NONPARAMETRIC PRE-PROCESSING METHODS AND INFERENCE TOOLS FOR ANALYZING TIME-OF-FLIGHT MASS SPECTROMETRY DATA

Anestis Antoniadis[*], *Sophie Lambert-Lacroix and Frédérique Letué*,
Laboratoire IMAG-LMC, University Joseph Fourier,
BP 53, 38041 Grenoble Cedex 9, France
and *Jérémie Bigot*
University Paul Sabatier, Toulouse, France.

Abstract

The objective of this paper is to contribute to the methodology available for extracting and analyzing signal content from protein mass spectrometry data. Data from MALDI-TOF or SELDI-TOF spectra require considerable signal pre-processing such as noise removal and baseline level error correction. After removing the noise by an invariant wavelet transform, we develop a background correction method based on penalized spline quantile regression and apply it to MALDI-TOF (matrix assisted laser deabsorbtion time-of-flight) spectra obtained from serum samples. The results show that the wavelet transform technique combined with nonparametric quantile regression can handle all kinds of background and low signal-to-background ratio spectra; it requires no prior knowledge about the spectra composition, no selection of suitable background correction points, and no mathematical assumption of the background distribution. We further present a multi-scale based novel spectra alignment methodology useful in a functional analysis of variance method for identifying proteins that are differentially expressed between different type tissues. Our approaches are compared with several existing approaches in the recent literature and are tested on simulated and some real data. The results indicate that the proposed schemes enable accurate diagnosis based on the over-expression of a small number of identified proteins with high sensitivity.

Key words: curve estimation, wavelets, regression quantiles, robust point-matching, P-splines smoothing, mean integrated square error, functional analysis of variance.

[*]Corresponding author. E-mail: antonia@imag.fr

1 Introduction

The proteins are the controllers of all cell functions and, as such, are closely connected with many diseases and metabolic processes. Microarray data has been successfully used to identify genes responsible for many diseases, especially cancer. However, although proteins are coded by genes, there is no one-to-one relationship between the protein and the mRNA due to different rates of translation. Hence studying mRNA expressions (microarrays) may be an indirect way to understand a disease etiology. This ineffectiveness of genomics caused a big shift of interest from genomics to proteomics, with the hope that proteomic studies may provide a more direct information for understanding the biological functions towards a disease profile and may help targeted drug therapy. An important tool used for protein identification and high throughput comparative profiling of disease and non-disease complex protein samples in proteomics is mass spectrometry (MS). With this technology it is possible to identify specific biomarkers related to a given metabolic process or disease from the lower molecular weight range of the circulating proteome from easily obtained biological fluids such as plasma or serum. Recent research has demonstrated that using such technology to generate protein expression profiles from lung cancer lystates is an alternative promising strategy in the search for new diagnostic and therapeutic molecular targets.

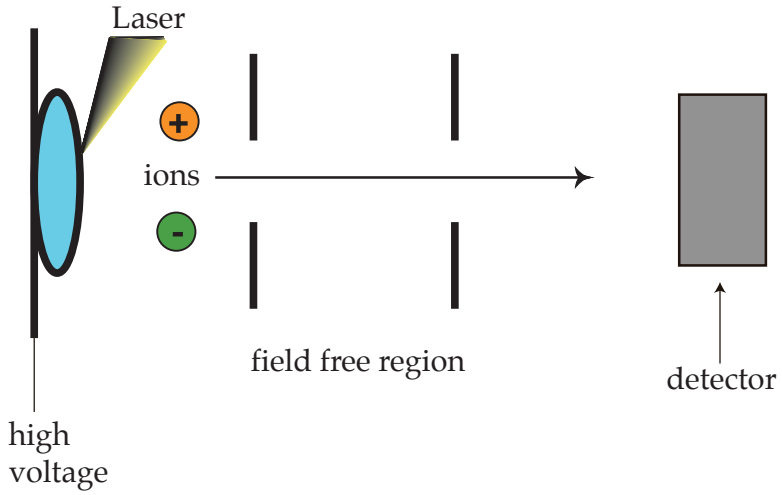


Figure 1.1: Simplified Schematic of MALDI-TOF Mass Spectrometry

There are at least two kinds of mass spectrometry instruments commonly applied to clinical and biological problems today, namely, Matrix-Assisted Laser Desorption and Ionization Time-Of-Flight (MALDI-TOF) and Surface-Enhanced Laser Desorption and Ionization Time-Of-Flight (SELDI-TOF) mass spectroscopy. The schematic setup of a linear MALDI-TOF instrument is shown in Figure 1.1. First, the biological samples are mixed with an organic compound that acts as a matrix to facilitate the desorption and ionization of compounds in the sample. The analyte molecules are distributed throughout the matrix so that they are completely isolated from each other. Some of the energy incident on the sample plate is absorbed by the matrix, causing rapid vibrational excitation. The analyte molecules can become ionized by simple protonation by the photo-excited matrix, leading to the formation of the typically singly charged ions. Some multiply charged ions are also formed, but this is rarer. The analyte ions are then accelerated by an electrostatic field to a common kinetic energy. If all the ions have the same kinetic energy, the ions with low mass to charge ratio (m/z) travel faster than those with higher m/z values, therefore, they are separated in the flight tube and the number of ions reaching the detector at the end of the flight tube is recorded as the intensity of the ions. For MALDI, normally the charge is equal to one or two. The SELDI-TOF system uses preactivated differential binding surfaces to achieve multidimensional chromatography and the protein-bound chips are then analyzed by a similar technique. Whatever is the technique used, the calibrated output is a mass spectrum characterized by numerous peaks, which correspond to individual proteins or protein fragments (polypeptides) present in the sample. The heights of the peaks represent the intensities or abundance of ions in the sample for a specific m/z value. These heights along with the m/z values represent the fingerprint of the sample. Hence by looking at the differential pattern of the spectra of samples one may detect the presence or absence of a metabolic process or a disease.

The above techniques result in a huge amount of data to be analyzed and generate a need for a rapid, efficient and fully automated method for matching and comparing MS spectra. The raw spectra acquired by TOF mass-spectrometers are generally a mixture of a real signal, noise of different characteristics and a varying baseline. Statistically, a possible model for a given MS spectrum is to represent it schematically by the equation

$$Y(m/z) = B(m/z) + NS(m/z) + \epsilon(m/z),$$

where $Y(m/z)$ is the observed intensity of the spectrum at mass to charge ratio m/z, $B(m/z)$ is the baseline representing a systematic relatively smooth artifact commonly seen in mass spectrometry data, $S(m/z)$ is the true signal of interest consisting of a sum of possible overlapping peaks, each corresponding to a particular biological molecule, N is a constant multiplicative normalization factor to adjust for possibly differing amounts of protein in each slide, and $\epsilon(m/z)$ is a white noise process arising primarily from electronic noise in the detector. Data preprocessing at the "spectrum level" is therefore extremely important for the quantitation of proteins from biological tissues and fluids and inadequate or incorrect preprocessing methods can result in data sets that exhibit substantial biases and make it difficult to reach meaningful biological conclusions.

While many powerful low-level processing methods have been introduced for analyzing mass spectrometry data, there is still room for improvement in this area. Complex interactions between baseline subtraction, normalization, noise estimation, and peak identification are related processes, so these steps should not be considered in isolation. In this paper, we propose several new preprocessing steps to be used before analyzing mass spectrometry (MS) data. These preprocessing steps include wavelet denoising, baseline correction and normalization along the mass/charge axis. While the benefits from denoising, baseline correction and normalization seem obvious, we also study a scale-space approach to automatically align multiple MS peak sets without manual parameter determination, by embedding intensity information into the alignment framework, thus generalizing current approaches that use only the m/z information during the alignment of peaks. Finally to avoid reliance on peak detection methods that are currently used for analyzing protein mass spectra, as in [8] and [35], we model the entire set of mass spectra as functions, and use functional analysis of variance methods (FANOVA) (see [2]) to identify characteristic differences across experimental conditions. From the FANOVA output, we also identify spectral regions that are differentially expressed across experimental conditions in a way that takes both statistical and clinical significance into account.

The rest of paper is organized as follows. Section 2 describes and introduces our MS-TOF data pre-processing and normalization methods and compares them to existing ones. Section 3 is devoted to a scale-space approach to automatically align multiple MS peak sets, while Section 4 summarizes key ideas and results of

FANOVA and demonstrates their use to analyze spectra from tumor samples and their comparison with normal ones. Section 5 discusses the results and points out possible extensions of the methodology. Some mathematical background and all proofs are given in the Appendix.

2 Denoising and baseline subtraction

2.1 Wavelet denoising

In the rest of this paper, to isolate and remove the noise component of a spectrum we will use a translation invariant discrete wavelet transform in a spirit similar to the UDWT filtering method of [13] for denoising SELDI-TOF spectra.

One reason for the popularity of the discrete wavelet transform (DWT), as formulated by [34] and [14] in times series analysis is that measured data from most processes are inherently multiscale in nature, owing to contributions from events occurring at different locations and with different localization in time and frequency. Consequently, data analysis and modeling methods that represent the measured variables at multiple scales are better suited for extracting information from measured data than methods that represent the variables at a single scale. This is why wavelets have recently received attention as a tool for preprocessing mass spectra (see e.g. [11, 41, 13]).

Donoho and Johnstone [19] considered the problem of estimating a signal in noise where all that is known about the signal is that it is spatially variable. They showed that wavelet-based "universal thresholding" exhibits certain asymptotic optimality properties – see [19] (p. 444) for details. Briefly, the following steps are used to denoise an observed n-length mass spectrum Y: the discrete wavelet transform is used to transform Y, certain subsets of coefficients are thresholded, then the inverse transform is applied to obtain the denoised signal. Such wavelet thresholding has become a standard technique used extensively in practice and available in many software packages. Some details on the various transform used in this paper are given in our Appendix (see also Section 3).

The orthogonal DWT is extremely efficient computationally, but it is not shift-invariant. Thus, its denoising performance can change drastically if the starting

position of the signal is shifted because the coefficients are not circularly shift equivariant, so that circularly shifting the observed signal by some amount will not circularly shift the discrete wavelet transform coefficients by the same amount. To try to alleviate this problem Coifman and Donoho [12] introduced the technique of "cycle spinning"; see also [46]. The idea of denoising via cycle spinning is to apply denoising not only to Y, but also to all possible unique circularly shifted versions of Y, and to average the results. As pointed out by Percival and Walden ([40], p. 429), this is equivalent to applying standard thresholding to the wavelet coefficients of the maximal overlap discrete wavelet transform or the undecimated discrete wavelet transform (UDWT), a transform we more briefly refer to as the stationary wavelet transform throughout.

It is important to recall that for Y of length n (we suppose that n is an integer multiple of 2^{J_0} for some integer J_0), the transform is overdetermined and produces a mean-zero wavelet coefficient sequence $\{\tilde{W}_{j,t}^{(Y)}, t = 0, \ldots, n-1\}$ at each level j of the transform. Such a sequence can be written as the length-n column vector $\tilde{\mathbf{W}}_j^{(Y)} = W_j Y$ where W_j is the level-j stationary wavelet transform matrix that maps Y to $\tilde{\mathbf{W}}_j^{(Y)}$. Coifman and Donoho [12] and Coombes *et al.* [13] found that "hard thresholding and translation invariance" combined gave both good visual characteristics and good quantitative characteristics; hence we have adopted universal and hard thresholding throughout for denoising MS spectra. As discussed in [40] (p. 429) cycle spinning can be implemented efficiently in terms of the stationary wavelet transform. Let $n_j = W_j \boldsymbol{\epsilon}$. Note that $j = 1$ corresponds to the finest resolution, and $j = J_0$ to the coarsest.Then the algorithm is:

1. Compute a level J_0 partial stationary wavelet transform giving coefficient vectors $\tilde{\mathbf{W}}_1^{(Y)}, \ldots, \tilde{\mathbf{W}}_{J_0}^{(Y)}$ and $\tilde{\mathbf{V}}_{J_0}^{(Y)}$, where $\tilde{\mathbf{V}}_{J_0}^{(Y)}$ denotes the vector of scaling coefficients at resolution J_0.

2. For each $j = 1, \ldots, J_0$ apply hard thresholding using the level-dependent universal threshold with $\sigma_{n_j}^2 = \sigma_\epsilon^2 / 2^j$, to obtain

$$\hat{W}_{j,t}^{(Y)} = \begin{cases} \tilde{W}_{j,t}^{(Y)}, & \text{if } \left|\tilde{W}_{j,t}^{(Y)}\right| > \sigma_{n_j}\sqrt{2\log n} \\ 0, & \text{otherwise.} \end{cases}$$

3. The denoised signal is then obtained by applying the inverse stationary wavelet transform to $\hat{\mathbf{W}}_1^{(Y)}, \ldots, \hat{\mathbf{W}}_{J_0}^{(Y)}$ and $\tilde{\mathbf{V}}_{J_0}^{(Y)}$.

Since σ_ϵ is unknow, it is estimated by the MAD scale estimate defined by

$$\hat{\sigma}_{\text{MAD}} = \frac{\sqrt{2}\,\text{median}\left\{|\tilde{W}_{1,0}^{(Y)}|, |\tilde{W}_{1,1}^{(Y)}|, \ldots, |\tilde{W}_{1,n-1}^{(Y)}|\right\}}{0.6745}.$$

After denoising in this manner, separating background from true signal is considerably easier, and peaks , if necessary, can be rapidly identified and precisely quantified.

2.2 Baseline correction by penalized quantile regression splines

As already noted, spectra frequently exhibit a decreasing “baseline" (with increasing m/z), that may be unrelated to constituent protein composition. Clearly, such nuisance variation must be accommodated before meaningful quantitative analysis can be conducted, since the background component is added to the real signal and overstates the intensities of peaks. Another serious problem caused by varying background is the difficulty with aligning spectra by maximizing a similarity measure between them (see [26]). Varying background present in warped spectra makes it difficult to properly calculate their similarity. Many numerical methods have been developed for estimation of varying background present in one-dimensional signals. Among these techniques are methods based on digital filters [50, 45]. Such filters usually introduce artefacts and simultaneously distort the real signal. Other approaches rely on automated peak rejection [44, 13]. These algorithms fit some functions to find regions of signal that consist only of the baseline without peaks of real signal. The functions being fitted may have different forms, e.g. polynomials, splines. The main disadvantages of peak-rejection approaches are difficulties related to identification of peak-free regions. On the other hand, threshold-based rejection of peaks gives good results when the baseline is relatively smooth [51], and fails for signals with significantly varying baseline.

Because of difficulties caused by automatic peak rejection other approaches have been designed to fit a baseline without detecting the peaks. In [1], the baseline is fitted with a low-order polynomial that prevents it from fitting the real signal peaks. For signals with many positive peaks, however, e.g. mass spectra, the baseline estimated in

this way has values which are too high. Subtraction of a background with values which are too high from a signal introduces significant distortions to the analyzed signal, i.e. the values for peaks are too low. Other approaches rely on statistical methods, such as maximum entropy [38]. There are also approaches based on baseline removal in the wavelet domain [33].

In this paper we focus on a method for background elimination based on penalized regression quantile splines and evaluate its potential as an automated approach. Although baseline drift correction is illustrated with respect to matrix-assisted laser desorption/ionization time-of-flight data, our approach has much wider application, since other types of spectral data suffer from baseline drift and, potentially, our technique can also assist with a variety of instrumentation (not necessarily in the bioinformatics domain) that suffers from baseline drift. Our method has also strong similarities with the method, proposed by [21] for background elimination in two-dimensional signals by asymmetric least squares splines regression (see Remark 2.1). Since the pathbreaking paper by [31], quantile regression methods have attracted considerable interest, basically in all domains of statistics: see the recent monograph in [30] for a review of regression and regression quantiles in a traditional setting of independent samples or time series data. To the best of our knowledge, and quite surprisingly so, quantile regression seldom has been considered in the context of baseline correction in MS spectra.

2.3 Mathematical presentation

We first describe the basic setup and background for quantile regression models. For the ease of notation we will denote in the following by x the m/z variable. In regression, the desired estimate of the regression function is not always given by a conditional mean, although this is most common. Sometimes one wants to obtain a good estimate that satisfies the property that a proportion τ of the conditional distribution of Y with respect to regressors will be above the estimate. For $\tau = 0.5$ this is an estimate of the median. What might be called median regression, is subsumed under the term quantile regression. In our context, most of the signal of interest in a spectrum lies above the baseline which is assumed to be slowly varying, and therefore it seems natural to estimate the baseline by using quantile regression with a small

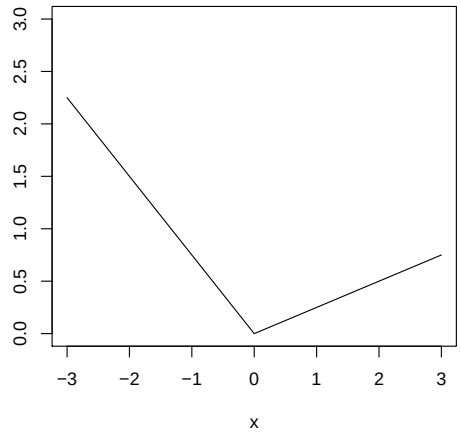


Figure 2.2: The check function with $\tau = 0.26$

value of τ. While there is no criterion for establishing when most of the data lie above the baseline, a cutoff of $\tau = 0.001$ works well. Since the baseline is assumed to be sufficiently smooth, our approach to estimating the background function $B(\cdot)$ in a flexible manner is to represent it as a linear combination of known basis functions $\{h_k,\ k = 1, \ldots, K\}$,

$$B(x) = \sum_{k=1}^{K} \beta_k h_k(x), \tag{1}$$

and then try to estimate the coefficients $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$. Usually the number K of basis functions used in the representation of B should be large in order to give a fairly flexible way for approximating B. Popular examples of such basis functions are wavelets and polynomial splines. A crucial problem with such representations is the choice of the number K of basis functions. A small K may result in a function space which is not flexible enough to capture the variability of the baseline, while a large number of basis functions may lead to serious overfitting and as a consequence to an underestimation of the intensities of the peaks. Traditional way of "smoothing" is through regularization which imposes a penalty on large fluctuations on the fitted curve and this is the approach we will concentrate in this paper. As such, our quantile regression based baseline-correction procedure may therefore be regarded as a method similar to the "peak rejection" approaches; there is, however, no need to detect peaks.

It will be based on a truncated power basis (P-splines) representation of the varying background signal $B(m/z)$ and a L_1-penalization of a regression quantile loss-function.

We start by presenting basic definitions and some background knowledge about regression splines, since not all the readers are familiar with these notions. Basic references are [15] and [17]. Polynomial regression splines are continuous piecewise polynomial functions where the definition of the function changes at a collection of knot points, which we write as $t_1 < \cdots < t_K$. Using the notation $z_+ = max(0, z)$, then, for an integer $p \geq 1$, the truncated power basis for polynomial of degree p regression splines with knots $t_1 < \cdots < t_K$ is

$$\{1, x, \ldots, x^p, (x - t_1)_+^p, \ldots, (x - t_K)_+^p\}.$$

Altough polynomial of degre p regression splines are continuous up to their $p - 1$th derivative, their derivatives of order p will not be in general differentiable at a knot point where the function is defined by different polynomial pieces to the immediate right and left of the knot. When representing an univariate function f as a linear combination of these basis functions as

$$f(x) = \sum_{k=0}^{p} \beta_k x^k + \sum_{j=1}^{K} \beta_{p+j} (x - t_j)_+^p,$$

it follows that each coefficient β_{p+j} is identified as a jump in the p-th derivative of f at the corresponding knot. Therefore coefficients in the truncated power basis are easy to interpret especially when tracking more or less abrupt changes in the regression curve.

The truncated power basis for polynomial of degree p regression splines with knots $t_1 < \cdots < t_K$ may be viewed as a given family of piecewise polynomial functions $\{h_j, j = 0, \ldots, p + K\}$. Assuming the initial location of the knots known, the $K + p + 1$ dimensional parameter vector, $\boldsymbol{\beta}$, describes the $K + p + 1$ necessary polynomial coefficients that parsimoniously represent the function B, i .e. we have $B(x) = H(x)\boldsymbol{\beta}$ where, for x given, $H(x)$ is the matrix whose columns are $h_j(x)$, for $j = 0, \ldots, K + p$. We will consider a two-stage knot selection scheme for adaptively fitting quantile regression splines to the background. An initial fixed large number of potential knots will be chosen at fixed quantiles of the x variable with the intention to have sufficient points at regions where the baseline curve shows rapid changes. Basis selection by non-smooth at zero penalties will then eliminate basis functions when they

are non necessary, retaining mainly basis functions whose support covers regions with sharp features.

Following now Koenker and Bassett [31], define the check function as

$$\rho_\tau(u) = \tau|u|I\{u > 0\} + (1-\tau)|u|I\{u \leq 0\},$$

where $I\{\cdot\}$ is an indicator function (see Figure 2.2). This check function highlights the basic difference between the conditional mean and the conditional quantile function. Here $\tau \in (0,1)$ indicates the quantile of interest. Our quantile spline estimator of B can then be given as the minimizer of

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^{K+p+1}} \sum_{i=1}^{n} \rho_\tau \left(\tilde{Y}(x_i) - H(x_i)\boldsymbol{\beta} \right) + \lambda \sum_{j=1}^{K+p} |\beta_j|.$$

Adapting some recent results from the literature on nonparametric function estimation, we show in the appendix that the resulting estimator adapts to the unknown smoothness of the underlying baseline function, as well as to its unknown identifiablity properties.

Like other nonparametric smoothing methods, the smoothing parameter λ plays a crucial role on determining the trade-off between the fidelity to the data and the penalty. When is too large, there is too much penalty placed on the estimate. As a consequence, the data is oversmoothed. On the other hand, when is too small, we tend to interpolate the data more and this will lead to undersmoothing and underestimation of peaks intensities. The main goal here is to pick a such that the distance between the resulting estimate and the true function is minimized. The major difficulty is that we do not observe the true baseline function. Therefore we cannot directly evaluate the distance. Instead, we should rely on some other proxies. Two commonly used criteria are the Schwarz information criterion [32] (SIC) and the generalized approximate cross-validation criterion [54] (GACV)

$$\text{GACV}(\lambda) = \frac{\rho_\tau \left(\tilde{Y}(x_i) - H(x_i)\hat{\boldsymbol{\beta}}_\lambda \right)}{n - df}$$

where df is a measure of the effective dimensionality of the fitted model. In our implementation we have used the SIC criterion.

Remark 2.1 *The baseline-correction procedure proposed by Eilers [21] may be regarded as a method similar to the quantile regression approach. His procedure is based on the Whittaker smoother [20], which minimizes the following cost function:*

$$\sum_{i=1}^{n} \nu_i \left(\tilde{Y}(x_i) - B(x_i)\right)^2 + \lambda \left(\Delta^d B(x_i)\right)^2, \tag{2}$$

where Y is the analyzed signal, B is a smooth approximation of Y (the baseline), d is the order of differences between adjacent values of B and ν are weights. Weights ν have high values in parts of the signal where the signal analyzed is allowed to affect estimation of the baseline. In all other regions of the signal, values of ν are zero. The positive parameter λ is the regularization parameter and controls the significance of the penalty term $\lambda(\Delta^d B(x_i))^2$, i.e. the higher the value of λ, the smoother the estimated baseline. Because of the asymmetry problem in baseline estimation, the weights should be chosen in a way that will enable "rejection" of the peaks. To achieve this, the weights are assigned as:

$$\nu_i = \begin{cases} p & \text{if} \quad \tilde{Y}(x_i) > B(x_i) \\ 1-p & \text{if} \quad \tilde{Y}(x_i) > B(x_i), \end{cases} \tag{3}$$

where $0 < p < 1$. The positive deviations from the estimated baseline (peaks) have low weights while the negative deviations (baseline) obtain high weights. There is, however, a problem of simultaneous determination of weights (ν) and baseline (B). Without the weights it is impossible to calculate the baseline and without the baseline it is impossible to determine the weights. This problem is solved iteratively, i.e., in the first iteration all weights get the same value, i.e. unity. Using these weights, a first estimate of the baseline is calculated. Iterating between calculation of the baseline and setting weights, gives an estimate of the baseline in the next iterations. The use of p close to zero and large λ enables baseline estimation. However, experience has shown that the above procedure requires many iterations to converge to a good baseline estimation and does not always converges.

2.4 Application and illustrations

The goal of this subsection is to illustrate the quantile regression based process of finding a baseline curve to real and simulated proteomics MALDI spectral data and to make some comparisons of the results with those obtained with the automated peak rejection method proposed recently in [13].

To estimate the baseline we have used two procedures; the first one implements quantile regression splines (QRS) without further constrains or knots removal while the second one uses a slight modification of a very attractive implementation of our ℓ_1-penalized version of quantile regression splines, namely the constrained B-spline smoothing method (COBS) of He and Ng [25] which extends earlier work of Koenker *et al.* [32]. The spline basis used to model the baseline is based on quadratic splines with usually a maximum number $K = 60$ of equally spaced quantile knots used in the representation of B. The final number of knots is selected via the AIC criterion.

Peaks detection in biological samples

We consider samples of nipple aspirate fluid (NAF) from breast cancer patients and from healthy women (for a complete description, see [13]). These data are available from the web site `http://bioinformatics.mdanderson.org/pubdata.html` and have been used by Coombes *et al.* [13] to look at the reproducibility of their method in detecting and identifying relevant peaks, since the 24 spectra of the NAF data were independently derived from the same starting material. The method of the above cited authors includes a processing step that determines which peaks found in individual spectra should be identified as representing the same biochemical substance across spectra.

In order to illustrate the effects of our quantile regression based baseline removal process, we have run the procedures of Coombes *et al.* [13], with the only difference being in our way for estimating the baseline (using QRS or COBS) instead of the monotone local minimum curve fitting procedure implemented by Coombes *et al.*. A comparison of the results obtained is illustrated by the histograms displayed in Figure 2.3. We found 174 distinct peaks across the 24 spectra when using either Coombes *et al.* procedure or the QRS quantile regression splines procedure for baseline removal. When using COBS, 181 distinct peaks has been found.

Of course, it is clear that the number of peaks found per spectrum is not, by itself, an adequate measure of the quality of a baseline removal algorithm. It is important to ascertain if the peaks being found by the algorithm correspond to real phenomena in the spectra. While we do not have knowledge of the "true" peaks in the spectra used, one can look at the reproducibility of the method, since we have 24 spectra

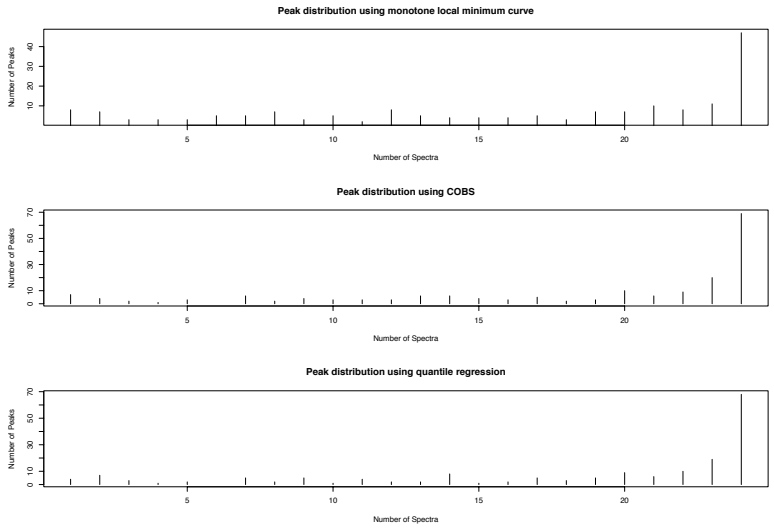


Figure 2.3: Histograms of the number of peaks found in multiple spectra.

independently derived from the same starting material. Table 1 gives the distribution of the number of peaks for the three methods. For example when using COBS, 69 of the 181 peaks were present in all 24 spectra. Moreover, 114 peaks were found in at least 20 spectra, 137 peaks were found in at least 15 spectra and 159 peaks were found in at least 10 spectra.

Table 1: `Distribution of the number of peaks.` Number of peaks present in all 24 spectra, at least 20, at least 15 and at least 10 spectra.

	all 24	at least 20	at least 15	at least 10
Monotone Local Minimum	47	83	106	130
QRS (with 60 knots)	68	112	128	145
COBS (with 60 initial knots)	69	114	137	159

These results tend to present substantial evidence that the removal of the baseline estimated by our algorithm finds most of the true, reproducible peaks in the data.

However, in order to also judge how well the baseline is estimated, we have conducted a small simulation study, described in the the next subsection.

Quality of the baseline estimate

To illustrate the quality of the baseline estimate, we need spectrum samples with a known baseline. To do so we have used the following procedure. Starting from samples of the NAF data set, we first estimate by our COBS splines procedure (with 60 initial knots) the baseline (denoted by B_0) of the average spectrum and the associated baseline corrected denoised spectrum (denoted by N_0S_0). Next we build 24 new spectra as noised replicates of $B_0 + N_0S_0$ with noise similar to that of the original sample. One may then consider that all spectra within this artificial data set have B_0 as a common baseline.

We then have run the three baseline estimation procedures over the simulated data. Both spline based procedures were used with 60 knots. For each method we compute the square average bias, the empirical variance and the mean square error at each point of the grid of m/z values and report their average value (over the grid). We give also the relative mean absolute error (see [9]) of reconstruction.

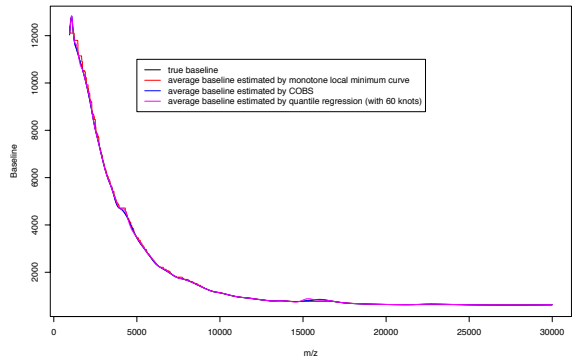


Figure 2.4: Comparison of estimated baseline on simulated data.

Table 2 reports the results obtained for the three methods. COBS appears to be the best method. The results obtained by QRS are slightly better than the ones of the estimation of the baseline by a monotone local minimum curve. Note that the

COBS procedure seems to be more variable with respect to the other two over the simulations, probably due to the fact that the knots are chosen adaptively at each run. However, the larger variance is compensated by a noticeable reduction in bias. Figure 2.4 illustrates the behavior of the three procedures. Note that the "true" baseline and the one estimated by COBS almost coincide.

Table 2: `Simulated example.` Comparison of the baseline estimates.

	bias^2	Var	MSE	RMAE
Monotone Local Minimum	1 346.78	8.40	1 347.63	0.0146
QRS (with 60 knots)	1 148.20	20.15	1 168.35	0.0063
COBS (with 60 initial knots)	317.15	67.66	384.81	0.0024

3 Wavelet transform and multi-scale robust spectra alignment

Biologically significant comparisons and conclusions are all based on the alignment results of spectra where the ultimate goal is to identify differentially expressed proteins in samples of diseased and healthy individuals. MS spectra alignment is difficult even after instrument calibration with internal markers because the mass errors vary with m/z in a nonlinear fashion as a result of experimental and instrumental complexity and data variation.

In this section, we introduce a new alignment algorithm that uses the wavelet transform. We formulate the problem of aligning two spectra as a wavelet based non-rigid registration problem and solve it using a robust point matching approach. To align multiple spectra, we propose and justify a nonlinear wavelet denoising averaging method to estimate a common spectrum as a standard using all individual spectra. Once the standard common spectrum is found, the multiple spectra alignment problem is simplified as pairwise alignment problem and is solved by using the robust point matching approach sequentially (i.e. each time we align only one spectra to the standard one).

Before introducing our framework, we briefly review few methods that have been

recently proposed for addressing the alignment problem for MS spectra. Randolph and Yasui [43] have also used wavelets to represent the MS data in a multiscale framework. Within their framework, using a specific peak detection procedure, they first align peaks at a dominant coarser scale from multiple samples and then align the remaining peaks at a finer scale. However one may question if representing peaks at multiple scales is biologically reasonable, i.e., if peaks at coarse scales really correspond to true peptides. A similar in spirit procedure has been also developed by Coombes *et al.* [13]. Johnson *et al.* [28] assume that the peak variation is less than the typical distance between peaks and use a closest point matching method in peak alignment. The applicability of their method is limited by the data quality and it cannot handle large peak variation or false positive peak detection results. A recent method, to which our method will latter be compared, is the nonparametric warping model with spline functions to align MS spectra proposed recently by Jeffries [27]. While the idea of using smoothing splines to model the warping function is interesting it is unclear if a smooth function with second order regularities is precise enough to describe the nonlinear shift of MS peaks encountered in practice. Tibshirani *et al.* [48] applied a hierarchical clustering method to construct a dendrogram of all peaks from multiple samples. They cut off the dendrogram using a predefined parameter and clustered the remaining branches into different groups. They then considered the centers of these groups as common peaks and aligned every peak set with respect to the common peaks. The implicit assumption behind their approach is that around each of the common peaks, the observed peaks from multiple samples obey a certain kind of distribution with the mean equal to the location of the common peak. The assumption agrees well with the motivation of peak alignment. However, the cut-off parameter and the final clustering results can be influenced by changing a few nodes in the dendrogram, while some noisy points or outliers (e.g., caused by false positive peak detection results) often cause such changes, as it is shown in the recent paper by Yu *et al.* [52]. To our knowledge, frameworks that are closest to the one we propose in this work are the recent approaches proposed by Saussen *et al.* [47] and Yu *et al.* [53]. Both use a robust point matching algorithm to solve the alignment problem but an implicit assumption in Saussen *et al.*'s approach is that there exists a one-to-one correspondence among peaks in multiple spectra while Yu *et al.* [53] use a super set method to calibrate the alignment. To end this subsection let us mention also that Sauve and Speed [3] also

discusses alignment methods that address proteomic data yet provide few details and no software for their implementation.

3.1 Averaging individual denoised spectra

The idea of using an average spectrum to get around the peak matching problem is not new and has been suggested by others (Morris *et al*. [36]). In their paper however, Morris and colleagues assume that all individual spectra are already well calibrated and justify their procedure by the law of large numbers. Our method takes into account the fact that the mass errors among individual spectra may vary with m/z in a nonlinear fashion and theoretically shows the benefits in denoising each individual spectrum before averaging. When spectra are well calibrated the two procedures are equivalent. Clearly, the average does not preserve the total intensity of individual peaks, but the point here is to reduce the redundance among multiple peaks that corresponding to the same peptide.

Experience shows that generally the m/z axis shift of peaks is relatively small, approximately around 0.1% to 0.2% of the m/z value. Such an assumption is actually implicit behind the motivation of peak alignment. Based on these facts, and for the sake of simplicity denoting by t the equidistant m/z values, re-scaled to vary within the interval $[0, 1]$, we assume that a reference spectrum is modeled as

$$Y_i^0 = f(t_i^0), \quad t_i^0 = i/n,\ i = 1, \ldots, n, \tag{4}$$

where $f(t) = B(t) + NS(t)$. Due to the variation of the m/z values, we observe a set of M spectra

$$Y_i^k = f(t_i^k) + \epsilon_i^k, \quad i = 1, \ldots, n,\ k = 1, \ldots, M, \tag{5}$$

where the ϵ_i^k are independent standard normal random variables with a common standard deviation σ and where the design points t_i^k are randomly shifted values of the reference values t_i^0, i.e.

$$t_i^k = t_i^0 + \delta_i^k, \quad i = 1, \ldots, n,$$

with $\boldsymbol{\delta}^k$ a random vector, independent of the signal noise such that

$$\mathbb{E}(\boldsymbol{\delta}^k) = 0,$$

$$\mathrm{var}(\delta_i^k) \leq C/n^2,$$
$$\delta_i^k - \delta_{i+1}^k \leq \frac{1}{n}, \quad i = 1, \ldots, n,$$

and

$$\delta_1 \geq -\frac{1}{n} \quad \delta_n \leq 0.$$

The above model for the random shifts is in agreement with the belief that the m/z shift of peaks in MS spectra is relatively small. The last two constraints on the components of δ^k are needed to ensure that the order of the signal points is not changed. Note that in such a model the peaks are moved by different amounts to the left or the right and the relative distance between peaks is changed as well.

Instead of using now a TI wavelet transform for denoising the observed signal, each observed signal will be analyzed via the discrete wavelet transform and the lack of invariance of the DWT will be actually an important part of our estimator of f. By obtaining different reconstructions via wavelet denoising for each observed signal $\mathbf{Y}^k$, $k = 1, \ldots, M$ and averaging over them, any dependence of peak placement will be removed or reduced.

Let ϕ and ψ represent the mother and mother wavelets and assume that they are compactly supported. Denoting by ϕ_{jl} and ψ_{jl} the translations and dilations of ϕ and ψ, the signal f can be expressed as an infinite series

$$f(t) = \sum_{\ell=0}^{2_0^j-1} \xi_{j_0,\ell}\phi_{j_0\ell}(t) + \sum_{j=j_0}^{\infty}\sum_{\ell=0}^{2^j-1} \theta_{j,\ell}\psi_{j\ell}(t),$$

where the coefficients $\xi_{j_0,\ell}$ represent the smooth part of f and the $\theta_{j,\ell}$ represent the detailed structure of f. Let W be the wavelet transform matrix corresponding to the choice of the wavelet function ϕ and ψ and denote by

$$\boldsymbol{\theta} = (\xi_{j_0,0}, \ldots, \xi_{j_0,2^{j_0}-1}, \theta_{j_0,1}, \ldots, \theta_{J-1,2^{J-1}-1})^T,$$

the vector of wavelet coefficients of f in (4). If n is a dyadic integer which we will assume hereafter, then the DWT estimate of $\boldsymbol{\theta}$ based on data $\mathbf{Y}^k$ is

$$\tilde{\boldsymbol{\theta}}^k = \frac{1}{\sqrt{n}} W\mathbf{Y}^k.$$

To get a denoised estimator of f one may use a term by term "soft" threshold rule

$$\eta(\tilde{\theta}_{j,\ell}, \lambda) = \frac{\tilde{\theta}_{j,\ell}}{|\tilde{\theta}_{j,\ell}|}\left(|\tilde{\theta}_{j,\ell}| - \lambda\right)_+,$$

where the threshold λ can be for example the universal threshold $\sigma\sqrt{2\log n}$. With this notation, our proposed estimate of the signal f based on the set of the observed signals $\mathbf{Y}^k$, $k = 1, \ldots, M$, can be symbolically written as

$$\hat{f} = \frac{1}{M}\sum_{k=1}^{M} W^{-1}\eta(W\mathbf{Y}^k, \lambda). \tag{6}$$

The optimal rate of convergence, when measured via the mean integrated squared error, in estimating a function in a Hölder space with unknown parameter α is $O(n^{-2\alpha/(2\alpha+1)})$ as can be see in the following theorem, whose proof is given in the appendix.

Theorem 3.1 *Suppose a signal such as the one given in expression (4) is in the Hölder ball $\Lambda^\alpha(R)$, $R > 0$, with parameter α and is observed as in (5) with the variances of the $\delta_i^k \leq C/n^2$. Let $\hat{f}$ be the estimator at (6). Let ψ have r vanishing moments. Then for $\alpha \in (0, r]$,*

$$\sup_{f\in\Lambda^\alpha(R)} \mathbb{E}\left(\|f - \hat{f}\|_2^2\right) \leq Cn^{-2\alpha/(2\alpha+1)}.$$

Thus, our estimator retains the optimal qualities of wavelet thresholding as if the observed data were observed on the reference grid of m/z values. The same is not true for the mean spectrum defined by Morris and his colleagues, which may be written in our notation as

$$\hat{f}_M = W^{-1}\eta(W\bar{\mathbf{Y}}, \lambda),$$

since the transform η is not linear.

To end this subsection, we have run some Monte Carlo simulations to compare these two denoising procedures on several test functions specifically chosen for their possession of jumps and steep changes. The functions are depicted in figure 3.5 together with a typical misaligned version of them. They have been scaled so that each has a standard deviation of 1.

To compare the reconstruction errors of the two estimators, Monte Carlo simulations were performed. For each function, a set of 24 misaligned version of it were generated and Gaussian noise was added at signal to noise ratio of 5. Sample sizes ranged from $n = 64$ to $n = 512$. The reconstruction errors were estimated from 50 Monte Carlo replications of this experiment. For each denoising procedure the soft SureShrink threshold rule with Symmlets of order 5 was used.

Averaging individual denoised spectra					Mean spectrum denoising				
n	bias^2	Var	MSE	RMAE	n	bias^2	Var	MSE	RMAE
Blip					Blip				
64	0.4712	0.0163	0.4875	0.0265	64	0.7809	0.0254	0.8063	0.0376
128	0.2293	0.0105	0.2398	0.0171	128	0.3029	0.0352	0.3380	0.0213
256	0.2696	0.0048	0.2745	0.0140	256	0.2161	0.0111	0.2272	0.0156
512	0.1465	0.0028	0.1492	0.0102	512	0.2092	0.0073	0.2166	0.0125
Corner					Corner				
64	0.1540	0.0116	0.1657	0.0157	64	0.2004	0.0158	0.2161	0.0179
128	0.0659	0.0057	0.0716	0.0099	128	0.0803	0.0105	0.0907	0.0110
256	0.0452	0.0036	0.0489	0.0079	256	0.0827	0.0065	0.0891	0.0112
512	0.0324	0.0022	0.0347	0.0061	512	0.0378	0.0061	0.0439	0.0066
Angles					Angles				
64	0.2329	0.0146	0.2475	0.0400	64	0.3709	0.0218	0.3927	0.0514
128	0.0878	0.0100	0.0978	0.0198	128	0.1132	0.0190	0.1322	0.0265
256	0.0643	0.0050	0.0694	0.0162	256	0.0769	0.0086	0.0855	0.0189
512	0.0559	0.0023	0.0582	0.0147	512	0.0658	0.0054	0.0712	0.0175
Waves					Waves				
64	0.2836	0.0284	0.3120	0.0335	64	0.6503	0.0153	0.6655	0.0517
128	0.1412	0.0174	0.1586	0.0220	128	0.6371	0.0083	0.6453	0.0469
256	0.0164	0.0099	0.0263	0.0076	256	0.0115	0.0173	0.0288	0.0066
512	0.0050	0.0050	0.0099	0.0041	512	0.0112	0.0103	0.0215	0.0064

Table 3: Simulation results for the two denoising procedures on 4 types of signals (Blip, Corner, Angles and Waves). Each signal has been randomly misaligned 24 times and an additive noise with signal to noise ratio equal to 5 was added on each misaligned signal. The experiment was replicated 50 times and the table displays the error statistics for each method and each type of signal. The left displays the results for the averaging individual denoised spectra methods while the right table summarizes the results obtained by mean spectra denoising.

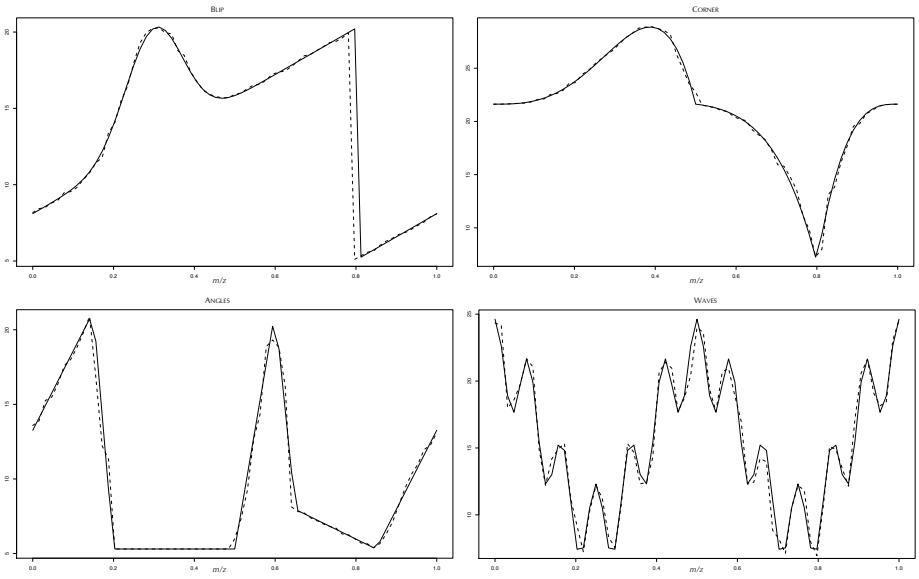


Figure 3.5: The four test functions (solid line) together with a typical misaligned version of them (dashed line) used in the simulation.

Results of these simulations are given in Table 3. As we can see, in all cases, the method we have proposed outperforms the mean spectrum denoising procedure in terms of mean squared error and relative absolute mean error statistics. Note also that the two estimators have similar errors when the sample size increases.

3.1.1 Alignment

Once a set of MS spectra is processed by the DWT algorithm described above, the resulting average denoised spectrum $\hat{f}$ will play the role of a "global" anchor to align them together. Before proceeding to the alignment, the mean spectrum will be baseline corrected using our regression quantile procedure and then each spectrum will be denoised and baseline corrected with the mean spectrum smooth baseline determined above. The alignment process is again based on the wavelet transform, avoiding the determination and quantification of individual peaks, as it is usually done, and can be decomposed into two steps: (1) alignment of two peak sets; (2) alignment of multiple peak sets. The details are described as follows.

Spectra alignment consists in finding, for each observed spectrum, a warping function in order to synchronize all spectra before any other statistical inferential

procedure. In what follows, the terms alignment, warping, registration or matching will also be used to refer to the synchronization of set of signals. Matching two functions can be done by aligning individual locations of corresponding structural points (or landmarks) from one curve to another. Previous approaches to landmark-based registration in a statistical setting include Kneip and Gasser [29], Gasser and Kneip [24], Ramsay and Li [42], Munoz Maldonado *et al*. [37] and Bigot [7]. For landmark-based matching one needs to detect the landmarks of a set of signals from discrete (noisy) observations. The estimation of the landmarks is usually complicated by the presence of noise whose fluctuations might give rise to spurious estimates which do not correspond to structural points of the unknown signals. Then, it is necessary to determine the landmarks that should be associated. This step is further complicated by the presence of outliers and by the fact that some landmarks of a given curve might have no counterpart in the other curves. In this subsection, we will use the scale-space approach proposed in Bigot [6] to estimate the landmarks of a noisy function. This method is based on the estimation of the significant zero-crossings of the continuous wavelet transform of a noisy signal, and on a new tool, the structural intensity, proposed in Bigot [6] to represent the landmarks of a signal via a probability density function. The main modes of the structural intensity correspond to the significant landmarks of the unknown signal. In a sense, the structural intensity can be viewed as a smoothing method which highlights the significant features of a signal observed with noise.

We will first consider the alignment of two given spectra by a fast and automatic method based on robust point matching of the structural intensities associated to the significant landmarks in the two curves. Its computational cost only depends on the number of landmarks and is therefore very low.

3.1.2 Scale-space estimation of the significant landmarks

We briefly recall here the scale-space approach proposed in Bigot [6], to automatically estimate the landmarks of an unknown signal. This approach is based on the detection of the significant zero-crossings of the continuous wavelet transform of a signal observed with noise.

Let $f \in L^2(\mathbb{R})$ and $\psi = (-1)^r \theta^{(r)}$ where $r \geq 1$ is the number of vanishing moments of the wavelet ψ, and θ is a smooth function with a fast decay such that

$\int_{-\infty}^{\infty}(\psi(u))^2 du = 1$. Then, by definition, the continuous wavelet transform of f at a given scale $s > 0$ is:

$$W_s(f)(x) = \int_{-\infty}^{+\infty} f(u)\psi_s(u-x)du \text{ for } x \in \mathbb{R},$$

where $\psi_s(u) = \frac{1}{\sqrt{s}}\psi(\frac{u}{s})$.The term *zero-crossings* is used to describe any point (z_0, s_0) in the time-scale space such that $z \mapsto W_{s_0}(f)(z)$ has exactly one zero at $z = z_0$ in a neighborhood of z_0. We will call *zero-crossings line* any connected curve $z(s)$ in the time-scale plane (x, s) along which all points are zero-crossings. Now, note that if f is C^r in an interval $[a, b]$, then for all $x \in]a, b[$

$$\lim_{s\to 0} \frac{W_s(f)(x)}{s^{r+1/2}} = Kf^{(r)}(x), \text{ where } K = \int_{-\infty}^{+\infty} \theta(u)du \neq 0. \tag{7}$$

Hence, equation (7) shows that at fine scales the zero-crossings of $W_s(f)(x)$ converge to the zeros of $f^{(r)}$ in $]a, b[$ (if any) . Thus, if the zero-crossings propagate up to fine scales, one can find the locations of the extrema (resp. the points of inflexion) of a function by following the propagation at small scales of the curves $z(s)$ when ψ has $r = 1$ (resp. $r = 2$) vanishing moment(s). This is the case when θ is a Gaussian since it is well known that scale-space representations derived from Gaussian guarantee that the zero-crossings lines are never interrupted when s goes to zero (see [6] for further references).

In Figure 3.6 (b), the zero-crossings of a simulated signal are computed for a Gaussian wavelet with $r = 1$ vanishing moment. One can see that the zero-crossings lines are never interrupted and converge to the extrema of the signal when the scale s goes to zero. When a smooth signal is observed with noise, its local extrema can be detected by estimating the significant zero-crossings of its continuous wavelet transform. A simple hypothesis testing procedure has been developed in [7] to estimate the zero-crossings lines of a signal at various scales (see [7] for further details). However, this procedure only gives a visual representation that indicates "where" the landmarks are located, but there is generally no analytical expression of these lines in a closed form. The structural intensity is a new tool introduced in [6] to identify the limits of the zero-crossings lines when they propagate to fine scales. The structural intensity method consists in using the zero-crossings of a signal at various scales to compute a "density" whose local maxima will be located at the landmarks of f. More

precisely, the structural intensity is defined to be:

$$G_z(x) = \sum_{i=1}^{\hat{p}} \int_{\hat{s}_i^1}^{\hat{s}_i^2} \frac{1}{s} \theta \left(\frac{x - \hat{z}_i(s)}{s} \right) ds, \tag{8}$$

where $\hat{p}$ is the number of estimated zero-crossings lines $\hat{z}_i(.), i = 1, \ldots, \hat{p}$ and $[\hat{s}_i^1, \hat{s}_i^2]$ represent the supports of these lines in the time-scale plane. The *landmarks of the unknown signal* f are then defined as the *local maxima* of $G_z(x)$ on $[0, 1]$ (we will usually refer to these local maxima as the modes of G_z). The structural intensity is therefore a tool to identify the limits of the lines $\hat{z}_i(.), i = 1, \ldots, \hat{p}$ in the time-scale plane. In Figure 3.6(c), one can see that the local maxima of the structural intensity correspond to the extrema of the signal (note that in Figure 3.6(c), the structural intensity is computed with the true zero-crossings). One can also remark that for estimating only the local maxima (resp. minima) of a signal f, one only keeps in the formula (8) the zero-crossings $z(s)$ such that $W_s(f)(z(s)) > 0$ (resp. < 0) for all $z \in [z(s) - \epsilon, z(s)[$ with ϵ sufficiently small. In Figure 3.6(e), we give an example of an estimation of the zero-crossings of a noisy signal (compare the quality of this estimation with Figure 3.6(b)). We also display in Figure 3.6(f) the structural intensity of the estimated zero-crossings that correspond to the local maxima of a function. One can see that the modes of this structural intensity correspond to the significant local maxima of the noisy signal shown in Figure 3.6(d).

3.1.3 Spectra registration via the alignment of the structural intensities

One of the first issues encountered by landmark-based matching methods is the correspondence problem between two sets of features. This step is usually performed manually which can be tedious and prone to error. Let G_{z1} and G_{z2} denote the structural intensities of the (estimated) zero-crossings of two signals f_1 and f_2. It has been observed in [7] that the features that one would align manually correspond to the modes of G_{z1} and G_{z2} whose shape and height are similar. To automatically solve this correspondence problem, Bigot [7] has proposed a new technique, called dynamic correspondence, to automatically compute a warping function that aligns the common modes of G_{z1} and G_{z2}. The computational cost of this approach depends only on the number of estimated landmarks which is usually very small, and dynamic correspondence is therefore a very fast alignment technique (see [7] for further details).

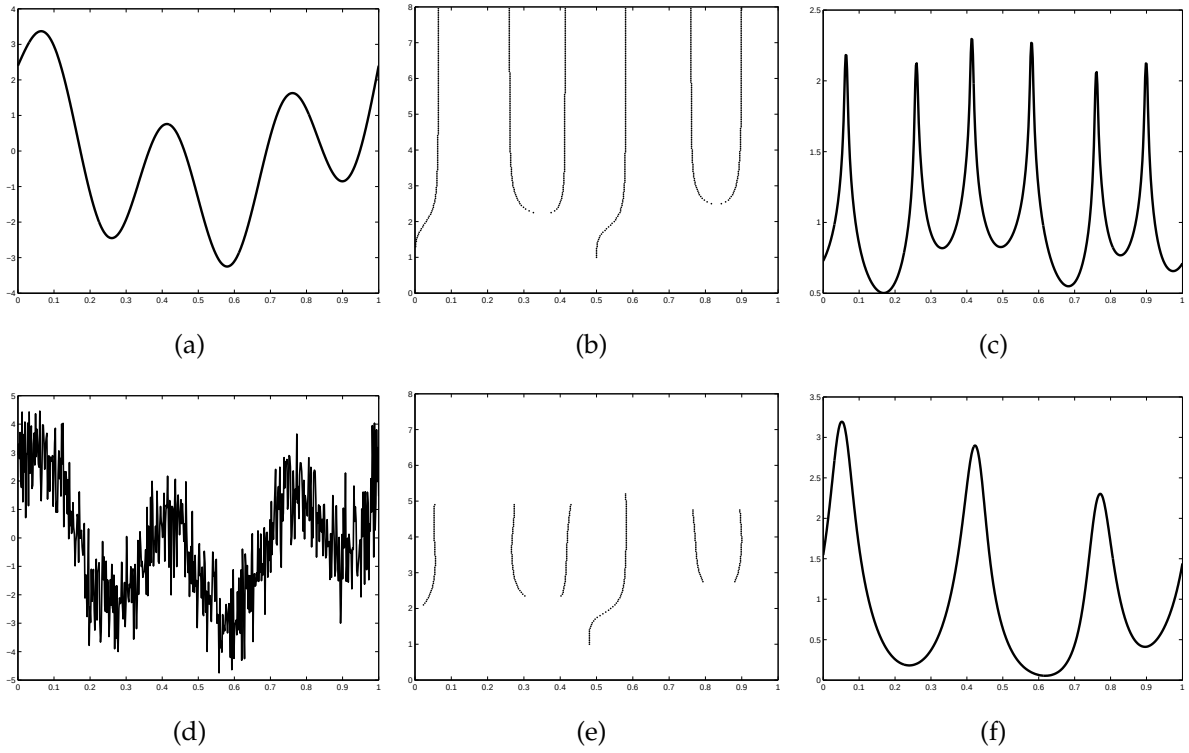


Figure 3.6: (a) Simulated signal with various extrema, (b) Zero-crossings computed for a Gaussian wavelet with $r = 1$ vanishing moment: the vertical axis gives $-\log_2(s)$, (c) Structural intensity of the zero-crossings, (d) Simulated signal + Gaussian noise, (e) Estimation of the zero-crossings, (f) Structural intensity of the estimated zero-crossings that correspond to local maxima.

Note that this approach handles the case when the two structural intensities do not have the same number of modes, and discards the landmarks of a given curve that do not have counterpart in the other curve. Dynamic correspondence is thus a robust point matching technique for one-dimensional curves. Once the average denoised spectrum $\hat{f}$ has been computed, we simply use dynamic correspondence to register each denoised and baseline corrected spectrum onto the "template" curve $\hat{f}$.

An illustrative example

Our purpose here is to test our scale-space alignment approach and to compare it with the smoothing spline nonparametric approach developed recently by Jeffries [27]. Here, we use replicate data coming from a real study to ensure that the sample

variation is minimal. These data are described in the recent paper by Jeffries and are available from `http://krisa.ninds.nih.gov/alignment/`. More precisely, as part of a larger study examining proteomic spectra from healthy individuals and those with multiple sclerosis, reference samples from a large pool of serum were included as part of a quality control procedure. As patient and control samples were processed, a few spectra were consistently drawn from this common, fixed reference pool and analyzed to alert investigators to deviations related to sample processing. Ideally, all the spectra from the reference samples should look very similar. Samples were processed on six distinct days using identical calibration procedures, personnel, equipment, and sample handling techniques. Samples from the first four days were processed within a single week while samples for the last two days were processed approximately 2 and 3 months later. We have chosen as reference spectrum a spectrum from the third day (4/2/2003) and a spectrum for the fifth day (5/20/2003) to be aligned. Both spectra are displayed in Figure 3.7. The graph indicates the data produced on the fifth day are not well aligned with those of the third.

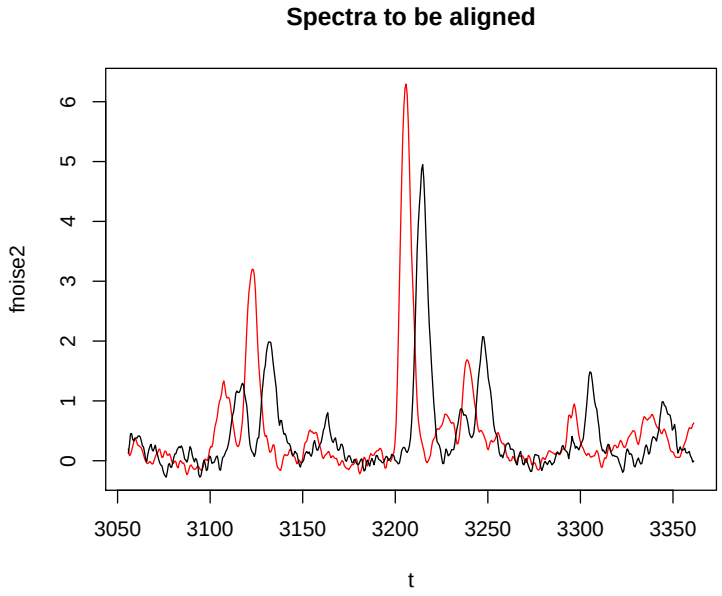


Figure 3.7: Two spectra requiring alignment shown on a restricted range of m/z values.

We compare our alignment algorithm to the algorithm proposed by [27], which, for the sake of completeness, we now briefly recall. As for our method its algorithm begins

with a single spectrum, a set of N peaks with associated m/z values (denoted as p_i), and a set of target m/z values for these peaks (denoted m_i). If $y_i = p_i/m_i$ denote the ratio of the original mass value to its target value, then given $\{p_1, m_1, \ldots, p_N, m_N\}$ and any positive value λ, the algorithm finds the unique function $f_\lambda(m)$ that minimizes

$$\sum_{i=1}^{N} (y_i - f(p_i))^2 + \lambda \int (\ddot{f}(p))^2 dp,$$

among all functions $f(m)$ with two continuous derivatives where m denotes any mass value in the range of interest. The parameter λ is determined by cross validation. Once the function f_λ is determined the data are recalibrated by computing the recalibrated $(m/z)_{recal}$ values associated to the ones of the reference spectrum $(m/z)_{ori}$ as

$$(m/z)_{recal} = \frac{(m/z)_{ori}}{f_\lambda((m/z)_{ori})}.$$

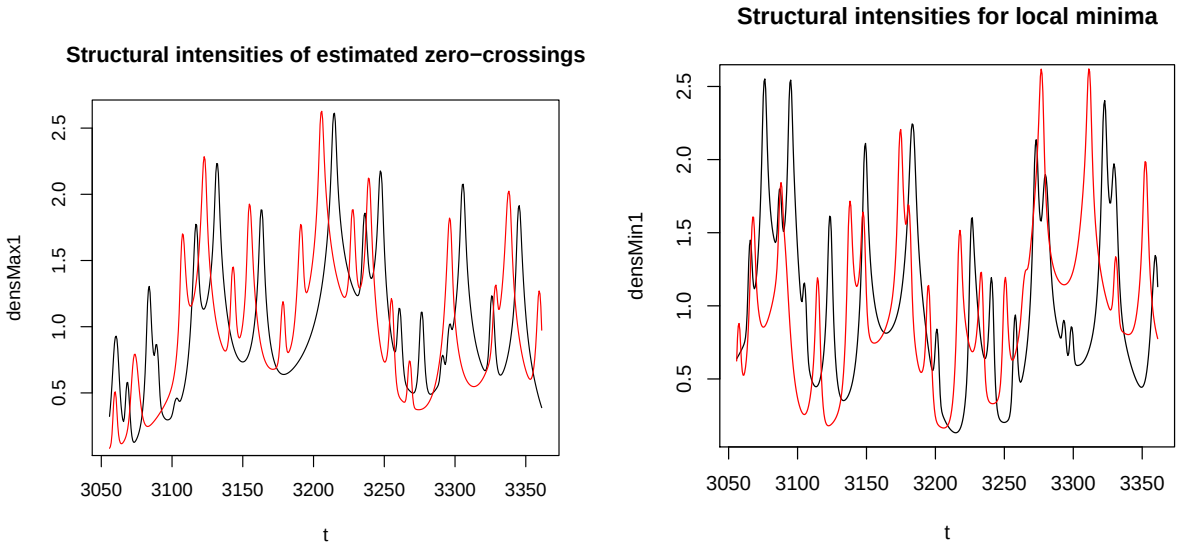
Linear interpolation of the recalibrated masses that are closest to the original mass is then used to obtain a new intensity for the original mass. Before proceeding to our comparison, note however that this method is only based on the m/z information.

We now present our illustrative example. The landmarks of both spectra were estimated by our detection algorithm of the significant zero-crossings of the continuous wavelet transform of the signals. Figure 3.8 displays the resulting structural intensities of the zero-crossings for the local maxima and local minima in both spectra leading to the appropriate landmarks for multi-scale alignment.

Once landmarks are estimated, we proceed to the alignment of both spectra by our robust point matching algorithm. The resulting warping function and the alignment obtained are displayed in Figure 3.9.

Finally to compare our method with the one by Jeffrie's we have also applied the smooth warping procedure of Jeffrie to the same set of spectra. The result is displayed in Figure 3.10 where one can clearly see that the peaks around the m/z values of 3220 Da and 3250 Da are not properly aligned.

To conclude, the above results show that our multi-scale based alignment method seems more robust against noise than the smoothing spline method. In addition, in our approach we use jointly both m/z information and intensity values in the alignment, which is not the case for most previous approaches, including the smoothing spline method but also Tibshirani's hierarchical clustering method.

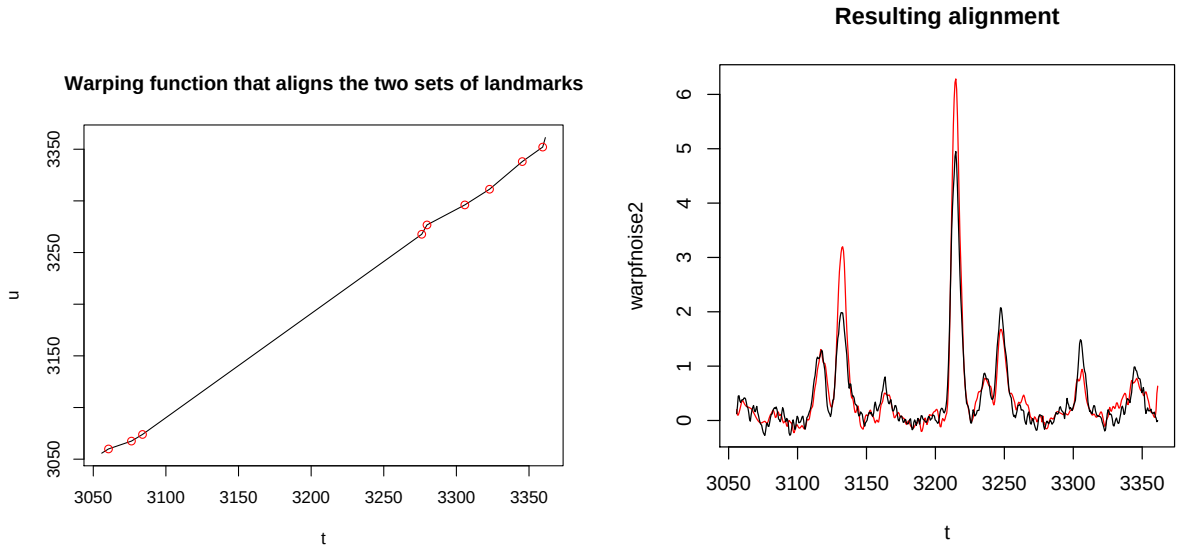


(a) (b)

Figure 3.8: Left: Structural intensity of the estimated zero-crossings that correspond to local maxima; Right: Structural intensity of the estimated zero-crossings that correspond to local maxima.

4 Wavelet based Functional analysis of Variance and significant features

One of the most frequently asked questions in proteome-wide studies involving MS data is how to find a list of biomarkers, defined as proteins differentially expressed between two groups of samples. Conventional methods first detect the peaks in protein spectrum for each sample after calibrations and then align these peaks across the samples. Next they find peaks related to a mass-to-charge ratio (m/z) that discriminate groups based on testing of peak intensities. There are two major concerns with this approach. First, for MALDI or SELDI-TOF MS data, the peak detecting methods are controversial because they are ad-hoc and the results can vary due to user defined parameters. Second, since a large number of proteins, potentially correlated with each other by unknown fashion, are tested simultaneously with a relatively small number of samples, it is expected to have a lot of false positives in detecting statistically significant biomarkers if one performs a series of standard univariate ANOVA tests



(a) (b)

Figure 3.9: Left: Warping function estimated using the landmarks (denoted as circles); Right: Resulting alignment of the two spectra.

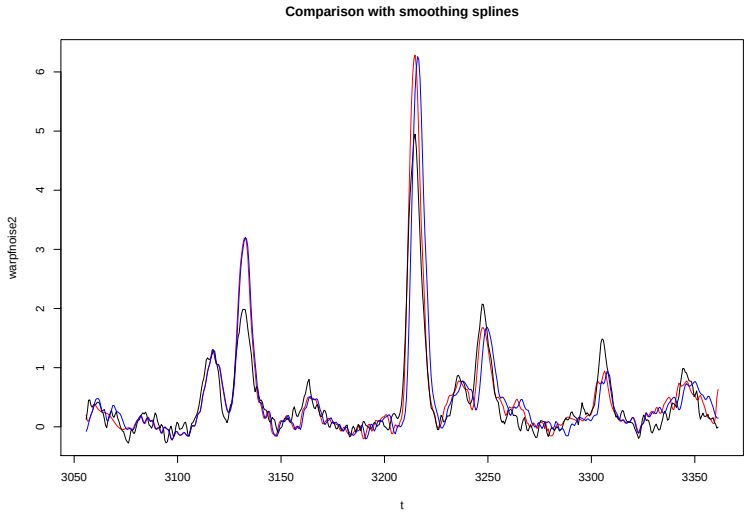


Figure 3.10: Comparison of alignments obtained by the two methods.

to compare a set of spectra at each specific biomarker due to a serious multiplicity problem given the usually large number of simultaneous tests. Ignoring multiplicity

leads to an uncontrolled overall Type I error while, for example, Bonferroni type procedures in which the p-value cutoff for each test is divided by the total number of hypotheses to be tested, are known to be overly conservative and to yield an extremely low power.

A more powerful and less ambiguous approach to high-dimensional hypothesis testing is to consider each spectrum as the basic unit in the analysis and to focus on testing for significant treatment effects in a functional data model. Such an approach is available from adaptations of Fan and Lin [23], and more recently of Abramovich et al. [2] discussing a functional analysis of variance (FANOVA) suitable for testing fixed effects. Since spectra are composed by many peaks, their sparse representation in the wavelet domain allows significant reduction in dimensionality of the original functional data and we will adopt hereafter the wavelet-based testing procedures of Abramovich *et al.* [2] of a zero signal in a "signal+white noise" model for testing in the fixed-effects FANOVA.

First we describe the following white noise(or diffusion) version of the FANOVA model we are going to consider hereafter. We will assume that the observed spectra, baseline-corrected, normalized and properly aligned by means of the methods described in the previous sections, are modeled as sample paths of a stochastic processes driven by

$$dY_{ij}(t) = S_i(t)\, dt + \epsilon\, d\mathcal{W}_{ij}(t), \quad i = 1, \ldots, r; j = 1, \ldots, n_i \quad t \in [0,1], \tag{9}$$

where $\epsilon > 0$ is the diffusion coefficient, r is a finite integer indicating the number of treatments to be compared, n_i the number of spectra recorded for treatment i, S_i the (unknown) mean spectrum in population i and $\mathcal{W}_{ij}$ are independent standard Wiener processes and where the m/z range has been re-scaled to $[0,1]$.

In practice, obviously, one always observes *discrete* data samples of size n with a noise variance σ^2, but from the well-known results of Brown and Low [10], under some general conditions, the corresponding discrete model is asymptotically equivalent to the white noise model (9) with $\epsilon = \sigma/\sqrt{n}$.

Each of the r average spectra S_i in model (9) admits the following unique decomposition

$$S_i(t) = m_0 + a_i + \mu(t) + \gamma_i(t) \quad i = 1, \ldots, r; \quad t \in [0,1], \tag{10}$$

where m_0 is a constant function (the *grand mean*), $\mu(t)$ is either zero or a non-constant function of t (the *main effect* of t, i.e. a common spectrum if no difference exists among treatments), a_i is either zero or a non-constant function of i (the *main effect* of i) and $\gamma_i(t)$ is either zero or a non-zero function which cannot be decomposed as a sum of a function of i and a function of t (the *interaction* component). The components of the decomposition (10) satisfy the following orthogonal (identifiability) conditions

$$\int_{[0,1]} \mu(t)\, dt = 0, \quad \sum_{i=1}^{r} a_i = 0, \tag{11}$$

$$\sum_{i=1}^{r} \gamma_i(t) = 0, \quad \int_{[0,1]} \gamma_i(t)\, dt = 0, \quad \forall\, i = 1, \ldots, r; \;\; t \in [0,1]. \tag{12}$$

As in the traditional fixed-effects ANOVA models, one is naturally interested in testing the significance of the main effects and the interactions in the fixed-effects FANOVA model (9)-(12). For each treatment i, $i = 1, \ldots, r$, averaging over the n_i observed paths in the FANOVA model (9)-(10) yields

$$d\overline{Y}_i(t) = m_0 + \mu(t) + a_i + \gamma_i(t)dt + \epsilon\, d\overline{\mathcal{W}}_i(t), \quad i = 1, \ldots, r; \quad t \in [0,1]. \tag{13}$$

where $\overline{\mathcal{W}}_i$ is the average of n_i independent standard Wiener processes on $[0,1]$. By the basic properties of the increments of a standard Wiener process on $[0,1]$, the stochastic processes $\{\overline{\mathcal{W}}_i;\; i = 1, \ldots, r\}$ are Wiener processes with covariances kernel $C(s,t) = \frac{1}{n_i} \min(s,t)$, and are still independent.

Integrating (13) with respect to t and using the identifiability conditions (11)-(12), we have

$$Y_i^* = m_0 + a_i + \epsilon\, \xi_i, \quad i = 1, \ldots, r, \quad \sum_{i=1}^{r} a_i = 0,$$

where $Y_i^* = \int_{[0,1]} d\overline{Y}_i(t)$ and ξ_i are independent $N(0, 1/n_i)$ random variables. This is the classical unbalanced one-way fixed-effects ANOVA model, so testing $(a_i = 0)$ (no differences in level) can be performed by standard techniques.

We are mainly interested in testing the zero-interactions. Hence, denoting hereafter the $L^2([0,1])$-norm by $||\cdot||_2$, we consider the alternative hypotheses to be of the form

$$H_1 \;\; : \;\; \gamma_i \in \mathcal{F}(\rho), \quad \text{at least for one} \quad i = 1, \ldots, r, \tag{14}$$

where $\mathcal{F}(\rho) = \{f \in L^2([0,1]) : ||f||_2 \geq \rho\}$ and ρ is a positive distance separating the alternative from the null hypothesis. Now, note that testing

$$H_0 : \gamma_i \equiv 0, \quad \forall\, i = 1, \ldots, r,$$

is equivalent to testing

$$H_0 : \theta_{jk}^{(i)} = 0, \quad \forall\, i = 1, \ldots, r, \quad j \geq 0; \quad k = 0, \ldots, 2^j - 1,$$

where $\theta_{jk}^{(i)}$ are the wavelet coefficients of γ_i. We can therefore apply the appropriate minimax tests of Abramovich et al. [2]. The interested reader is referred to the later paper for the implementation details. Let us just mention that the above wavelet formulation provides an efficient way to make meaningful inference on the fixed-effects and that the properties of the test are nonasymptotic with an optimal power (small ρ) over various classes of alternatives for the γ_i's simultaneously.

These above focus on functional hypothesis testing for fixed effect functions. This is clearly of interest, but is not the only relevant question with mass spectrometry proteomics where the primary goal is not to only to decide whether there are any systematic differences in the mean curves for different groups of patients, but also to identify which regions of the curves demonstrate differences. These specific regions can subsequently be mapped to individual proteins that may serve as useful biomarkers in medical applications. There is some recent and ongoing related work on this subject. More precisely, Park *et al.* [39] propose a new and simple algorithm based on permutation method to visualize the possible range of difference in protein abundance between groups with statistical significance while guarding against false positives simultaneously by constructing confidence bands of the contrast between groups. They also define a new concept for peaks (biomarkers) based on the proposed confidence band method. Once a significant difference is assessed by the FANOVA procedure, one then may use their procedure to find the relevant biomarkers.

To end this section, let us finally remark that since spectra are usually acquired from certain subjects, a mixed-effects model is necessary to generalize inferences to the population from which the subjects were selected. Researchers have started to derive and apply wavelet based mixed-effects models to mass spectrometry data recently (see [35, 4]), but we will not discuss them further in this paper.

Application to Petricoin's ovarian SELDI-TOF MS dataset

We analyzed the latest SELDI-TOF MS data from the ovarian cancer study available in the Clinical Proteomics Programs Databank with our method. This set of data

consists of serum profiles of 162 subjects with ovarian cancer and 91 non-cancer control subjects. For each subject, a set of data consisting of intensities at 15,154 distinct m/z values ranging from 0.0000786 to 19,995.513 was available for analysis. This data set was constructed using Ciphergen WCX2 ProteinChip Arrays. Preparation of chips for sample analysis was performed robotically and the raw data, without baseline subtraction, was posted for download. We used the normalization method, baseline removal and alignment procedures outlined in the previous section before proceeding to testing. We analyzed the ovarian cancer data set with 8192 m/z data points within the m/z range of $I = [1,500\ m/z, 20,000\ m/z]$ for 91 normal and 162 tumor samples. The intensity measures within the range below 1500 m/z were discarded due to the effects of matrix.

Figure 4.11 below displays the wavelet based estimated model components μ and γ_1 for Petricoin's normalized Ovarian MS data set within the region of [4001 m/z, 12192 m/z]. To perform the estimation we have used Daubechies wavelets of order 6 (3 zero moments).

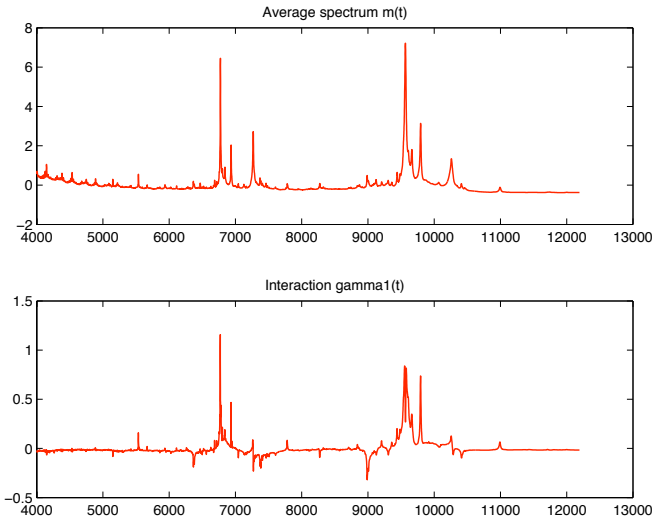


Figure 4.11: Zoom-in figure within the region [4001 m/z, 12192 m/z] for Petricoin's normalized Ovarian MS data set; mean $\mu(t)$ [toppanel] and the group effect $\gamma_1(t)$ for the cancer group[bottom panel].

Assuming that the noise variance σ^2 in each group (cancer and normal) is of the same amplitude, in order to perform the tests suggested by Abramovich *et al.* [2], an

estimate of the variance of the noise present in the data is necessary. The common practice, adopted here also, is to robustly estimate σ^2 by the median of absolute deviation of wavelet coefficients at the highest resolution level divided by 0.6745 (see [18]). This is done for each individual spectrum. The resulting estimates were then averaged. Note that in this case the estimate of σ^2 is *independent* from the test statistics since the later ones do not involve coefficients from the finest level.

Figure 4.11 gives some ideas where we would expect the test to reject. One can see that while the mean curve $\mu(t)$ and the group effect curve $\gamma_1(t)$, progress similarly, they show different patterns in amplitude at several m/z locations and therefore it is natural to use some kind of local test which is exactly the purpose of the adaptive wavelet based FANOVA procedure of Abramovich *et al.*. In our analysis (using $j(s) = 1$ and $j_\eta = 11$ for the testing procedures), the null hypothesis $H_0 : \gamma_i(t) = 0,\ i = 1, 2$ is rejected by the adaptive version of the appropriate FANOVA testing procedure, the corresponding value of the test statistic being 20.8892 to be compared with the threshold 0.1488.

Acknowledgements

This work was supported by the 'IAP Research Network P5/24', the "Région Rhône-Alpes" and the 'EC-HPRN-CT-2002-00286 Breaking Complexity Network'.

Mathematical appendix

We briefly recall first some relevant facts about the wavelet series expansion and the discrete wavelet transform that we have used in the paper.

4.1 The wavelet series expansion and the discrete wavelet transform

Throughout the paper we have assumed that we are working within an orthonormal basis generated by dilations and translations of a compactly supported scaling function, $\phi(t)$, and a compactly supported mother wavelet, $\psi(t)$, associated with an r-regular ($r \geq 0$) multiresolution analysis of $\left(L^2[0,1], \langle \cdot, \cdot \rangle\right)$, the space of squared-integrable functions on $[0,1]$ endowed with the inner product $\langle f, g \rangle = \int_{[0,1]} f(t)g(t)\, dt$.

For simplicity in exposition, we have worked with periodic wavelet bases on $[0,1]$ (see, e.g., [34], Section 7.5.1), letting

$$\phi_{jk}^{\mathrm{p}}(t) = \sum_{l\in\mathbb{Z}} \phi_{jk}(t-l) \quad \text{and} \quad \psi_{jk}^{\mathrm{p}}(t) = \sum_{l\in\mathbb{Z}} \psi_{jk}(t-l), \quad \text{for} \quad t \in [0,1],$$

where $\phi_{jk}(t) = 2^{j/2}\phi(2^j t - k)$ and $\psi_{jk}(t) = 2^{j/2}\psi(2^j t - k)$. For any given primary resolution level $j_0 \geq 0$, the collection

$$\{\phi_{j_0k}^{\mathrm{p}},\ k = 0,1,\ldots,2^{j_0}-1;\ \psi_{jk}^{\mathrm{p}},\ j \geq j_0;\ k = 0,1,\ldots,2^j-1\}$$

is then an orthonormal basis of $L^2[0,1]$. The superscript "p" has been suppressed from the notation for convenience. Despite the poor behavior of periodic wavelets near the boundaries, where they create high amplitude wavelet coefficients, they are commonly used because the numerical implementation is particular simple. Therefore, for any $f \in L^2[0,1]$, we denote by $\xi_{j_0k} = \langle f, \phi_{j_0k}\rangle$ $(k = 0,1,\ldots,2^{j_0}-1)$ the scaling coefficients and by $\theta_{jk} = \langle f, \psi_{jk}\rangle$ $(j \geq j_0;\ k = 0,1,\ldots,2^j-1)$ the wavelet coefficients of f for the orthonormal periodic wavelet basis defined above; the function f is then expressed in the form

$$f(t) = \sum_{k=0}^{2^{j_0}-1} \xi_{j_0k}\phi_{j_0k}(t) + \sum_{j=j_0}^{\infty}\sum_{k=0}^{2^j-1} \theta_{jk}\psi_{jk}(t), \quad t \in [0,1].$$

In statistical settings and signal processing, we are more usually concerned with discretely sampled, rather than continuous, functions. It is then the wavelet analogy to the discrete Fourier transform which is of primary interest and this is referred to as the discrete wavelet transform (DWT). Given a vector of function values $\mathbf{f} = (f(t_1),...,f(t_n))^T$ at equally spaced points t_i, the discrete wavelet transform of $\mathbf{f}$ is given by $\mathbf{d} = W_{n\times n}\mathbf{f}$, where $\mathbf{d}$ is an $n \times 1$ vector comprising both discrete scaling coefficients, c_{j_0k}, and discrete wavelet coefficients, d_{jk}, and $W_{n\times n}$ is an orthogonal $n \times n$ matrix associated with the orthonormal periodic wavelet basis chosen. The c_{j_0k} and d_{jk} are related to their continuous counterparts ξ_{j_0k} and θ_{jk} (with an approximation error of order n^{-1}) via the relationships $c_{j_0k} \approx \sqrt{n}\,\xi_{j_0k}$ and $d_{jk} \approx \sqrt{n}\,\theta_{jk}$. Note that, because of orthogonality of $W_{n\times n}$, the inverse DWT (IDWT) is simply given by $\mathbf{f} = W_{n\times n}^{\mathrm{T}}\mathbf{d}$, where $W_{n\times n}^{\mathrm{T}}$ denotes the transpose of $W_{n\times n}$. If $n = 2^J$ for some positive integer J, the DWT and IDWT can be performed through a computationally fast algorithm (see, e.g., [34], Section 7.3.1) that requires only order n operations.

4.2 Proofs and notes

For sake of brevity, we provide outlines of the proofs of the theoretical results obtained in the paper.

Notes on the optimality of ℓ_1 penalized regression splines

Optimality of our ℓ_1 penalized regression splines procedure follows from general results of van de Geer [49] on penalized regression once her conditions L, A, B and C are satisfied (refer to this paper for more details on these conditions). In order to prove that these assumptions hold in our case, we will make some standard assumptions under our setup and then check that the above conditions hold.

We will denote by Y the response variable and by X the regressor. We will assume that $\mathbb{E}(Y^2) < \infty$ and the following assumptions on the distribution of the random variable X representing the observed m/z recorded values:

Assumption 1 *The random variable X has a distribution Q that admits some density q with respect to the Lebesgue measure on $[0,1]$. Furthermore, there exists some positive ε such that, for any $x \in [0,1]$,*

$$\varepsilon \leq q(x) \leq 1/\varepsilon.$$

In particular, note that $||1/q||_\infty \leq 1/\varepsilon$.

Under our setup, we deal with the family

$$\mathcal{F} = \left\{ f_\beta(x) = \sum_{k=0}^{p} \beta_k x^k + \sum_{j=1}^{K} \beta_{p+j} (x - t_j)_+^p, \beta \in \mathbb{R}^{p+1+K} \right\}.$$

It is well known (see for instance [22]) that this family can also be written in terms of B-splines denoted hereafter by B_j:

$$\mathcal{F} = \left\{ f_\theta(x) = \sum_{j=1}^{p+1+K} \theta_j B_j(x,p), \theta \in \mathbb{R}^{p+1+K} \right\}.$$

We first recall below some relevant results on such B-splines approximations. Barron and Sheu ([5]: remark 2) have shown that, if the knots $t_j, 1 \leq j \leq K$, are equally spaced on the interval $[0,1]$, then for any function $g \in \mathcal{F}$,

$$||g||_\infty \leq (p+1)\sqrt{K+1}||g||_2.$$

Let f be an element of the Sobolev space W_2^r with $1 \leq r \leq p+1$, and let f_m be its best L_2-approximation in $\mathcal{F}$. Then de Boor and Fix [16] have shown that

$$||f - f_m||_2 = O\left(\frac{1}{m^r}\right), \; ||f - f_m||_\infty = O\left(\frac{1}{m^{r-1/2}}\right),$$

where $m = p + 1 + K$ is the dimension of $\mathcal{F}$.

Our second assumption concerns the Gram matrix associated to the B-splines basis.

Assumption 2 *Define the vector $B = (B_1, \ldots, B_m)^T$. The $m \times m$ matrix*

$$\Sigma = \int_0^1 B(x)B^T(x)q(x)dx$$

is positive definite with smallest eigenvalue $\rho^2 > 0$.

Considering the loss function $\gamma_f(x, y) = \gamma(f(x), y) = \rho_\tau(y - f(x))$, we have

$$\gamma(f, y) = \tau(y - f) - (y - f)\,\mathbb{1}_{\{y-f<0\}}.$$

Denote by

$$\overline{f} = argmin_{f \in \mathcal{F}} \mathbb{E}\left(\gamma_f(X, Y)\right),$$

the target function for the loss γ_f. Moreover, we will also assume that

Assumption 3 *The conditional cumulative distribution function of Y given X, denoted by $F^x_{Y/X}(\cdot)$, is differentiable in a neighborhood of $z = \overline{f}(X)$, Q-a.s. with a derivative at least $1/C_0$ Q-a.s.*

Let us now check that conditions of [49] hold.

Condition L

For a given $y \in \mathbb{R}$, $\gamma(f, y)$ is clearly convex as a function of f. Moreover,

$$|\gamma(f_\theta(x), y) - \gamma(f_{\tilde{\theta}}(x), y)| =$$

$$\begin{cases} \tau|(f_{\tilde{\theta}}(x) - f_\theta(x))|, & \text{if } y - f_\theta(x) > 0 \text{ and } y - f_{\tilde{\theta}}(x) > 0, \\ (1-\tau)|(f_{\tilde{\theta}}(x) - f_\theta(x))|, & \text{if } y - f_\theta(x) \leq 0 \text{ and } y - f_{\tilde{\theta}}(x) \leq 0, \\ |\tau(f_{\tilde{\theta}}(x) - f_\theta(x)) + y - f_{\tilde{\theta}}(x)|, & \text{if } y - f_\theta(x) > 0 \text{ and } y - f_{\tilde{\theta}}(x) \leq 0, \\ |\tau(f_{\tilde{\theta}}(x) - f_\theta(x)) - y + f_{\tilde{\theta}}(x)|, & \text{if } y - f_\theta(x) \leq 0 \text{ and } y - f_{\tilde{\theta}}(x) > 0. \end{cases}$$

In the last but one case, $f_\theta(x) < y \leq f_{\tilde{\theta}}(x)$, so that

$$f_\theta(x) - f_{\tilde{\theta}}(x) \leq y - f_{\tilde{\theta}}(x) \leq \tau(f_{\tilde{\theta}}(x) - f_\theta(x)) + y - f_{\tilde{\theta}}(x) \leq \tau(f_{\tilde{\theta}}(x) - f_\theta(x)).$$

The last case can be treated in a similar way and finally, we obtain,

$$|\gamma(f_\theta(x), y) - \gamma(f_{\tilde{\theta}}(x), y)| \leq |f_{\tilde{\theta}}(x) - f_\theta(x)|,$$

so that the Lipschitz property (L) is verified.

Condition A

We suppose here that the knots $t_j, 1 \leq j \leq K$ are equally spaced on the interval $[0, 1]$. By our assumption 1, we have for $l = 1, \ldots, p + K + 1$,

$$\sigma_l^2 = \mathbb{E}\, B_l(X, p)^2 \geq \varepsilon ||B_l||_2^2.$$

Then, using Barron and Sheu's [5] results, we see that

$$K_n = \max_{l=1,\ldots p+1+K} \frac{||B_l||_\infty}{\sigma_l} \leq \frac{(p+1)\sqrt{K+1}}{\sqrt{\varepsilon}}.$$

Condition A is satisfied as soon as the degree p and the number of knots K remain bounded as n tends to infinity.

Condition B

We have, for any $f \in \mathbb{R}$

$$\begin{aligned}
\mathbb{E}(\gamma(f, Y)|X) &= \mathbb{E}(\tau(Y - f) - (Y - f)\, \mathbb{1}\!\!\mathrm{l}_{\{Y-f<0\}}\, |X) \\
&= \tau\, \mathbb{E}(Y|X) - \tau f + \mathbb{E}((f - Y)\, \mathbb{1}\!\!\mathrm{l}_{\{Y-f<0\}}\, |X) \\
&= \tau\, \mathbb{E}(Y|X) - \tau f + \int_0^{+\infty} \mathbb{P}\left((f - Y)\, \mathbb{1}\!\!\mathrm{l}_{\{Y-f<0\}} > t|X\right) dt
\end{aligned}$$

since the random variable $(f - Y)\, \mathbb{1}\!\!\mathrm{l}_{\{Y-f<0\}}$ is nonnegative. Hence,

$$\begin{aligned}
\mathbb{E}(\gamma(f, Y)|X) &= \tau\, \mathbb{E}(Y|X) - \tau f + \int_0^{+\infty} \mathbb{P}\left(Y - f < 0 \cap f - Y > t|X\right) dt \\
&= \tau\, \mathbb{E}(Y|X) - \tau f + \int_0^{+\infty} \mathbb{P}\left(Y < f - t|X\right) dt \\
&= \tau\, \mathbb{E}(Y|X) - \tau f + \int_{-\infty}^{f} F_{Y/X}^x(u) du.
\end{aligned}$$

By our Assumption 3, this function is clearly twice differentiable and at least $1/C_0$ on a neighbourhood of $\bar{f}(X)$, Q-a.s., which suffices to prove that condition B of van de Geer is fulfilled with $G(u) = \frac{u^2}{2C_0}$.

Assumption C

By the Cauchy-Schwarz inequality, we have

$$\sum_{k=1}^{K+p+1} \sigma_k|\theta_k - \tilde{\theta}_k| \leq \sqrt{\sum_{k=1}^{K+p+1} \sigma_k^2}\sqrt{\sum_{k=1}^{K+p+1} |\theta_k - \tilde{\theta}_k|^2} = \sqrt{\sum_{k=1}^{K+p+1} \sigma_k^2}||\theta - \tilde{\theta}||_2.$$

Furthermore, $f_\theta(x) = B^T(x)\theta$, so that by our assumption 2,

$$||f_\theta - f_{\tilde{\theta}}||_2^2 = \int_0^1 (\theta - \tilde{\theta})^T B(x)B^T(x)(\theta - \tilde{\theta})q(x)dx = (\theta - \tilde{\theta})^T\Sigma(\theta - \tilde{\theta}) \geq \rho^2||\theta - \tilde{\theta}||_2^2.$$

Hence,

$$\sum_{k\in\mathcal{K}} \sigma_k|\theta_k - \tilde{\theta}_k| \leq \frac{\sqrt{\sum_{k\in\mathcal{K}} \sigma_k^2}}{\rho}||f_\theta - f_{\tilde{\theta}}||_2.$$

Condition C then holds with $D(\mathcal{K}) = \sum_{k\in\mathcal{K}} \sigma_k^2/\rho^2$.
The above conclude the optimality of our ℓ_1 penalized quantile regression method.

Proof of Theorem 3.1.

Let f^* be the approximation of f given in (4) at resolution $J = \log_2(n)$. Since f belongs to the Hölder ball $\Lambda^\alpha(R)$ with $0 < \alpha \leq r$, we know that (see [14])

$$\|f - f^*\|_2^2 \leq Cn^{-2\alpha/(2\alpha+1)},$$

so it will be sufficient to compare the estimator $\hat{f}$ with f^*. For some fixed integer $j_0 < J$, and for each $k = 1, \ldots, M$, the multiresolution properties of wavelets gives

$$\begin{aligned}
\tilde{f}_k(t) &= \frac{1}{\sqrt{n}}\sum_{i=1}^n Y_i^k \phi_{J,i}(t) \\
&= \sum_{i=1}^n \xi_{J,i}\phi_{J,i}(t) + \sum_{i=1}^n \left(\frac{1}{\sqrt{n}}f(t_i^k) - \xi_{J,i}\right)\phi_{J,i}(t) \\
&\quad + \sum_{i=1}^n \frac{1}{\sqrt{n}}\sigma\epsilon_i^k\phi_{J,i}(t),
\end{aligned}$$

which may be written as

$$\begin{aligned}
\tilde{f}_k(t) &= \sum_{\ell=0}^{2^j_0-1}\left(\xi_{j_0,\ell} + \xi_{j_0,\ell}^{1,k} + \xi_{j_0,\ell}^{2,k}\right)\phi_{j_0,\ell}(t) \\
&\quad + \sum_{j=j_0}^{J-1}\sum_{\ell=0}^{2^j-1}\left(\theta_{j,\ell} + \theta_{j,\ell}^{1,k} + \theta_{j,\ell}^{2,k}\right)\psi_{j,\ell}(t)
\end{aligned}$$

where $\xi_{j_0,\ell}$ and $\theta_{j,\ell}$ are the coefficients for f^*, $\xi^{1,k}_{j_0,\ell}$ and $\theta^{1,k}_{j,\ell}$ are the coefficients for $\sum_{i=1}^n \left(n^{-1/2} f(t_i^k) - \xi_{J,i}\right) \phi_{J,i}(t)$, and $\xi^{2,k}_{j_0,\ell}$ and $\theta^{2,k}_{j,\ell}$ are the coefficients for $\sum_{i=1}^n \frac{1}{\sqrt{n}} \sigma \epsilon_i^k \phi_{J,i}(t)$. Set

$$\hat{\xi}_{j_0,\ell} = \xi_{j_0,\ell} + \xi^{1,k}_{j_0,\ell} + \xi^{2,k}_{j_0,\ell}, \quad \ell = 0, \ldots, 2^{J_0} - 1$$

and

$$\tilde{\theta}_{j,\ell} = \theta_{j,\ell} + \theta^{1,k}_{j,\ell} + \theta^{2,k}_{j,\ell}, \quad \ell = 0, \ldots, 2^j - 1.$$

Conditional on δ^k and by the orthogonality of the DWT, $\hat{\xi}_{j_0,\ell} \sim N(\xi_{j_0,\ell} + \xi^{1,k}_{j_0,\ell}, \sigma^2/n)$ and $\tilde{\theta}_{j,\ell} \sim N(\theta_{j,\ell} + \theta^{1,k}_{j,\ell}, \sigma^2/n)$. Thresholding is applied to the details coefficients $\tilde{\boldsymbol{\theta}}$. Given δ^k and the orthonormality of the wavelet basis,

$$\mathbb{E}\left(\|\hat{f} - f^*\|_2^2\right) = \mathbb{E}\left(\sum_{\ell=0}^{2^{j_0}-1} (\hat{\xi}_{j_0,\ell} - \xi_{j_0,\ell})^2\right) + \mathbb{E}\left(\sum_{j=j_0}^{J-1} \sum_{\ell=0}^{2^j-1} (\hat{\theta}_{j,\ell} - \theta_{j,\ell})^2\right).$$

For the first term on the right in the above equation, given δ^k,

$$\mathbb{E}\left(\hat{\xi}_{j_0,\ell} - \xi_{j_0,\ell}\right)^2 \leq 2\mathbb{E}\left(\hat{\xi}_{j_0,\ell} - (\xi_{j_0,\ell} + \xi^{1,k}_{j_0,\ell})\right)^2 + 2\mathbb{E}(\xi^{1,k}_{j_0,\ell})^2 \leq 2\sigma^2/n + 2\mathbb{E}(\xi^{1,k}_{j_0,\ell})^2.$$

In a similar way, for the second term

$$\mathbb{E}\left(\hat{\theta}_{j,\ell} - \theta_{j,\ell}\right)^2 \leq 2\mathbb{E}\left(\hat{\theta}_{j,\ell} - (\theta_{j,\ell} + \theta^{1,k}_{j,\ell})\right)^2 + 2\mathbb{E}(\theta^{1,k}_{j,\ell})^2.$$

To bound the first portion of this, it is only necessary to note that all the conditions are met here for usual theorems on thresholding with wavelets. For example, if the rule η is SureShrink or a block shrinkage rule, the rate is less than or equal to $Cn^{-2\alpha/(2\alpha+1)}$. For a rule such as VisuShrink, this would be have an additional factor of $\log n$ in it. In either case, the estimator is of the same order of convergence as the DWT. For sake of argument, assume that η is a rule admitting the fast rate (no $\log n$ penalty). Then, given δ^k,

$$\mathbb{E}\left(\|\hat{f} - f^*\|_2^2\right) \leq Cn^{-2\alpha/(2\alpha+1)} + C\mathbb{E}\left(\sum_{\ell=0}^{2^{j_0}-1} (\xi^{1,k}_{j_0,\ell})^2\right) + C\mathbb{E}\left(\sum_{j=j_0}^{J-1} \sum_{\ell=0}^{2^j-1} (\theta^{1,k}_{j,\ell})^2\right).$$

To bound the summation portion of this equation, note first that

$$\mathbb{E}\left(\sum_{\ell=0}^{2^{j_0}-1} (\xi^{1,k}_{j_0,\ell})^2\right) + \mathbb{E}\left(\sum_{j=j_0}^{J-1} \sum_{\ell=0}^{2^j-1} (\theta^{1,k}_{j,\ell})^2\right) = \mathbb{E}\sum_{i=1}^n \left(\int [f(t_i^k) - f(t)] \phi_{J,i}(t) dt\right)^2.$$

Assuming that the wavelet ϕ has support $[0,s]$, the scaling function $\phi_{J,i}$ has support $[i/n,(i+s)/n]$ for $i=1,\ldots,n-s$ and $[0,(i+s-n)/n]\cup[i/n,1]$ for $i=n-s+1,\ldots,n$. Using this and the fact that f is α-Hölder, then for $i=1,\ldots,n-s$,

$$\begin{aligned}
\left(\int [f(t_i^k)-f(t)]\phi_{J,i}(t)dt\right)^2 &= \left(\int_{i/n}^{(i+s)/n} [f(t_i^k)-f(t)]\phi_{J,i}(t)dt\right)^2 \\
&\leq C\left(\int_{i/n}^{(i+s)/n} |t_i^k-t|^{\min(\alpha,1)}\phi_{J,i}(t)dt\right)^2 \\
&= C\left(\int_{i/n}^{(i+s)/n} |i/n+\delta_i^k-t|^{\min(\alpha,1)}\phi_{J,i}(t)dt\right)^2 \\
&\leq C\left(\int_{i/n}^{(i+s)/n} (|i/n-t|^{\min(\alpha,1)}+|\delta_i^k|^{\min(\alpha,1)})\phi_{J,i}(t)dt\right)^2 \\
&\leq C(n^{-2\min(\alpha,1)}+|\delta_i^k|^{\min(\alpha,1)})2^{-J}.
\end{aligned}$$

Using similar arguments, it is easy to show that, for $i=n-s+1,\ldots,n$ we have

$$\left(\int [f(t_i^k)-f(t)]\phi_{J,i}(t)dt\right)^2 \leq C2^J = Cn^{-1}$$

Therefore,

$$\mathbb{E}\sum_{i=1}^{n}\left(\frac{1}{\sqrt{n}}f(t_i^k)-\xi_{J,i}\right)^2 \leq Cn^{-2\min(\alpha,1)}+C\frac{1}{n}\sum_{i=1}^{n-s}\mathrm{var}(\delta_i^k)^{\min(\alpha,1)}.$$

Using now the fact that the variances of the δ_i^k's are less than or equal to Cn^{-2} , then the bound becomes $Cn^{-2\min(\alpha,1)}$ and the overall bound of the estimate's $\hat{f}_k$ error is less than $Cn^{-2\alpha/(2\alpha+1)}$. We now have

$$\mathbb{E}\|\hat{f}-f\|_2^2 = \mathbb{E}\left(\frac{1}{M}\sum_{k=1}^{M}(\hat{f}_k-f)\|_2^2\right) \leq \frac{1}{M}\max_{k=1,\ldots,M}\mathbb{E}(\|\hat{f}_k-f\|_2^2) = O(M^{-1}n^{-2\alpha/(2\alpha+1)}),$$

which proves the Theorem.

References

[1] H. Abbink Spaink, T. Lub, R. Otjes, and H. Smith. Baseline correction for second-harmonic detection with funable diode lasers. *Anal. Chim. Acta*, 183:141–151, 1986.

[2] F. Abramovich, A. Antoniadis, T. Sapatinas, and B. Vidakovic. Optimal testing in a fixed-effects functional analysis of variance model. *International Journal of Wavelets, Multiresolution and Information Processing*, 2:323–349, 2004.

[3] S. A.C. and T. Speed. Normalization, baseline correction and alignment of. high-throughput mass spectrometry data. In *Proceedings Gensips*, 2004. in press.

[4] A. Antoniadis and T. Sapatinas. Estimation and inference in functional mixed-effects models. Technical Report TR-15-2004, Department of Mathematics and Statistics, University of Cyprus, Cyprus, 2004.

[5] A. Barron and C. Sheu. Approximations of density functions by sequences of exponential families. *The Annals of Statistics*, 19(3):1347–1369, 1991.

[6] J. Bigot. A scale-space approach with wavelets to singularity estimation. *ESAIM: PS*, 9:143–164, 2005.

[7] J. Bigot. Landmark-based registration of curves via the continuous wavelet transform. *Journal of Computational and Graphical Statistics*, 15(3):542–564, September 2006.

[8] D. Billheimer. Functional data analysis of protein expression in matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. *Journal of Agricultural, Biological, and Environmental Statistics*, pages 271–290, 2006.

[9] L. Breiman and S. Peters. Comparing automatic smoothers (a public service enterprise). *Int. Statist. Rev.*, 60:271–290, 1992.

[10] L. D. Brown and M. G. Low. Asymptotic equivalence of nonparametric regression and white noise. *Annals of Statistics*, 24:2384–2398, 1996.

[11] F. Chao and A. Leung. *Application of wavelet transform in processing chromatographic data*. Walczak B (ed.) Wavelets in Chemistry, Elsevier Science, 2000.

[12] R. Coifman and D. Donoho. Translation invariant de-noising. *Lecture Notes in Statistics*, 103:125–150, 1995.

[13] K. Coombes, S. Tsavachidis, J. Morris, K. A. Baggerly, and R. Kobayashi. Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra using the undecimated discrete wavelet transform. *Proteomics*, 41:4107–4117, 2005.

[14] I. Daubechie. *Ten lectures on wavelets*. SIAM, 1992.

[15] C. De Boor. *A pratical guide to splines*. Vol. 27 of Applied Mathematical Sciences, Springer-Verlag, New-York, 1978.

[16] C. De Boor and G. Fix. Spline approximation by quasiinterpolants. *J. Approximation Theory*, 8:19–45, 1973.

[17] P. Dierckx. *Curve and surface fitting with splines*. Clarendon, Oxford, 1993.

[18] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, 1994.

[19] D. Donoho and I. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26:879–921, 1998.

[20] P. Eilers. A perfect smoother. *Analytical Chemistry*, 75:3631–3636, 2003.

[21] P. Eilers. Parametric time warping. *Analytical Chemistry*, 76(2):404–411, 2004.

[22] P. Eilers and B. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, 11(2):89–121, 1996.

[23] J. Fan and S. Lin. Test of significance when data are curves. *Journal of American Statistical Association*, 93:1007–1021, 1998.

[24] T. Gasser and A. Kneip. Searching for structure in curve samples. *Journal of the American Statistical Association*, 90(432):1179–1188, 1995.

[25] X. He and P. Ng. Cobs: Qualitatively constrained smoothing via linear programming. *Computational Statistics*, pages 879–921, 2006.

[26] C. Heipke. Overview of image matching techniques. In *OEEPE - Applications of Digital Photogrammetric Work- stations, Proceedings, Lausanne, Switzerland*, pages 173–191, 1996.

[27] N. Jeffries. Algorithms for alignment of mass spectrometry proteomic data. *Bioinformatics*, 21(14):3066–3073, 2005.

[28] K. Johnson, B. Wright, K. Jarman, and R. Synovec. High-speed peak matching algorithm for retention time alignment of gas chromatographic data. *Journal of Chromatography A*, 996:141–155, 2003.

[29] A. Kneip and T. Gasser. Statistical tools to analyze data representing a sample of curves. *Annals of Statistics*, 20(3):1266–1305, 1992.

[30] R. Koenker. *Quantile Regression, Econometric Society Monograph Series*. Cambridge University Press, 2005.

[31] R. Koenker and G. Basset. Regression quantiles. *Econometrica*, 1:33–50, 1978.

[32] R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81(4):673–680, 1994.

[33] X. Ma and Z. Zhang. Application of wavelet transform to background correction in inductively coupled plasma atomic emission spectrometry. *Anal. Chim. Acta*, 485(2):233–239, 2003.

[34] S. Mallat. *A Wavelet Tour of Signal Processing. 2nd ed.* San Diego: Academic Press, 1999.

[35] J. Morris and R. Carroll. Wavelet-based functional mixed models. *Journal of the Royal Statistical Society, Series B*, 68, 2006. to appear.

[36] J. Morris, K. Coombes, J. Koomen, K. Baggerly, and R. Kobayashi. Feature extraction and quantification for mass spectrometry data in biomedical applications using the mean spectrum. *Bioinformatics*, 21(9):1764–1775, 2005.

[37] Y. Munoz Maldonado, J. Staniswalis, L. Irwin, and D. Byers. A similarity analysis of curves. *The Canadian Journal of Statistics*, 30(3):373–381, 2002.

[38] J. Padayachee, V. Prozesky, W. von der Linden, M. Nkwinika, and V. Dose. Bayesian pixe background subtraction. *Nucl. Instrum. Methods Phys. Res. B*, 150:129–135, 1999.

[39] Y. Park, S. Downing, C. Li, W. Hahn, P. Kantoff, and L. Wei. Simultaneous and exact interval estimates for the contrast of two groups based on an extremely high dimensional response variable: Application to mass spec data analysis. Technical Report 29, Harvard University Harvard University Biostatistics Working Paper Series, 2006.

[40] D. B. Percival and A. T. Walden. *Wavelet Methods for Time Series Analysis*. Cambridge University Press, Cambridge, UK, 2000.

[41] Y. Qu, B. Adam, M. Thornquist, J. Potter, M. Thompson, Y. Yasui, J. Davis, P. Schellhammer, L. Cazares, M. Clements, G. Wright, and F. Z. Multiscale processing of mass spectrometry data. *Biometrics*, 59:143–151, 2003.

[42] J. Ramsay and X. Li. Curve registration. *Journal of the Royal Statistical Society, Series B*, 60:351–363, 1998.

[43] T. W. Randolph and Y. Yasui. Multiscale processing of mass spectrometry data. *Biometrics*, 2005. in press.

[44] A. Rouh, M. Delsuc, G. Bertrand, and J. Lallemand. The use of classification in baseline correction of ft nmr spectra. *J Magn Reson Ser A*, 102:357–359, 1993.

[45] A. Ruckstuhl, M. Jacobson, R. Field, and J. Dodd. Baseline subtraction using robust local regression estimation. *Journal of Quantitative Spectroscopy and Radiative Transfer*, 68(2):179–193, 2001.

[46] S. Sardy, D. B. Percival, A. Bruce, H. Gao, and W. Stuetzle. Wavelet denoising for unequally spaced data. *Statistics and Computing*, 9(1):65–75, 1999.

[47] B. Saussen, M. Kirchner, H. Steen, J. Jebanathirajah, and F. Hamprecht. The rpm package: aligning lc/ms mass spectra with r. In *Interdisciplinary Center for Scientific Computing, University of Heidelberg, Germany UseR2006*, 2006.

[48] R. Tibshirani, T. Hastie, B. Narasimhan, S. Soltys, G. Shi, A. Koong, and Q. Le. Sample classification from protein mass spectrometry, by peak probability contrasts. *Bioinformatics*, 20(17):3034–3044, 2004.

[49] S. A. van de Geer. High-dimensional generalized linear models and the lasso. Technical Report 133, Seminar für Statistik, Eidgenössissche Technische Hochscule (ETH), 2006.

[50] E. van Veen and M. de Loos-Vollebregt. Application of mathematical procedures to background correction and multivariate analysis in inductively coupled plasma-optical emission spectrometry. *Spectrochimica Acta Part B: Atomic Spectroscopy*, 53(5):639–669, 1998.

[51] W. W. Dietrich, C. Rüdel, and M. Neumann. Fast and precise automatic baseline correction of one- and two-dimensional nmr spectra. *J. Magn. Reson.*, 91:1–11, 1991.

[52] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao. Detecting and aligning peaks in analyzing maldi mass spectrometry data. *Computational Biology and Chemistry*, 30:27–38, 2006.

[53] W. Yu and H. Zhao. Aligning spectral peaks in mass spectrometry data with a robust point matching approach. In *In 52nd ASMS Conference on Mass Spectrometry and Allied Topics, Nashville, TN, May*, pages 23–27, 2004.

[54] M. Yuan. Gacv for quantile smoothing splines. *Computational Statistics and Data Analysis*, 50(3):813–829, 2006.