated according to model (5), in which $\epsilon_i(t), i = 1, \ldots, 20, t = 1, \ldots, 100$, are independent and identically distributed (iid) as $N(0, \sigma^2)$. And we generate the m/z's from the second-order Markov chain with the given transition probabilities. As an illustration, four simulated spectra are plotted in Fig. 3. Random shifts in the m/z's can be seen in this realization. Apparently, alignment of the spectra is needed for estimating the true spectra and for any statistical inference. Note that the random shifts will be more evident if the number of m/z values is larger.

We run 500 simulation replicates using the settings as described. To evaluate the conditional expectations in the expression of estimates, we generate 100 copies of m/z starting from 5500 with second-difference jumping at the scale of 0.005Da for each spectral curve by importance sampling method. To estimate the true curve, we use the Gaussian kernel with bandwidth set at $h^2 = 5$ for getting a spiky estimate and $h^2 = 20$ for a slightly smoother estimate. These values correspond to 0.11% and 0.21% sliding windows, respectively.

The fitted curves and the true log-intensity values are shown in Fig. 4. It shows that the algorithm can produce

Table 2. Regression parameter estimates

			$h^2 = 5$		$h^2 = 20$	
ID	α	β	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
1	0	1	0.000(0.000)	1.000(0.000)	0.000(0.000)	1.000(0.000)
2	0	1	0.000(0.014)	0.999(0.048)	−0.001(0.015)	1.000(0.051)
3	0	1	−0.003(0.014)	0.988(0.048)	−0.003(0.014)	0.998(0.051)
4	0	1	−0.003(0.014)	0.993(0.046)	−0.002(0.014)	0.997(0.049)
5	−0.25	1.25	−0.251(0.016)	1.248(0.053)	−0.25(0.017)	1.252(0.055)
6	−0.25	1.25	−0.251(0.016)	1.248(0.053)	−0.25(0.017)	1.258(0.06)
7	0.25	1.25	0.246(0.015)	1.236(0.053)	0.247(0.018)	1.247(0.058)
8	0.25	1.25	0.247(0.015)	1.240(0.054)	0.248(0.017)	1.251(0.056)
9	−0.5	1.5	−0.500(0.017)	1.498(0.061)	−0.501(0.020)	1.504(0.066)
10	−0.5	1.5	−0.500(0.017)	1.497(0.055)	−0.499(0.020)	1.505(0.062)
11	0.5	1.5	0.494(0.018)	1.484(0.060)	0.496(0.021)	1.496(0.066)
12	0.5	1.5	0.497(0.017)	1.487(0.062)	0.498(0.021)	1.497(0.064)
13	−0.75	1.75	−0.750(0.020)	1.748(0.068)	−0.750(0.023)	1.757(0.073)
14	−0.75	1.75	−0.751(0.019)	1.746(0.067)	−0.751(0.024)	1.757(0.076)
15	0.75	1.75	0.743(0.020)	1.728(0.066)	0.745(0.024)	1.749(0.076)
16	0.75	1.75	0.746(0.020)	1.734(0.065)	0.747(0.023)	1.744(0.074)
17	−1	2	−1.001(0.021)	1.997(0.078)	−1.000(0.026)	2.007(0.083)
18	−1	2	−1.002(0.021)	1.992(0.076)	−1.001(0.028)	2.007(0.085)
19	1	2	0.993(0.022)	1.978(0.072)	0.995(0.027)	1.997(0.081)
20	1	2	0.996(0.022)	1.979(0.073)	0.996(0.025)	1.997(0.079)

*Standard deviations are in parentheses

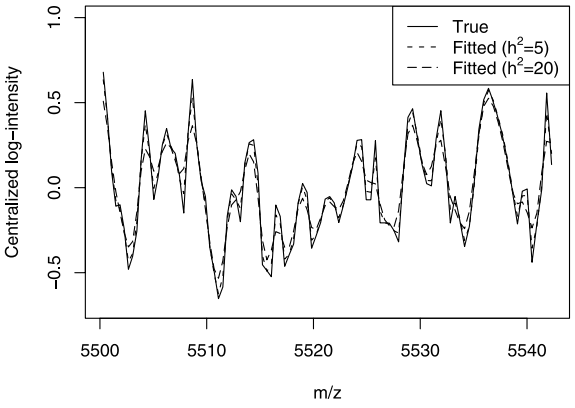

Figure 4. The true value and the fitted value of the intensities versus m/z.

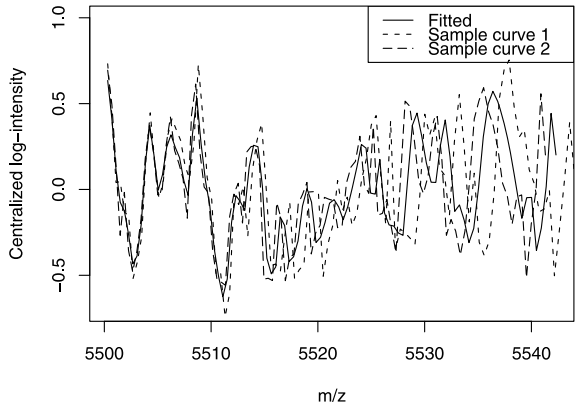

Figure 5. Two sample spectra and estimated spectra when $h^2 = 5$.

rather accurate estimate of the true curve. The signals represented by the spikes are retained in the estimated curves for small choice of bandwidths ($h^2 = 5$). Furthermore, larger bandwidth ($h^2 = 20$) leads to smoother estimate.

The true and the mean estimated values α_i, β_i, and their variances are shown in Table 2. We can see that the parameter estimates are close to the true values and that there is no systematic bias. However, the variances of the $\hat{\beta}$ tend to increase as true value β increases, this is caused by the standardization in step (b) of the algorithm and the greater variabilities of the intensities for large β. Variances of $\hat{\alpha}$ have the same trend since it is a linear function of $\hat{\beta}$. The estimated value of σ is 0.124 ($sd = 0.0016$) for $h^2 = 5$ and 0.193 ($sd = 0.0017$) for $h^2 = 20$. These values tend to overestimate the true variance of noises, mainly due to the local smoothing effect.

Figure 5 illustrates the effect of alignment. The dotted lines are two sample spectra curves with random errors in both x-axis and y-axis, while the solid line is the fitted curve. To make the alignment more evident, we normalize the two sample spectra curves to the same level and scale by subtracting the estimated α and rescaling by estimated β. The fitted curve looks very similar to the two sample curves in the shape and they only differs in the m/z values. This implies alignment have been achieved. Therefore the IMS al-

Table 3. Regression parameter estimates ($h^2 = 5$)

			Sample 1		Sample 2	
ID	α	β	$\hat{\alpha}$	$\hat{\beta}$	$\hat{\alpha}$	$\hat{\beta}$
1	0	1	0.000(0.000)*	1.000(0.000)	0.000(0.000)	1.000(0.000)
2	0	1	0.000(0.014)	1.004(0.049)	0.002(0.015)	0.996(0.041)
3	−0.25	1.25	−0.255(0.017)	1.243(0.058)	−0.244(0.016)	1.255(0.048)
4	0.25	1.25	0.245(0.019)	1.251(0.058)	0.255(0.016)	1.258(0.050)
5	−0.5	1.5	−0.502(0.019)	1.505(0.054)	−0.497(0.015)	1.487(0.056)
6	0.5	1.5	0.498(0.019)	1.509(0.060)	0.502(0.017)	1.487(0.055)
7	−0.75	1.75	−0.756(0.021)	1.742(0.072)	−0.742(0.021)	1.750(0.065)
8	0.75	1.75	0.746(0.022)	1.747(0.065)	0.758(0.018)	1.746(0.068)
9	−1	2	−1.001(0.022)	2.004(0.076)	−0.996(0.023)	1.986(0.073)
10	1	2	0.996(0.023)	2.010(0.077)	1.002(0.022)	1.990(0.075)

*Standard deviations are in parentheses

gorithm does the alignment automatically by averaging over the random shift effect.

We have simulated $N = 100$ m/z values with the integrated Markovian shifts and demonstrated that the IMS algorithm can estimate the reference spectrum and the location and scale parameters rather accurately. The amount of signal information retained in the estimated curves depends on the bandwidth specification. The reference curve is chosen to be rather spiky. Since keeping the spike information is important for subsequent analysis, the curve have been estimated using small bandwidths in the alignment procedure.

3.2 Two-sample case

The following contains some simulation results for the two-sample case. The true regression parameters (α, β) are shown in Table 3, the true standard deviation for the error terms is taken to be $\sigma = 0.1$. The two true curves $m_1(\cdot)$ and $m_2(t)$ for the two groups are displayed in the top panel of Fig. 6. The peak patterns are similar at most of the m/z values for these two samples, though the intensity values differ for some of the m/z values. The major difference lies at m/z value around 5530 where group 1 has a peak and group 2 is flat. We repeatedly generate 10 curves from each group, and implement 500 replicates of simulation. The bottom panel of Fig. 6 displays two observed curves for each group (thin lines), the fitted curves using $h^2 = 5$ for the two groups are the two thick lines, which are seen to be very accurate estimates of the true curves as shown in the top panel. Variabilities in both m/z and intensities can be clearly seen from the observed curves in both groups. Especially at the right end, the two observed curves in each group are shifted prominently.

To keep the signal information we have used a small bandwidth $h^2 = 5$, which corresponds to a 0.11% sliding window in the common peak alignment algorithm. This is because we have used Gaussian kernel, the 99% confidence interval has width $2 \times 2.576h = 11.52$ which is about $\pm 0.11\%$ of the m/z values ranging between 5500 to 5540 we had simulated.

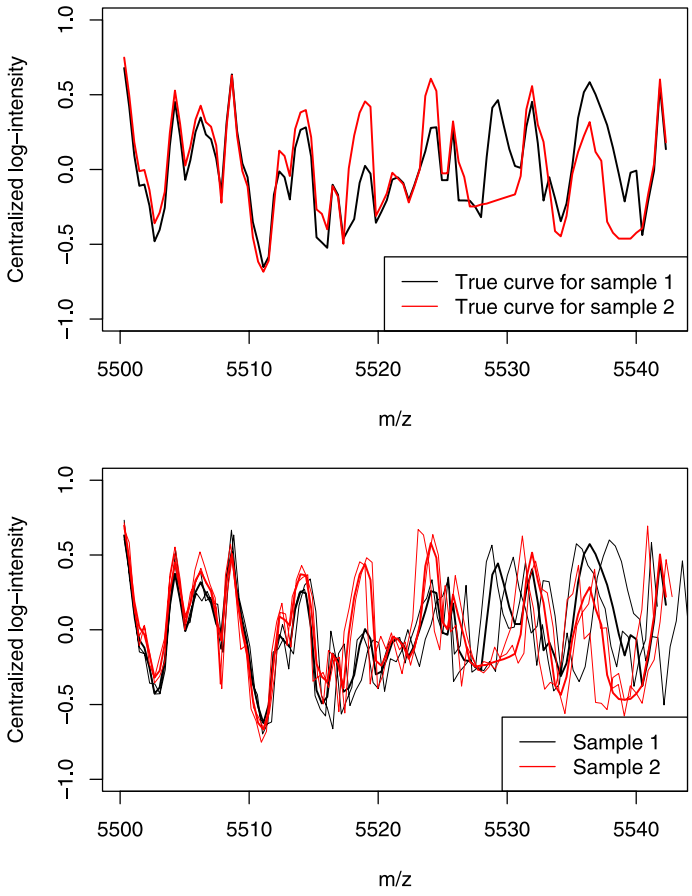

Figure 6. Two-sample curve estimation with IMS alignment (Top panel: true curves; bottom panel: observed (thin lines) and fitted (thick lines, $h^2 = 5$) curves for the two groups).

Figure 7 compares the fitted curves and the true curves for the two samples. It can be seen that the estimated curve for each sample fits the true curve almost perfectly, while the sample/group differences are kept. These fitted curves are obtained from randomly shifted curves. Table 3 lists the

Table 4. Estimates of $\alpha_i(\times 10^{-4})$ and β_i of the twenty curves

ID	1	2	3	4	5	6	7	8	9	10
$\hat{\alpha}_i$	0.000	0.033	4.640	4.457	2.920	0.085	4.949	2.754	0.062	4.653
$\hat{\beta}_i$	1.000	1.164	1.259	1.222	1.272	1.416	1.351	1.204	1.303	1.262
ID	11	12	13	14	15	16	17	18	19	20
$\hat{\alpha}_i$	4.617	3.223	0.026	4.721	4.923	4.358	−5.555	4.907	0.032	4.299
$\hat{\beta}_i$	1.264	1.395	1.128	1.291	1.332	1.196	1.372	1.328	1.158	1.180

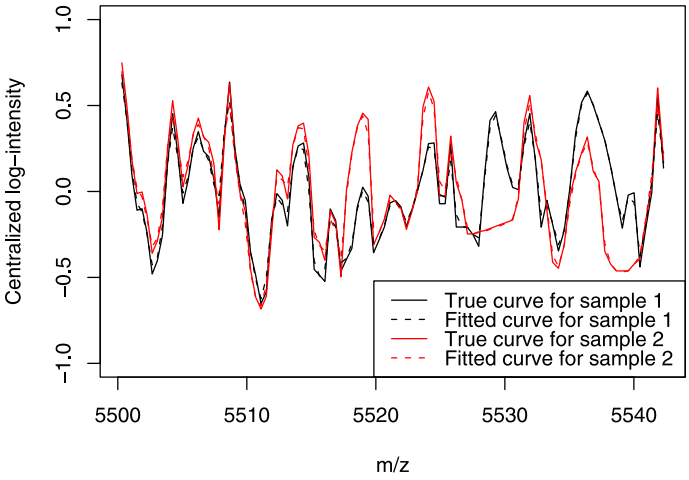

Figure 7. Observed and fitted curves for two-sample (thick lines: fitted curves; thin lines: observed curves with random shifts).

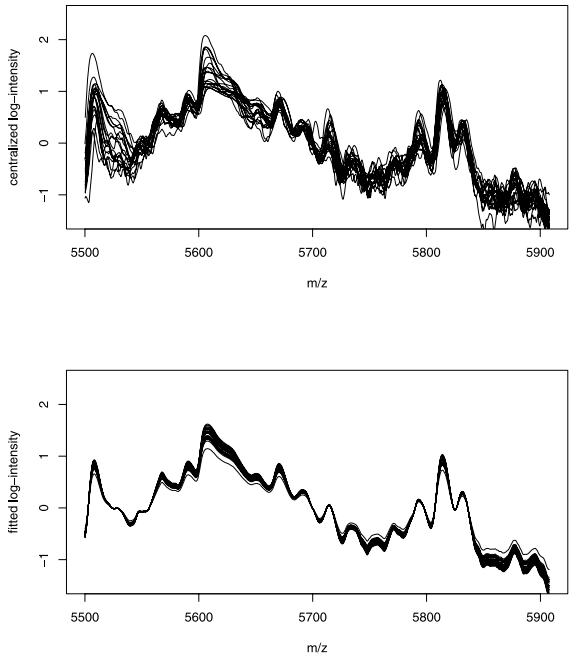

Figure 8. The original centralized log-intensities (upper row) and the fitted values (lower row) for 20 liver cancer patients.

estimated regression parameters for $h^2 = 5$. All the estimates are close to the true values. The standard deviation of the noise is estimated to be $\hat{\sigma} = 0.123$ with standard error 0.0026.

It is worth noting that the proposed IMS model is in fact a marginal model for log-intensities. The dependence structure of neighboring intensities is hard, if not impossible, to capture, and the normality assumption may not be true in a strict sense in practice. In some sense, the independence and normality assumptions in our model are only used to derive an effective algorithm. Validity of this approach is analogous to the validity of least squares method when the latter is applied to data that may be neither independent nor normally distributed.

3.3 A real data example

The data set we use is from a liver cancer study conducted in Changzheng Hospital, Shanghai, in 2004. Normalization, baseline subtraction had been done by using version 3.1.1 of Ciphergen ProteinChip Software (Biosystems, Inc.). For the purpose of demonstrating our proposed method, we randomly choose 20 of the MS spectra from the liver cancer group, and implement the alignment algorithm on these spectra for 1000 m/z values.

The data show great variabilities as seen in Fig. 8 (upper panel) for 1000 m/z values. These variabilities arise from both the errors in intensity values and misalignment. Clearly, the spectra need to be aligned first. We designate the first spectrum as the reference curve in the IMS method, and choose the bandwidth $h^2 = 5$. The estimated α and β values of the twenty spectra are as in Table 4. We can see from the table that the values of β_i range from 1 to 1.416, which is consistent with the approximately 50% variation from true intensities as claimed by the vendor. Figure 8 (bottom panel) displays the centralized fitted log-intensities, alignment can be spotted in the fitted values. Furthermore, all the characterized peaks have been extracted out while the small fluctuations are smoothed out.

Since unaligned spectrum curves presumably exhibit larger variation across multiple spectra at each observed m/z, the coefficients of variation (cv), which is the ratio of standard deviation and mean, of the log-intensities is a proper measure of alignment algorithms (Jeffries, 2005). We emphasize that the variation in the original curves consists

Table 5. Quantiles of coefficients of variation

Quantiles	25%	50%	75%
cv_1 (original data)	0.201	0.307	0.716
cv_2 (aligned data)	0.081	0.081	0.081
cv_1/cv_2	2.482	3.790	8.840

of both the variations from noises and from misalignment of the curves. To assess the accuracy of the alignment, we calculate the cv at each of the 1000 m/z's across the 20 original spectra and the aligned ones. See Table 5 for summaries of the cv values. Note that our method assumes that all the 20 spectra share the same shape curve $m(\cdot)$ because they belong to the same group and all the variations of the aligned spectra rooted from variations of the estimated β_i's. Therefore, the cv of the aligned spectra is the cv of the estimated β's, which is 0.081. The median of the cv's is 0.307 for the original data and 0.081 for aligned data by the IMS algorithm. The median ratio of the two sets of cv's is 3.790, implying that our algorithm can in general reduce variation by a factor of $1 - 1/3.790 = 73.6\%$. The data show great variabilities as seen in Fig. 8 (upper panel) for 1000 m/z values. These variabilities arise from both the errors in intensity values and misalignment. Clearly, the spectra need to be aligned first. We designate the first spectrum as the reference curve in the IMS method, and choose the bandwidth $h^2 = 5$. The estimated α and β values of the twenty spectra are as in Table 4. We can see from the table that the values of β_i range from 1 to 1.416, which is consistent with the approximately 50% variation from true intensities as claimed by the vendor. Figure 8 (bottom panel) displays the centralized fitted log-intensities, alignment can be spotted in the fitted values. Furthermore, all the characterized peaks have been extracted out while the small fluctuations are smoothed out.

Since unaligned spectrum curves presumably exhibit larger variation across multiple spectra at each observed m/z, the coefficients of variation (cv), which is the ratio of standard deviation and mean, of the log-intensities is a proper measure of alignment algorithms (Jeffries, 2005). We emphasize that the variation in the original curves consists of both the variations from noises and from misalignment of the curves. To assess the accuracy of the alignment, we calculate the cv at each of the 1000 m/z's across the 20 original spectra and the aligned ones. See Table 5 for summaries of the cv values. Note that our method assumes that all the 20 spectra share the same shape curve $m(\cdot)$ because they belong to the same group and all the variations of the aligned spectra rooted from variations of the estimated β_i's. Therefore, the cv of the aligned spectra is the cv of the estimated β's, which is 0.081. The median of the cv's is 0.307 for the original data and 0.081 for aligned data by the IMS algorithm. The median ratio of the two sets of cv's is 3.790, implying that our algorithm can in general reduce variation by a factor of $1 - 1/3.790 = 73.6\%$.

4. DISCUSSION

Data preprocessing such as normalization and baseline subtraction is usually the first step for analyzing MS data. Besides internal/external calibration, more delicate spectral alignment is needed in order to carry out data analysis. The IMS method proposed in this paper appears to be capable of achieving better alignment of the whole curves, not merely peaks. A key ingredient is the integrated Markov chain model for the random shifts in m/z values. Our experience with TOF MS data indicates that this model does capture important features of MS spectral data. The peak differences among several groups can be efficiently detected by the IMS model for multi-sample problems.

The IMS algorithm computes estimates iteratively, alternating between the nonparametric component and parametric component of the model. The distribution of the integrated Markov chain of shifts is obtained from the observed m/z values, and the random shifting effect is averaged out naturally by the conditional expectation of shifts. In this connection, the IMS algorithm is a self-modelling alignment (warping) method, making use of the natural connection of shifts with intensity magnitudes through computing expectations with respect to the conditional distribution of the random shift given the corresponding intensity information.

Results of simulation studies and real data analysis show that the proposed IMS method is rather satisfactory in aligning multiple TOF MS spectra. Random shifts resulted from accumulations of minor random jumps of second-order differences of m/z are accounted for by the integrated Markov chain shifting effects. There is a substantial reduction in the variability (from both the random errors in intensities and the misalignment) among multiple spectra as measured by coefficients of variation. As shown in the real data example, even if the data have been normalized in preprocessing step, inclusion of the baseline (α_i) and scale (β_i) parameters is still needed in order to account for individual baseline variabilities.

A large bandwidth choice smoothes out the curves but may cause losses in signals or peaks which is regarded as informative in mass spectrometry. We propose using 0.04%–0.08% bandwidths when Gaussian kernel function is used in smoothing, which is compatible with the common 0.1%–0.2% sliding window method. In this sense, our method can achieve similar alignment effect with the sliding window method. This bandwidth selection rule has been shown to be efficient in keeping signal information. In some applications alignment of the spectra instead of peaks may be required for subsequent data analysis, but peak picking methods may overlook some small peaks that are informative for differentiating different groups. An advantage of the proposed IMS method is that it can simultaneously perform alignment of entire spectra curves and local smoothing, rather than just alignment of peaks.

We have assumed that the errors in the proposed model are independent and normally distributed. Apparently,

nearby intensities are likely to be correlated with one another. Rationale for this assumption can be seen from a marginal modelling viewpoint, for which dependent data is modelled marginally with a working independence assumption. Useful methods of this kind include the generalized estimating equation method (GEE, Liang and Zeger, 1986) for modelling longitudinal data and the Naive Bayesian Classifier (NBC) in machine learning. The approach may be best understood via the least squares (LS) method, which is most efficient if the errors are independent and identically normally distributed. But LS method is still valid if the independence or the normality assumption is dropped as long as the mean functions are correctly specified (misspecification of variance-covariances only causes some efficiency losses). We have used similar marginal methods in building the random shift model. The parameter and curve estimates are correct (unbiased) when the mean function is correctly specified. The validity of the proposed model can also be seen from the simulation studies.

We have demonstrated estimation of curves with random shifts mainly for the one-sample problem. We have also considered application of the IMS algorithm in multi-sample problem by introducing different shape functions for different samples. It is shown that the algorithm can effectively align multiple spectrum curves and detect the differences among different samples. Adjustment of other factors can be done by adding extra covariate variables in the semiparametric random effect model. This issue will be pursued elsewhere.

ACKNOWLEDGEMENTS

We would like to thank Liang Zhu and Cheng Wu at Shanghai Changzheng Hospital for providing and explaining the data. This research was supported by Grants from the Chinese Academy of Sciences and National Natural Science Foundations (Yang) and the US National Science Foundation and National Institutes of Health (Ying).

APPENDIX A. DERIVATIONS OF THE NPMLE ESTIMATES IN THE ONE-SAMPLE MODEL

To simplify expression, we denote $D_i(u) \equiv x_i(u) - s_i(u)$ and $\mathbf{D}_i = (D_i(u), u = 1, \ldots, N)$. The likelihood function for the spectra can be written as

$$(7)\quad L(m,\theta) = \prod_{i=1}^{n} f_{\mathbf{y_i}}(\mathbf{y}_i) = \prod_{i=1}^{n} \int f_{\mathbf{y_i}|\mathbf{s}_i}(\mathbf{y}_i) f_{\mathbf{s}}(\mathbf{s}_i) d\mathbf{s}_i = \prod_{i=1}^{n} \int \frac{\exp\left(-\frac{||\mathbf{y}_i - \alpha_i - \beta_i m(\mathbf{D}_i)||^2}{2\sigma^2}\right) f_{\mathbf{s}_i}(\mathbf{s}_i) d\mathbf{s}_i}{(2\pi\sigma^2)^{\frac{N}{2}}},$$

where $\mathbf{y}_i = (y_i(u), u = 1, \ldots, N)$, $\mathbf{x}_i = (x_i(u), u = 1, \ldots, N)$, $m(\mathbf{D}_i) = (m(x_i(u) - s_i(u)), u = 1, \ldots, N)$. Let $f_{\mathbf{s}_i|\mathbf{y}_i}(\mathbf{s_i})$ be the conditional probability function of $\mathbf{s}_i$ given $\mathbf{y}_i$:

$$f_{\mathbf{s}_i|\mathbf{y}_i}(\mathbf{s}_i) = \frac{f_{\mathbf{y}_i|\mathbf{s}_i}(\mathbf{y}_i) f_{\mathbf{s}_i}(\mathbf{s}_i)}{\int f_{\mathbf{y}_i|\mathbf{s}_i}(\mathbf{y}_i) f_{\mathbf{s}_i}(\mathbf{s}_i) d\mathbf{s}_i},$$

where

$$f_{\mathbf{y}_i|\mathbf{s}_i}(\mathbf{y}_i) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}||\mathbf{y}_i - \alpha_i - \beta_i m(\mathbf{D}_i)||^2\right).$$

We then proceed with estimating the unknown quantities as follows.

Given m, we maximize the profile likelihood over the remaining parameters to get

$$(8)\qquad \hat{\beta}_i = \frac{\sum_{u=1}^{N} E_{s_i(u)|\mathbf{y}_i}[m(D_i(u)) - \bar{m}_i] y_i(u)}{\sum_{u=1}^{N} E_{s_i(u)|\mathbf{y}_i}[m(D_i(u)) - \bar{m}_i]^2},$$

$$(9)\qquad \hat{\alpha}_i = \bar{y}_i - \bar{m}_i \hat{\beta}_i,$$

and

$$(10)\quad \hat{\sigma}^2 = \frac{1}{nN} \sum_{i=1}^{n} \sum_{u=1}^{N} E_{s_i(u)|\mathbf{y}_i}(y_i(u) - \alpha_i - \beta_i m(D_i(u)))^2,$$

where $\bar{y}_i = \sum_{u=1}^{N} y_i(u)/N$, $\bar{m}_i = E_{\mathbf{s}_i|\mathbf{y}_i}[\sum_{u=1}^{N} m(D_i(u))/N]$. As we mentioned before, for all parameters to be identifiable, we need to set

$$\hat{\beta}_i = \hat{\beta}_i/\hat{\beta}_1, \qquad \hat{\alpha}_i = \hat{\alpha}_i - \hat{\alpha}_1,\ i = 1, \ldots, n.$$

To derive the score function for m, we use the Hadmard derivatives for functionals. Given an arbitrary direction $h(\cdot)$, the score for $m(\cdot)$

$$\begin{aligned}
\frac{\partial \log L}{\partial m}(h) &= \sum_{i=1}^{n} \frac{1}{L_i(m,\theta)} \frac{\partial L_i(m + uh, \theta)}{\partial u}|_{u=0} \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{n} \frac{\int \beta_i \langle h(\mathbf{D}_i), \mathbf{y}_i - \alpha_i - \beta_i m(\mathbf{D}_i)\rangle f_{\mathbf{y}_i|\mathbf{s}_i}(\mathbf{y}_i) f_{\mathbf{s}_i}(\mathbf{s}_i) d\mathbf{s}_i}{\int f_{\mathbf{y}_i|\mathbf{s}_i}(\mathbf{y}_i) f_{\mathbf{s}_i}(\mathbf{s}_i) d\mathbf{s}_i} \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{n} E_{\mathbf{s}_i|\mathbf{y}_i} \beta_i \langle h(\mathbf{D}_i), \mathbf{y}_i - \alpha_i - \beta_i m(\mathbf{D}_i)\rangle \\
&= -\frac{1}{\sigma^2} \sum_{i=1}^{n} \sum_{u=1}^{N} \beta_i E_{s_i(u)|\mathbf{y}_i} \left[y_i(u) - \alpha_i - \beta_i m(D_i(u))\right] h\left(D_i(u)\right).
\end{aligned}$$

Since this is linear in h, we may choose various h in the score to obtain a series of equations. In particular, letting $h(r) = \delta_t(r)$, the Dirac function, in the score equation $\frac{\partial \log L}{\partial m}(h) = 0$, we get

$$(11)\quad \hat{m}(t) = \frac{\sum_{i=1}^{n} \sum_{u=1}^{N} \beta_i (y_i(u) - \alpha_i) f_{s_i(u)|\mathbf{y}_i}(x_i(u) - t)}{\sum_{i=1}^{n} \sum_{u=1}^{N} \beta_i^2 f_{s_i(u)|\mathbf{y}_i}(x_i(u) - t)},$$

which can be shown to be the solution to $\frac{\partial \log L}{\partial m}(h) = 0$ for any h.

APPENDIX B. KERNEL APPROXIMATION TO THE NPMLE IN THE ONE-SAMPLE MODEL

Calculation of the conditional density $f_{s_i(u)|\mathbf{y}_i}(x_i(u)-t)$ is rather involved and inaccurate. One may approximate it by a kernel method. Let $K(\cdot)$ be a kernel density function such that $K \geq 0$, $\int K(u)du = 1$, symmetric and attains the maximum at 0. For example, the kernel function can be taken as the Gaussian density function $K(x) = e^{-\frac{x^2}{2}}$. The NPMLE for $m(\cdot)$ in (11) can be approximated through taking $h_t(u) = K(|u-t|/h)$ in the score equation for an arbitrarily small $h > 0$. Intuitively as $h \to 0$, $h_t(u)$ converges to the Dirac function $\delta_t(u)$, i.e., as $h \to 0$,

$$\int g(x)K(|x-t|/h)dx \to g(t) = \int g(x)\delta_t(x)dx,$$

if g is continuous. This may be best understood when the kernel is taken to be the Gaussian density function, for which h is the standard deviation, and as the standard deviation approaches 0, mass concentrates around the center/mean t, and the expectation $Eg(x)$ approaches $g(t)$. Then

(12)
$$\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i E_{\mathbf{s}_i|\mathbf{y}_i}\{(y_i(u)-\alpha_i-\beta_i m(D_i(u))h_t(D_i(u)))\}$$
$$\approx \sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i(y_i(u)-\alpha_i-\beta_i m(t))E_{\mathbf{s}_i|\mathbf{y}_i}K\left(\frac{|D_i(u)-t|}{h}\right).$$

From (11) and (12), we obtain the following approximate estimate of m

$$\hat{m}_h(t) = \frac{\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i(y_i(u)-\alpha_i)E_{\mathbf{s}_i|\mathbf{y}_i}K(\frac{|D_i(u)-t|}{h})}{\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i^2 E_{\mathbf{s}_i|\mathbf{y}_i}K(\frac{|D_i(u)-t|}{h})}.$$

Remark: Approximation (12) can be improved by noting

$$\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i E_{\mathbf{s}_i|\mathbf{y}_i}\left\{(y_i(u)-\alpha_i-\beta_i m^{(k)}(D_i(u)))h_t(D_i(u))\right\}$$
$$\approx \sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i E_{\mathbf{s}_i|\mathbf{y}_i}\left\{(y_i(u)-\alpha_i-\beta_i m^{(k)}(t))K\left(\frac{|D_i(u)-t|}{h}\right)\right\}$$
$$+\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i E_{\mathbf{s}_i|\mathbf{y}_i}\left\{\left[\beta_i m^{(k-1)}(t)-\beta_i m^{(k-1)}(D_i(u))\right]\right.$$
$$\left.\times K\left(\frac{|D_i(u)-t|}{h}\right)\right\}.$$

This results in the following iterative process for calculating the NPMLE:

$$\hat{m}^{(k)}(t) = \frac{\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i(y_i(u)-\alpha_i)E_{\mathbf{s}_i|\mathbf{y}_i}K\left(\frac{s_i(u)-(x_i(u)-t)}{h}\right)}{\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i^2 E_{\mathbf{s}_i|\mathbf{y}_i}K\left(\frac{s_i(u)-(x_i(u)-t)}{h}\right)}$$
$$+\frac{\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i^2 E_{\mathbf{s}_i|\mathbf{y}_i}\left\{\Delta^{(k-1)}(t,u)K\left(\frac{s_i(u)-(x_i(u)-t)}{h}\right)\right\}}{\sum_{i=1}^{n}\sum_{u=1}^{N}\beta_i^2 E_{\mathbf{s}_i|\mathbf{y}_i}K\left(\frac{s_i(u)-(x_i(u)-t)}{h}\right)},$$

where $\Delta^{(k-1)}(t,u) = \hat{m}^{(k-1)}(t) - \hat{m}^{(k-1)}(x_i(u)-s_i(u))$.

This is what we have used in the algorithm. We found that with this iterative modification for estimating $m(\cdot)$, the fitting process tends to converge more quickly and the resulting estimates tend to be more accurate.

Received 9 April 2009

REFERENCES

[1] Baggerly, K. A., Morris, J. S., Wang, J., Gold, D., Xiao, L.-C., and Coombes, K. R. (2003). A comprehensive approach to the analysis of MALDI-TOF proteomics spectra from serum samples. *Proteomics* **3** 1667–1672.

[2] Baggerly, K. A., Morris, J. S., and Coombes, K. R. (2004). Reproducibility of SELDI-TOF Protein Patterns in Serum: comparing datasets from different Experiments. *Bioinformatics* **20** 777–785.

[3] Brumback, L. C. and Lindstrom, M. J. (2004). Self-modelling with flexible, random time transformations. *Biometrics* **60** 461–470. MR2066281

[4] Diamandis, E. P. (2004). Mass Spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. *Mol Cell Proteomics* **3** 367–378.

[5] Gervini, D. and Gasser, T. (2004). Self-modelling warping functions. *J R Statist Soc B* **66** 959–971. MR2102475

[6] Guilhaus, M. (1995). Principles and instrumentation in time-of-flight mass spectrometry. *J Mass Spectrometry* **30** 1519–1532.

[7] Jeffries, N. (2005). Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics* **21** 3066–3073.

[8] Kneip, A. and Engel, J. (1995). Model estimation in nonlinear regression under shape invariance. *Ann. Statist.* **23** 551–570. MR1332581

[9] Kneip, A. and Gasser, T. (1992). Statistical tools to analyze data representing a sample of curves. *Ann. Statist.* **20** 1266–1305. MR1186250

[10] Kneip, A., Li, X., MacGibbon, K. B., and Ramsay, J. O. (2000). Curve registration by local regression. *Can. J. Statist.* **28** 19–29. MR1789833

[11] Lawton, W. H., Sylvestre, E. A., and Maggio, M. S. (1972). Self-modelling for nonlinear regression. *Technometrics* **14** 513–532.

[12] Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika* **73** 13–22. MR0836430

[13] Ramsay, J. O. and Li, X. (1998). Curve registration. *J R Statist Soc B* **60** 351–363. MR1616045

[14] Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer. MR2168993

[15] Ronn, B. B. (2001). Nonparametric maximum likelihood estimation for shifted curves. *J R Statist Soc B* **63** 243–259. MR1841413

[16] Tibshirani, R., Hastie, R., Narasimhan, B., Soltys, S., Shi, G., Koong, A., and Le, Q. T. (2004). Sample classification from

protein mass spectrometry by peak probability contrasts. *Bioinformatics* **20** 3034–3044.
[17] Wang, K. and Gasser, T. (1999). Synchronizing sample curves nonparametrically. *Ann. Statist.* **27** 439–460. MR1714722
[18] Wolski, W. E., Lalowski, M., Jungblut, P., and Reinert, K. (2005). Calibration of mass spectrometric peptide mass fingerprint data without specific external or internal calibrants. *BMC Bioinformatics* **6** 203.
[19] Yasui, Y., Pepe, M., Thompson, M., Adam, B.-L., Wright, G., Qu, Y., Potter, J., 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.

Yang Feng
Department of Operations Research and Financial Engineering
Princeton University
Princeton, NJ 08544
USA
E-mail address: yangfeng@princeton.edu

Weiping Ma
Department of Mathematics
Fudan University
Shanghai 200433
China
E-mail address: weipingma1984@gmail.com

Zhanfeng Wang
Department of Statistics and Finance
University of Science and Technology of China
Hefei, Anhui 230026
China
E-mail address: zfw@mail.ustc.edu.cn

Zhiliang Ying
Department of Statistics
Columbia University
New York, NY 10027
USA
E-mail address: zying@stat.columbia.edu

Yaning Yang
Department of Statistics and Finance
University of Science and Technology of China
Hefei, Anhui 230026
China
E-mail address: ynyang@ustc.edu.cn