NONLINEAR STRUCTURAL FUNCTIONAL MODELS

Michael R. Wierzbicki

A DISSERTATION

in

Epidemiology and Biostatistics

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2013

Supervisor of Dissertation

Signature _____

Wensheng Guo, Professor of Biostatistics

Graduate Group Chairperson

Signature _____

Daniel F. Heitjan, Professor of Biostatistics

Dissertation Committee

Warren B. Bilker, Professor of Biostatistics

Sarah J. Ratcliffe, Associate Professor of Biostatistics

Bruce I. Turetsky, Associate Professor of Psychiatry

NONLINEAR STRUCTURAL FUNCTIONAL MODELS

ACKNOWLEDGMENT

I would like to thank my advisor, Wensheng Guo, for his guidance, advice, and patience the past three years. Your mentorship and teaching have made me a much better statistician and scientist. I also thank my committee members: Warren Bilker, I thank you for advising me and for bring me onto the Mental Health Training Grant [No. 5T32MH065218] which has provided invaluable experience (and also financial support). Sarah Ratcliffe, I thank you for introducing me to the world of functional data, your advice, and time you spent advising and teaching me. Bruce Turetsky, I thank you for providing the outside perspective that I would not have without you.

Thank you to the Biostatistics faculty and staff. In particular, I thank Kevin Lynch for his guidance and advice during my last two years at Penn; Phyllis Gimotty for her guidance during my first year; Clay Wells for all your help and computing wizardry; Ann Facciolo, Cathy Vallejo, and Marissa Fox for all that you do for the department. Also thanks to Jonas Ellenberg, Susan Ellenberg, Benjamin French, Dick Landis, Nandita Mitra, Reneé Moore, and Thomas Ten Have.

A very special thanks to: Drs. Bill Baker, Bob Fray, David Moffett, David Penniston, and Wesley Wong.

I also thank the students of the various schools and departments I have crossed paths with during my many years of schooling.

I thank my family and friends, past and present, for their love and support throughout the years. Last, but certainly not least, I thank my significantly better half: my wife, Brittainy.

ABSTRACT

NONLINEAR STRUCTURAL FUNCTIONAL MODELS

Michael R. Wierzbicki

Wensheng Guo

A common objective in functional data analyses is the registration of data curves and estimation of the locations of their salient structures, such as spikes or local extrema. Existing methods separate curve modeling and structure estimation into disjoint steps, optimize different criteria for estimation, or recast the problem into the testing framework. Moreover, curve registration is often implemented in a pre-processing step. The aim of this dissertation is to ameliorate the shortcomings of existing methods through the development of unified nonlinear modeling procedures for the analysis of structural functional data. A general model-based framework is proposed to unify registration and estimation of curves and their structures. In particular, this work focuses on three specific research problems. First, a Sparse Semi-parametric Nonlinear Model (SSNM) is proposed to jointly register curves, perform model selection, and estimate the features of sparsely-structured functional data. The SSNM is fitted to chromatographic data from a study of the composition of Chinese rhubarb. Next, the SSNM is extended to the nonlinear mixed effects setting to enable the comparison of sparse structures across group-averaged curves. The model is utilized to compare compositions of medicinal herbs collected from two groups of production sites. Finally, a Piecewise Monotonic B-spline Model (PMBM) is proposed to estimate the locations of local extrema in a curve. The PMBM is applied to MRI data from a study of gray matter growth in the brain.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER 1

Introduction

The prevalence of functional data has increased greatly in the last few decades. Functional data are commonplace in a variety of research areas including pharmacological, biomedical, and environmental fields, and have motivated a plethora of statistical research in their display, exploration, and analysis. The starting point of most functional data analyses is the assumption that the data vector is comprised of discrete observations arising from some underlying continuous curve, and the curve itself is treated as the basic unit of data analysis. In such analyses, a common objective is to register curves and locate and draw inferences on their salient structures, such as spikes or local extrema. Numerous methodologies for the analysis of functional data and the estimation of their structures exist which accommodate complex designs and various correlation structures (e.g. Ke and Wang, 2001; Guo, 2002; Ramsay and Silverman, 2005; Morris and Carroll, 2006). However, in drawing inferences on structures, existing methods model a curve and estimate its structures in disjoint steps, optimize different criteria for curve and structure estimation, or recast the problem into the testing framework. Moreover, curve registration is treated as a pre-processing step and implemented separately from the modeling procedure. The unified, nonlinear, model-based estimation procedures developed in this dissertation address the shortcomings of existing approaches in the analysis of structured functional data.

This dissertation is motivated by three data applications. The first two applications concern chromatographic data of medicinal herbs. Medicinal herbs have been used in Eastern nations for the treatment of a variety of ailments and diseases for thousands of years (Duke, 2002). The chemical composition of medicinal herbs is complex, com-

prising of numerous compounds, and it is widely accepted that the therapeutic effects of herbs are due to multiple compounds in conjunction with each other. Determining the composition of medicinal herbs is the first step towards elucidating their active composition and their development into standardized pharmaceuticals. The need for composition identification in medicinal herb research has prompted the rise in popularity of a tool used in analytical chemistry, namely High Performance Liquid Chromatography.

High Performance Liquid Chromatography (HPLC) is a tool for the separation and detection of compounds in biological mixtures, such as herbs (Snyder and Kirkland, 1979). Chromatographic experiments analyze an herb sample and output a curve, termed chromatogram, characterized by spikes over experiment time corresponding to detected compounds. The HPLC process is described in more detail in Chapter 2. As the particular set of compounds is unique to an herb and spike locations can be used to identify compounds, chromatograms provide a visual representation, or fingerprint, of the herb. Therefore, obtaining the fingerprint of an herb is vital for active composition exploration.

Certain characteristics of medicinal herbs and chromatography preclude the direct identification of herb compositions via HPLC. First, the exact chemical composition of an herb differs based on its particular species and origin; properties of the seed and field used to grow the herbs; and the particular processes used in the growing, harvesting, and storage of the herbs (Leung and Cheng, 2008). This lack of proper quality control results in the existence of many variations of a single herb. Second, constructing statistical models for the estimation and comparison of chromatographic fingerprints is difficult due to the sparse, spiky nature of the curves. In addition, across different experimental conditions, the location of spikes can be

shifted, preventing the establishment of a standardized fingerprint. The inability to establish a standardized fingerprint from multiple experiments has a financial implication in regards to compound identification, as well. Unknown compounds can be identified by analyzing samples of known compounds under identical conditions and comparing the reference spike timings to those in the study chromatogram. The misalignment prevents the comparison of reference spike timings across conditions and requires the purchase of many reference compounds to be analyzed under every experimental condition. As known compounds samples are expensive, the use of HPLC in large-scale medicinal herb studies is not practical. The first two studies considered in this dissertation address the complications presented by medicinal herbs and chromatography.

The first study aims to establish a fingerprint of the herb, Rheum palmatum, or Chinese rhubarb. Samples of Chinese rhubarb of identical composition are analyzed via HPLC under a set of different experimental conditions. The varying conditions induce transformations of the underlying fingerprint shape and, as a result, curve spikes are unaligned across experiments.

Chapter 2 develops a Sparse Semiparametric Nonlinear Model (SSNM) for the registration of sparsely-structured functional data, such as chromatographic data. Data-driven basis expansion is used to model the common shape of curves while a parametric time warping function registers individual curves. Penalized weighted least squares with the Adaptive Lasso penalty provides a unified criterion for registration, model selection, and estimation. The unified criterion results in a unique solution and allows the study of sampling properties, as opposed to existing methods which can guarantee neither. A back-fitting algorithm is proposed for estimation and sampling properties of model estimators are proved. The performance of SSNMs is assessed

through a simulation study and the SSNM is fitted to the rhubarb chromatographic data and a standardized fingerprint is established. Furthermore, through the use of the SSNM, known compounds need only be analyzed under a single condition, greatly reducing the cost of compound identification.

The aim of the second study is to compare the compositions of samples of the herb, Andrographis paniculata, collected from two groups of production sites: sites that adhere to Good Agricultural Practices (GAP) and pharmacies, or market sites, whose compliance with GAP is unknown. Chapter 3 discusses GAP in further detail. Identifying discordant compounds between the two groups of sites can aid in the quality control of herbs produced by the market sites. The site-level warpings of the fingerprint shapes prevent the establishment of and comparison between GAP-compliant and market fingerprints.

Chapter 3 extends the SSNM to accommodate nested designs, such as longitudinal data settings, by developing a Sparse Semiparametric Nonlinear Mixed Effects Model (SSNMM) for the registration and comparison of grouped functional data with sparse structures. Similar to the the SSNM, group-averaged curve shapes are modeled using data-driven basis expansion. To correctly account for the sources of variability, subject-specific deviations in the group-averaged curves are modeled using parametrically-specified random effects. Penalized likelihood with the Adaptive Lasso provides a unified criterion for joint registration, model selection, and estimation. A back-fitting algorithm using the Laplace Approximation is proposed for estimation and the sampling properties of model estimators, whose variation account for the joint estimation of the shape functions, fixed and random effects, and variance components, are proved. The performance of the SSNMM is assessed through a simulation study and the SSNMM is fitted to the Andrographis paniculata chromatographic data and

identifies compounds which are not common between the GAP-compliant and market sites.

The third and final motivating example is a cross-sectional study of the growth of gray matter, an important neurological tissue, in the prefrontal cortex of the brain. Previous research has shown that gray matter volume increases until some point in adolescence and then decreases into adulthood. Identifying the unknown age at which gray matter volume stops increasing has implications in predictions of neurological development in children and the classification and diagnosing of neurological conditions. In this study, Magnetic Resonance Images (MRIs) were obtained on 107 subjects aged one month to 25 years and volumetric measurements of gray matter were collected in multiple regions of the brain.

Chapter 4 develops a Piecewise Monotonic B-spline regression Model (PMBM) for the estimation of local extrema locations in a multi-modal curve. The model-based approach enables joint estimation of the curve shape and extrema timings via optimization of a unified nonlinear least squares criterion. As a result, the procedure developed in this chapter allows flexible modeling of a data curve and enables statistical inference to be drawn on the locations of its extrema. The performance of the PMBM is assessed by comparing its performance to an alternative method using smoothing splines in a simulation study. The PMBM is fitted to the volumetric MRI data set and point and interval estimates of peak gray matter volume are obtained in multiple regions of the brain.

CHAPTER 2

Sparse Semiparametric Nonlinear Model with Application to Chromatographic Fingerprints

2.1. Introduction

Traditional Chinese Herbal Medicines (TCHMs) have been used for thousands of years for the treatment and prevention of a wide range of medical ailments and diseases (Duke, 2002). TCHMs are comprised of substances extracted from the roots, stems, or leaves of plants and herbs by boiling (Liang et al., 2004). The structure of TCHMs is complex and their therapeutic effects are due to a combination of multiple compounds. Furthermore, the preparation process is difficult to replicate exactly, inhibiting proper quality control. Thus efficacy research, classification, and the development of standardized medications is problematic and has been the focus of much research in recent years.

The first step in TCHM research is identifying compounds in an herbal medicine. Chromatography is a standard technique used in the exploration of biological sample composition and has become a key tool in herbal medication research. A chromatographic experiment outputs a curve displaying an intensity measurement over experiment time characterized by a number of sharp, narrow spikes, where each spike corresponds to a compound in the sample. The location in time of a spike, called retention time, can be used to identify the compound. For instance, a sample of a known compound can be analyzed under the identical experimental condition and if its retention time is the same as one of the spikes in the study chromatogram, this is evidence that the particular spike corresponds to the known compound.

Therefore the chromatographic curve, termed chromatogram, provides a visual representation of the composition of a study sample. As the combination of compounds is unique to a TCHM, its chromatogram serves as a fingerprint of the medicine. The main difficulty in fingerprinting is that, due to variations in experimental conditions, spikes are often shifted in time across experiments. The incomparability of spikes across experiments prevents the establishment of an overall standardized fingerprint. In addition, due to the shifting of retention times, any known samples used to identify compounds must be analyzed under every experimental condition, which can become cost-prohibitive. Hence retention time warping poses a large obstacle in the practical application of chromatographic experiments in TCHM research.

Recent statistical methods proposed to address the time warping seek to align spikes across experiments. Alignment is performed through parametric models of the warping function or dynamic time warping of the chromatograms (Kassidas et al., 1998; van Nederkassel et al., 2006). Generally, alignment is performed as a pre-processing step and the mean curve is obtained separately on the aligned curves. The disadvantages of such two-step approaches have been noted previously (Morris et al., 2008). Two-step methods do not take the variability in the alignment step into account which leads to downward attenuation of subsequent parameter estimates and optimizing different objective functions for each step does not guarantee overall convergence. In addition, these methods usually focus on a few large spikes and the smaller spikes are ignored. Curve registration offers a framework to model chromatograms and warping functions jointly (Ramsay and Li, 1998). By assuming time has been warped by some unknown monotonic function and estimating the warping function jointly with the standardized fingerprint, estimation and alignment of chromatograms are accomplished in a unified fashion.

In regards to modeling the fingerprint shape, wavelets are a particular family of functions that have become popular due to their ability to estimate curves with local features, such as sharp spikes. This ability has prompted their use in modeling mass spectrometry (MS) data which is similar in nature to chromatographic data (Randolph and Yasui, 2006; Morris et al., 2008). In particular, Morris et al. (2008) considered proteomic MS data and applied the discrete wavelet transform (DWT) to the curves and fit a linear functional mixed effects model to the transformed data. Their methodology can accommodate functional effects and can be used in nested designs. However, their methods are restrictive in that the DWT is a linear transformation and their model is linear in the mixed effects. In our setting, the warping function parameters enter into the model nonlinearly, thus we cannot employ the DWT. Instead wavelet basis expansion can be utilized to supply a finite-dimensional representation of the curves that can accommodate nonlinear parameters.

A salient feature of chromatograms is that only a small fraction of each curve is true signal. While the curves are comprised of thousands of data points, there are relatively few narrow spikes along with long, flat regions. This sparse structure suggests that their basis representation is also sparse, in that most of the basis functions correspond to the flat regions of the curves and the true values of their coefficients are zero. It is imperative, then, to identify the subset of nonzero coefficients and include only their corresponding basis functions in the model. In other words, model selection should be performed on the basis functions. Imposing an ℓ_1 penalty on the basis functions shrinks irrelevant coefficients to 0, in essense dropping them from the model while retaining and estimating the important variables. Thus model selection is performed on the basis functions as well as nonparametric estimation, as the dimension of the basis parameter vector is chosen by the data.

The semiparametric nonlinear regression models of Wang and Ke (2009) is a large class of flexible models that can accommodate nonlinear parameters, such as warping function parameters. However, Wang and Ke (2009) use smoothing splines to estimate the mean function. They induce smoothness in the functional estimate via an ℓ_2 penalty, which does not perform model selection as all variables are included in the model. Furthermore, the estimation algorithms used for ℓ_2-penalized methods differ from those for ℓ_1-penalized methods.

In this chapter we propose a sparse semiparametric nonlinear model for the registration of chromatograms and establishment of a standardized chromatographic fingerprint. We assume the chromatograms arise from a common spiky function and employ basis expansion by the Battle-Lemarié spline wavelets to model this function. We assume each curve has been distorted by a smooth, monotonic, nonlinear warping function, and model the warping functions by monotonic parametric smooth functions. We impose the Adaptive Least Absolute Shrinkage and Selection Operator (Adaptive Lasso) penalty on the coefficients of the wavelet basis functions. Penalized weighted least squares provides a unified criterion to simultaneously select wavelet basis parameters and estimate the standardized fingerprint and warping functions, guaranteeing a unique solution and enabling us to study sampling properties of the estimates (Fan and Li, 2001; Zou, 2006). We propose a computationally efficient back-fitting algorithm for the estimation of model parameters which is equivalent to a blockwise coordinate descent algorithm and utilizes existing algorithms.

Estimating the warping functions enables recovery of information across experiments in establishing a standardized fingerprint. By averaging across samples, the consistent spikes will be preserved even if their amplitudes are small, and the inconsistent spikes will be averaged out even if their amplitudes are large. This is key as the large

spikes do not necessarily represent the active set of compounds. As the number of curves grows larger, a consistent estimate of the standardized fingerprint is obtained. The search for active compounds can then be focused on the common compounds reflected in the standardized fingerprint. Furthermore, if the time scale of a particular experiment is used as the reference scale to which all experiments are warped, any known compounds used to identify compounds in the study chromatograms need be analyzed under only the reference condition as opposed to every condition. Thus comparisons of retention times of a large number of known compounds to those in the study chromatograms is much more financially possible than the current practice.

The use of the adaptive lasso results in root-n consistency, asymptotic normality and variable selection consistency of the model estimates, known as the oracle property (Donoho and Johnstone, 1994). Statistical inference can be made on the curves and warping functions and the variance estimates reflect the variability in their joint estimation.

We construct chromatographic experiments to demonstrate the application of our procedure. We obtain chromatograms of samples of the herbal medicine, rhubarb, through High Performance Liquid Chromatography (HPLC) under a set of uniquely calibrated experimental settings, chosen to induce time retention warping across settings. More details about rhubarb, HPLC, and the design of our experiments are given in Section 2.2. We further assess our procedure via simulation.

The remainder of the chapter is organized as follows: We discuss rhubarb, HPLC, and the design of our experiments in Section 2.2. In Section 2.3 we present the sparse semiparametric nonlinear model. In Section 2.4 we discuss estimation and properties of the estimates. Application of the model to fingerprint data and simulations are presented in Sections 2.5 and 2.6, respectively. We conclude with discussion in

Section 2.7.

2.2. Data

Rhubarb is a medicinal plant that has been used since at least 250 A.D. for the treatment and prevention of number of medical conditions including cancer, constipation, fever, and inflammations (Peigen et al., 1984; Duke, 2002). The rhizomes and roots of the plant are typically used for its medical applications. The structure of rhubarb is quite complex, with more than one hundred compounds identified across six species of the plant, and its medicinal properties are still not fully understood (Ye et al., 2007). As rhubarb is one of the more popular and widely-used TCHMs, there is much interest in the exploration of the medicine. Before identification and quantification of the active compounds in rhubarb is possible, its chemical composition must be determined.

High Performance Liquid Chromatography (HPLC) is a particular technique for separating compounds of a biological sample. HPLC dissolves a sample into a liquid solution of two solvents, called the mobile phase. The relative composition of the two solvents is varied at a controlled rate over time, where the levels and timings together are termed the gradient program. The sample and mobile phase are pumped through a column containing sorbent materials, called the stationary phase. As the sample passes through the column, the compounds separate from each other. Due to the unique properties of the individual compounds and the mobile and stationary phases, the compounds travel through the column at different rates and leave the column at different times (Snyder and Kirkland, 1979). A detector, such as an Ultraviolet (UV) detector, records two measurements: the time a compound leaves the column, or retention time, and an intensity measurement.

UV detectors measure UV absorbance, which is a function of concentration and molar absorptivity of each compound. The amplitude of the resulting spike in the chromatogram increases with compound concentration and molar absorptivity (Meyer, 2010). Thus spike amplitude is not indicative of the importance of the corresponding compound in the medicine's therapeutic effects, but instead provides information regarding a combination of the amount of the compound in the sample and its structural properties.

The retention time of a compound is a function of various conditions such as column length, temperature, and stationary and mobile phase volumes. As a compound's particular behavior in the column is due to its unique structure, and under identical conditions remains unchanged, its retention time provides an indicator to its identity (Meyer, 2010). If after a large number of known compounds are analyzed and a retention time match has not been found for every compound in the study sample, then mass spectrometry (MS) can be used to elucidate the ionic structure of any remaining unknown compounds. MS ionizes a compound through some mechanism, such as electrospray ionization, and then measures ion abundance and the mass to charge ratio of the ions. Plotting ion abundance over mass to charge ratio provides an ion chromatogram of the compound and can be used to identify it.

Between HPLC experiments, a number of experimental factors can vary, altering retention times of compounds. For example, unique calibrations of different HPLC equipment can result in slight differences in column temperature, gradient program, and the rate of flow of the mobile phase. These differences lead to variation in the separation and velocity of compounds which alters retention time and thus, spike timings, across experiments. In this chapter we wish to emulate this phenomenon.

We consider the setting in which we have samples of rhubarb with identical compo-

sitions. The samples are analyzed using HPLC under eight experimental conditions set to induce changes in retention times of sample compounds. Under the same experimental conditions, the resulting chromatograms are identical apart from measurement error. Across experiments, the differing conditions induce nonlinear shifting of spikes.

Table 2.1: Gradient program and flow rate for the eight experiments. Shown are the proportion of 0.1% phosphoric acid aqueous solution (A) and acetonitrile (B) in the mobile phase and the timings of the changing of relative proportions along with the flow rate of the mobile phase. The italicized values indicate the altered parameters in comparison to those in condition 1.

	1		2		3		4	
Time (min)	A	B	A	B	A	B	A	B
0	82	18	*83*	*17*	*83*	*17*	*83*	*17*
2	72	28	72	28	72	28	72	28
6.8	50	50	50	50	50	50	50	50
9.6	41	59	41	59	41	59	41	59
13	0	100	0	100	0	100	0	100
15	82	18	*83*	*17*	*83*	*17*	*83*	*17*
Flow rate (mL/min)	0.21		0.21		*0.18*		*0.20*	
	5		6		7		8[a]	
Time (min)	A	B	A	B	A	B	A	B
0	*83*	*17*	*83*	*17*	*85*	*15*	*85*	*15*
2	72	28	72	28	72	28	72	28
6.8	*55*	*45*	*60*	*40*	50	50	50	50
9.6	41	59	41	59	41	59	41	59
13	0	100	0	100	0	100	0	100
15	*83*	*17*	*83*	*17*	*85*	*15*	*85*	*15*
Flow rate (mL/min)	0.21		0.21		0.21		0.21	

[a] Time at which last two concentration changes occur at 12 and 14 minutes, respectively

For each of the eight experimental conditions, we analyze three samples of rhubarb via HPLC. The conditions are constructed by varying the gradient program and flow rate as described in Table 2.1. Each sample is run using the ACQUITY BEH C_{18} column held at 30 degrees Celsius. The wavelength of the UV detector is set at 260 nm. The mobile phase is comprised of 0.1% phosphoric acid aqueous solution and acetonitrile.

The observed total retention times ranged from 14.0021 to 16.0025 minutes. The observed vectors for conditions 1–8 were of respective lengths {19200, 18000, 18000, 18000, 18000, 18000, 18000, 16800}. Figure 2.1 displays the chromatograms for the 24 samples. To illustrate the ability to identify compounds across different conditions using the fingerprint of a set of known compounds under one single condition, we also analyzed a mixture of seven compounds known to be in rhubarb under condition 1 (Ye et al., 2007) and the resulting chromatogram is displayed in Figure 2.1. The 7 known compounds are (1) gallic acid, (2) catechin, (3) aloe-emodin, (4) rhein, (5) emodin, (6) chrysophanol, and (7) physcion.

2.3. Model

Let y_{ijk} be observation k from chromatogram j in experimental condition i at time point t_{ijk}, where we assume without loss of generality that $t_{ijk} \in [0, 1]$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, m$, $k = 1, 2, \ldots, L_i$. Our model is

$$y_{ijk} = f(\tau_{ijk}) + e_{ijk}, \qquad \tau_{ijk} = g(t_{ijk}, \mathbf{b}_i) \tag{2.1}$$

where f is an unknown spiky function, τ_{ijk} is the warped time relating to t_{ijk} via some smooth monotonic function g, indexed by unknown condition-level parameter vector $\mathbf{b}_i$, and $\mathbf{e}_{ij} = (e_{ij1}, \ldots, e_{ijL_i})^T$ is the vector of measurment errors for chromatogram j with mean 0 and covariance matrix $\sigma^2 V_i(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of unknown parameters and σ^2 is an unknown scale parameter. V_i depends on i through its dimension. In this chapter we assume equal number of experiments per condition and each experiment within a condition are observed on the same time vector, however these assumptions can be relaxed.

The common shape function f is expanded into the Battle-Lemarié spline wavelet

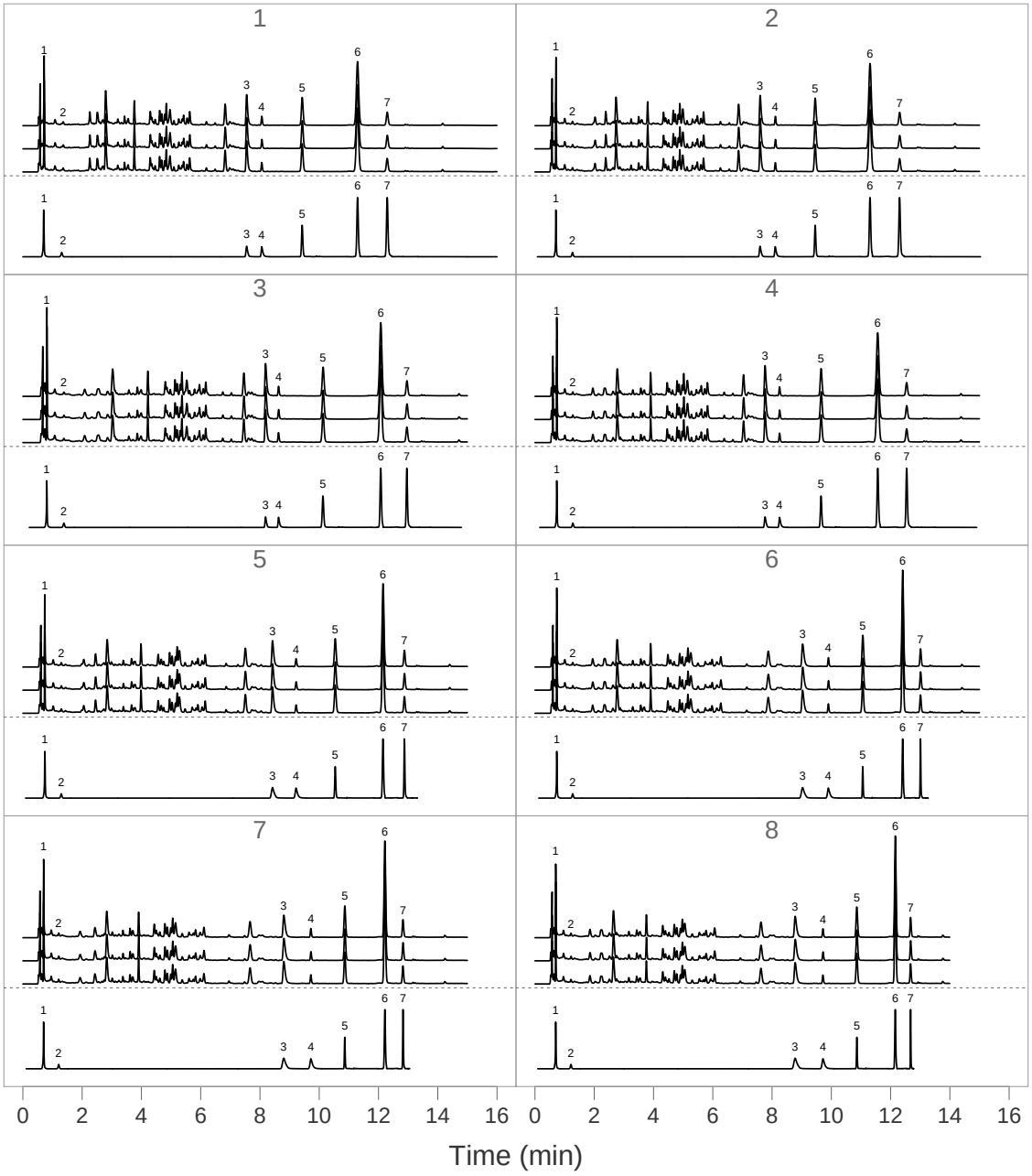

Figure 2.1: Top portion of each panel: chromatograms for one of the eight experimental conditions. The three samples within each condition are displayed with an arbitrary vertical shift. The set of known compounds are denoted as: (1) gallic acid, (2) catechin, (3) aloe-emodin, (4) rhein, (5) emodin, (6) chrysophanol, and (7) physcion. Bottom portion of each panel: the chromatogram from known compound sample under Condition 1 transformed to the time in the given condition using the corresponding estimated warping function.

basis, that is $f(\tau_{ijk}) = \sum_{s=1}^{p} x_s(\tau_{ijk})\beta_s$, where $x_s(\tau_{ijk})$ is the sth spline wavelet basis function evaluated at τ_{ijk} and β_s is the corresponding unknown basis function parameter. In wavelet bases, the x_s estimate oscillations in the observed signal at particular dyadic scales and integer translates for different s. Battle-Lemarié spline wavelets possess a number of attractive properties such as orthogonality, regularity, vanishing moments, symmetry, and a closed-form expression (Daubechies, 1992; Unser, 1997).

Some common examples of parametric models for the warping functions include: a linear combination of polynomial functions $g(t_{ijk}, \boldsymbol{b}_i) = \sum_{q=0}^{Q} b_q t_{ijk}^q$. By setting $b_q = 0$ for $q > 1$ we get the location and scale functions as in shape-invariant and self-modeling regression models (Lawton et al., 1972). A more flexible example is a linear combination of B-splines: $g(t_{ijk}, \boldsymbol{b}_i) = B_{q,\boldsymbol{r}}(t_{ijk})\boldsymbol{b}_i$, where $B_{q,\boldsymbol{r}}(t_{ijk})$ is the $1 \times Q$ design matrix of B-spline basis functions of order q and knot sequence $\boldsymbol{r}$ evaluated at t_{ijk} (Brumback and Lindstrom, 2004). We focus on estimation using B-splines with a uniform knot sequence in the current chapter.

The time warping functions are constrained to be monotonic so that no time point in the original scale is mapped to more than one time point in the warped scale. B-splines are a convenient choice of basis as imposing monotonicity in the warping functions can be accomplished by imposing monotonicity in the B-spline coefficients (Schumaker, 2007, chap. 4.9). Numerically, we can either invoke inequality constraints for each i or reparametrize the sequential increments of the coefficients by an exponential transformation resulting in a sequence of increasing B-spline coefficients. That is, for condition i, letting b_{il} be the lth parameter for condition i, we set

$$b_{i1} = b_{i1}, \qquad b_{i2} = b_{i1} + \exp(\delta_{i2}), \qquad \cdots, \qquad b_{iQ} = b_{i(Q-1)} + \exp(\delta_{iQ})$$

where $\delta_{ij} \in (-\infty, \infty)$, $j = 2, \ldots, Q$. As with most wavelet bases, a boundary constraint must also be placed on the wavelet basis, such as periodicity or symmetry. As such, a further constraint must be placed on the B-spline coefficients to maintain identifiability of the warped time vector. For example, assuming periodic wavelets, we can constrain b_{i1} to be in $[-0.5, 0.5]$ to ensure identifiability. This can be accomplished by letting $b_{i1} = \exp(\delta_{i1})/(1 + \exp(\delta_{i1})) - 0.5$ where $\delta_{i1} \in (-\infty, \infty)$. While there is a monotonic constraint among the B-spline coefficients, we do not need to consider explicit constraints in estimating $\delta_{i1}, \cdots, \delta_{iQ}$.

To set a particular experimental condition as the reference time scale, we set the warping function parameters of one condition to correspond to the 45 degree line, which denotes no warping. For example, for cubic B-splines with k uniform knots, the coefficients for the reference condition are equal to $\{0, 1/(3k-3), \ldots, 3z/(3k-3), \ldots, (3k-4)/(3k-3), 1\}$, for $z = 1, \ldots k-2$. Without loss of generality, we select condition 1 as the reference condition.

Thus (2.1) can be expressed as:

$$y_{ijk} = \sum_{s=1}^{p} x_s\{\tau_{ijk}\}\beta_s + e_{ijk}, \qquad \tau_{ijk} = B_{q,r}(t_{ijk})\mathbf{b}_i \tag{2.2}$$

As we wish to restrict the number of parameters to be less than the number of data points per curve, p should be less than $\min(L_i) - Q(n-1)$. Thus the total number of parameters to be estimated is less than $\min(L_i)$.

2.4. Estimation

Let $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijn_i})^T$, $\mathbf{t}_{ij} = (t_{ij1}, \ldots, t_{ijn_i})^T$, and $\mathbf{b}_i = (\mathbf{b}_{i1}^T, \ldots, \mathbf{b}_{im})^T$. The penalized weighted least squares criterion is:

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}\sum_{j=1}^{m}\left(\mathbf{y}_{ij} - \sum_{s=1}^{p} x_s\{g(\mathbf{t}_{ij}, \mathbf{b}_i)\}\beta_s\right)^T V_i^{-1}(\boldsymbol{\theta})\left(\mathbf{y}_{ij} - \sum_{s=1}^{p} x_s\{g(\mathbf{t}_{ij}, \mathbf{b}_i)\}\beta_q\right) + \lambda\sum_{j=1}^{p}\hat{w}_j(|\beta_j|) \tag{2.3}$$

where $\hat{w}_j$ are data-driven weights, and λ is a tuning parameter for the Adaptive Lasso penalty which controls the amount of shrinkage. We use $\hat{w}_j = |\hat{\beta}_j^{\text{NLS}}|^{-1}$ where $\hat{\beta}_j^{\text{NLS}}$ is the unpenalized, nonlinear weighted least squares estimator of β_j as it is a root-n consistent estimator of $\hat{\beta}_j$ (Jennrich, 1969).

Letting $\mathbf{b} = (\mathbf{b}_2^T, \ldots, \mathbf{b}_n^T)^T$, computation can be greatly simplified by setting $\lambda^* = \sigma^2\lambda$, and noting that for fixed $\mathbf{b} = \hat{\mathbf{b}}$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, solving (2.3) for $\boldsymbol{\beta}$ simplifies to a penalized least squares problem with the Adaptive Lasso penalty:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{\left\|\mathbf{y}_{ij}^* - \sum_{s=1}^{p} x_s^*\{g(\mathbf{t}_{ij}, \hat{\mathbf{b}}_i)\}\beta_s\right\|_2^2 + \lambda^*\sum_{j=1}^{p}\hat{w}_j(|\beta_j|)\right\} \tag{2.4}$$

where $\mathbf{y}_{ij}^* = V_i^{-1/2}(\hat{\boldsymbol{\theta}})\mathbf{y}_{ij}$ and $x_s(\cdot)^* = V_i^{-1/2}(\hat{\boldsymbol{\theta}})_s x_s(\cdot)$ where $V_i^{-1/2}(\boldsymbol{\theta})_s$ is the sth row of $V_i^{-1/2}(\boldsymbol{\theta})$. Minimizing (2.4) can be accomplished using standard algorithms for ℓ_1-penalized regression (Zou, 2006). For fixed $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, using the reparameterization of $\mathbf{b}_i$ described in the previous section, solving (2.3) for $\mathbf{b}$ is equivalent to solving the following unpenalized nonlinear least squares problem for

$\boldsymbol{\delta} = (\delta_{21}, \ldots, \delta_{2Q}, \ldots, \delta_{n1}, \ldots, \delta_{nQ})^T$:

$$\hat{\boldsymbol{\delta}} = \arg\min_{\boldsymbol{\delta}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| \mathbf{y}_{ij}^* - \sum_{s=1}^{p} x_s^*(g\{\mathbf{t}_{ij}, \boldsymbol{\delta}_i)\}\hat{\beta}_s \right\|_2^2 \tag{2.5}$$

Similarly, for fixed $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$ and $\mathbf{b} = \hat{\mathbf{b}}$, solving (2.3) for $\boldsymbol{\theta}$ is also an unpenalized nonlinear least squares problem:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| V_i^{-1/2}(\boldsymbol{\theta}) \left(\mathbf{y}_{ij} - \sum_{s=1}^{p} x_s(g\{\mathbf{t}_{ij}, \hat{\mathbf{b}}_i)\}\hat{\beta}_s \right) \right\|_2^2 \tag{2.6}$$

Hence we propose the following iterative back-fitting algorithm to estimate $\boldsymbol{\beta}$, $\mathbf{b}$, and $\boldsymbol{\theta}$:

At iteration k:

1. Fix the $\mathbf{b}$ and $\boldsymbol{\theta}$ at their estimates from iteration $(k-1)$, denoted $\mathbf{b}^{(k-1)}$ and $\boldsymbol{\theta}^{(k-1)}$. Solve (2.4) for $\boldsymbol{\beta}$, obtaining $\hat{\boldsymbol{\beta}}^{(k-1)}$.

2. Fix $\boldsymbol{\beta}$ at $\hat{\boldsymbol{\beta}}^{(k-1)}$ from step 1 and $\boldsymbol{\theta}$ at $\boldsymbol{\theta}^{(k-1)}$. Solve (2.5) for $\boldsymbol{\delta}$, obtaining $\hat{\boldsymbol{\delta}}^{(k-1)}$ and $\hat{\mathbf{b}}^{(k-1)}$.

3. Fix $\boldsymbol{\beta}$ at $\hat{\boldsymbol{\beta}}^{(k-1)}$ and $\mathbf{b}$ at $\hat{\mathbf{b}}^{(k-1)}$ from steps 1 and 2, respectively. Solve (2.6) for $\boldsymbol{\theta}$, obtaining $\hat{\boldsymbol{\theta}}^{(k-1)}$.

Iterate steps 1–3 until convergence. At convergence, estimate σ^2 via

$$\hat{\sigma}^2 = \frac{1}{m\sum_{i=1}^{n} L_i} \sum_{i=1}^{n} \sum_{j=1}^{m} \left\| V_i^{-1/2}(\hat{\boldsymbol{\theta}}) \left(\mathbf{y}_{ij} - \sum_{s=1}^{p} x_s\{g(\mathbf{t}_{ij}, \hat{\mathbf{b}}_i)\}\hat{\beta}_s \right) \right\|_2^2 \tag{2.7}$$

The algorithm as a whole minimizes the unified criterion, (2.3), and by using line searches to solve (2.4), (2.5), and (2.6), is equivalent to a blockwise coordinate descent

algorithm (Tseng, 2001).

Convergence of the back-fitting algorithm depends on good initial estimates. Specifically, the algorithm requires good initial estimates of the warping function parameters. One method for obtaining initial estimates is to set $\lambda = 0$ and solve (2.3). In this situation, (2.3) is an unpenalized nonlinear least squares problem. As standard nonlinear least squares procedures are iterative, initial estimates of the warping function parameters are necessary for this step, as well. These can be obtained by first estimating the warping functions via dynamic time warping (Kassidas et al., 1998; Giorgino, 2009) and then modeling the estimated warping functions using monotonic B-splines.

It is important to note that since the $\mathbf{b}_i$ are updated at each iteration, the wavelet basis design matrix also changes at each iteration. The adaptive weights, $|\hat{\boldsymbol{\beta}}^{\text{NLS}}|^{-1}$, are calculated at the initial estimation step using the unpenalized estimates and are kept constant throughout the estimation procedure.

The estimates obtained from (2.3) depend on the particular value of the tuning parameter, λ, as the proposed back-fitting algorithm is for a fixed λ. Numerous techniques to choose the optimal value of λ exist with varying degrees of properties. In penalized least squares problems with the smoothly clipped absolute deviation (Fan and Li, 2001) chosen as the penalty, it has been shown that BIC consistently chooses the correct model (Wang et al., 2007). This results follows when using the Adaptive Lasso. We use the BIC to choose the optimal value of λ.

Our proposed method does not require normality. In practice, when applying our procedure to chromatographic data, we have found that the tails of the distribution of the residuals are fatter than those of normally distributed residuals. The robustness of

the Lasso to fat-tailed errors has been studied previously. Finite sample performance is affected by fat tails, however Fan et al. (2013) showed that Lasso estimates retain sign consistency if the signal is large enough. Bunea and Gupta (2010) and Sang and Sun (2012) consider Lasso and SCAD estimates under correlated data, including auto-regressive errors, and show that their asymptotic properties are still valid. As chromatograms exhibit a large signal-to-noise ratio and contain thousands of data points, heavy-tailed errors should not pose a problem in practice.

The model estimates possess attractive sampling properties. As in Zou (2006), we assume, without loss of generality, that there is a $p_0 < p$ such that $|\beta_k| > 0$ for $k \leq p_0$ and $\beta_k = 0$ for $p_0 < k \leq p$. Thus the true active set of $\boldsymbol{\beta}$, $\mathcal{A}$, is $\{1, 2, \ldots p_0\}$. Let $\hat{\mathcal{A}}_n = \{j : \hat{\beta}_j \neq 0\}$ be the estimated active set of $\boldsymbol{\beta}$, where $\hat{\beta}_j$ is the estimate of β_j from the adaptive lasso criterion.

Theorem 1. *Let* $\boldsymbol{\psi} = (\boldsymbol{\theta}^T, \boldsymbol{b}^T, \boldsymbol{\beta}^T)^T$ *and* $\boldsymbol{\psi}_{\mathcal{A}} = (\boldsymbol{\theta}^T, \boldsymbol{b}^T, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$. *Under the regularity conditions described in Fan and Li (2001); Zhang and Lu (2007); Bunea and Gupta (2010), the estimates obtained by minimizing* (2.3) *satisfy the following:*

1. $\lim_{n\to\infty} P(\hat{\mathcal{A}}_n = \mathcal{A}) = 1$.

2. $\sqrt{n}(\hat{\boldsymbol{\psi}}_{\mathcal{A}} - \boldsymbol{\psi}_{\mathcal{A}}) \to \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{C}^{-1})$.

where $\frac{1}{n}\sum_i f(\boldsymbol{\psi}_{\mathcal{A}}) V^{-1}(\boldsymbol{\theta}) f(\boldsymbol{\psi}_{\mathcal{A}})^T \to \boldsymbol{C}$.

The proof of Theorem 1 follows very closely to the corresponding proofs in Fan and Li (2001); Zhang and Lu (2007); Bunea and Gupta (2010), so it is omitted. Theorem 1 shows that the adaptive lasso estimators obtained from (2.3) are variable selection consistent, root-n consistent, and asymptotically normal. Thus, the estimators possess the oracle property (Donoho and Johnstone, 1994). The root-n consistency is achieved due in part to the parametric rate of the estimates of the warping function

parameter coefficients. The asymptotic normality of the estimates enables the construction of asymptotic confidence intervals for the estimated fingerprint and warping functions which reflect the combined variability in estimating $\boldsymbol{\beta}$, $\boldsymbol{b}$, and $\boldsymbol{\theta}$ jointly. For example, by taking a Taylor expansion of f about the true value of $\boldsymbol{\psi}$, we estimate the approximate asymptotic variance for the estimated fingerprint as $\hat{\sigma}^2 s(\hat{\boldsymbol{\psi}})\boldsymbol{C}^{-1}s(\hat{\boldsymbol{\psi}})^T$, where $s(\hat{\boldsymbol{\psi}}) = \partial f(\hat{\boldsymbol{\psi}})/\partial\boldsymbol{\psi}$.

2.5. Application to Data Example

We applied (2.2) to the rhubarb chromatographic dataset described in Section 2.2. For computational efficiency, we subsampled the data such that the data vectors for the eight conditions were of respective lengths {2743, 2572, 2572, 2572, 2572, 2572, 2572, 2400}. The observed time vectors were mapped to $[0, 1]$ by setting the maximum observed time, 16.0025 min., to 1.

Spline wavelets can be evaluated at any arbitrary time point, however the computation involves nested summations of many terms, so for computational speed, we computed the design matrix at arbitrary time points using the following procedure: a reference design matrix is evaluated on a grid of 2^{17} time points in $[0, 1]$. Design points are computed via a table-look-up and for time values not in the reference design matrix, interpolation between adjacent points is used. We chose the finest wavelet scale to be 10, resulting in 2048 basis functions where the $2^i, 2^i + 1, \ldots, (2^{i+1} - 1)$th functions correspond to the wavelet basis functions at scale i. The finest level wavelet scale is chosen based on the data so that the resolution is fine enough to capture the narrowest spikes. We assume periodic boundary conditions for convenience.

To model the warping functions, we used monotonic cubic B-splines with 14 uniform knots. From our simulation presented in the subsequent section, uniform knots

appears to work well for a moderate amount of warping. We assume a first-order auto-regressive correlation structure, thus $\boldsymbol{\theta}$ is a scalar, denoted θ, and represents the correlation between points one unit of time apart.

The algorithm was written in MATLAB and run on an Intel Xeon CPU E7-4860. For each value of λ, the algorithm converged in less than five outer iterations, and finished, on average, in 10 minutes. The adaptive lasso parameter λ, found using BIC and a golden search algorithm, was 0.489. The estimated variance of the measurement error was 1.24×10^{-3} and the estimate for θ was 0.698.

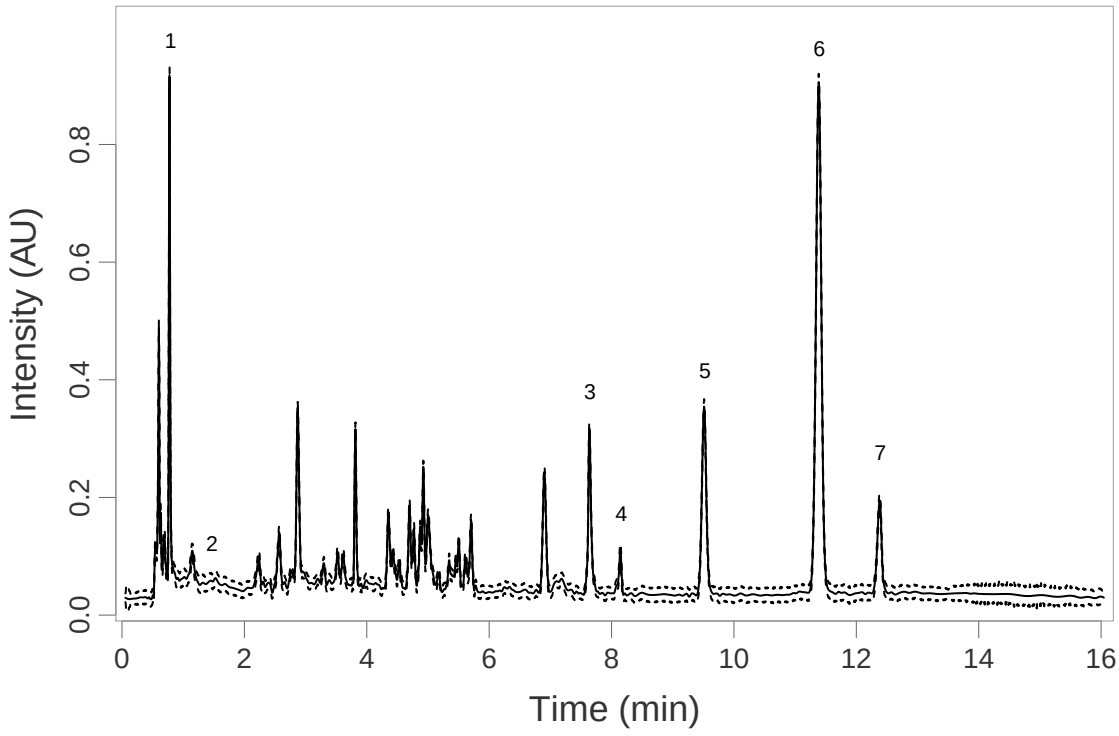

Figure 2.2: Population averaged fingerprint (solid) from the estimates along with 95% confidence bands (dashed). Identified compounds are denoted: (1) gallic acid, (2) catechin, (3) aloe-emodin, (4) rhein, (5) emodin, (6) chrysophanol, and (7) physcion.

The population averaged estimate of the standardized chromatographic fingerprint along with pointwise 95% confidence intervals are displayed in Figure 2.2. The time

scale corresponds to the real time scale under condition 1. The population-averaged curve captures the shape of the individual chromatograms quite well, resulted from a good alignment across conditions. As the known compounds were analyzed under condition 1, spike locations between the reference chromatogram and the estimated fingerprint can be compared. The seven spikes that correspond to the known compounds are labeled in Figure 2.2.

To identify the seven known compounds in other conditions, we use the inverse estimated warping functions to warp the reference chromatogram in condition 1 to the time scale of each other condition. The estimated reference chromatograms are displayed in Figure 2.1 along with the raw data. By comparing with the estimated reference chromatograms, the seven known compounds are identified in each condition. These also show that we are able to estimate each warping function accurately.

The final estimate of the fingerprint shape contained 427 nonzero wavelet coefficients, while the remaining 1621 were shrunk to 0. This highlights the sparse nature of the fingerprint shape as a little more than 79% of the wavelet basis functions are irrelevant. The confidence intervals for the fingerprint can be used to determine whether a spike is artificial or real based on whether the corresponding confidence interval contains 0 or not. For example, the amplitude of the spike corresponding to catechin is small however the confidence interval at the peak of the spike is $[0.014, 0.044]$, suggesting it is present in rhubarb, which coincides with previous research (Ye et al., 2007).

The groups of tightly-packed spikes located between 4 and 6 minutes in the fingerprint can either correspond to single compounds or multiple compounds which were not well separated in the HPLC column. If we wished to identify the compounds in this region of the fingerprint, parameters of the experiment would have to be adjusted so

that these particular compounds are better separated and travel through the column at more distinct velocities. This would result in more spacing among spikes in this region of the chromatograms.

Figure 2.3 displays the estimated warping function for each of the eight conditions. Confidence intervals were computed for the warping functions however due to the low variance of the B-spline coefficients ($\text{Var}(b_i) \sim \mathcal{O}(10^{-4})$), the bands are visually indistinguishable from the estimated warping functions. The low order of variance is due to the large effective sample size.

The estimated warping functions represent the distortion in retention time of conditions 2–8 in comparison to condition 1. The relative proportions of phosphoric acid and acetonitrile differs by only 1% at time 0 between conditions 1 and 2, and this difference caused little to no warping, as evidenced by the estimated warping function of condition 2. Conversely, a 3% change, as in condition 7, results in substantial warping after minute 4. The estimated warping functions for conditions 3 and 4 in comparison to those for conditions 5–8 suggest that flow rate does not have quite as large of an impact on retention time as the composition of the mobile phase.

2.6. Simulation

A vast number of mathematical functions have been proposed to simulate chromatographic data (Di Marco and Bombi, 2001). One such example is the Laplace distribution function. We simulated the true fingerprint shape by overlaying a fixed number of spikes, where each spike was generated via $f(t) = a \exp(-|t - c|/b)/(2b)$ where a controls the amplitude of the spike, c the location of the maximum of the spike, and b is a scale parameter, affecting the width and tails of the spike. We used twelve spikes with locations randomly chosen from $[0.1, 0.9]$. The amplitudes of the eight spikes

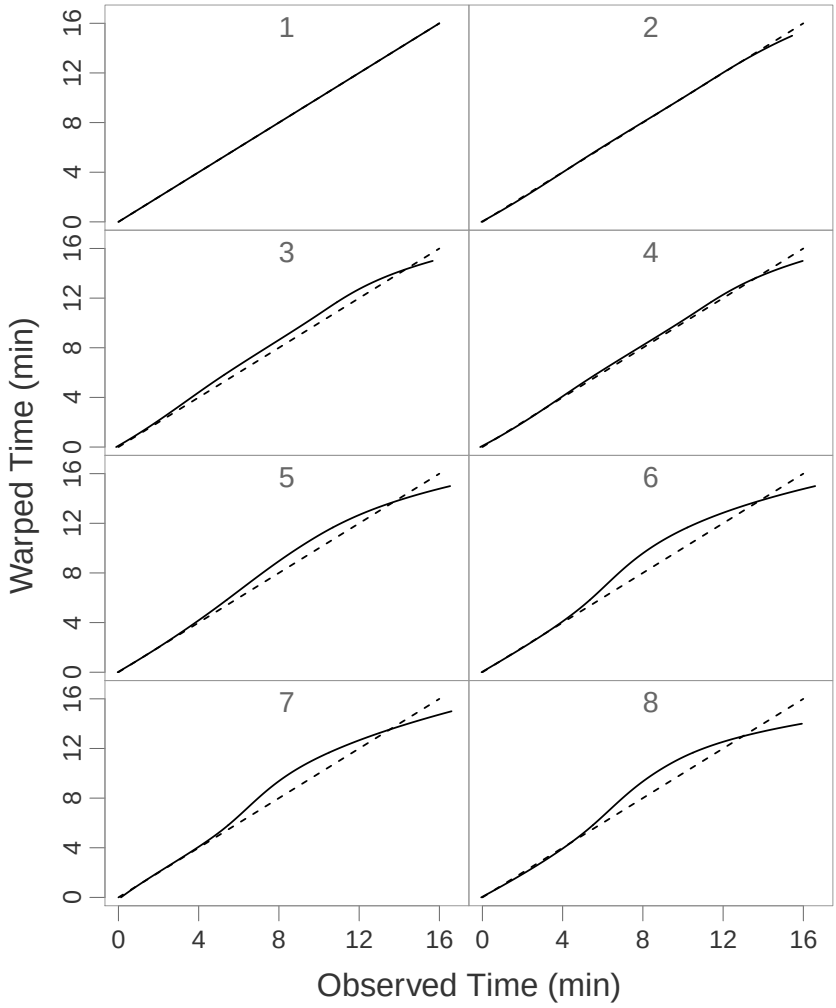

Figure 2.3: The estimated warping functions (solid) along with the 45 degree line denoting no warping for each condition (dashed).

were generated from a $N(20, 25)$ distribution and the scale parameters were sampled with replacement from the set $\{0.45, 0.50, \ldots, 0.75\}$.

To simulate time warping similar to those seen in the rhubarb data, the time warping functions were generated by the following logistic function, $g(t) = [1 + \exp(-14t + 7))]^{-1}$. The point-wise mean between $g(t)$ and the 45 degree line, denoting no warping, can be iteratively calculated to generate warping functions with similar shape, but less and less degrees of warping. If we let z be the number of iterated averages between

the function and the 45 degree line, then the warping function deviates less from the 45 degree line as z increases and converges to the 45 degree line as $z \to \infty$. The inverse logistic function was used to generate one warping function with a mirrored shape to the rest. For notational brevity, a negative value of z corresponds to iterated averaging the 45 degree line with the inverse logistic function z times.

A set of four warping functions were generated using $z = \{\infty, 4, 3, -3\}$, representing four different experimental conditions, and the warping functions were estimated using three, four, and five uniform knots. For each condition we generated 4 fingerprints sampled on 1000 equispaced time points in $[0, 1]$. We added t-distributed noise, with 9 degrees of freedom, to each curve. The t-distribution was used as it has longer tails than the normal distribution. Figure 2.4 displays an example of a noisy fingerprint from each condition.

For comparison, we also fit a two-step model where the first step used dynamic time warping to align the curves and the second applied a wavelet thresholding to the aligned curves and averaged across the curves to obtain the population averaged fingerprint. The dynamic time warping step was accomplished using the 'dtw' package in R (Giorgino, 2009) and the wavelet thresholding was performed using the 'wavethresh' package in R (Nason, 2008). We simulated 100 datasets and performed the four methods on each set.

The proposed model was fit using $2^9 = 512$ cubic Battle-Lemarié spline wavelet basis functions to model the shape and cubic B-splines to model the warping function. For each value of λ, our algorithm converged in less than five outer iterations and finished, on average, in 40 seconds. We also used 512 cubic Battle-Lemarié spline wavelets for the two-step method and the Bayesian approach of Abramovich et al. (1998) was used to select λ.

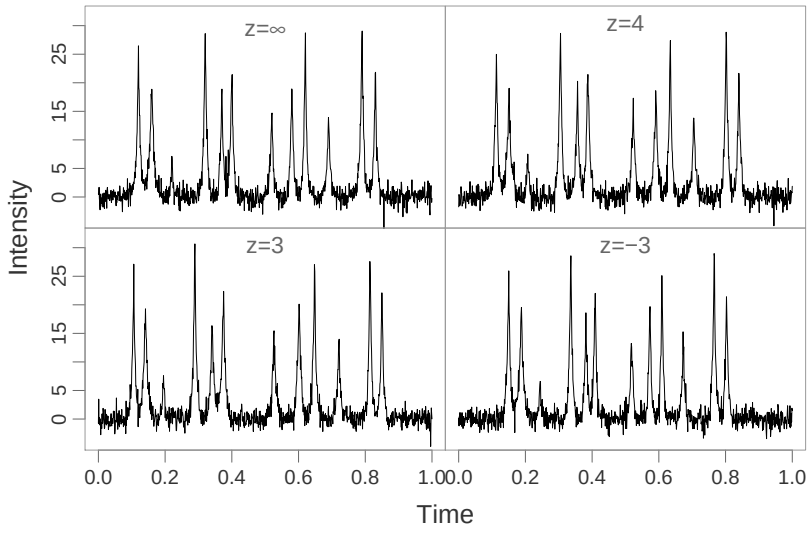

Figure 2.4: Simulated warped curves with 12 peaks and additive t-distributed noise. The curve was generated using the Laplace distribution function and the warping functions were generated by logistic and inverse-logistic functions.

To assess the fits, boxplots of the mean squared error (MSE) between the true and estimated fingerprints were used, which are displayed in Figure 2.5. First, the proposed 4- and 5-knot models performed the best. The 3-knot model performed the worst of the four methods. This is due to the fact that three knots does not provide enough flexibility in capturing the shape of the warping functions. The boxplots suggest that by estimating the warping function using B-splines with at least four uniform knots, the degree of the warping does not appear to affect the MSE in any substantial way. The two-step approach does not perform as well as the proposed 4- and 5-knot models, though it outperforms the 3-knot model.

Figure 2.6 displays the mean of the estimated warping functions across all 100 simulations for the proposed 4-knot model overlaid on the true warping functions. The proposed model with four knots estimates the true warping functions well except at the ends of the tails, where there is slight deviation. Though we only consider uniform knots and there are numerous other possible parametric models for the warping

function, our simulation shows that a large amount of flexibility is not required for quality performance of the modeling procedure.

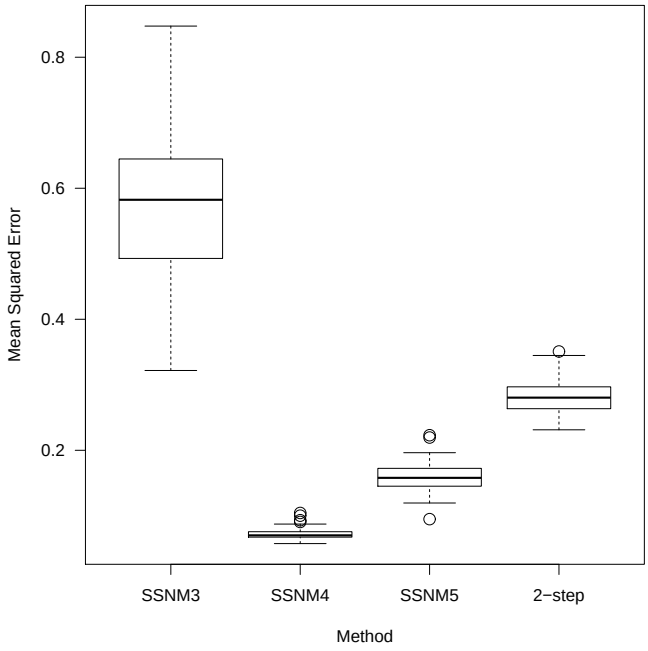

Figure 2.5: Boxplots of the mean square error of the estimated fingerprints for the proposed procedure using 3-, 4-, and 5-knots and the two-step procedure using dynamic time warping and wavelet thresholding.

2.7. Discussion

We have proposed a sparse semiparametric nonlinear model for the registration of chromatographic data and establishment of a standardized fingerprint. The common shape of the chromatograms is modeled via data-driven basis expansion. Curve registration is accomplished by parametric modeling of the time warping functions. As chromatograms are sparsely-structured curves, we impose the adaptive lasso penalty on the common shape basis function parameters to induce sparsity in the estimated fingerprint. A unified penalized weighted least squares criterion is minimized to simultaneously register chromatograms, perform model selection on the fingerprint

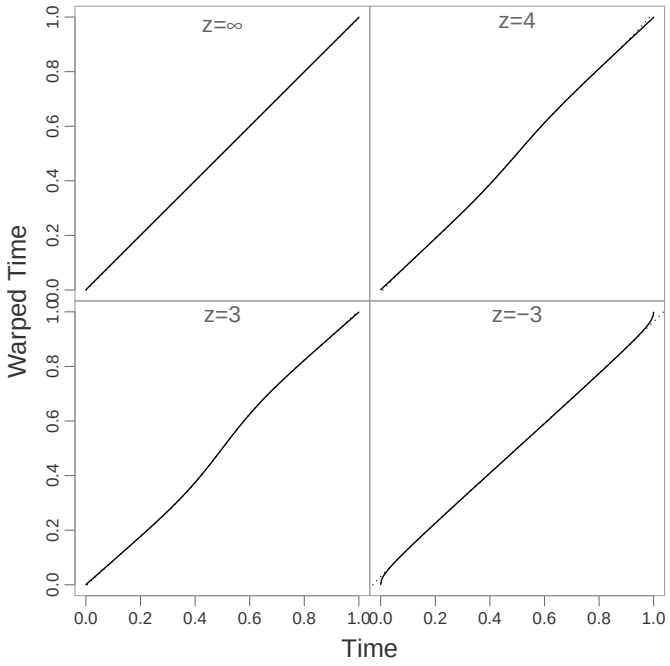

Figure 2.6: True (solid) and mean of 100 estimated warping functions (dashed) using 4 uniform knots.

shape, and estimate the model parameters. From the penalized criterion naturally arises a back-fitting algorithm which is equivalent to a blockwise coordinate descent algorithm. The unified criterion results in a unique solution and allows the study of asymptotic properties, as opposed to two-step methods which can guarantee neither. Our resulting estimators possess the oracle property.

We applied our model to a chromatographic dataset of the TCHM, rhubarb. Our model was effective in establishing a standardized fingerprint of rhubarb based on chromatograms arising from different experimental conditions. A sample of known compounds was analyzed under a single condition to identify compounds in the fingerprint. We demonstrated that known compounds need only be analyzed under a single condition, substantially decreasing both time and cost of chromatographic experiments.

We use BIC to select the tuning parameter λ however other tuning parameter selection techniques such as AIC, GCV, and Stein's unbiased risk exist, and in our simulations, performed similarly. We have focused the discussion on modeling the common shape using a linear combination of wavelet basis function and the warping functions using monotonic B-splines, however the proposed model can be easily extended to include multiple population average curves or in the situation that the observations relate to the population-average curves through some known nonlinear function.

CHAPTER 3

Semiparametric Nonlinear Mixed Effects Models for Sparsely-Structured Functional Data

3.1. Introduction

Functional data arise in numerous applications in pharmacological, environmental, and biomedical research including chromatographic, weather pattern, biomarker, and growth hormone studies. The cornerstone of functional data analyses is viewing the observed data as samples of unknown functions and modeling the data as curves. A common objective in functional data analyses is aligning data curves and estimating and comparing their structures, such as spikes.

Our research is motivated by a study of the chemical composition of the medicinal herb, Andrographis paniculata. Medicinal herbs have been used for the treatment of a number of medical ailments for thousands of years (Duke, 2002; Chao and Lin, 2010). As the exact chemical composition of an herb differs based on its particular species and origin, properties of the seed and field used to grow the herbs, and the particular processes used in the growing, harvesting, and storage of the herbs (Leung and Cheng, 2008), tools and methodologies for composition determination and the identification of compounds not shared between, say, production sites are vital for the quality control of herbal medications.

High Performance Liquid Chromatography (HPLC), which is a set of techniques to separate compounds of biological mixtures, has played a large role in medicinal herb research. HPLC analyzes an herb sample and outputs a curve, termed chromatogram,

characterized by a number of spikes, where each spike corresponds to a particular compound in the herb. Since the location of spikes in the curve can be used to identify the compounds and the chemical composition is unique to an herb, chromatograms provide a fingerprint of an herb. Consider Figure 3.1 which displays chromatograms of samples of Andrographis paniculata collected from two types of production sites: five sites which comply with Good Agricultural Practices (GAP) and five pharmacies whose GAP compliance unknown and analyzed using HPLC. The objective of the study is to compare the compositions of samples between the two groups, GAP and pharmacy, or market sites, and identify compounds present in one group of herbs but not the other. The presence of any discordant compounds will suggest deficiencies in the production processes of the market sites. Further details regarding GAP are described in Section 3.4. Through HPLC, comparing compositions between GAP and market herbs amounts to comparing their group-averaged chromatographic fingerprints.

However, a number of characteristics of chromatograms evident in Figure 3.1 impede the direct establishment of chromatographic fingerprints and their comparison. First, spiky curves, such as chromatograms, are not well estimated by parametric functions. This common issue in functional data analyses can be addressed by expanding the curve into some basis system such as splines or wavelets (Ramsay and Silverman, 2005). Second, due to differing experimental and environmental conditions, the timings of spikes in chromatograms often vary across experiments. The site-level transformations of the curves induce misalignment of spikes and prevent the establishment and comparison of group-averaged curves. Curve registration is a set of methodologies in which features of interest are aligned by warping the individual curve argument, often time, to a common scale (Ramsay and Li, 1998). This

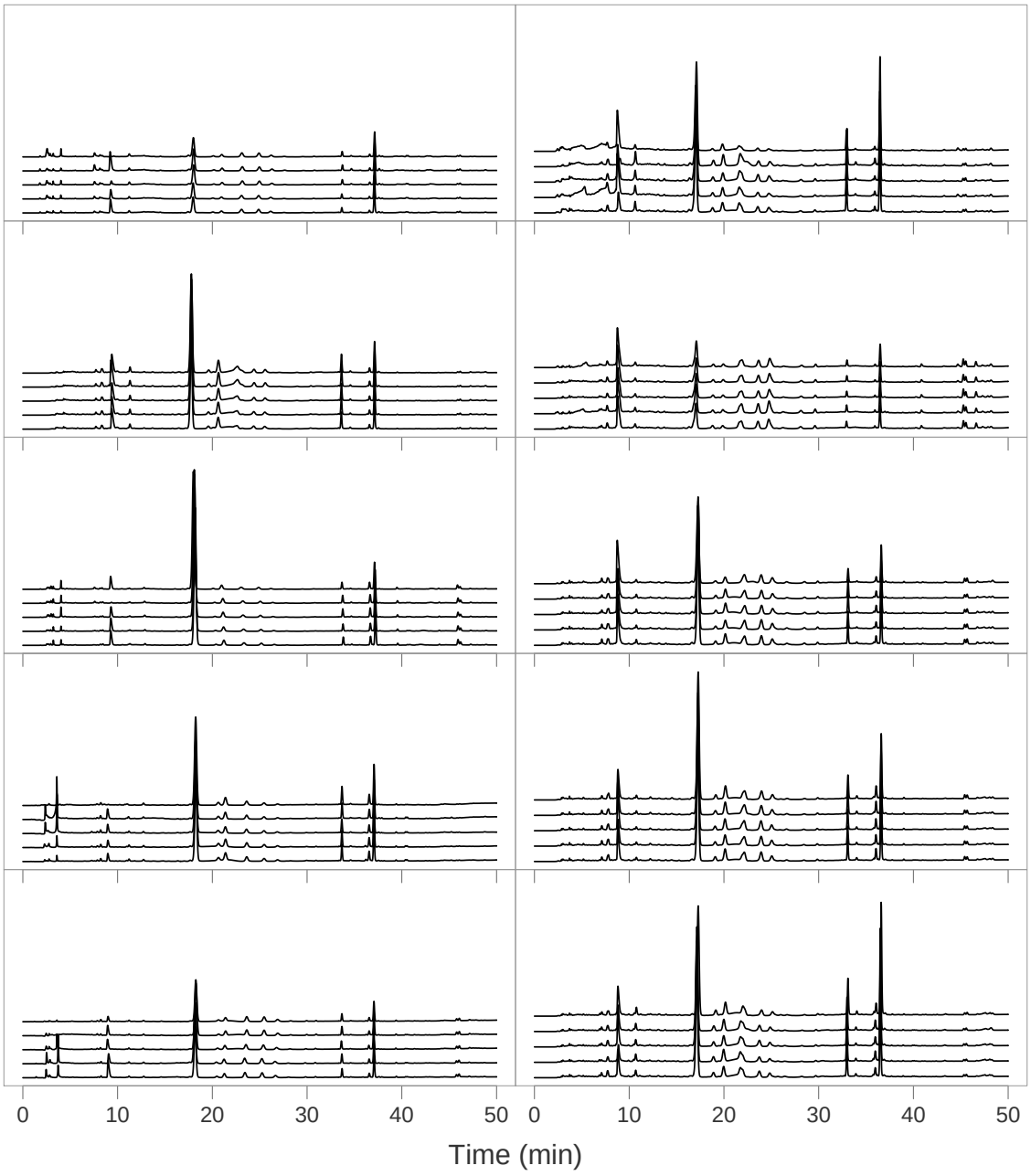

Figure 3.1: Left panel: chromatograms for the five GAP-adherent production sites. Right panel: chromatograms for the five market sites. Within each site, five samples were analyzed and their chromatograms are displayed with vertical shifts.

is accomplished by modeling the unknown, monotonic time warping function within the basis expansion of the fingerprint shapes, as in shape-invariant and self-modeling regression models (SEMOR) models (e.g. Lawton et al., 1972; Ke and Wang, 2001; Brumback and Lindstrom, 2004). Since there are a large number of production sites under study and the number of parameters in SEMOR models increases with the number of sites. Thus, along with the fact that we wish to compare group-averaged curves while taking into account the correct sources of variability, warping function parameters should be modeled as random (Ke and Wang, 2001; Guo, 2002). Thus, curve registration offers a flexible, nonlinear framework to align and estimate spikes.

Finally, thousands of data points are collected per chromatogram but the curves exhibit only a small number of sharp, narrow spikes. The large sections of the curves between spikes are flat and do not contain meaningful signal. Due to their sparse nature, estimating curves via basis expansion results in the irrelevancy of a large number of basis functions as they correspond to the flat section and are merely estimating noise. As their true coefficient value is 0, these superfluous variables need to be excluded to avoid overfitting. As such, identifying, or selecting, important functions while leaving out irrelevant functions is a necessary component to modeling sparsely-structured functional data. As we wish to select the functions whose true coefficient values are not equal to zero, this falls within the model selection framework.

There exist numerous methods for the analysis of chromatographic data. For example, Morris et al. (2008) model mass spectrometry data, which is similar in nature to chromatographic data, using a wavelet-based functional mixed effects model (Morris and Carroll, 2006). The model is an extension of the functional mixed effects models of Guo (2002) to the sparsely-structured setting and can accommodate a wide range of shape functions and subject-specific deviations and extend to multiple

levels of nested structures. However the models of Guo (2002); Morris and Carroll (2006); Morris et al. (2008) are linear in the mixed effects and so data which exhibit subject-specific transformations of the common shape functions cannot be adequately modeled within these frameworks. Much like most current methods for chromatographic data, Morris et al. (2008) circumvents this issue by aligning curves in a pre-processing step. By aligning individual curves and estimating mean curves in separate steps, the variability in the alignment step is ignored in the second step. Thus variance estimates, which do not account for both phase and measurement error variability, will be incorrect and valid statistical inference cannot be drawn. A model which allows for the mixed effects to enter into the model nonlinearly is needed.

Ke and Wang (2001) proposed a general class of semiparameteric nonlinear mixed effects models (SNMM), which includes extensions of nonlinear mixed effects and SEMOR models. SNMMs can handle numerous curve shapes encountered in practice and allow for nonlinearity in the mixed effects specification. Ke and Wang (2001) propose estimating the mean functions nonparametrically with parametrically-specified covariates using a double-penalized likelihood criterion. They propose an iterative back-fitting estimation algorithm where the first step estimates the shape function via smoothing splines and the second estimates fixed and random effects and variance components via a nonlinear mixed effects model through linearization of the likelihood about the random effects. However, the second step of the estimation algorithm in Ke and Wang (2001) ignores the fact that the random effects are allowed to enter into the shape functions (Lin and Zhang, 2001; Elmi et al., 2011). As shown in Lin and Zhang (2001), their estimation procedure is equivalent to a two-step procedure in which each step fits a separate mixed effects model. As the two mixed effects model do not correspond to a unified criterion, convergence to the true parameters

is not guaranteed and variance estimates will be incorrect (Elmi et al., 2011). Using a unified criterion for both steps and establishing sampling properties of model estimates in this framework will result in a procedure which guarantees convergence and variance estimates that reflect the joint estimation of the shape function and any population-level and subject-specific covariates, enabling correct statistical inference. Furthermore, Ke and Wang (2001) use an ℓ_2 penalty on the common curve shapes which induces smoothness in its solutions. All covariates are included in its solutions, thus model selection is not performed and a sparse solution is not obtained.

Penalized likelihood with a sparsity penalty is a popular framework for inducing sparsity in model estimates and results in simultaneous model selection and estimation (Fan and Li, 2001). A number of sparsity penalties with varying properties have been studied extensively, including the Least Absolute Shrinkage and Selection Operator (Lasso), Adaptive Lasso, and the Smoothly Clipped Absolute Deviation (SCAD) penalties (Tibshirani, 1996; Zou, 2006; Fan and Li, 2001). These penalties shrink irrelevant covariates continuously to zero and retain and estimate the nonzero covariates. The use of the Adaptive Lasso and SCAD penalties result in procedures which asymptotically select the correct submodel and estimates that are consistent at the optimal rate and asymptotically normal, often termed the oracle property (Donoho and Johnstone, 1994; Fan and Li, 2001; Zou, 2006). The consistency and asymptotic normality of model estimates enable the establishment of statistical inference on data curves. Arribas-Gil et al. (2013) proposed Lasso-type estimators for SNMMs as an extension of Ke and Wang (2001), replacing the ℓ_2 penalty with Lasso-type penalties. Their procedure has the same drawbacks as in Ke and Wang (2001) as reliance of the shape function on model parameters is ignored in the first step of their iterative algorithm. Model selection and sparsity have also been extended to the generalized

linear mixed-effects setting (Schelldorfer and Bühlmann, 2011) and the functional data setting (James et al., 2009; Ferraty et al., 2010). However generalized linear models do not allow for random effects to enter into the mean functions and the procedures of (James et al., 2009) and (Ferraty et al., 2010) only consider linear models or models whose design points are fixed. Thus the applications of these methods is limited in the same way as Guo (2002) and Morris and Carroll (2006).

In this chapter we propose a class of sparse semiparametric nonlinear mixed effects models (SSNMM) for the registration and comparison of sparsely-structured functional data. We assume subjects within groups share an underlying sparsely-structured function and exhibit subject-specific transformations in the group-level shape. We use data-driven basis expansion to estimate the shape functions. Sparsity is induced in the shape estimates through the use of the Adaptive Lasso. The parametrically-specified subject-specific transformations of the mean curves enter into the argument of the basis functions, enabling curve registration to be performed. Furthermore, to correctly account for the sources of variability, the transformation effects are modeled as random. Penalized marginal likelihood provides a unified criterion for estimation and model selection, guarantees convergence, and enables the establishment of statistical inference with correct variance estimates. Due to the nonlinearity in the random effects, the marginal likelihood involves an integral with no closed form. We implement the Laplace Approximation to the marginal likelihood for the estimation of model parameters. We prove that, in our setting, our Adaptive Lasso procedure with the Laplace approximation results in estimates that possess the oracle property. We propose a computationally-efficient back-fitting algorithm which allows for the utilization of existing algorithms. The performance of the SSNMM is assessed through simulation and its application to the chromatographic data set.

The remainder of the chapter is organized as follows: in Section 3.2, we introduce and describe SSNMMs. In Section 3.3 we discuss the estimation and algorithmic details of our procedure and the sampling properties of model estimates. In Section 3.4 we discuss medicinal herbs, GAP, and apply the proposed model to the chromatographic data set. In Section 3.5, we assess our procedure through simulation. We conclude with discussion in Section 3.6.

3.2. Model

The class of Sparse Semiparametric Nonlinear Mixed Effects Models (SSNMM) is defined as

$$\begin{aligned} y_{ijk} &= f_i(\tau_{ijk}) + e_{ijk}, \quad i = 1, \ldots, l; \quad j = 1, \ldots, m_i; \quad k = 1, \ldots n \\ \tau_{ijk} &= h(\mathbf{b}_{ij}; t_{ijk}), \\ \mathbf{b}_{ij} &\sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{D}(\boldsymbol{\theta})), \quad \mathbf{e}_{ij} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{V}_{ij}(\boldsymbol{\theta})), \quad \mathbf{b}_{ij} \perp \mathbf{e}_{ij} \end{aligned} \tag{3.1}$$

where y_{ijk} is response k from subject j in group i at time point t_{ijk}, f_i are unknown sparsely-structured functions, τ_{ijk} is a warped time point relating to the t_{ijk} via some unknown, smooth, monotonic function h, $\mathbf{b}_{ij}$ is a $q \times 1$ vector of random effects associated with subject j in group i, σ^2 and $\boldsymbol{\theta}$ are variance component parameters, and $\mathbf{e}_{ij} = (e_{ij1}, \ldots, e_{ijn_j})^T$ are vectors of measurement errors. $V_{ij}(\boldsymbol{\theta})$ depends on i and j through its dimension. In this chapter we assume each curve has the same number of observations and has been observed on the same grid of time points, but these assumptions can be relaxed. For notational brevity, we will suppress the reliance of $h(\mathbf{b}_{ij}; t_{ijk})$ on the random effects and time and denote it by h_{ij}.

The form of (3.1) is similar to that of the class of SEMOR models defined in Lawton

et al. (1972) and extended in Brumback and Lindstrom (2004) and Ke and Wang (2001) with the allowance of l group-specific unknown functions. Subjects within a group share a common curve and exhibit subject-specific deviations in the shape through the unknown warping function h_{ij}. The monotonicity constraint on h_{ij} is to ensure identifiability since without it, t_{ijk} can potentially be mapped to more than one τ_{ijk} and h_{ij} would not be invertible.

The unknown shape functions are estimated using basis expansion, that is, $f_i(\tau_{ijk}) = X_i(\tau_{ijk})\boldsymbol{\beta}_i$, where X_i is the basis function design matrix for group i evaluated at τ_{ijk} and $\boldsymbol{\beta}_i$ its associated unknown parameter vector. Expanding f_i into some finite-dimensional basis is in contrast to Ke and Wang (2001) who assumed the functions lie in an infinite-dimensional space and required specification of the particular space.

The warping function are modeled as $h_{ij} = t_{ijk} + h^*_{ij}$, where h^*_{ij} represents the deviation from t_{ijk} for subject j in group i. Expanding h^*_{ij} as a linear combination of B-splines, $Z_{ij}(t_{ijk})\mathbf{b}_{ij}$, where $Z_{ij}(t_{ijk})$ is the design matrix for the B-spline basis and $\mathbf{b}_{ij}$ is the associated parameter vector, provides a flexible model for the subject-specific deviations. This process is often called time synchronization or curve registration (Ramsay and Li, 1998). We assume that the deviations in the group-averaged shapes are small enough such that the monotonicity of h_{ij} is dominated by the t_{ijk} term. That is, the values of $\mathbf{b}_{ij}$ do not need to be constrained to ensure monotonicity.

We set the time scale of one of the subjects in one group as the reference time scale to which all other subjects will be warped. As the reference time scale is not warped, its B-spline coefficient vector, $\mathbf{b}_{ij}$, is fixed and equal to the zero vector. Without loss of generality, we choose $i, j = 1$ as the reference.

Thus the specific form (3.1) we focus on in the current chapter is

$$y_{ijk} = X_i(\tau_{ijk})\boldsymbol{\beta}_i + e_{ijk}, \qquad \tau_{ijk} = t_{ijk} + Z_{ij}(t_{ijk})\mathbf{b}_{ij} \tag{3.2}$$

and our model has fixed effect parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_l^T)^T$ and random effect parameter vector $\mathbf{b} = (\mathbf{b}_{12}^T, \ldots, \mathbf{b}_{1m_1}^T, \ldots, \mathbf{b}_{l1}^T, \ldots, \mathbf{b}_{lm_l}^T)^T$.

3.3. Estimation

3.3.1. Penalized Likelihood with the Adaptive Lasso

Let $\mathbf{y}_{ij} = (y_{ij1}, \ldots, y_{ijn_j})^T$, $\mathbf{y}_i = (\mathbf{y}_{i1}^T, \ldots, \mathbf{y}_{im_i}^T)^T$, $\mathbf{t}_{ij} = (t_{ij1}, \ldots, t_{ijn_j})^T$, $\mathbf{t}_i = (\mathbf{t}_{i1}^T, \ldots, \mathbf{t}_{im_i}^T)^T$, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_l^T)^T$, $\mathbf{b}_i = (\mathbf{b}_{i1}, \ldots, \mathbf{b}_{im_i})^T$, $V_i = \text{diag}(V_{i1}, \ldots, V_{im_i})$, $\tilde{D} = \text{diag}(D, \ldots, D)$, and $N_i = \sum_{j=1}^{m_i} n$. The marginal log-likelihood can be expressed as:

$$\ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \sum_{i=1}^{l} \log \left((2\pi\sigma^2)^{-(m_i q + N_i)/2} |V_i(\boldsymbol{\theta})|^{-1/2} |\tilde{D}|^{-1/2} \int \exp\left[-g(\mathbf{b}_i)/(2\sigma^2) \right] d\mathbf{b}_i \right) \tag{3.3}$$

where

$$g(\mathbf{b}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}) = (\mathbf{y}_i - X_i(\mathbf{b}_i; \mathbf{t}_i)\boldsymbol{\beta}_i)^T V_i(\boldsymbol{\theta})^{-1} (\mathbf{y}_i - X_i(\mathbf{b}_i; \mathbf{t}_i)\boldsymbol{\beta}_i) + \mathbf{b}_i^T \tilde{D}^{-1} \mathbf{b}_i \tag{3.4}$$

To induce sparsity in the estimate of the mean functions, we employ penalized likelihood with the Adaptive Lasso penalty. Letting p_i be the dimension of $\boldsymbol{\beta}_i$, the penalized likelihood criterion is then

$$\ell_P(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \ell(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) - \lambda \sum_{i=1}^{l} \sum_{h=1}^{p_i} \hat{w}_{ih} |\beta_{ih}| \tag{3.5}$$

The tuning parameter, λ, controls the amount of shrinkage on the basis function

coefficients for the unknown functions, f_i. As λ increases, more coefficients are shrunk to 0 and removed from the final model. Conversely, as λ decreases to 0, less coefficients are shrunk to 0 and more are retained and estimated in the final model. The use of the adaptive weights, $\hat{w}_{ih}$, proposed by Zou (2006), result in estimates which possess attractive sampling properties that the estimates using the classical Lasso do not necessarily possess. We use the weights $\hat{w}_{ih} = |\tilde{\beta}_{ih}|^{-1}$, where $\tilde{\beta}_{ih}$ denotes the unpenalized maximum likelihood estimate of β_{ih}.

There exist multiple methods to approximating integrals numerically (Pinheiro and Bates, 2000). In this chapter we consider the use of the Laplace Approximation to the integral in (3.3). Let $\hat{\mathbf{b}}_i = \arg\min_{\mathbf{b}_i} g(\mathbf{b}_i)$. As shown by Vonesh (1996), using the Laplace Approximation to approximate (3.3) is equivalent to approximating the log-likelihood for group i as:

$$\begin{aligned} 2\tilde{\ell}_i(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = &-(m_i q + N_i)\log(2\pi\sigma^2) \\ &- \log(|\Lambda_i(\boldsymbol{\theta})|) - (\mathbf{w}_i - X_i(\hat{\mathbf{b}}_i)\boldsymbol{\beta}_i)^T \Lambda_i(\boldsymbol{\theta})^{-1}(\mathbf{w}_i - X_i(\hat{\mathbf{b}}_i)\boldsymbol{\beta}_i)/\sigma^2 \end{aligned}$$

where $\mathbf{w}_i = \mathbf{y}_i + \tilde{Z}_i\hat{\mathbf{b}}_i$, $\tilde{Z}_i = (\partial f_i/\partial \mathbf{b}_i^T)|_{\boldsymbol{\beta}_i, \hat{\mathbf{b}}_i}$, and $\Lambda_i(\boldsymbol{\theta}) = \tilde{Z}_i\tilde{D}(\boldsymbol{\theta})\tilde{Z}_i^T + V_i(\boldsymbol{\theta})$. We suppress the dependence of X_i on $\mathbf{t}_i$ and the dependence of Z_i on $\boldsymbol{\beta}_i$ and $\mathbf{t}_i$ for notational brevity. Thus the approximate penalized log-likelihood is:

$$\tilde{\ell}_P(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \sum_{i=1}^{l} \tilde{\ell}_i(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) - \lambda \sum_{i=1}^{l}\sum_{h=1}^{p_i} \frac{|\beta_{ih}|}{|\tilde{\beta}_{ih}|} \tag{3.6}$$

3.3.2. Back-fitting Algorithm

Maximization of (3.6) can be accomplished via a coordinate descent algorithm, where we partition the full vector of parameters and maximize with respect to each partition

while holding the other parameters fixed. We propose the following iterative backfitting algorithm to estimate model parameters: Let $\hat{\boldsymbol{\alpha}}^{(k)}$ denote the estimate of $\boldsymbol{\alpha}$ obtained at iteration k. At iteration k,

1. Setting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k-1)}$ and $\mathbf{b} = \hat{\mathbf{b}}^{(k-1)}$, maximize (3.6) with respect to $\boldsymbol{\beta}$ to obtain $\hat{\boldsymbol{\beta}}^{(k)}$.

2. Setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(k)}$ and $\mathbf{b} = \hat{\mathbf{b}}^{(k-1)}$, maximize (3.6) with respect to $\boldsymbol{\theta}$ to obtain $\hat{\boldsymbol{\theta}}^{(k)}$.

3. Setting $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(k)}$ and $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)}$, minimize (3.4) to obtain $\hat{\mathbf{b}}^{(k)}$.

Iterate steps 1–3 until convergence.

Performing the maximization in this manner leads to simplification in the computations as each step can be solved using standard software. In particular, noting the form of (3.2), the model is linear in $\boldsymbol{\beta}$. Moreover, letting $\lambda^* = \sigma^2\lambda$, $\mathbf{y}_i^* = V_i^{-1/2}(\hat{\boldsymbol{\theta}}^{(k-1)})\mathbf{y}_i$, and $X_i^* = V_i^{-1/2}(\hat{\boldsymbol{\theta}}^{(k-1)})X_i$, optimizing (3.6) for $\boldsymbol{\beta}_i$ for each group is equivalent to solving the classical penalized least squares problem with the Adaptive Lasso:

$$\hat{\boldsymbol{\beta}}_i^{(k)} = \arg\max_{\boldsymbol{\beta}_i} \left\{ (\mathbf{y}_i^* - X_i^*(\hat{\mathbf{b}}_i^{(k-1)}, \mathbf{t}_i)\boldsymbol{\beta}_i)^T(\mathbf{y}_i^* - X_i^*(\hat{\mathbf{b}}_i^{(k-1)}, \mathbf{t}_i)\boldsymbol{\beta}_i) - \lambda^* \sum_{h=1}^{p_i} \frac{|\beta_{ih}|}{|\tilde{\beta}_{ih}|} \right\} \tag{3.7}$$

and can be solved using standard Lasso software (Zou, 2006; Bunea and Gupta, 2010).

As the penalty does not involve the variance components, optimizing (3.6) for $\boldsymbol{\theta}$ is equivalent to optimizing the unpenalized approximate log-likelihood with respect to $\boldsymbol{\theta}$. Noting that, since the model is linear in $\boldsymbol{\beta}$, then $\partial f_i/\partial \boldsymbol{\beta}_i^T = X_i(\mathbf{b}_i)$, $\boldsymbol{\theta}$ can be

estimated using the linear-mixed effects step of the Lindstrom and Bates algorithm:

$$\hat{\boldsymbol{\theta}}^{(k)} = \arg\min_{\boldsymbol{\theta}} \sum_{i=1}^{l} \{\log|\Lambda_i(\boldsymbol{\theta})| + (\mathbf{w}_i - X_i(\hat{\mathbf{b}}_i^{(k-1)})\hat{\boldsymbol{\beta}}_i^{(k)})^T \Lambda_i^{-1}(\boldsymbol{\theta})(\mathbf{w}_i - X_i(\hat{\mathbf{b}}_i^{(k-1)})\hat{\boldsymbol{\beta}}_i^{(k)})\Big\} \tag{3.8}$$

where $\mathbf{w}_i$ is as defined above. Alternatively, one can use the second-order estimating equations proposed by Vonesh (1996). As pointed out in Vonesh (1996), the linear mixed effects step of the Lindstrom and Bates algorithm and his estimating equations approach differ in that the linear mixed effects step ignores the dependence of $\tilde{Z}_i$ on $\boldsymbol{\beta}$. Thus the resulting estimates of $\boldsymbol{\theta}$ using (3.8) are conditional maximum likelihood estimates (Vonesh, 1996). In this chapter, we focus on using (3.8) to update $\boldsymbol{\theta}$.

Similar to the above, the penalty term does not depend on $\mathbf{b}$, either. Thus we can estimate the warping function parameters as we would in the unpenalized setting, that is $\mathbf{b}_{ij}$ is estimated via

$$\hat{\mathbf{b}}_i^{(k)} = \arg\min_{\mathbf{b}_i} g(\mathbf{b}_i, \hat{\boldsymbol{\beta}}_i^{(k)}, \hat{\boldsymbol{\theta}}^{(k)}) \tag{3.9}$$

That is, we calculate the posterior mean of $\mathbf{b}_i$ for each group. This can be accomplished using the pseudo-data step in the Lindstrom and Bates algorithm (Lindstrom and Bates, 1990), with the modification of fixing $\boldsymbol{\beta}$, as described in Vonesh (1996).

We propose the following iterative back-fitting algorithm to estimate model parameters: At iteration k,

1. Setting $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k-1)}$ and $\mathbf{b} = \mathbf{b}^{(k-1)}$, solve (3.7) to obtain $\boldsymbol{\beta}^{(k)}$.

2. Setting $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$ and $\mathbf{b} = \mathbf{b}^{(k-1)}$, solve (3.8) to obtain $\boldsymbol{\theta}^{(k)}$.

3. Setting $\boldsymbol{\beta} = \boldsymbol{\beta}^{(k)}$ and $\boldsymbol{\theta} = \boldsymbol{\theta}^{(k)}$, solve (3.9) to obtain $\mathbf{b}^{(k)}$.

Iterate steps 1–3 until convergence. At convergence, estimate σ^2 via:

$$\hat{\sigma}^2 = \frac{1}{\sum_{i=1}^{l} N_i} \sum_{i=1}^{l} (\mathbf{y}_i - X_i(\hat{\mathbf{b}}_i, \mathbf{t}_i)\hat{\boldsymbol{\beta}}_i)^T V_i(\hat{\boldsymbol{\theta}})^{-1} (\mathbf{y}_i - X_i(\hat{\mathbf{b}}_i, \mathbf{t}_i)\hat{\boldsymbol{\beta}}_i)$$

The proposed algorithm provides a computationally efficient procedure to optimize (3.3). By expanding f_i into a linear combination of basis functions and penalizing the coefficients of the basis functions, the sparsity penalty does not involve any additional parameters, and double integration is not needed. Using our method, convergence is guaranteed and resulting variance estimates will be correct, enabling valid statistical inference.

Convergence of the proposed algorithm relies on good initial estimates. Starting values can be calculated by setting λ to 0 and solving the resulting unpenalized nonlinear mixed effects model. For some curve types, such as those with sharp, narrow spikes, we have found that iterating this process while updating $\mathbf{b}$ two or three times provides good initial estimates. The unpenalized nonlinear mixed effects models require initial values for the warping functions and these can be obtained by guesses through visual inspection.

The proposed algorithm is for fixed λ. Choice of tuning parameter is an important step in the estimation procedure and numerous criteria for selecting λ exist with varying finite-sample and asymptotic properties. We choose λ using the following BIC-like criterion:

$$\Gamma = -2\tilde{\ell}_p + d \log \left(\sum_{i=1}^{l} N_i \right) \tag{3.10}$$

where d is the total number of parameters, both fixed and random, in the final model. Further investigation of the optimal choice of λ is beyond the scope of the current chapter.

As the Laplace Approximation has been shown to be a special case of penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993; Wolfinger, 1993), procedures utilizing the approximation can be subject to limitations of PQL. In particular, PQL can result in biased variance estimates (Breslow and Lin, 1995; Lin and Breslow, 1996) in certain situations, however has good performance when there are a large number of observations per subject, as the approximation is $\mathcal{O}(\min(N_i)^{-1})$ for each group (Vonesh, 1996). We are considering sparsely-structured functional data and a dense enough grid of observations is required to capture the sparse features. Thus we assume each curve is observed on a large number of time points and the use of PQL is justified in our setting and should perform well. If a more accurate approximation of the likelihood is needed, Adaptive Gaussian Quadrature (AGQ) can be utilized. The Laplace Approximation is a special case of AGQ, thus our procedure readily extends to include more accurate approximations (Pinheiro and Bates, 2000).

3.3.3. Sampling Properties

In this section we study the sampling properties of the estimates resulting from our procedure. For this section, we let β_k be the kth element of the full basis function parameter vector $\boldsymbol{\beta}$ and let p be the dimension of $\boldsymbol{\beta}$. As in Zou (2006), we assume without loss of generality that there is a $p_0 < p$ such that $|\beta_k| > 0$ for $k \leq p_0$ and $\beta_k = 0$ for $p_0 < k \leq p$. Thus the true active set of $\boldsymbol{\beta}$, $\mathcal{A}$, is $\{1, 2, \ldots p_0\}$, and denote the estimated active set of $\boldsymbol{\beta}$ as $\hat{\mathcal{A}}$. Let $\boldsymbol{\psi} = (\mathbf{b}^T, \boldsymbol{\theta}^T, \boldsymbol{\beta}^T)^T$, and $\boldsymbol{\psi}_{\mathcal{A}} = (\mathbf{b}^T, \boldsymbol{\theta}^T, \boldsymbol{\beta}_{\mathcal{A}}^T)^T$.

Theorem 2. *Under the regularity conditions described in the Appendix, the estimates obtained by minimizing* (3.6) *satisfy the following:*

1. $\lim_{n\to\infty} P(\hat{\mathcal{A}}_n = \mathcal{A}) = 1$.

2. $\sqrt{n}(\hat{\boldsymbol{\psi}}_{\hat{\mathcal{A}}} - \boldsymbol{\psi}_{\mathcal{A}}) \to \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}^{-1}(\boldsymbol{\psi}))$.

where $\mathbf{I}(\boldsymbol{\psi})$ *is the Fisher information knowing* $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\mathcal{A}}^T, \mathbf{0}^T)^T$.

Theorem 1 shows that the adaptive lasso estimators obtained from (3.6) are variable selection consistent, root-n consistent, and asymptotically normal, and thus possess the oracle property (Donoho and Johnstone, 1994). The asymptotic normality of the estimates enables the construction of asymptotic confidence intervals for the group-averaged curves and warping functions which reflect the combined variability in estimating $\boldsymbol{\beta}$, $\mathbf{b}$, and $\boldsymbol{\theta}$ jointly.

Specifically, pointwise approximate confidence intervals can be obtained by approximating f_i via a Taylor expansion about the estimated BLUPs. Specifically, the marginal distribution of $\mathbf{y}_i$ is approximated as

$$\mathbf{y}_i \dot{\sim} \mathcal{N}\left(X_i(\hat{\mathbf{b}}_i)\boldsymbol{\beta} - Z_i\hat{\mathbf{b}}_i, \sigma^2 V(\boldsymbol{\theta})\right)$$

and so,

$$\operatorname{Var} \mathbf{y}_i \approx X_i H_{\boldsymbol{\beta}} X_i^T + Z_i H_{\mathbf{b}_i} Z_i^T$$

where $H_{\boldsymbol{\beta}}$ is the Hessian of the log-likelihood with respect to $\boldsymbol{\beta}$ and $H_{\mathbf{b}_i}$ is the second derivative of (3.4) with respect to $\mathbf{b}_i$ (Lindstrom and Bates, 1990).

3.4. Application to Data Example

To demonstrate the application of our procedure, we apply it to the Andrographis paniculata chromatographic data set. Five samples were collected from each of five production sites which adhere to GAP and five samples were collected from each of

five market sites. All 50 chromatograms were observed on 3001 time points between 0 and 50 minutes.

The dependence of the exact chemical composition of herbs on production processes prompted the development of guidelines regarding Good Agricultural and Practices (GAP) for medicinal herbs. Multiple State Departments of China and the World Health Organization (WHO) drafted series of documents detailing rules and regulations for the quality control of herbal medicines. (Leung and Cheng, 2008). The documents developed by the WHO can be accessed at `http://apps.who.int/medicinedocs/en/d/Js4928e.html`. Ideally, production sites that follow GAP manufacture standardized herbs which can be used to study their properties and medical applications.

In the current study, we wish to compare the compositions of the samples between GAP-adhering and market sites, and identify compounds which appear in one set of samples, but not the other. As the GAP sites produce what can be considered gold-standard forms of the herbs, identifying discordant compounds will aid in diagnosing issues in production practices in the market sites. Furthermore, if important compounds are missing from the market sites, the therapeutic effects of market-produced herbs will be less than advertised.

Assuming the samples arising from GAP sites share a fingerprint shape and the market sites share a separate fingerprint shape, we fit the SSNMM to the data to compare the group-averaged fingerprints, while taking into account the salient features of chromatographic data. We assume there is site-specific warping of the group-averaged fingerprints. Letting $\mathbf{y}_{ijk}$ be the kth chromatogram from site j in group i, we fit the

following model:

$$\mathbf{y}_{ijk} = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1(\boldsymbol{\tau}_{ijk}) \\ \boldsymbol{\alpha}_2(\boldsymbol{\tau}_{ijk}) \end{pmatrix} + \mathbf{e}_{ijk}, \qquad \tau_{ijk} = t_{ijk} + Z_{ij}(t_{ijk})\mathbf{b}_{ij} \tag{3.11}$$

where $i = 1, 2$, $j = 1, \ldots 5$, $k = 1, \ldots, 5$, $\boldsymbol{\alpha}_q(\boldsymbol{\tau}_{ijk}) = X_q(\tau_{ijk})\boldsymbol{\beta}_q$, $X_q(\cdot)$ is the Battle-Lemarié spline wavelet design matrix, $Z_{ij}(\cdot)$ is the cubic B-splines design matrix. $2^{10} = 1024$ wavelet basis functions were used and periodicity was assumed for convenience. Six uniform knots were used for the warping functions. A compound symmetric correlation structure was assumed for the B-spline random effects. The tuning parameter for the SSNMM was chosen using the BIC-like criterion (3.10). Through (3.11), $\boldsymbol{\alpha}_1(\cdot)$ represents the fingerprint for the GAP-compliant sites and $\boldsymbol{\alpha}_2(\cdot)$ represents the difference between the GAP-compliant and market fingerprints.

Initial estimates were obtained via the process described in the previous section. The algorithm was written in MATLAB and run on an Intel Xeon CPU E7-4860. As their was a shift of spikes between GAP and market chromatograms of about 0.75 minutes, a rightwards horizontal shift of 0.75 minutes was applied to the market chromatograms as a pre-processing step.

The estimate of the GAP fingerprint includes 180 wavelet basis functions and the market fingerprint estimate includes 154 functions. The estimated residual variance was 0.09 and the chosen value of λ was 2.78. Figure 3.2 displays the group-averaged chromatograms for the GAP and market sites and Figure 3.3 displays the subject-specific estimates of one chromatogram from two GAP sites and two market sites plotted on the observed time scale to demonstrate the registration of the curves through the warping function. Figure 3.4 displays the estimated warping functions minus the observed time vector. The curves in the plot correspond to the estimated

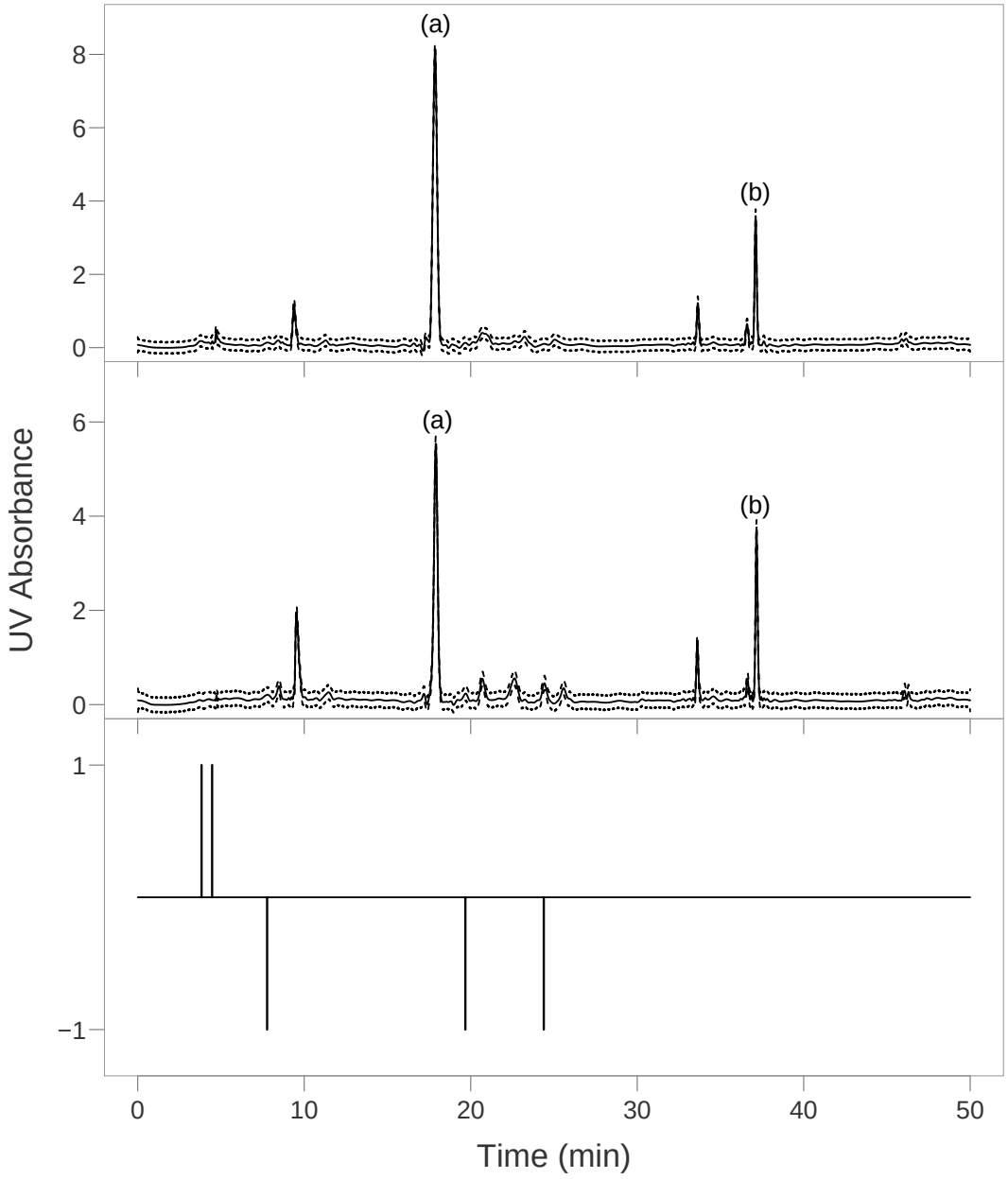

Figure 3.2: Top panel: Group-averaged estimate and confidence interval of the chromatographic fingerprint for the GAP sites. Middle panel: Group-averaged estimate and confidence interval of the chromatographic fingerprint for the market sites. The labels in the top two panels denote the location of the known compounds, andrographolide and dehydroandrographolide, respectively. Bottom panel: Locations of discordant compounds which are present in one fingerprint but not the other.

deviation from the unwarped time scale. The large deviation observed in the GAP sites between 0 and 10 minutes is due to the large misalignment of the small peaks in that time window.

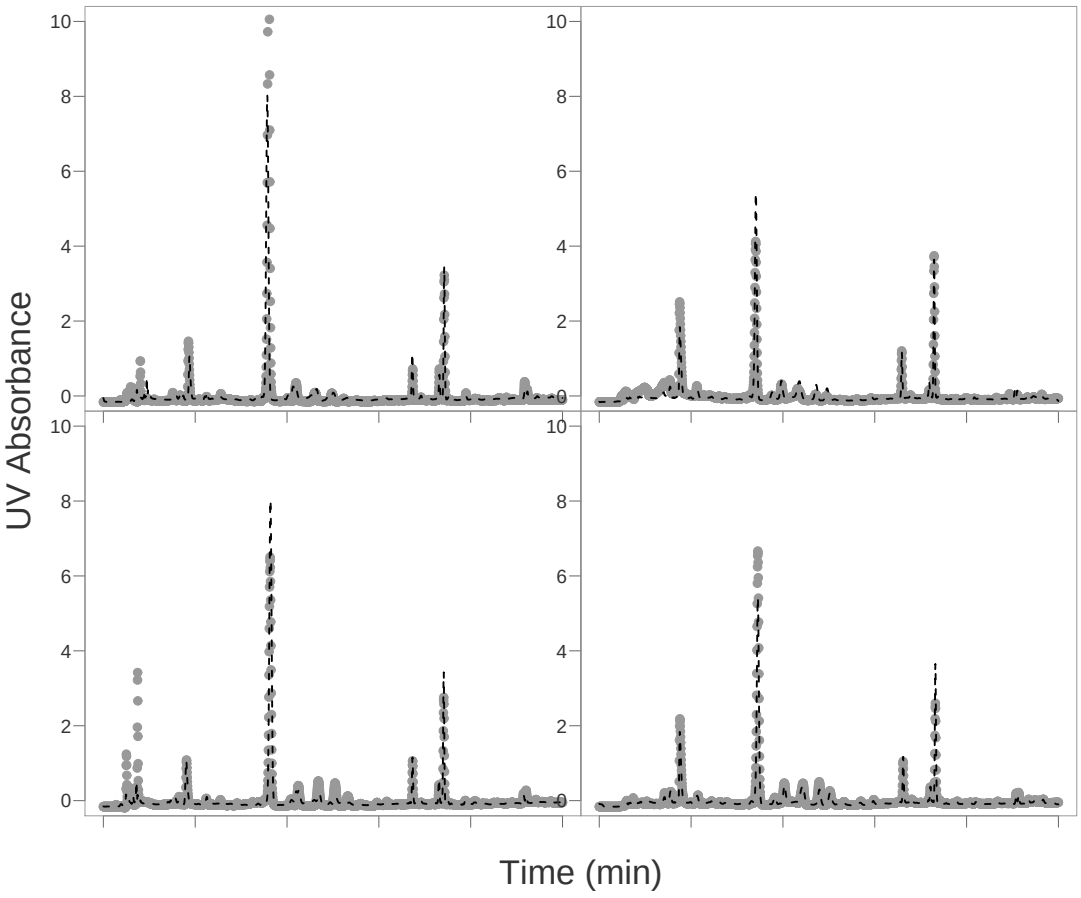

Figure 3.3: Left panel: Data and subject specific estimates for a chromatogram from two GAP sites. Right panel: Data and subject specific estimates for a chromatogram from two market sites. Subject-specific curves are plotted on the observed scale using their inverse warping functions.

Figure 3.2 also displays the difference in compositions between the two groups where the difference was calculated as follows: First the confidence bounds for the group-averaged fingerprints were calculated using the result of Theorem 1. Spikes whose confidence bounds did not cover 0 were retained, otherwise they were shrunk to 0. The retained spikes were compared across groups where, if spikes contained any

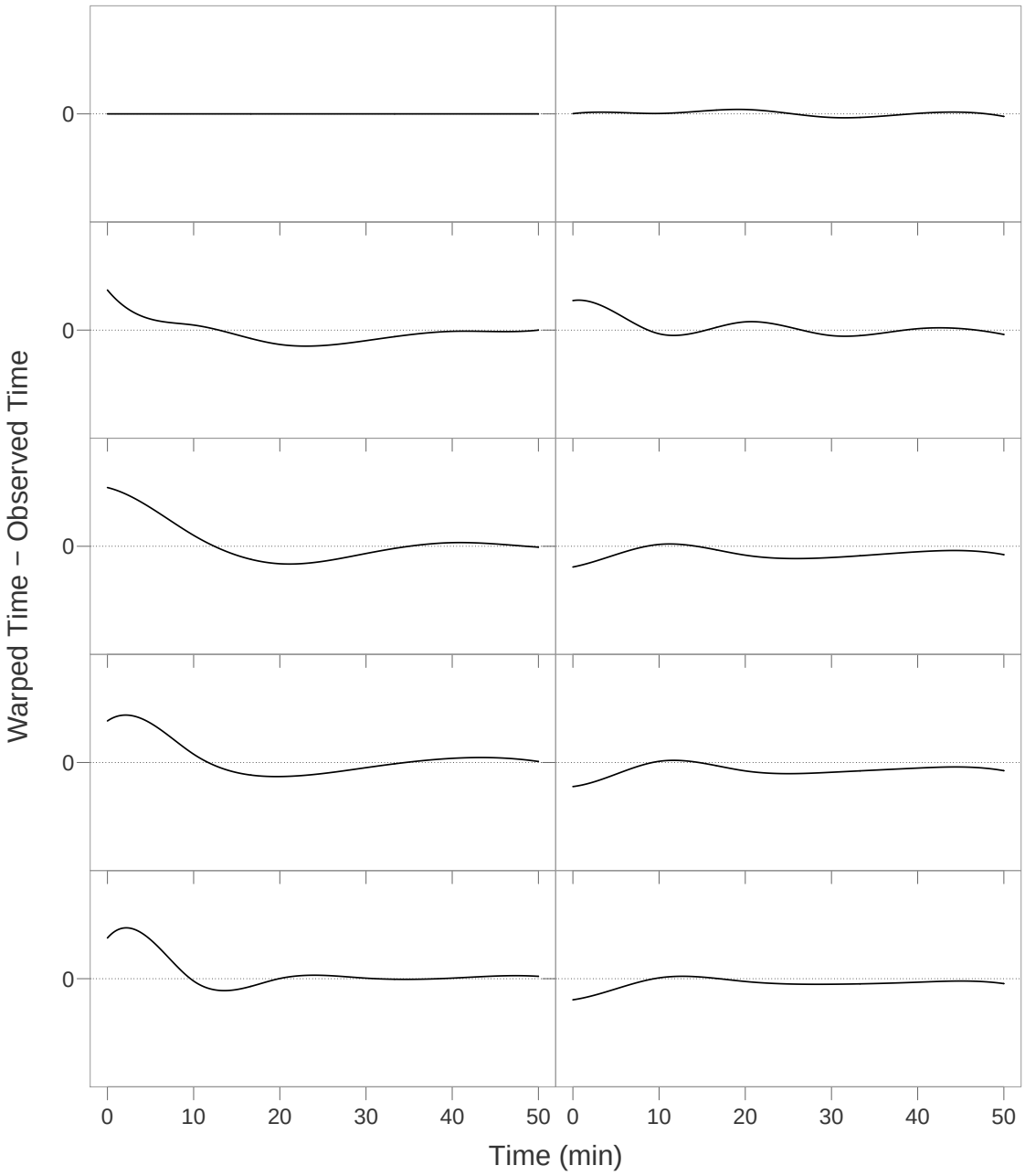

Figure 3.4: Left panel: Estimated deviation from the unwarped time scale for the GAP sites. Right panel: Estimated deviation from the unwarped time scale for the market sites.

overlap in their support, they were deemed to correspond to the same compound, otherwise they correspond to different compounds. The plot in Figure 3.2 displays indicator functions where, if the compound is present only in the GAP fingerprint, a point at $y = 1$ is plotted at the time of the spike. Likewise, the point $y = -1$ is plotted at times corresponding to spikes present only in the market fingerprint. We see that there are two compounds present in the GAP fingerprint which are absent from the market fingerprint, and three compounds that are present only in the market fingerprint. The two missing compounds in the market herbs may result in lessened effectiveness of the herbs in their medicinal applications. The extra compounds are unnecessary, however their presence can be dangerous in that the compounds may interact with other medications a subject is taking with the herb or induce an allergic reaction. Identification of these compounds is thus vital.

Finally, the locations of two compounds known to be in Andrographic paniculata, andrographolide and dehydroandrographolide, are denoted in Figure 3.2. The two compounds are highly abundant in the herb and have been found to possess bioactive properties (Chao and Lin, 2010). The compounds are present in both fingerprints, suggesting that the market-produced herbs contain these important compounds.

3.5. Simulation

Sparsely-structured functional data can manifest as a variety of shape functions. One particular function with a sparse structure is a curve characterized by a number of sharp spikes. To generate spiky curves, we used the Laplace distribution function, also known as the double-exponential function. The true spiky function was generated by overlaying a fixed number of spikes, where each spike arose from the model $f(t) = a\exp(-|t-c|/b)/(2b)$ where a controls the amplitude of the spike, c is the timing

of the maximum of the spike, and b is a scale parameter, affecting the width and tails the spike. We simulated two shape functions, f_1 and f_2, where each function is comprised of a series of 9 spikes. The functions were generated according to the following:

$$f_1(t) = \sum_{i=1}^{9} a_{1,i} \exp\left(\frac{-|t - c_{1,i}|}{b_{1,i}}\right)$$
$$f_2(t) = \sum_{i=1}^{9} a_{2,i} \exp\left(\frac{-|t - c_{2,i}|}{b_{2,i}}\right)$$

where $t \in [0, 1]$, $\mathbf{c}_1 = (c_{1,1}, \ldots, c_{9,1})^T$ and $\mathbf{c}_2 = (c_{1,2}, \ldots, c_{9,2})^T$ were chosen such that $(c_{1,1}, \ldots, c_{8,1})^T = (c_{1,2}, \ldots, c_{8,2})^T$ and are equally spaced apart in $[0.1, 0.9]$. The remaining spike in each vector was chosen randomly. The additional spikes were $c_{9,1} = 0.25$ and $c_{9,2} = 0.48$. Figure 3.5 displays the true data curves.

The amplitude parameters, $\mathbf{a}_1 = (a_{1,1}, \ldots, a_{9,1})^T$ and $\mathbf{a}_2 = (a_{1,2}, \ldots, a_{9,2})^T$ were generated from a $\mathcal{N}(20, 25)$ distribution and the amplitudes of the spikes common between the two functions were set to be the same. The scale parameters, $\mathbf{b}_1 = (b_{1,1}, \ldots, b_{9,1})^T$ and $\mathbf{b}_2 = (b_{1,2}, \ldots, b_{9,2})^T$, were selected from a $\mathcal{N}(0.5, 0.0025)$ distribution, where spikes common between the two functions share the same scale parameter. The warping functions were generated using cubic B-splines with 3 uniform knots and $\mathbf{b}_{ij} \sim \mathcal{N}(0, \sigma^2 V(\boldsymbol{\theta}))$. We generated a compound-symmetric correlation structure for $V(\boldsymbol{\theta})$ with the 2×1 parameter vector $\boldsymbol{\theta}$. Normally distributed noise with mean 0 and variance σ^2 was added to each curve. We used a 2×2 factorial design with the following parameter choices: $\boldsymbol{\theta} = (.001, .0001)^T$, $(.002, .0002)^T$ and $\sigma^2 = 1$, 4. For each setting we simulated 5 subjects per group. For each subject, 3 replicate curves were simulated. The curves were evaluated on 1000 equispaced time points in $[0, 1]$. As an example, one observed curve from each group for the setting

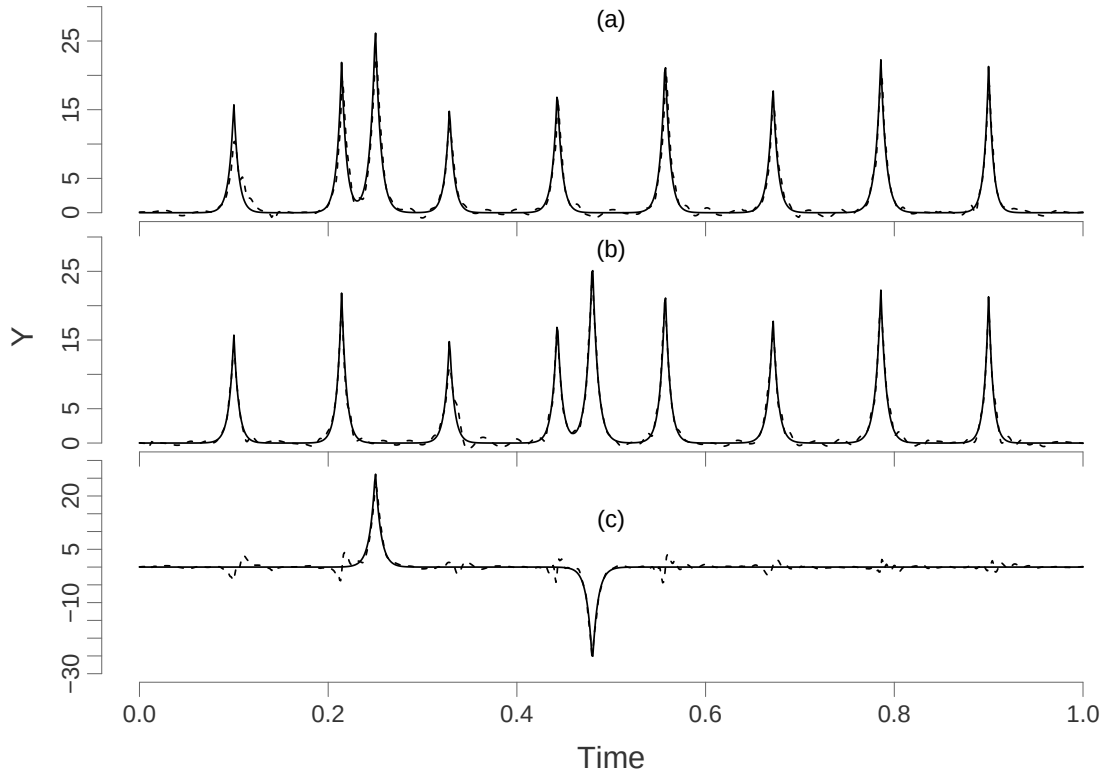

Figure 3.5: Examples of true simulated functions for groups 1 and 2 along with the difference between the curves (solid lines) and estimates from the SSNMM (dashed lines).

$\boldsymbol{\theta} = (.002, .0002)^T$, $\sigma^2 = 4$ is displayed in Figure 3.6.

We fit the SSNMM using Battle-Lemarié spline wavelet basis functions for the shape function and cubic B-splines with 3 uniform knots for the warping functions. The algorithm was written in MATLAB and run on an Intel Xeon CPU E7-4860. For comparison, we also fit a two-step procedure where the first step aligns the curves using dynamic time warping (Giorgino, 2009) and the second fits the wavelet mixed effects model of Morris and Carroll (2006). The dynamic time warping was performed using the 'dtw' package in R (Giorgino, 2009) and the wavelet mixed effects model was fit using the WFMM software provided by Morris and Carroll (2006). For both

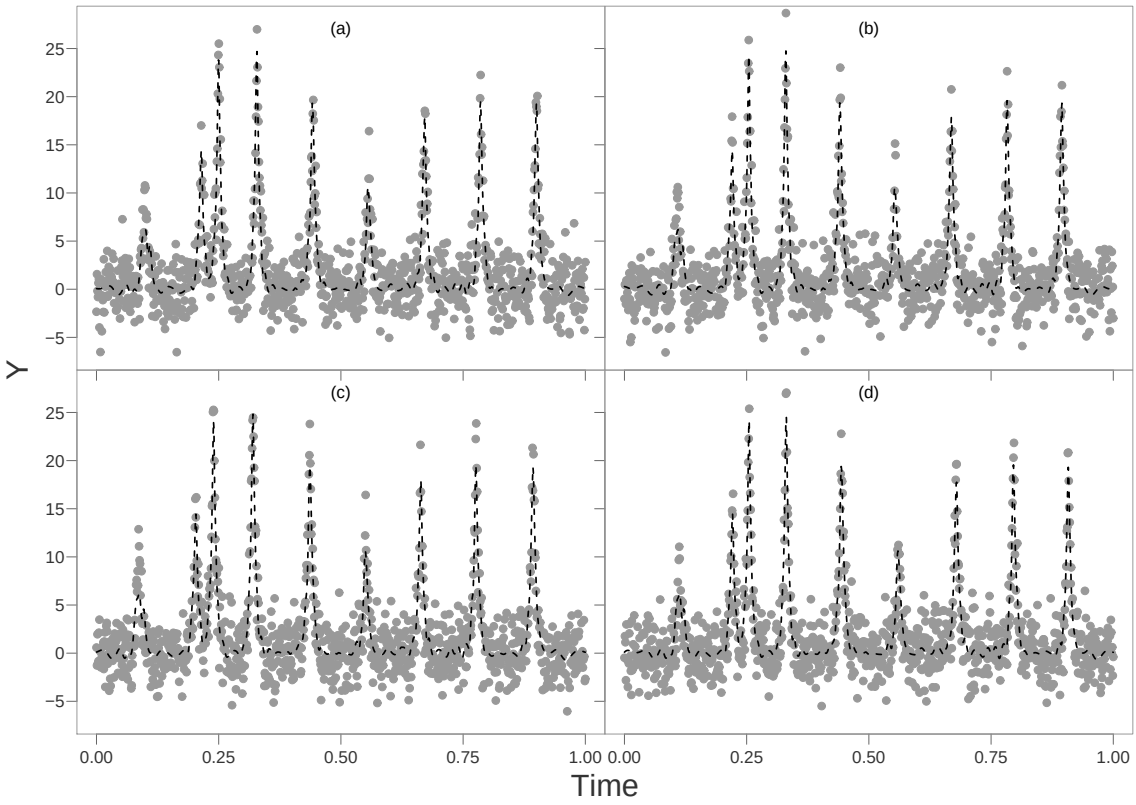

Figure 3.6: One set of observed data points from each group for the setting $\boldsymbol{\theta} = (.002, .0002)^T$, $\sigma^2 = 4$. The estimated subject-specific curve (dashed line) from the SSNMM is also displayed for each setting.

our proposed procedure and the two-step method, $2^9 = 512$ wavelet basis functions were used and periodicity was assumed for convenience. The tuning parameter for the SSNMM was chosen using the BIC-like criterion (3.10). For each value of λ, the proposed procedure typically converged in less than 10 iterations. Details on the computation of the wavelet mixed effects model can be found on the WFMM software website.

To assess the fits, we consider the difference between the two group-averaged curves as seen in Figure 3.5. Figure 3.7 displays boxplots of the mean squared error between the true and estimated group difference. The SSNMM performs better than the

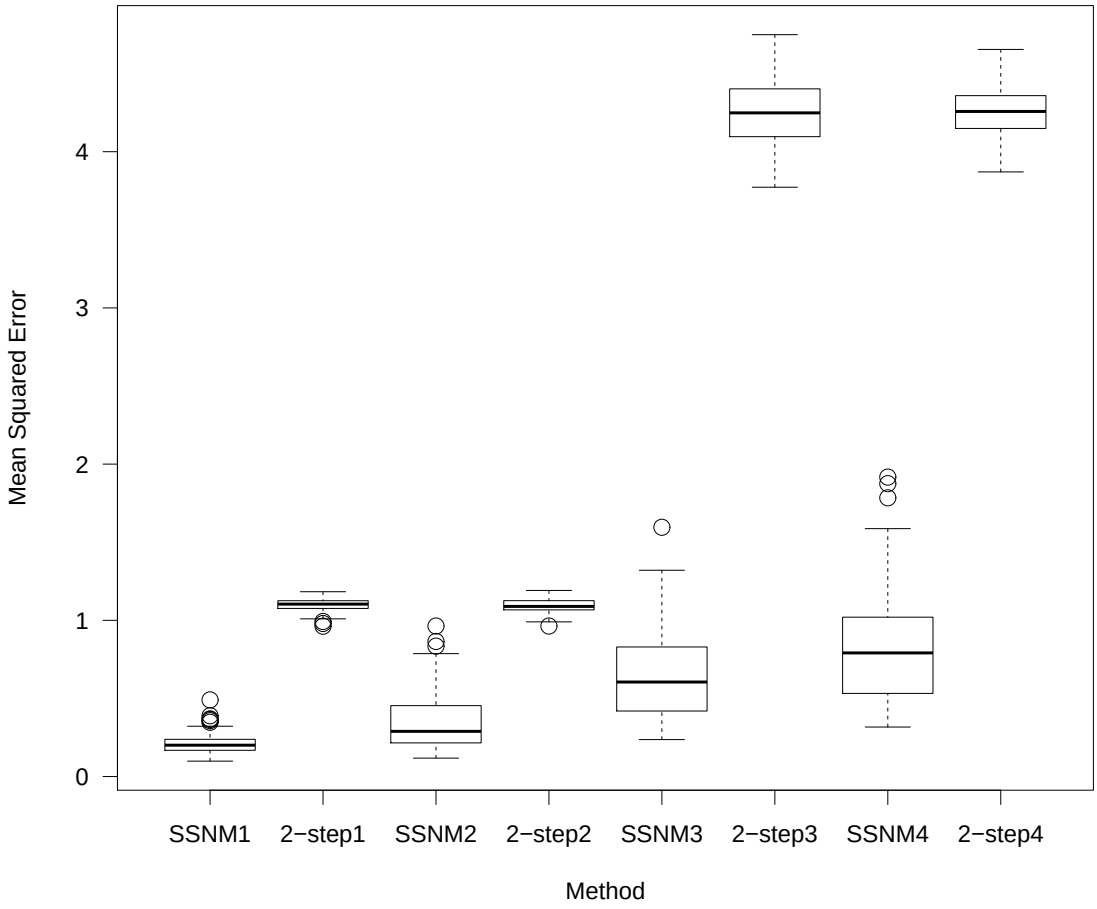

Figure 3.7: Boxplots of the mean square error of the estimated group-averaged difference for the proposed procedure and the two-step procedure using dynamic time warping and WFMM. The settings are as follows: 1: $\sigma^2 = 1$, $\boldsymbol{\theta} = (.001, .0001)^T$; 2: $\sigma^2 = 1$, $\boldsymbol{\theta} = (.002, .0002)^T$; 3: $\sigma^2 = 4$, $\boldsymbol{\theta} = (.001, .0001)^T$; 4: $\sigma^2 = 4$, $\boldsymbol{\theta} = (.002, .0002)^T$.

two-step method in all four settings. In aligning the curves, dynamic time warping will not only align the spikes but will also attempt to align the noise. In doing so, artificial peaks and plateaus are created in the aligned curves. The mixed effects model step regards these artificial features as true signal and so under-thresholds the estimates. This finding is further supported by the fact that when analyzing the same curves under the setting of no warping, the wavelet mixed effects model of Morris and Carroll (2006) thresholds correctly and results in estimates much closer to the

truth. Our results also suggest that the level of warping does not affect the fit of either the SSNMM or the two-step method.

3.6. Discussion

In this chapter we proposed a class of Sparse Semiparametric Nonlinear Mixed Effects Models (SSNMM) for the alignment, estimation, and comparison of sparsely-structured functional data. SSNMMs are general models that allow for the comparison of multiple group-averaged profiles while taking into account subject-specific deviations in the mean shapes through curve registration. We proposed a unified framework using penalized likelihood with the Adaptive Lasso penalty, enabling simultaneous model selection and estimation and statistical inference of mean curves and subject-level warping functions. Furthermore, by optimizing a unified criterion, convergence is guaranteed. The procedure was assessed via simulation and demonstrated its application to a chromatographic data set by identifying compounds not shared between two estimated fingerprints despite the presence of site-specific warpings in the group-level shapes.

Estimation involved the use of the Laplace Approximation to evaluate the marginal likelihood. Under the setting of large numbers of data points per curve, the approximate penalized likelihood leads to estimates that possess the oracle property. In settings where the asymptotic rate of the number of data points needs to be relaxed, Adaptive Gaussian Quadrature (AGQ), of which our method is a special case, can instead be utilized. As the approximation using AGQ is $\mathcal{O}(N_i^{-\lfloor r/3+1 \rfloor})$ where $\lfloor x \rfloor$ denotes the floor function and r is the number of quadrature points, the rate of N_i can be increasingly relaxed as r increases. Extending our procedure to include AGQ would greatly increase the computational complexity, however.

We focused on SSNMMs of the form (3.2), however SSNMMs fall under the more general class of models of the form

$$y_{ijk} = \alpha(\mathbf{a}, \mathbf{b}_{ij}; t_{ijk}) + \delta(\mathbf{a}, \mathbf{b}_{ij}; t_{ijk}) f_i(\gamma(\mathbf{a}, \mathbf{b}_{ij}; t_{ijk})) + e_{ijk}$$

where α, δ, and γ are known functions and $\mathbf{a}$ and $\mathbf{b}$ are fixed and random parameter vectors, respectively. Ke and Wang (2001) focused on such models due to their prevalence in practice and, since f_i enters into the model linearly, our procedures readily extend to handle this class of models.

CHAPTER 4

Piecewise Monotonic B-spline Model for estimating local extrema

4.1. Introduction

In numerous clinical research problems, it is of interest to identify the timings of peaks and troughs in a data curve. Consider Figure 4.1, which displays the volume of the neurological tissue, gray matter, in different regions of the brain measured from magnetic resonance images (MRI) of 107 healthy individuals aged one month to 25 years. Gray matter volume appears to grow rapidly in early childhood and then, at some point in time, begins to decrease into adulthood. Determining the unknown point at which gray matter volume begins to decrease and comparing the peak timings across regions of the brain and across subjects or groups has applications in prediction, classification, and diagnostics, thus an estimate of the location of the volume peak is important and useful.

There has been much research into the exploration of extrema in data curves. A number of proposed methods focus on testing the presence and determining the number of peaks in a curves (e.g. Silverman, 1981; Heckman, 1992; Chaudhuri and Marron, 1999; Fisher and Marron, 2001), often referred to as mode- or bump-hunting. However, as in the volumetric MRI data, the number of extrema is some times known a priori and the objective is estimating and drawing inference on their locations in time. Methods that address this objective primarily rely on nonparametric methods, such as kernel estimators (Muller, 1989; Ziegler, 2002), smoothing splines and a Monte Carlo estimate of the extrema's posterior distribution (Silverman, 1985), and

nonparametric least squares estimators (Shoung and Zhang, 2001).

In particular, Muller (1989) proposed an adaptive kernel estimator of a unique peak by determining optimal bandwidths for the estimation of the size and location of the peak. The estimator targets the first derivative and an estimate of its zero crossing is obtained. Using the estimated first derivative to calculate point and interval estimates of the peak timing is not unlike statistical calibration, or inverse regression, in which the regression model for the dependent variable (for example, the first derivative) is inverted and used to make inference on an independent variable (for example, the time at which the first derivative equals zero). However, as his procedure relies on kernel estimators, it is not straightforward to implement in practice. Moreover, as with numerous proposed peak estimation methods, the curve is assumed to possess a single mode, and so the extension to multiple peaks and troughs is unknown.

As peaks and troughs can be identified through the first derivative of a curve, estimating the first derivative can lead to procedures which allow for multiple extrema. For example, viewing smoothing splines from the Bayesian context, it is straightforward to calculate the posterior distribution of derivatives of estimated curves (Silverman, 1985). However, smoothing splines do not take into account the fact that a curve possesses local extrema which are of direct interest. Models that explicitly assume the presence of features which are targets of inference result in more correct inference on the features than models that ignore their presence and then estimate the features in a subsequent step.

Taking the presence of peaks and troughs into account in a modeling procedure can be accomplished by considering the nature of curves that possess local extrema. For any continuous curve, a peak is a point preceded by a monotonically increasing part of the curve and followed by a monotonically decreasing part. A similar definition follows

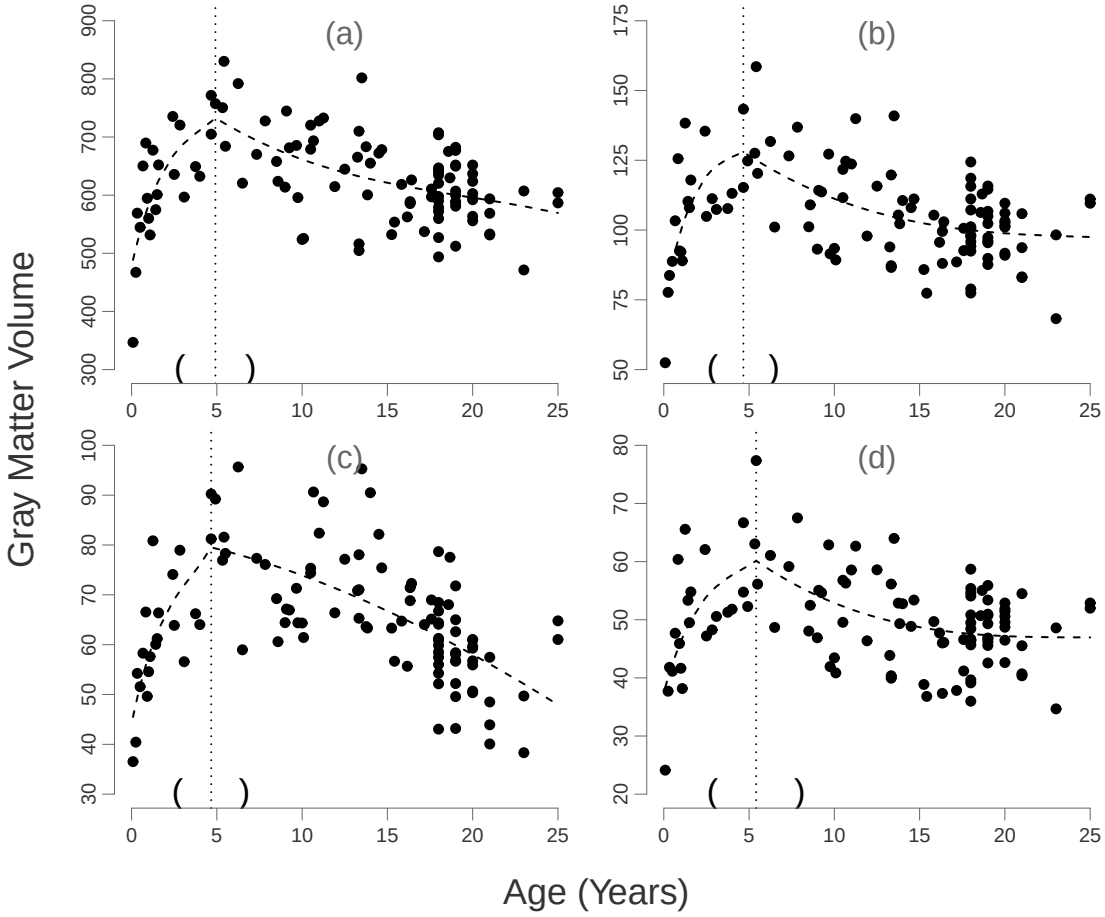

Figure 4.1: Gray matter measured on 107 healthy individuals in the following regions of the brain: (a) entire brain; (b) entire prefrontal region; (b) dorsal prefrontal region; (d) orbital prefrontal region. Estimated curves (dashed line), timing of peak (vertical dotted line), and confidence interval for peak timing (open parentheses) are displayed for each region.

for a trough. Thus a curve with multiple local extrema is a piecewise monotonic curve where each piece meets continuously at the extrema. Noting this fact, we can utilize methods developed for monotone functions to model such curves. Modeling monotonic functions has received a great deal of attention in the statistical literature and a large number of methodologies have been proposed, such as monotonic B-splines (Kelly and Rice, 1990; He and Shi, 1998), integrated regression splines and

differential equations (Ramsay, 1988, 1998), kernel estimators and the pool adjacent violator algorithm (Mammen, 1991), and isotonic regression methods (Wu et al., 2001). Extending these methods to piecewise monotonic functions provides a simple and elegant model-based approach to extrema estimation.

In this chapter we propose a piecewise monotonic regression model for the estimation of curves and the timings of their peaks and troughs. The curves are partitioned into their monotonic intervals which connect extrema and the portion of the curve in each interval is modeled using monotonic B-splines. A set of simple constraints on the first derivatives induce monotonicity of the estimates within each interval and another set constrains the monotonic pieces to meet continuously at each extrema. A least squares criterion is optimized, enabling joint estimation of the curve shape and the timings of its extrema. The model-based approach and unified framework enable the study of sampling properties and statistical inference to be drawn on the timings of the extrema. We assess the procedure by comparing its performance to an alternative method using smoothing splines via simulation and apply it to the volumetric MRI data.

The chapter is organized as follows: In Section 4.2 we present the piecewise monotonic B-spline model. In Section 4.3 we discuss estimation and inference. The description of the alternative method and its comparison to the proposed procedure through simulation is presented in Section 4.4. The proposed procedure is applied to the MRI data set in Section 4.5. We conclude with discussion in Section 4.6.

4.2. Piecewise Monotonic B-Spline Model

4.2.1. Notation and assumptions

Let y_i be the ith independent observation at time t_i, where we assume without loss of generality that $t_i \in [0,1]$, $i = 1, 2, \ldots, n$. We assume the observations arise from some continuous function, η, associated with an m-vector of unknown extrema, denoted $\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in (0,1)$ is an alternating sequence of local maxima and minima. That is, if θ_j is a local maximum, then θ_{j-1} and θ_{j+1} are local minima. Without loss of generality, we assume θ_1 is a maximum. As η is assumed to be continuous, it is monotonic in between extrema. Letting $\theta_0 = 0$ and $\theta_{m+1} = 1$, the observations have the following form:

$$y_i = \eta(t_i, \boldsymbol{\theta}) + e_i \tag{4.1}$$

where $\eta(t_i, \theta)$ is such that

$$\begin{aligned} \eta^{(1)}(t_i) \geq 0, \quad \theta_{j-1} < t_i \leq \theta_j, \quad j \text{ odd} \\ \eta^{(1)}(t_i) \leq 0, \quad \theta_{j-1} < t_i \leq \theta_j, \quad j \text{ even} \end{aligned}$$

for $j = 1, \ldots, m+1$, where $\eta^{(1)} = \partial\eta/\partial t$ and $e_i \sim N(0, \sigma^2)$.

4.2.2. Model specification

We model the observations by partitioning (4.1) into its monotonic pieces and estimate each piece using a monotonic B-spline basis expansion. The Piecewise Monotonic B-Spline Model (PMBM) is of the form:

$$\begin{aligned} y_i &= f(t_i, \boldsymbol{\theta}, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{m+1}) + e_i \\ &= \sum_{j=1}^{m+1} I(\theta_{j-1} < t_i \leq \theta_j) X_j(t_i) \boldsymbol{\beta}_j + e_i \end{aligned} \tag{4.2}$$

where

$$\begin{aligned} \beta_{j1} \leq \beta_{j2} \leq \cdots \leq \beta_{jp_j}, & \qquad j \text{ odd} \\ \beta_{j1} \geq \beta_{j2} \geq \cdots \geq \beta_{jp_j}, & \qquad j \text{ even} \\ \textstyle\sum_{l=K_j+3}^{p_j} x_{jl}(\theta_j)\beta_{jl} = \sum_{l=1}^{r-1} x_{(j+1)l}(\theta_j)\beta_{(j+1)l} & \qquad j = 1, \ldots, m-1 \end{aligned}$$

where $I(\cdot)$ is the indicator function, $X_j(t_i) = (x_{j1}(t_i), \ldots, x_{jp_j}(t_i))$ is the $1 \times p_j$ B-spline design matrix for the interval $(\theta_{j-1}, \theta_j]$ with knot sequence $\mathbf{k}_j$, and $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jp_j})^T$ is its associated unknown parameter vector. For B-splines of order r with knot sequence of length K_j, the number of basis functions is $p_j = r + K_j - 1$. Each knot sequence includes fixed, known knots within its interval and knots placed at the unknown endpoints θ_{j-1} and θ_j. Thus adjacent intervals both possess a knot at their meeting time point. As knots are placed at the extrema, estimating the location of the knots at each endpoint will result in estimates of the location of the extrema.

The B-spline expansion in (4.2) for the interval $(\theta_{j-1}, \theta_j]$ proceeds by mapping the interval to $(0, 1]$ and constructing the B-spline design matrix on the interval $(0, 1]$ evaluated at the remapped time points whose observed values lie in $(\theta_{j-1}, \theta_j]$. To force monotonicity in the interval, B-splines are a convenient choice as constraining the elements of $\boldsymbol{\beta}_j$ to be monotonic ensures a monotonic estimate (Schumaker, 2007, chap. 4.9). This is accomplished by either reparameterizing the $\boldsymbol{\beta}_j$ or through linear inequality constraints. For convenience, we use the latter method.

The monotonic pieces must also meet continuously at each extrema. That is, for $j = 1, \ldots, m-1$,

$$X_j(\theta_j)\boldsymbol{\beta}_j = X_{j+1}(\theta_j)\boldsymbol{\beta}_{j+1}$$

For r-order B-splines, only the first and last $r-1$ basis functions have nonzero value at $t=0$ and $t=1$, respectively. Therefore constraining the jth and $(j+1)$th B-spline estimates to meet continuously at θ_j is accomplished through constraints on $(r-1)+(r-1)=2r-2$ coefficients.

4.3. Estimation and Inference

4.3.1. Estimation procedure

Let $\boldsymbol{\beta}=(\boldsymbol{\beta}_1^T,\ldots,\boldsymbol{\beta}_{m+1}^T)^T$. The unknown parameters $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ are estimated via least squares. The least squares criterion is defined as

$$\ell(\boldsymbol{\theta},\boldsymbol{\beta})=\sum_{i=1}^{n}\{y_i-f(t_i,\boldsymbol{\theta},\boldsymbol{\beta})\}^2 \tag{4.3}$$

and the estimator of $(\boldsymbol{\theta}^T,\boldsymbol{\beta}^T)^T$ is the minimizer of (4.3).

As $\boldsymbol{\theta}$ enters the model nonlinearly, estimating $(\boldsymbol{\theta}^T,\boldsymbol{\beta}^T)^T$ is a constrained nonlinear least squares problem. Estimation is simplified by profiling $\boldsymbol{\beta}$ out from (4.3). Noting the form of (4.2), for fixed $\boldsymbol{\theta}$, maximization of (4.3) is simply a linear least squares problem with linear constraints on the coefficients. That is, for a given value of $\boldsymbol{\theta}$, solving (4.3) for $\boldsymbol{\beta}$ is accomplished via:

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\theta})=\arg\min_{\boldsymbol{\beta}}\sum_{i=1}^{n}\left(y_i-\sum_{j=1}^{m+1}I(\theta_{j-1}<t_i\le\theta_j)X_j(t_i)\boldsymbol{\beta}_j\right)^2 \tag{4.4}$$

subject to

$$\begin{aligned} A_j\boldsymbol{\beta}_j&\le\mathbf{0}, \quad j \text{ odd}\\ A_j\boldsymbol{\beta}_j&\ge\mathbf{0}, \quad j \text{ even}\\ D\boldsymbol{\beta}&=\mathbf{0}\end{aligned}$$

where

$$A_j = \begin{pmatrix} 1 & -1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & \cdots & \cdots & \cdots & 1 & -1 \end{pmatrix}$$

and

$$D = (D_1^T, \ldots, D_m^T)^T$$
$$D_j = \left(\mathbf{0}_{p_1}^T, \ldots, \mathbf{0}_{p_{j-1}}^T, X_j^T(\theta_j), -X_{j+1}^T(\theta_j), \mathbf{0}_{p_{j+2}}^T, \ldots, \mathbf{0}_{p_m}^T\right)$$

where $\mathbf{0}_d$ denotes the vector of zeros of dimension $d \times 1$. The inequality constraint matrices, A_j, are of dimension $(p_j - 1) \times p_j$ and ensures monotonicity of the elements of $\boldsymbol{\beta_j}$ for each interval and the equality constraint matrices, D_j are of dimension $1 \times p$ and ensure the monotonic pieces meet continuously at each extrema.

Nonlinear least squares optimization with linear inequality and equality constraints can be more problematic and computationally intensive than its unconstrained counterpart. However in the applications we consider in the current chapter, neither high order of B-spline nor large numbers of knots are necessary to adequately model the data curves. Thus the dimension of $\boldsymbol{\beta}_j$ is low for each j and the constrained optimization does not result in issues in convergence or computation time.

The above gives the profiled least squares criterion:

$$\ell_P(\boldsymbol{\theta}) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{m+1} I(\theta_{j-1} < t_i \leq \theta_j) X_j(t_i) \hat{\boldsymbol{\beta}}_j(\boldsymbol{\theta}) \right)^2 \tag{4.5}$$

Thus minimizing the least squares criterion (4.3) is equivalent to minimizing (4.5)

with respect to $\boldsymbol{\theta}$ using some standard nonlinear least squares optimization procedure, such as an interior-point algorithm. The estimate of $\boldsymbol{\beta}$ is found by setting $\boldsymbol{\theta}$ to its final estimate, $\hat{\boldsymbol{\theta}}$, and solving (4.4) to obtain $\hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}})$. Finally, σ^2 is estimated by

$$\hat{\sigma}^2 = (n - p - m)^{-1} \sum_{i=1}^{n} \{y_i - f(t_i, \hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\beta}}(\hat{\boldsymbol{\theta}}))\}^2$$

where $p = \sum_{j=1}^{m+1} p_j$.

As the procedure is a nonlinear optimization, good initial values of $\boldsymbol{\theta}$ are essential for convergence of the algorithm. Through our simulations, we have found that for many curves, providing a good initial guess is sufficient.

4.3.2. Statistical Inference

Since we optimize a least squares criterion for $\boldsymbol{\theta}$, its sampling properties can be studied using conventional nonlinear least squares theory. Assuming standard regularity conditions for asymptotic normality of least squares estimates (Seber and Wild, 2003), we have that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \to N\{\mathbf{0}, \sigma^2 C^{-1}\} \tag{4.6}$$

where $C = (df/d\boldsymbol{\theta}_0)^T (df/d\boldsymbol{\theta}_0)$. With this result, we can construct approximate confidence intervals for the extrema. Specifically, using $\hat{C} = (df/d\hat{\boldsymbol{\theta}})^T (df/d\hat{\boldsymbol{\theta}})$ to estimate C, the approximate $100(1-\alpha)\%$ confidence interval for θ_j is

$$\left(\hat{\theta}_j - t^{\alpha/2}_{n-p-m} \hat{\sigma} \hat{C}^{-1/2}_{jj},\ \hat{\theta}_j + t^{\alpha/2}_{n-p-m} \hat{\sigma} \hat{C}^{-1/2}_{jj}\right) \tag{4.7}$$

where $t^{\alpha/2}_{n-p-m}$ is the $(\alpha/2)$th quantile of the t-distribution with $n - p - m$ degrees of freedom and $\hat{C}^{-1/2}_{jj}$ denotes the (j, j)th element of $\hat{C}^{-1/2}$. In practice, interval bounds may extend outside the possible values of the extrema timing (for example less than

0 or greater than 1). Bounds are truncated to the endpoints if they extend outside the possible range.

As the curve shape and the location of the peaks and troughs are estimated within a unified framework, the variance estimates take into account their joint estimation. Thus correct estimates of the variance are used and valid statistical inference is made.

4.4. Simulation

In this section we present an alternative method for estimating local extrema in a curve and compare it to the proposed procedure.

4.4.1. Alternative Procedure

Smoothing splines are a popular tool for estimating functions which cannot be adequately described using parametric models and offer a convenient alternative for estimating curve extrema. Wahba (1978) showed that smoothing splines can be viewed from a Bayesian stochastic model and Wecker and Ansley (1983) demonstrated that this model can be formulated in a state-space form. In particular, the state-space form for the quintic (order 6) smoothing spline is

$$\begin{aligned} y_i &= (1\ 0\ 0)\mathbf{F}(t_i) + e_i \\ \mathbf{F}(t_i) &= \mathbf{T}(t_i, t_{i-1})\mathbf{F}(t_{i-1}) + \mathbf{U}(t_i, t_{i-1}) \\ \mathbf{U}(t_i, t_{i-1}) &\sim N\{\mathbf{0}, \mathbf{W}(t_i, t_{i-1})\} \end{aligned}$$

where $\mathbf{F}(t) = (F^{(1)}(t), \ldots, F^{(3)}(t))^T$ with jth element

$$F^{(j)}(t) = \sum_{k=0}^{j-1} f^k(a)\frac{(t-a)^k}{k!} + \lambda^{1/2}\int_a^t \frac{(t-h)^{j-1}}{(j-1)!}dW(h)$$

where $W(u)$ is a Wiener process, λ is the smoothing parameter, and $\mathbf{T}(t_i, t_{i-1})$ is the 3×3 upper triangular matrix and $\mathbf{U}(t_i, t_{i-1})$ the 3×3 covariance matrix whose (j,k)th element is

$$\mathbf{T}(t_i, t_{i-1})_{jk} = \frac{(t_i - t_{i-1})^{j-k}}{(j-k)!}, \qquad \mathbf{W}_{jk}(t_i, t_{i-1}) = \lambda \frac{(t_i - t_{i-1})^{7-j-k}}{(7-j-k)(3-j)!(3-k)!}$$

Note that $\mathbf{F}(t)$ contains $f(t)$ and its first and second derivative.

Given a diffuse prior on the coefficients, the smoothing spline estimate, $\hat{f}(t)$, is the posterior mean of the stochastic model, $\mathrm{E}(F^{(1)}(t) \mid \mathbf{y})$, where $\mathbf{y} = (y_1, \ldots, y_n)^T$ and the variance of the smoothing spline estimate, $\mathrm{Var}(\hat{f}(t))$ is the posterior variance, $\mathrm{Var}(F^{(1)}(t) \mid \mathbf{y})$, (Wahba, 1978, 1983; Silverman, 1985).

The state-space form enables the direct estimation of the smoothing spline estimate of the first derivative and the variance of the estimate. The definition of the estimates follow similarly from the Bayesian framework, that is, $\hat{f}'(t) = \mathrm{E}(F^{(2)}(t) \mid \mathbf{y})$ and $\mathrm{Var}(\hat{f}'(t)) = \mathrm{Var}(F^{(2)}(t) \mid \mathbf{y})$. Estimates can be efficiently calculated in $O(n)$ operations using the well-known Kalman filtering and smoothing algorithms (Durbin and Koopman, 2001).

Thus, we propose to estimate of the timing of each extrema with the zero-crossing of $\hat{f}'(t)$, that is $\{t \mid \hat{f}'(t) = 0\}$. Similarly, the upper and lower confidence bounds of the timing estimate are estimated as $\{t \mid \hat{f}'_u(t) = 0\}$ and $\{t \mid \hat{f}'_l(t) = 0\}$, respectively, where $\hat{f}'_u(t) = \hat{f}'(t) + z_{1-\alpha}\sqrt{\mathrm{Var}(\hat{f}'(t))}$ is the upper confidence bound for $\hat{f}'(t)$ and $\hat{f}'_l(t) = \hat{f}'(t) - z_{1-\alpha}\sqrt{\mathrm{Var}(\hat{f}'(t))}$ is the lower confidence bound. Thus, a confidence interval for each extrema takes the form:

$$\left(\hat{f}'^{-1}_l(0), \hat{f}'^{-1}_u(0)\right) \tag{4.8}$$

Using the zero-crossings of the posterior estimates, as in (4.8) of the first derivative is, in essence, a statistical calibration step and a similar idea as the procedure of Muller (1989). However the smoothing spline method is more straightforward and easier to implement.

4.4.2. Simulation of Piecewise Sinusoidal Curve

We compare our modeling procedure to the alternative method via simulation where we estimate a function with one local maximum and one local minimum. We generated the following piecewise sinusoidal function:

$$y_i = \begin{cases} \sin(6t_i) & 0 \leq t_i \leq \frac{\pi}{12} \\ \cos\left\{t_i - \frac{\pi}{12}\right\} & \frac{\pi}{12} < t_i \leq \frac{13\pi}{12} \\ \cos\left\{3(t_i - \frac{13\pi}{12})\right\} & \frac{13\pi}{12} < t_i \leq \frac{7\pi}{6} \end{cases}$$

The maximum occurs at $t = \pi/12$ and the minimum occurs at $t = 13\pi/12$. We sampled the function on 300 time points equispaced in $[0, 7\pi/6]$ where Normally distributed noise with mean 0 and variance 0.1^2 was added to the curve. This process was simulated 500 times. Figure 4.2 displays the true function along with a noisy example.

For each simulation, we analyzed the data using the PMBM with quintic splines as well as the alternative smoothing spline model (SSM) with quintic splines for comparison. As the SSM requires estimation of the first derivative, the order of the spline should be large enough such that the first derivative has a certain degree of smoothness. It is important to note that the PMBM does not require estimation of the first derivative. For the PMBM, the number of knots in the three monotonic sections were chosen to be $\{3, 7, 3\}$. Both procedure were written in MATLAB and

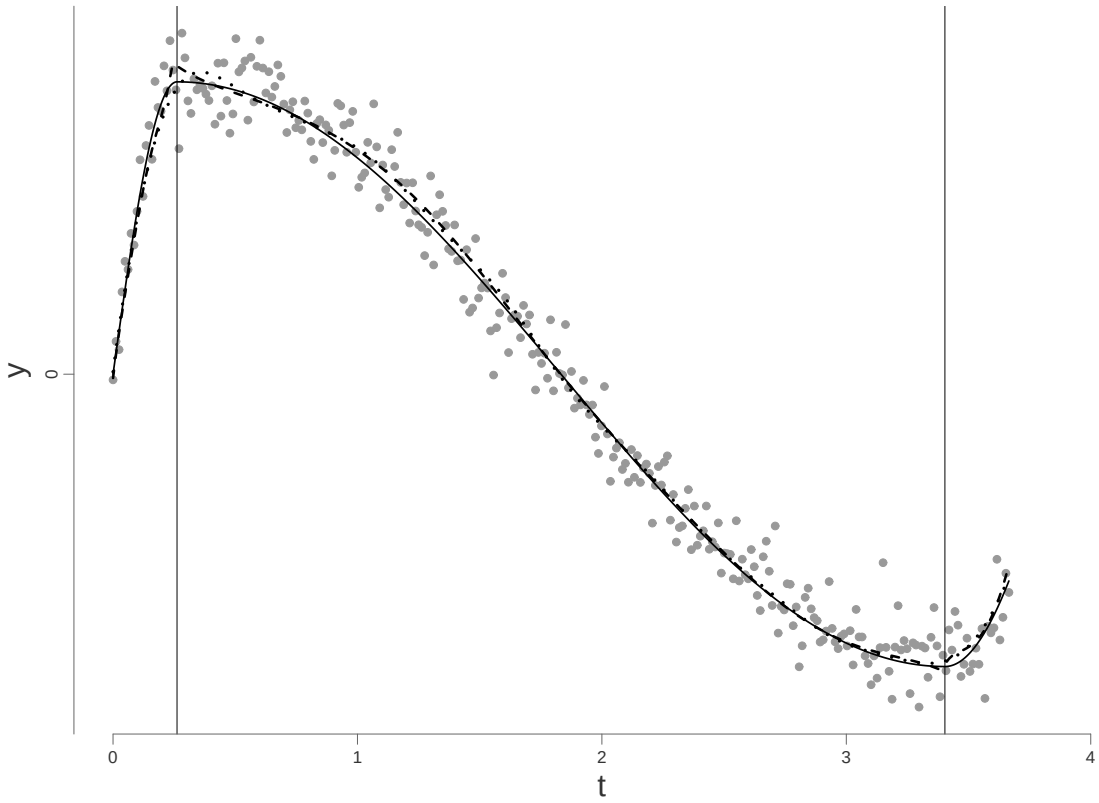

Figure 4.2: True piecewise sinusoidal function (solid line) with observed data points, estimate using PMBM (dashed line), and estimated using SSM (dotted line). Vertical lines denote the true locations of the extrema.

run on an Intel i5-2400. On average, the PMBM took 1.3 seconds to complete and the SSM took 3 seconds.

To assess the fits, we considered the bias in the estimated timings of the extrema and the width of the confidence intervals of the extrema locations for the 500 runs. Figure 4.3 displays the corresponding boxplots for the two methods. The PMBM results in smaller biases than SSM for the maximum location, which tended to overestimate it. This is due to the asymmetry in the rate of the change of the curve before and after the maximum. Smoothing splines oversmooth the curve near the peak, resulting in an estimated peak location to the right of the true location, as seen in Figure 4.2.

By modeling each piece with its own basis expansion, the PMBM can accommodate such sharp slope differences between pieces. We see a similar phenomenon with the minimum, but the SSM underestimated the location of the true minimum. The rate of change of the curve is more symmetric before and after the minimum, thus the absolute bias of the minimum location is slightly smaller than that of the maximum, but still larger than the bias using the PMBM.

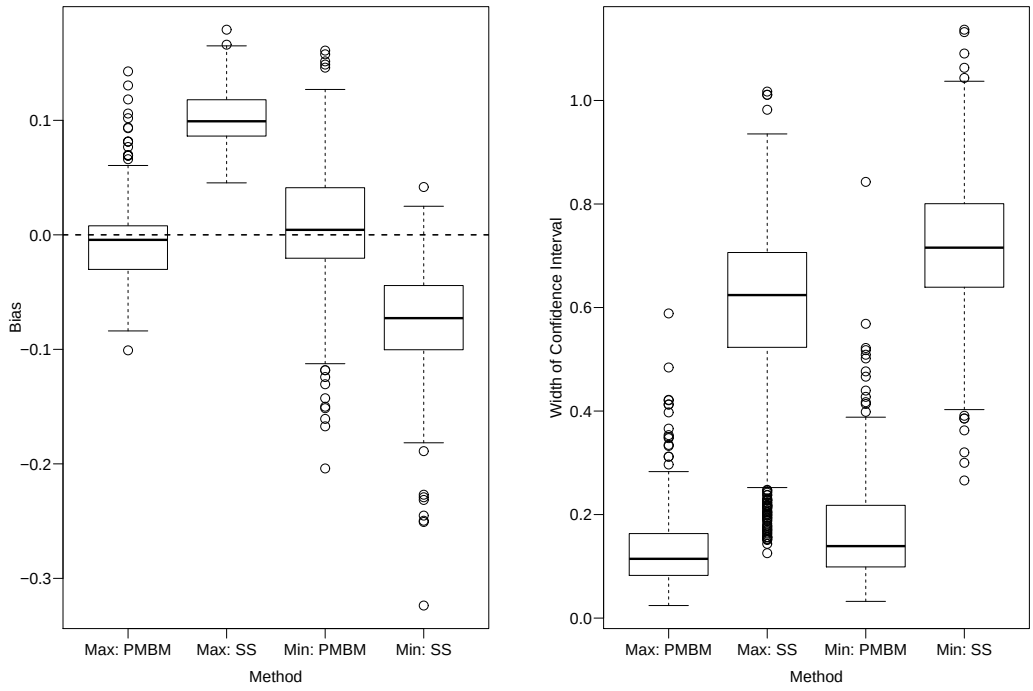

Figure 4.3: Left panel: Boxplots of the bias of the extrema location estimates from the 500 runs of the proposed procedure and the alternative smoothing spline procedure. Right panel: Boxplots of the confidence intervals widths of the extrema locations from 500 runs of the two prcedures.

The widths of the confidence intervals for the SSM are wider for both extrema than those of the PMBM. As variance estimates of smoothing splines are large near boundaries and the extrema are close to $t = 0$ and $t = 7\pi/6$, the variance of the first derivative is inflated, whereas the PMBM avoids this problem. From these findings,

we determine that the PMBM performs well in estimating the timings of extrema despite the shape of the curve or the proximity of extrema to endpoints. The SSM provides a simple and elegant alternative to the PMBM for peak estimation, however its use is not recommended for curves with extrema near boundaries or with widely asymmetric extrema.

4.5. Application to MRI Data Set

In this section, we analyze the volumetric MRI data set. MRIs were obtained on 107 healthy individuals aged one month to 25 years and volumetric measurements were calculated in several regions of the brain. Details on the recruitment, MRI acquisition and processing, and explicit definitions of the different regions of the brain are described in the unpublished report Matsui et al. (2013). We consider total gray matter volume in the entire brain, the entire prefrontal cortex, and the dorsal and orbital subregions of the prefrontal cortex. The frontal regions of the brain play important roles in functions such as emotion responsiveness and social and personality development (Stuss and Benson, 1986; Stuss and Alexander, 2000) and the prefrontal cortex controls, among other functions, inhibition and attention (Gur et al., 2000; Fuster, 2008). Understanding the morphology of gray matter in these regions has multiple important applications in brain development research. For example, understanding volume growth trajectories and determining the typical age of peak volume can be used in monitoring and predicting a child's cognitive and behavioral development. In addition, by collecting volumetric MRI data in subjects with neurological disorders, such as schizophrenia, we can study the utility of volume growth timing comparisons in regards to classifying and diagnosing patients.

Figure 4.1 displays the observed data for the entire brain, entire prefrontal, and dorsal

and orbital subregions of the prefrontal region. Previous research has shown that gray matter volume increases rapidly in early childhood and then begins to decrease some time after adolescence (Durston et al., 2001; Matsuzawa et al., 2001). This fact is evident in the plots, however the time at which the maximal volume occurs appears to happen early in childhood. As the peak occurs early in time and thus near the lower boundary on the time axis, using the SSM would not be appropriate, as described in the previous section. We applied the proposed procedure to the data set to estimate the gray matter volume peaks in each region. Cubic B-splines with two knots were used for both pre- and post-peak intervals.

The fitted estimates along with the peak estimates and confidence intervals from the proposed procedure are shown in Figure 4.1. The estimated peaks (and CI), in years, for the entire brain, entire prefrontal, prefrontal dorsal, and prefrontal orbital regions are, respectively, 4.92 (2.83, 7.01), 4.67 (2.85, 6.48), 4.67 (2.70, 6.63), and 5.42 (2.86, 7.98). These findings suggest that gray matter volume begins to decrease much earlier in life than what is currently believed and so our current hypotheses regarding gray matter growth need to be investigated.

4.6. Conclusion

In this chapter we proposed a piecewise monotonic B-spline model for the analysis of a multi-modal curve and the estimation of its local extrema. Our procedure provides a simple and elegant methodology to simultaneously estimate a curve's shape and its extrema and enables the construction of confidence intervals of the timings of the extrema with variance estimates that reflect the joint estimation of the curve shape and their peaks and troughs. The proposed procedure was applied to a volumetric MRI data set to estimate modes of growth curves.

We also introduced an alternative method for extrema estimation using smoothing splines. Through our simulation, we determined that the two procedures provide simple and elegant methods for estimating the location of extrema, however the proposed procedure can be safely applied to more curve shapes than the alternative procedure. It is important to also note that the alternative procedure does not explicitly assume the curve contains extrema of direct inferential interest. Thus, estimates of multi-modal curves using the alternative procedure may contain fewer extrema than the curves. The proposed procedure does not suffer from this limitation.

CHAPTER 5

Discussion

The estimation procedures presented in this dissertation were motivated by three specific data applications, however they were developed for a general setting in which the applications can be viewed as special cases. At the heart of the problems considered in this work is the registration of functional data curves and the estimation of their salient structures. This dissertation addresses these objectives by providing unified approaches to the analysis of structured functional data.

The estimation procedures in Chapter 2 and Chapter 3 can be readily extended to address other sparsely-structured functional data problems in neurological, genetic, and pharmacological research. In particular, numerous neurological and psychological studies implement medical imaging procedures, such as Magnetic Resonance Imaging or Functional MRIs. The images collected from these procedures can be viewed as three-dimensional functional data. A common clinical objective is to identify the small regions of the brain which respond to particular stimuli, where the response manifests as sharpely-edged regions in the image. This is difficult to achieve in practice as the edges of the regions are sharp, there may be only a few regions of activity in the images, and the exact topology of the brain varies slightly subject to subject. Thus locating and comparing edges in medical images across subjects or groups requires a modeling procedure which registers and estimates three-dimensional images with sparse structures. Extensions of the SSNM and SSNMM can provide a flexible unified framework for this problem.

A main limitation of the procedures developed in Chapter 2 and Chapter 3 is the com-

putational burden, or execution time, of the estimation algorithm. For moderately-sized functional data, that is, curves with a similar number of observations as the Chinese rhubarb data set in Chapter 2, the burden is not an issue. However for data sets with upwards of tens of thousands of data points per curve, which is not extraordinary for imaging or genetic data, modifications to the estimation algorithm are necessary for practical application of the procedures to these data settings. A possible solution to the large execution time is to utilize the attractive computational properties of wavelets. Wavelets are widely popular due to the $O(n)$ execution time of the Discrete Wavelet Transform (DWT) (Daubechies, 1992), however due to the nonlinear nature of the SSNM and SSNMM, the DWT cannot be directly applied. Investigating whether the DWT can be utilized in some clever way may be a productive endeavor.

For the two medicinal herb applications considered, the height of the chromatogram spikes was not of interest and the SSNM and SSNMM reflected this agnosticism towards spike height by assuming the height of a compound's spike was consistent across chromatograms. For these applications, this is an adequate assumption, however the assumption can be construed as a limitation for other related data problems. Accommodating amplitude variation as well as phase variation is another potential extension to these models.

The procedure developed in Chapter 4 can be extended to enable the comparison of extrema timings across groups to assess whether curves exhibit a certain peak or trough at a different time based on group membership. Another immediate extension of the procedure is to accommodate longitudinal data. In fact, the procedures described in Chapter 2 and Chapter 3 can be used as a guide to develop a piecewise monotonic B-Spline model to register curves and compare local extrema across

groups in the presence of subject-specific deviations in the group-level functional shapes. Collecting longitudinal volumetric MRI data on subjects and applying this model to compare peak volume growths across gender or schizophrenia status would provide an important step in neurological development research.

The PMBM shares some limitations with smoothing spline models. In particular, if a peak or trough is flat because the curve approaches the extremum very slowly, there may be bias in the estimated location. However these types of extrema are difficult to estimate using any methodology. The PMBM also cannot test the presence of an extremum. For many data applications, the estimation of the location of an assumed extremum is a primary objective, however it is not uncommon to wish to determine whether a peak or trough exists at a certain point in a curve. Extending the PMBM to allow for testing of an extremum is conceptually natural as it entails determining whether a particular θ_j should be retained or dropped from the model. It is worthwhile to investigate how model selection can be implemented within the PMBM framework.

APPENDIX

Proof of Theorem 2

We assume the regularity conditions of Fan and Li (2001) and Vonesh (1996) with the following modification: $\lambda_n \to \infty$, $n^{-1/2}\lambda_n \to 0$, and $\max\{n^{-1/2}, \min(N_i)^{-1}\} = n^{-1/2}$. We follow steps similar to the proofs of Theorem 1 in Fan and Li (2001) and Zhang and Lu (2007). Let $s_n(\boldsymbol{\psi}) = \partial\ell/\partial\boldsymbol{\psi}$ and $\nabla s_n(\boldsymbol{\psi}) = \partial s_n/\partial\boldsymbol{\psi}^T$ where $\boldsymbol{\psi} = \{\mathbf{b}^T, \boldsymbol{\theta}^T, \boldsymbol{\beta}^T\}^T$.

A.1. Consistency

We want to show that for any given $\epsilon > 0$, there exists a large constant $C > 0$ such that

$$P\left\{\sup_{\|\mathbf{u}\|=C} \tilde{\ell}_P(\boldsymbol{\psi}_0 + n^{-1/2}\mathbf{u}) < \ell_P(\boldsymbol{\psi}_0)\right\} \geq 1 - \epsilon \tag{A.1}$$

This implies that there exists a local maximizer of $\tilde{\ell}_P(\boldsymbol{\psi})$ in the ball $\{\boldsymbol{\psi}_0 + n^{-1/2}\mathbf{u}\}$ with probability at least $1-\epsilon$, and thus the maximizer, $\hat{\boldsymbol{\psi}}$ satisfies $\|\hat{\boldsymbol{\psi}} - \boldsymbol{\psi}_0\| = O_p(n^{-1/2})$.

Applying the Laplace approximation to ℓ_P, as in Vonesh (1996), the contribution of the ith group to $\tilde{\ell}_P(\boldsymbol{\psi})$ can be written as

$$\tilde{\ell}_{Pi}(\boldsymbol{\psi}) = \ell_i(\boldsymbol{\psi}) + O(N_i^{-1}) - \lambda_n \sum_{j=1}^{p_i} \frac{|\beta_{ij}|}{|\tilde{\beta}_{ij}|}$$

where $\ell_i(\boldsymbol{\psi})$ is the ith summand of

$$\ell(\boldsymbol{\psi}) = \sum_{i=1}^{l} \left\{(\mathbf{y}_i - f_i(\boldsymbol{\beta}))^T V_i(\boldsymbol{\theta})^{-1}(\mathbf{y}_i - f_i(\boldsymbol{\beta})) + \log|V_i(\boldsymbol{\theta})|\right\}$$

Thus, $\tilde{\ell}_P(\boldsymbol{\psi})$ can be written as

$$\tilde{\ell}_P(\boldsymbol{\psi}) = \ell(\boldsymbol{\psi}) + O(n \min(N_i)^{-1}) - \lambda_n \sum_{i=1}^{l} \sum_{j=1}^{p_i} \frac{|\beta_{ij}|}{|\tilde{\beta}_{ij}|}$$

Note that we have $n^{-1/2} s_n(\boldsymbol{\psi}_0) = O_p(1)$ and $\nabla s_n(\boldsymbol{\psi}_0) = I(\boldsymbol{\psi}_0) + o_p(1)$. We have

$$\begin{aligned}
D_n(\mathbf{u}) &= \frac{1}{n}\left\{\tilde{\ell}_P(\boldsymbol{\psi}_0 + n^{-1/2}\mathbf{u}) - \ell_P(\boldsymbol{\psi}_0)\right\} \\
&= \frac{1}{n}\left\{\ell_P(\boldsymbol{\psi}_0 + n^{-1/2}\mathbf{u}) + O(n \min(N_i)^{-1}) - \ell_P(\boldsymbol{\psi}_0)\right\} \\
&\leq \frac{1}{n}\left\{\ell(\boldsymbol{\psi}_0 + n^{-1/2}\mathbf{u}) - \ell(\boldsymbol{\psi}_0)\right\} + O(\min(N_i)^{-1}) \\
&\quad - n^{-1}\lambda_n \sum_{i=1}^{l} \sum_{j=1}^{p_{0i}} \left(\frac{|\beta_{ij0} + n^{-1/2}u_{ij}|}{|\tilde{\beta}_{ij}|} - \frac{|\beta_{ij0}|}{|\tilde{\beta}_{ij}|} \right) \\
&\leq \frac{1}{n}\left\{\ell(\boldsymbol{\psi}_0 + n^{-1/2}\mathbf{u}) - \ell(\boldsymbol{\psi}_0)\right\} + O(\min(N_i)^{-1}) + n^{-3/2}\lambda_n \sum_{i=1}^{l} \sum_{j=1}^{p_{0i}} \frac{|u_{ij}|}{|\tilde{\beta}_{ij}|} \\
&= \frac{1}{n} s_n(\boldsymbol{\psi}_0) n^{-1/2}\mathbf{u} - \frac{1}{2n}\mathbf{u}^T \left(n^{-1}\nabla s_n(\boldsymbol{\psi}_0)\right)\mathbf{u} + \frac{1}{n}\mathbf{u}^T o_p(1)\mathbf{u} + O(\min(N_i)^{-1}) \\
&\quad + n^{-3/2}\lambda_n \sum_{i=1}^{l} \sum_{j=1}^{p_{0i}} \frac{|u_{ij}|}{|\tilde{\beta}_{ij}|} \\
&= -\frac{1}{2n}\mathbf{u}^T \left\{I(\boldsymbol{\psi}_0) + o_p(1)\right\}\mathbf{u} + \frac{1}{n} O_p(1) \sum_{i=1}^{l} \sum_{j=1}^{p_{0i}} |u_{ij}| + O(\min(N_i)^{-1}) \\
&\quad + n^{-3/2}\lambda_n \sum_{i=1}^{l} \sum_{j=1}^{p_{0i}} \frac{|u_{ij}|}{|\tilde{\beta}_{ij}|}
\end{aligned} \tag{A.2}$$

Since the unpenalized approximate maximum likelihood estimator $\tilde{\boldsymbol{\beta}}$ satisfies $\|\tilde{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_p(\max\{n^{-1/2}, \min(N_i)^{-1}\})$ (from Vonesh (1996)), we have, for $1 \leq j \leq p_{0i}$,

$i = 1, \ldots, l$,

$$\begin{aligned}\frac{1}{|\tilde{\beta}_{ij}|} &= \frac{1}{|\beta_{ij0}|} - \frac{\text{sgn}(\beta_{ij0})}{\beta_{ij0}^2}(\tilde{\beta}_{ij0} - \beta_{ij0}) + o_p(|\tilde{\beta}_{ij0} - \beta_{ij0}|) \\ &= \frac{1}{|\beta_{ij0}|} + \max\{n^{-1/2}, \min(N_i)^{-1}\}O_p(1)\end{aligned}$$

By assumption, $\max\{n^{-1/2}, \min(N_i)^{-1}\} = n^{-1/2}$ and $n^{-1/2}\lambda_n = O_p(1)$, so we have,

$$\begin{aligned}n^{-3/2}\lambda_n \sum_{i=1}^{l}\sum_{j=1}^{p_{0i}} \frac{|u_{ij}|}{|\tilde{\beta}_{ij}|} &= n^{-3/2}\lambda_n \sum_{i=1}^{l}\sum_{j=1}^{p_{0i}} \left(\frac{|u_{ij}|}{|\beta_{ij0}|} + |u_{ij}| \max\{n^{-1/2}, \min(N_i)^{-1}\}O_p(1)\right) \\ &\leq Cn^{-1}\lambda_n \max\{n^{-1/2}, \min(N_i)^{-1}\}O_p(1) = Cn^{-1}(n^{-1/2}\lambda_n)O_p(1) \\ &= Cn^{-1}O_p(1)\end{aligned}$$

Thus, in (A.2), if C is sufficiently large, the first term is of the order C^2n^{-1}, the second and fourth term are of the order Cn^{-1}, and by assumption, the third term is of the order n^{-1}. Thus (A.2) is dominated by the first term, which is negative, and so (A.1) holds.

A.2. Variable selection consistency

For the next two proofs, we return to denoting the full basis function parameter vector as $\boldsymbol{\beta}$ where without loss of generality, we assume $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$, where $\boldsymbol{\beta}_1$ is the $p_0 \times 1$ subvector of truly nonzero coeffcients and $\boldsymbol{\beta}_2$ is the $(p - p_0) \times 1$ subvector of truly zero coefficients. Let $\boldsymbol{\psi}_1 = (\mathbf{b}^T, \boldsymbol{\theta}^T, \boldsymbol{\beta}_1^T)^T$. We will show that, for any sequence $\boldsymbol{\psi}_1$ satisfying $\|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_{10}\| = O_p(n^{-1/2})$ and for any constant C,

$$\ell_P(\boldsymbol{\psi}_1, 0) = \max_{\|\boldsymbol{\beta}_2\| \leq Cn^{-1/2}} \ell_P(\boldsymbol{\psi}_1, \boldsymbol{\beta}_2)$$

This will show that $\hat{\boldsymbol{\beta}}_{2n} = 0$. To do this, we show that, with probability tending to 1, for any $\boldsymbol{\psi}_1$ satisfying $\|\boldsymbol{\psi}_1 - \boldsymbol{\psi}_{10}\| = O_p(n^{-1/2})$, $\partial \ell_P(\boldsymbol{\psi})/\partial \beta_d$ and β_d have different signs for $\beta_d \in (-Cn^{-1/2}, Cn^{1/2})$ with $i = p_0 + 1, \ldots, p$. For $j = p_0 + 1, \ldots, p$, we have

$$\begin{aligned}\frac{\partial \ell_P(\boldsymbol{\psi})}{\partial \beta_d} &= \frac{\partial \ell}{\partial \beta_d} + O(n \min(N_i)^{-1}) - \lambda_n \frac{\operatorname{sgn}(\beta_d)}{|\tilde{\beta}_d|} \\ &= O_p(n^{1/2}) + O(n \min(N_i)^{-1}) - \\ &\qquad \lambda_n \min\{n^{1/2}, \min(N_i)\} \frac{\operatorname{sgn}(\beta_d)}{|\min\{n^{1/2}, \min(N_i)\}\tilde{\beta}_d|}\end{aligned}$$

Finally, since $\min\{n^{1/2}, \min(N_i)\}\tilde{\beta}_d = O_p(1)$, $\min\{n^{1/2}, \min(N_i)\} = n^{1/2}$, and $n^{1/2} \min(N_i)^{-1} = O_p(1)$, we have that

$$\frac{\partial \ell_P(\boldsymbol{\psi})}{\partial \beta_d} = n^{1/2} \left\{ O_p(1) + O_p(1) - \lambda_n \frac{\operatorname{sgn}(\beta_d)}{|O_p(1)|} \right\}$$

Since $\lambda_n \to \infty$, the sign of $\partial \ell_P(\boldsymbol{\psi})/\partial \beta_d$ is completely determined by the sign of β_d when n is large, and they always have opposite signs.

A.3. Asymptotic Normality

From Theorem 1, we can show that there exists a root-n consistent maximizer $\hat{\boldsymbol{\psi}}_{1n}$ of $\ell_P(\boldsymbol{\psi}_1, 0)$. Let $s_{1n}(\boldsymbol{\psi})$ be the first p_0 elements of $s_n(\boldsymbol{\psi})$, where $s_n(\boldsymbol{\psi}) = \partial \ell / \partial \boldsymbol{\psi}$, and let $\hat{I}_{11}(\boldsymbol{\psi})$ be the first $p_0 \times p_0$ submatrix of $\nabla s_n(\boldsymbol{\psi})$, where $\nabla s_n(\boldsymbol{\psi}) = \partial s_n / \partial \boldsymbol{\psi}^T$. We

have

$$
\begin{aligned}
0 = & \left. \frac{\partial \ell_P(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}_1} \right|_{\boldsymbol{\psi}=\{\hat{\boldsymbol{\psi}}_{1n}^T, 0^T\}^T} \\
= & \left. \frac{\partial \ell(\boldsymbol{\psi})}{\partial \boldsymbol{\psi}_1} \right|_{\boldsymbol{\psi}=\{\hat{\boldsymbol{\psi}}_{1n}^T, 0^T\}^T} + O(n \min(N_i)^{-1}) - \lambda_n \left(\frac{\operatorname{sgn}(\hat{\beta}_1)}{\tilde{\beta}_1}, \ldots, \frac{\operatorname{sgn}(\hat{\beta}_{p_0})}{\tilde{\beta}_{p_0}} \right)^T \\
= & \, s_{1n}(\boldsymbol{\psi}_0) - \hat{I}_{11}(\boldsymbol{\psi}^*)(\hat{\boldsymbol{\psi}}_{1n} - \boldsymbol{\psi}_{10}) + O(n \min(N_i)^{-1}) \\
& - \lambda_n \left(\frac{\operatorname{sgn}(\hat{\beta}_{10})}{\tilde{\beta}_1}, \ldots, \frac{\operatorname{sgn}(\hat{\beta}_{p_0 0})}{\tilde{\beta}_{p_0}} \right)^T
\end{aligned}
$$

where $\boldsymbol{\psi}^*$ is between $\hat{\boldsymbol{\psi}}_n$ and $\boldsymbol{\psi}_0$, and the last equation is implied by $\operatorname{sgn}(\hat{\beta}_{jn}) = \operatorname{sgn}(\beta_{j0})$ when n is large. We have that $n^{-1/2} s_{1n}(\boldsymbol{\psi}_0) \to \mathcal{N}(0, I_1(\boldsymbol{\psi}_{10}))$ in distribution and $n^{-1}\hat{I}_{11}(\boldsymbol{\psi}^*) \to I_1(\boldsymbol{\psi}_{10})$ in probability as $n \to \infty$. If $n^{-1/2}\lambda_n \to 0$ and we have

$$
\sqrt{n}(\hat{\boldsymbol{\psi}}_{1n} - \boldsymbol{\psi}_{10}) = I_1^{-1}(\boldsymbol{\psi}_{10}) \left\{ n^{-1/2} s_{1n}(\boldsymbol{\psi}_0) + O(n^{1/2} \min(N_i)^{-1}) \right\} + o_p(1)
$$

By assumption, $n^{1/2} \min(N_i)^{-1} = O_p(1)$, thus by Slutsky's Theorem, we have

$$
\sqrt{n}(\hat{\boldsymbol{\psi}}_{1n} - \boldsymbol{\psi}_{10}) \to \mathcal{N}(0, I_1^{-1}(\boldsymbol{\psi}_{10}))
$$

in distribution as $n \to \infty$.

BIBLIOGRAPHY

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society: Series B*, 60(4):725–749.

Arribas-Gil, A., Bertin, K., Meza, C., and Rivoirard, V. (2013). Lasso-type estimators for semiparametric nonlinear mixed-effects models estimation. *Statistics and Computing*.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25.

Breslow, N. and Lin, X. (1995). Bias correction in generalised linear mixed models with a single component of dispersion. *Biometrika*, 82(1):81–91.

Brumback, L. C. and Lindstrom, M. J. (2004). Self modeling with flexible, random time transformations. *Biometrics*, 60(2):461–70.

Bunea, F. and Gupta, S. (2010). A study of the asymptotic properties of Lasso for correlated data. Technical Report, Dept. of Statistics, Florida State University.

Chao, W. and Lin, B. (2010). Isolation and identification of bioactive compounds in Andrographis paniculata (Chuanxinlian). *Chinese Medicine*, 5(17):1–15.

Chaudhuri, P. and Marron, J. (1999). SiZer for exploration of structures in curves. *Journal of the American Statistical Association*, 94(447):807–823.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics.

Di Marco, V. B. and Bombi, G. G. (2001). Mathematical functions for the representation of chromatographic peaks. *Journal of chromatography. A*, 931(1-2):1–30.

Donoho, D. and Johnstone, I. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3):425–455.

Duke, J. (2002). *Handbook of Medicinal Herbs*. CRC Press, 2nd edition.

Durbin, J. and Koopman, S. (2001). *Time Series Analysis by State Space Models*. Oxford University Press.

Durston, S., Hulshoff Pol, H. E., Casey, B. J., Giedd, J. N., Buitelaar, J. K., and van Engeland, H. (2001). Anatomical MRI of the developing human brain: what have

we learned? *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(9):1012–20.

Elmi, A., Ratcliffe, S. J., Parry, S., and Guo, W. (2011). A B-Spline Based Semiparametric Nonlinear Mixed Effects Model. *Journal of Computational and Graphical Statistics*, 20(2):492–509.

Fan, J., Fan, Y., and Barut, E. (2013). Adaptive Robust Variable Selection. *arXiv preprint arXiv:1205.4795*.

Fan, J. and Li, R. (2001). Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. *Journal of the American Statistical Association*, 96(456):1348–1360.

Ferraty, F., Hall, P., and Vieu, P. (2010). Most-predictive design points for functional data predictors. *Biometrika*, 97(4):807–824.

Fisher, N. and Marron, J. (2001). Mode testing via the excess mass estimate. *Biometrika*, 88(2):499–517.

Fuster, J. (2008). *The Prefrontal Cortex*. Academic Press, 4 edition.

Giorgino, T. (2009). Computing and visualizing dynamic time warping alignments in R: The dtw package. *Journal of Statistical Software*, 31(7).

Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58(1):121–128.

Gur, R. E., Cowell, P. E., Latshaw, A., Turetsky, B. I., Grossman, R. I., Arnold, S. E., Bilker, W. B., and Gur, R. C. (2000). Reduced dorsal and orbital prefrontal gray matter volumes in schizophrenia. *Archives of general psychiatry*, 57(8):761–8.

He, X. and Shi, P. (1998). Monotone B-Spline Smoothing. *Journal of the American Statistical Association*, 93(442):643–650.

Heckman, N. E. (1992). Bump hunting in regression analysis. *Statistics & Probability Letters*, 14(2):141–152.

James, G. M., Wang, J., and Zhu, J. (2009). Functional linear regression thats interpretable. *The Annals of Statistics*, 37(5A):2083–2108.

Jennrich, R. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, 40(2):633–643.

Kassidas, A., MacGregor, J. F., and Taylor, P. a. (1998). Synchronization of batch trajectories using dynamic time warping. *AIChE Journal*, 44(4):864–875.

Ke, C. and Wang, Y. (2001). Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*, 96(456):1272–1298.

Kelly, C. and Rice, J. (1990). Monotone smoothing with application to dose-response curves and the assessment of synergism. *Biometrics*, 46(4):1071–85.

Lawton, W., Sylvestre, E., and Maggio, M. (1972). Self modeling nonlinear regression. *Technometrics*, 14(3):513–532.

Leung, P. and Cheng, K. (2008). Good Agricultural Practice (GAP) -Does It Ensure a Perfect Supply of Medicinal Herbs for Research and Drug Development? *International Journal of Applied Research in Natural Products*, 1(2):1–8.

Liang, Y.-Z., Xie, P., and Chan, K. (2004). Quality control of herbal medicines. *Journal of chromatography. B*, 812(1-2):53–70.

Lin, X. and Breslow, N. (1996). Bias correction in generalized linear mixed models with multiple components of dispersion. *Journal of the American Statistical Association*, 91(435):1007–1016.

Lin, X. and Zhang, D. (2001). Comment on: Semiparametric nonlinear mixed-effects models and their applications. *Journal of the American Statistical Association*, 96(456):1288–1291.

Lindstrom, M. J. and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics*, 46(3):673–687.

Mammen, E. (1991). Estimating a smooth monotone regression function. *The Annals of Statistics*, 19(2):724–740.

Matsui, M., C, T., Niu, L., Noguchi, K., Yagi, S., Ichida, F., Miyawaki, T., Bilker, W., Wierzbicki, M., and Gur, R. (2013). Age-related volumetric changes of prefrontal gray and white matter from healthy infants to adults.

Matsuzawa, J., Matsui, M., Konishi, T., Noguchi, K., Gur, R. C., Bilker, W., and Miyawaki, T. (2001). Age-related volumetric changes of brain gray and white matter in healthy infants and children. *Cerebral Cortex*, 11(4):335–342.

Meyer, V. R. (2010). *Practical High-Performance Liquid Chromatography*. John Wiley & Sons, Inc., 5th edition.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. a., and Coombes, K. R.

(2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64(2):479–89.

Morris, J. S. and Carroll, R. J. (2006). Wavelet-based functional mixed models. *Journal of the Royal Statistical Society. Series B, Statistical methodology*, 68(2):179–199.

Muller, H. (1989). Adaptive nonparametric peak estimation. *The Annals of Statistics*, 17(3):1053–1069.

Nason, G. (2008). *Wavelet Methods in Statistics with R*. Springer.

Peigen, X., Liyi, H., and Liwei, W. (1984). Ethnopharmacologic study of Chinese rhubarb. *Journal of Ethnopharmacology*, 10(3):275–93.

Pinheiro, J. and Bates, D. (2000). *Mixed effects models in S and S-PLUS*. Springer, New York.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science*, 3(4):425–461.

Ramsay, J. and Silverman, B. W. (2005). *Functional data analysis*. Springer, New York, 2 edition.

Ramsay, J. O. (1998). Estimating smooth monotone functions. *Journal of the Royal Statistical Society: Series B*, 60(2):365–375.

Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(2):351–363.

Randolph, T. W. and Yasui, Y. (2006). Multiscale Processing of Mass Spectrometry Data. *Biometrics*, 62(2):589–597.

Sang, H. and Sun, Y. (2012). Simultaneous sparse model selection and coefficient estimation for heavy-tailed autoregressive processes. *arXiv Pre-print*, pages 1–23. arXiv:1112.2682v2.

Schelldorfer, J. and Bühlmann, P. (2011). GLMMLasso: An Algorithm for High-Dimensional Generalized Linear Mixed Models Using L1-Penalization. *Arxiv preprint arXiv:1109.4003*, pages 1–20.

Schumaker, L. (2007). *Spline Functions: Basic Theory*. Cambridge University Press, 3rd edition.

Seber, G. and Wild, C. (2003). *Nonlinear Regression*. John Wiley & Sons.

Shoung, J. and Zhang, C. (2001). Least squares estimators of the mode of a unimodal regression function. *The Annals of Statistics*, 29(3):648–665.

Silverman, B. (1981). Using kernel density estimates to investigate multimodality. *Journal of the Royal Statistical Society. Series B*, 43(1):97–99.

Silverman, B. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society Series B*, 47(1):1–52.

Snyder, L. and Kirkland, J. (1979). *Introduction to Modern Liquid Chromatography*. John Wiley & Sons, Inc., 2nd edition.

Stuss, D. and Benson, D. (1986). *The Frontal Lobes*. Raven Press.

Stuss, D. T. and Alexander, M. P. (2000). Executive functions and the frontal lobes: a conceptual view. *Psychological research*, 63(3-4):289–98.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B*, 58(1):267–288.

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, 109(3):475–494.

Unser, M. (1997). Ten good reasons for using spline wavelets. *Proceedings of SPIE*, 3169:422–431.

van Nederkassel, A., Daszykowski, M., Eilers, P., and Heyden, Y. V. (2006). A comparison of three algorithms for chromatograms alignment. *Journal of Chromatography A*, 1118(2):199–210.

Vonesh, E. (1996). A note on the use of Laplace's approximation for nonlinear mixed-effects models. *Biometrika*, 83(2):447–452.

Wahba, G. (1978). Improper priors, spline smoothing and the problem of guarding against model errors in regression. *Journal of the Royal Statistical Society. Series B*, 40(3):364–372.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B*, pages 133–150.

Wang, H., Li, R., and Tsai, C.-L. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568.

Wang, Y. and Ke, C. (2009). Smoothing Spline Semiparametric Nonlinear Regression Models. *Journal of Computational and Graphical Statistics*, pages 1–27.

Wecker, W. and Ansley, C. (1983). The signal extraction approach to nonlinear regression and spline smoothing. *Journal of the American Statistical Association*, 78(381):81–89.

Wolfinger, R. (1993). Laplace's Approximation for Nonlinear Mixed Models. *Biometrika*, 80(4):791.

Wu, W., Woodroofe, M., and Mentz, G. (2001). Isotonic regression: Another look at the changepoint problem. *Biometrika*, 88(3):793–804.

Ye, M., Han, J., Chen, H., Zheng, J., and Guo, D. (2007). Analysis of phenolic compounds in rhubarbs using liquid chromatography coupled with electrospray ionization mass spectrometry. *Journal of the American Society for Mass Spectrometry*, 18(1):82–91.

Zhang, H. H. and Lu, W. (2007). Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Ziegler, K. (2002). On nonparametric kernel estimation of the mode of the regression function in the random design model. *Journal of Nonparametric Statistics*, 14(6):749–774.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.