BAYESIAN NONPARAMETRIC DIFFERENTIAL EQUATION MODELS FOR FUNCTIONS

Matthew W. Wheeler

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Biostatistics.

Chapel Hill
2013

Approved by:

Dr. Amy H. Herring

Dr. David B. Dunson

Dr. Eric Bair

Dr. James A. Swenberg

Dr. Hongtu Zhu

Abstract

MATTHEW W. WHEELER: Bayesian Nonparametric Differential Equation Models for Functions
(Under the direction of Dr. Amy H. Herring and Dr. David B. Dunson)

Bayesian nonparametric methods develop priors over a large class of functions that essentially allow any continuous function to be modeled. Though these methods are flexible, they are black box approaches that do not explicitly incorporate additional information on the shape of the curve. In many contexts, though the exact parametric form of the curve is unknown, additional scientific information is available in the form of differential operators. This dissertation develops nonparametric priors over function spaces that are specified by differential operators. Here two novel approaches to nonparametric function estimation are considered. In the first approach the prior is specified by a linear differential equation. The Mechanistic Hierarchical Gaussian process defines a prior over functions consistent with a differential operator. The method is applied to muscle force tracings in a functional ANOVA context, and is shown to adequately describe the between subject variability often seen in such tracings. In the second case a novel spline based approach is considered. Here prior information is specifies the maximum number of extrema (changepoints) for an arbitrary function located on an open set in $\mathbb{R}$. The Local Extrema (LX) spline models the first derivative of the curve and puts a prior over the maximum number of changepoints. This method is applied to animal toxicology studies, human health surveys, and seasonal data; and it is shown to remove artifactual bumps common to other nonparametric methods. It is further shown to superior in terms of estimated squared error loss in simulation

studies.

To my wife Kimberly and all of those sleepless nights you wanted to kill me.

To my parents who taught me right from wrong.

Finally to Hank: “And then came the revolution.”

Acknowledgments

I would first and foremost thank my wife Kimberly. If it were not for your prayers, love, support, and your insistence on staying in North Carolina this dissertation would have never been written. You are the love of my life, and I thank God everyday for you.

I would also like to thank my advisors David and Amy. Your guidance and insights were invaluable. David, quite honestly, if it was not for you being such a hard ass, I would have never improved. My writing might still suck, but it sucks much much less. Thanks for putting up with me for the past three years.

A special thank you goes to John Bailer. If it were not for an hour long conversation between Upham and Gaskill in the Fall of '06 I would have never applied to UNC, and this would have never happened.

I would also like single out Dustin and Leann Long. Dustin for putting up with a papist – I still say the reformation was a dumb idea – and Leann for studying with me all those hours leading up to our qualifiers as well as lending me your husband to talk about obscure topics no one gives a shit about. You both are pretty damn good people – for Methodists.

Also, I would like to acknowledge all those that supported me at NIOSH. Especially Chris for being a good boss that enabled this to happen. Sudha and Brent words cannot thank you enough for getting that dataset, without it I would have thrown in the towel. You are both pretty good cheerleaders too!

Also I would like to take a moment and apologize to all I pissed off/put off with my gruff exterior. For me, this dissertation was more about expelling past demons than professional advancement, and exorcisms are always difficult business. They are even more difficult when you are surrounded by people, who through no fault of their own, remind you of all of the people who told you you were stupid, ignorant, and backwards most of your life. It was never my fault (nor my families) that I came from a non-college educated working class family whom by many of the PC intelligentsia would be considered uneducated bigots in 'polite conversation'. Further, I should not be at fault if I saw no reason to prefer the new culture I have lived in reluctantly for the past 5 years over the one I grew up in. However, it was not your fault that you might have been caught up in the war that has raged in my head the past 5 years, and for that I apologize.

Finally a must acknowledge my three children Megan, Jeni, and Claire. I did this for you. If you don't fight for what you believe in and pick yourself up when you fall who will? This document is a testimony to that fact and you should not forget this. Finally, Claire, though you are much too young to read this, I thank you for that 3 a.m. feeding where I first thought about the LX-spline. Your giggles made finishing this document a little more bearable. -AMF

Table of Contents

List of Tables

List of Figures

Chapter 1

Literature Review

Bayesian data analysis proceeds by positing that, given a sequence of data $Y = (y_1, \ldots, y_n)$, one can learn about the underlying generating mechanism through a series of simplifying assumptions. This is done by assuming data arise from a sampling model $P(Y|\Theta)$ controlled by parameter vector Θ. In a Bayesian analysis one assumes that Θ is a random quantity, and that its uncertainty can be quantified *a priori* by $P(\Theta)$, a probability measure over possible values of Θ. The quantity $P(\Theta)$ is prior knowledge on Θ, and we wish to update this prior belief given new information. Learning is accomplished through the use of Bayes Rule

$$P(\Theta|Y) = \frac{P(Y|\Theta)P(\Theta)}{\int P(Y|\Theta)P(\Theta)d\Theta},$$

which updates the distribution for Θ in the presence of new information Y.

Bayesian analyses often proceed by assuming that the data vector Y comes from a known distribution, and that Θ enters into this distribution with parametric form known *a priori*. For example linear regression assumes $Y = X\Theta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I_n)$. This implies that $Y \sim N(X\Theta, \sigma^2 I_n)$, which forces explicit structure on the mean and all higher level moments of Y. Given typical prior assumptions, that is $\theta \sim N(a, b)$ and $\sigma^{-2} \sim Ga(c, d)$, on puts strong prior structure on the system under

study. Strong assumptions may not be warranted, and may not fully encapsulate the uncertainties in the system of interest. In the above example multiple assumptions may be called into question. First the normality assumption may be unrealistic as it assumes that the data arise from a unimodal distribution having relatively light tails. Also the linearity assumption may also be called into question as it is overly restrictive of the functional form. Such analyses may lead to unrealistic inference.

An alternative to such restrictive assumptions is to develop priors which are more reflective of the uncertainty in the system of interest. Such approaches put priors over a rich class of both probability measures and function spaces that better reflect the uncertainties in the system. For example, the Dirichlet prior (Ferguson 1973; 1974) and other stick breaking priors (Sethuraman 1994; Ishwaran and James 2001) can be used to define priors over the space of probability measures. As these priors are almost surely discrete, they are frequently used in combination with mixing kernels such as the Dirichlet mixing process (DPM) (Lo 1984), which can then be used to develop priors over the space of distribution functions that do not assume a specific parametric form on $P(Y|\Theta)$.

Similarly, priors over functional forms can be developed to circumvent the use of simplifying assumptions such as linearity in the mean response, i.e., $E[Y] = X\Theta$. Here versatile priors, such as the Gaussian process (GP) (Rasmussen and C. 2006), can be used to define a prior over a large set of smooth functions in $\mathbb{R}^p$. Such an approach, when combined with the DPM approach, allows one to define priors having large support over possible generating mechanisms. These approaches may be better in encapsulating the uncertainties in the system under study.

In what follows many aspects of non-parametric Bayesian inference are reviewed. Section (1.1) reviews stick breaking priors such as the Dirichlet process, and section (1.2) reviews non-parametric Bayesian regression methods.

1.1 The Dirichlet process prior and other stick-breaking priors

Much of the recent work in Bayesian non-parametrics has focused on the use of the Dirichlet process (Ferguson 1973; 1974; Sethuraman 1994), and other stick-breaking priors (Ishwaran and James 2001). Given a complete and separable metric space $(\mathbf{\Theta}, \mathcal{B})$, a stick-breaking prior G defines a prior over $\mathcal{P}$, the collection of probability measures on $(\mathbf{\Theta}, \mathcal{B})$. In other words, if one defines $\mathcal{C}$ to be the smallest $\sigma-$field generated by sets of the form $\{P : P(\theta) < k\}$ where $\theta \in \mathcal{B}$, $P \in \mathcal{P}$ and $k \in [0, 1]$, then the stick-breaking prior G, defines a probability measure over $(\mathcal{P}, \mathcal{C})$. Given any finite measurable partition $\{\theta_1, \theta_2, \ldots, \theta_L\}$ of $\mathbf{\Theta}$, G defines a prior probability measure over sets of the form $\{P(\theta_1), P(\theta_2), \ldots, P(\theta_L)\}$. Such priors allow one to learn about an arbitrary probability measure P given Y using Bayesian methods.

Stick-breaking priors are unique in that they admit a specific construction on G. A prior G has a stick-breaking representation if and only if

$$G(\cdot) = \sum_{h=1}^{L} w_h \delta_{\theta_h}(\cdot), \tag{1.1}$$

where $\delta_{\theta_h}(\cdot)$ is a discrete measure concentrated at $\theta_h \in \mathbf{\Theta}$, and $\{w_h\}_{h=1}$ are weights such that $0 \leq w_h \leq 1$ and $\sum_h^L w_h = 1$. Each weight is constructed from a set of random variables $\{V_h\}_{h=1}^L$ defined on $(0, 1)$ where $V_h \sim H$, and H is a known probability measure. The stick breaking construction defines $\{w_h\}_{h=1}^L$ through

$$w_1 = V_1 \tag{1.2}$$

and

$$w_k = V_k \prod_{h=1}^{k-1} (1 - V_h). \tag{1.3}$$

In this construction the first few weights (i.e $w_1, w_2, \ldots$ etc.) receive a large portion of the prior mass, with each subsequent weight receiving a geometrically diminishing probability. This implies only a few atoms $\theta_i \in \Theta$ receive a large prior probability of being selected. Note that the total number of atoms L may be finite, or countably infinite. This construction in (1.1) is completed by noting that $\{\theta_h\}_{h=1}^{L}$ are independently drawn from a base line measure G_0 and are independent from the weights.

In applications the V_k are often taken as independent $Beta(a_k, b_k)$ random variables. Both the Dirichlet (Ferguson 1973) and Pittman-Yor processes (Pitman 1996; Pitman and Yor 1997) can be shown to be stick-breaking processes as in (1.1). By taking $L = \infty$, $a_k = 1$, and letting $b_k = b$ for all k, one arrives at the Dirichlet process (Sethuraman 1994). Also by taking $L = \infty$, setting $a_k = 1 - a$ and $b_k = b + ka$), for $0 \leq a < 1$ and $b > -a$, one constructs the Pittman-Yor $\mathcal{PY}(a, b)$ process (Pitman 1995). The V_k are not necessarily limited to beta random variables, and other possibilities have been explored (e.g., Rodriguez and Dunson (2011)). Rodriguez and Dunson (2011) defined the probit stick-breaking process using standard normal random variables, and constructed weights using $V_k = \Phi(a_k)$ where $a_k \sim N(0, 1)$.

The stick-breaking construction is almost surely discrete, which limits its usefulness for most applications. Instead of being used as a prior for observed data, it is frequently employed as a prior over weights in mixture modeling. That is, given some parametric density g, a prior over possible data generating mechanisms is specified as

$$f_G(y) = \int g(y; \theta) dG(\theta). \tag{1.4}$$

Such a mixture distribution was originally proposed by Lo (1984) with G is defined as a Dirichlet process. This approach defines a rich prior over a variety of distributions and can be used in for continuous density estimation as well as repeated measures data.

Direct estimation of the posterior distribution of such stick-breaking mixture models is unavailable in closed form, and various MCMC methods have been developed to sample from the posterior distribution. These methods generally can be divided into two categories. The first marginalizes over the stick-breaking process relating the process to the Polya urn model. The second approach samples the full conditional distribution. Here the weights and the unique atoms $\{\theta_h^*\}_{h=1}^L$ are sampled conditionally on the other terms of the model.

Generalized Polya-Urn Sampling

This sampling method is related to the Polya urn model that Blackwell and MacQueen (1973) connected to the Dirichlet process. It was later shown by Pitman (1996) that when $V_k \sim Beta(a_k, b_k)$ the stick-breaking process can be characterized in term of a generalized Polya-Urn mechanism, and can be used when the V_k are drawn from a $Beta(a_k, b_k)$ distribution. For clarity the sampling method is first described in relation to the Dirichlet process and then generalized in relation Pittman-Yor model.

In the Polya-Urn model colored balls are drawn from an urn in succession. After each draw, the drawn ball is put back in the urn along with C balls of the same color. Once a ball is drawn there is an increased probability of it being drawn in the future. Blackwell and MacQueen (1973) noted that by marginalizing over the Dirichlet process one arrives at the Polya urn model. As draws from this model can be shown to be exchangeable, any draw in the process can be taken conditionally with respect to the other draws.

Given a Dirichlet process $G \sim \mathcal{DP}(bG_0)$ where b is the weight parameter, and G_0 is the base measure, the atoms of G can be sampled using the following construction. Let $\theta_1, \theta_2, \ldots$ be the successive ordered draws from G Blackwell and MacQueen (1973)

showed that the conditional distribution drawing θ_i given the previous draws is

$$\theta_i|\theta_1, \theta_2, \ldots, \theta_{i-1} \sim \frac{b}{b+i-1}G_0 + \sum_{k=1}^{i-1} \frac{1}{b+i-1}\delta_{\theta_k}(\cdot).$$

Here Individual atoms can be thought of as being drawn from the urn in succession. For each draw there is a uniform probability of the next atom drawn as being any one of the previous draws, and a positive probability proportional to G_0 of the next draw being drawn from the base measure. As there may be ties (i.e., $\theta_i = \theta_j$ for $i \neq j$) one can equivalently define $\theta_1^*, \theta_2^*, \ldots, \theta_L^*$ as the L unique atoms that have been drawn from the urn. Letting defining $m_1, m_2, \ldots, m_L$ be the number of times each atom has been drawn, the probabilities specified above can be re-expressed as

$$\theta_i^*|\theta_1^*, \ldots, \theta_{i-1}^*, \theta_{i+1}^*, \ldots, \theta_L^* \sim \frac{b}{b+i-1}G_0 + \sum_{k=1}^{i-1} \frac{m_k}{b+i-1}\delta_{\theta_k^*}(\cdot).$$

With this representation it one can see that successive draws of the same atom increases the probability the atom is drawn in the future. Observations tend to cluster around distinct atoms often resulting in fewer atoms than observations. This clustering of draws can be seen as a feature of the stick-breaking process where there exists a high probability of the next atom being drawn from one of a small number of atoms.

Such conditional probabilities can be generalized to any $\mathcal{PY}(a, b, G_0)$ process. The probability of the current draw, conditional on the previous draws, is

$$Pr(\theta_i = \theta_j^*|\theta_1^*, \ldots, \theta_L^*) = \begin{cases} \frac{m_j - a}{b+i-1} & j \leq L \\ \\ \frac{b+aL}{b+i-1}G_0 & otherwise \end{cases} \tag{1.5}$$

The Polya-urn scheme can be used to sample from mixture distribution as in (1.4). This connection was first utilized by Escobar (1994) and Escobar and West (1995) to

formulate MCMC sampling methods for Dirichlet mixture processes, under the conjugate assumption that $g(\cdot)$ and G_0 are both Gaussian. It can further be applied to priors on G_0.

I focus on an general algorithm that includes both the conjugate and non-conjugate cases. Given observed data vector $(y_1, y_2, \ldots, y_n)'$, we wish to sample $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_n)'$, which are the n latent quantities distributed G, which is a vector of quantities defining the relation $y_i|\theta_i \sim g(y_i; \theta_i)$. Letting $\{\theta_k^*\}_{k=1}^L$ to be the set of unique draws from the urn, and $\boldsymbol{\theta}_{-}i$ the vector $\boldsymbol{\theta}$ without entry i, the algorithm proceeds for any $\mathcal{PY}(a, b, G_0)$ as follows:

1. For each i, $i = 1, \ldots, n$ draw θ_i from

$$\theta_i|\boldsymbol{\theta}_{-i} \sim \frac{b + aL}{b + i - 1} q_0 G_0 + \sum_{k=1}^{i-1} \frac{m_k - a}{b + i - 1} g(y_i; \theta_k^*) \delta_{\theta_k^*}(\cdot). \tag{1.6}$$

   where $q_0 = \int g(y_i; \theta) dG_0(\theta)$. It is possible that θ_i is the only member in the cluster implying that the set $\{\theta_k^*\}_{k=1}^L$ should be recomputed for each draw.

2. For each k, $k = 1, \ldots, L$ and each y_i allocated to cluster k draw θ_k^* from

$$\theta_k^* \propto G_0(\theta_k^*) \prod_{\{y_i: \theta_i = \theta_k^*\}} g(y_i; \theta_k^*) \tag{1.7}$$

In cases where G_0 is non-conjugate with the kernel $g(\cdot; \theta)$ the integral representation of q_0 may be intractable. Various computational methods to ease (MacEachern and Müller 1998; Neal 2000) have been developed.

The Polya urn sampler has a tendency to mix slowly, and it may take many iterations before any new θ^*'s are generated. MacEachern (1994) proposed an acceleration method to increase the efficiency of the Gibbs sampler (Gelfand 1990; Geman and Geman 1993). Here one introduces membership variables $\{\xi_i\}_{i=1}^n$, such that $\xi_i = k$ if subject

i is assigned to θ_k^*. One proceeds by first sampling the augmented cluster membership variable and then updating $\{\theta_k^*\}_{k=1}^L$ given this membership variable.

Conditional Methods

As mixing for the Polya-urn sampler is often poor and non-conjugate sampling can be difficult, various methods have been developed to sample from the posterior conditional on knowledge of the weights. These methods, which are often termed conditional methods, frequently provide better mixing than methods based upon the Poly-urn scheme. The first of such methods described are the block Gibbs sampling methods of Ishwaran and Zarepour (2000) and Ishwaran and James (2001).

These methods approximate the infinite stick breaking process G through a finite dimensional truncation of the posterior distribution. Given the proper truncation level, these methods define a prior that can be shown to be arbitrarily close to the desired countably infinite stick breaking process. Let L be the number of elements in the truncation. To create an truncation of an infinite stick breaking process one discards the $w_{L+1}, w_{L+2}, \cdots$ weights by setting $w_L = 1 - w_1 - w_2 - \cdots w_{L-1}$. This construction can be shown to have a marginal density $\mu_L(Y)$ that is arbitrarily close to $\mu_\infty(Y)$ for large L. Define $\|\cdot\|$ to be the $\mathcal{L}_1$ distance, then (Ishwaran and James 2001) showed that

$$\|\mu_L(Y) - \mu_\infty(Y)\| \leq 4\left(1 - E\left[\left(\sum_{k=1}^{L-1} p_k\right)^n\right]\right). \tag{1.8}$$

This implies that if the stick breaking weights are constructed such that $E\left[\left(\sum_{k=1}^{L-1} w_k\right)^n\right] \to 1$ as $L \to \infty$ there should be little difference between the finite truncation model and the countably infinite stick breaking process. It can be shown for both the Dirichlet and Pittman-Yor processes that accurate truncations exist. For the $\mathcal{PY}(a,b)$ process

one has

$$\|\mu_L(Y) - \mu_\infty(Y)\| \leq 4(1 - E[1 - (\sum_{k=L}^{\infty} w_k)^n]) \tag{1.9}$$

and for the Dirichlet process this expression simplifies to

$$\|\mu_L(Y) - \mu_\infty(Y)\| \sim 4\ n\ \exp(-(L-1)/b). \tag{1.10}$$

As a consequence on can create a finite truncation that is virtually indistinguishable from the infinite stick-breaking prior when L is moderately large.

As sampling from G_L is computationally simpler than sampling from G, block sampling can accurately approximate the infinite stick breaking process. We describe the algorithm in terms of $V_k \sim Beta(a_k, b_k)$, noting that for general $V_k \sim H$ slight modifications are needed. The algorithm introduces a latent variable ξ_i for each observation. Here $\xi_i = k$ if and only if observation y_i is allocated to cluster k.

1. Sample $\boldsymbol{\theta}^*|Y, \boldsymbol{\xi}$: For each K, such that $1 \leq k \leq L$ sample from the density

$$\boldsymbol{\theta}_k^*|Y, \boldsymbol{\xi} \propto G_0(\theta_k) \prod_{i:\xi_i=k} g(y_i|\theta_k^*)$$

2. Sample $\boldsymbol{\xi}|\boldsymbol{\theta}^*, p$: For each $i = 1, \ldots, n$ sample ξ_i from

$$\xi_i|\theta^*, \boldsymbol{\xi}, Y \sim Multinomial(p_{1i}, \cdots, p_{Li}) \tag{1.11}$$

   where $p_{ki} \propto p_k g(y_i|\theta_{ki})$

3. Sample $V_k|\boldsymbol{\xi}$: As the stick-breaking probabilities are conditionally-conjugate to

the multinomial distribution we have

$$V_k \sim Beta(a_k + M_k, b_k + \sum_{l=k+1}^{L} M_l), \tag{1.12}$$

where $M_k = \sum_{i=1}^{n} 1(\xi_i \geq k)$. Given $V_1, \ldots, V_{L-1}$ one can calculate the $p_k's$ as in (1.2) and (1.3).

As the value of the finite truncation level is chosen *a priori* some caution is needed when using a block sampler. Values of L that are too small may lead to inference from a posterior that does not closely approximate the infinite stick-breaking process. Conversely values of L that are too large unnecessarily increase the computational burden.

Other methods have been developed to avoid the truncation problem, which allow sampling from the exact distriution G. Papaspiliopoulos and Roberts (2008) proposed one such method. This algorithm modifies the above by letting L change across MCMC iterations. Again letting L be the current number of atoms in the sampler a retrospective sampler introduces an auxiliary variable $U_i \sim Uniform(0, 1)$ setting $\xi_i = j$ if $\sum_{k=1}^{j-1} p_k < U_i < \sum_{k=1}^{j-1} p_k$, with more weights/atoms introduced if $\sum_{k=1}^{L} p_k < U_i$. This method allows one to sample from a countably infinite stick breaking process using only a finite number of atoms at any given iteration. As the method requires maintaining a detailed balance condition, it is non-trivial in many cases, and, consequently, Walker (2007) developed an equivalent method for sampling mixture models formed from infinite stick-breaking processes that is much simpler computationally.

1.1.1 Extensions to stick-breaking process

The stick-breaking process is a versatile prior over probability measures distributions, and can be used in many situations to develop rich prior distributions. It is

however defined assuming the base measures for the atoms and weights are independent. As there are various situations where one may want to pool information across repeated observations considerable work has been devoted to extending stick-breaking to situations where the atoms and/or weights are dependent MacEachern (1999).

In defining priors over rich function spaces Gelfand et al. (2005) developed the functional Dirichlet process. This process modeled spatial data over some compact domain $\mathcal{D}$. The functional Dirichlet process puts a non-parametric prior, such as those described in (1.2), on the base measure G_0, and puts a rich non-parametric prior over function spaces.

The functional Dirichlet process induced global clustering for each observation y_i. Other methods have been developed to add dependence in the weights induce local clustering of observations. Duan et al. (2007) and Petrone et al. (2009) extended the functional Dirichlet process to allow for local clustering. In Petrone et al. (2009) the weight corresponding to selecting atom at location s_i is spatially dependent. This results in defining f_i as a patchwork of functions made up of a set of global species $\{f_i^*\}_{i=1}^L$. Closely related to these approaches is that of Nguyen and Gelfand (2011) who developed the Dirichlet labeling process for clustering functional data.

1.2 Regression Methods

Consider modeling the function $f : \mathcal{X} \to \mathbb{R}^p$, $p \geq 1$, where $\mathcal{X}$ is an index set. Simple parametric assumptions on the form of f may fail to adequately characterize the curve of interest. Various methods exist that define priors over a large class of smooth functions. In what follows we consider two closely related approaches: basis function approximation and Gaussian Processes (GP).

Basis approximations assume $f(\cdot)$ can be approximated through a linear combination of functions, i.e.,

$$f(x) = \sum_{j=1}^{J} \theta_j b_j(x), \tag{1.13}$$

with $x \in \mathcal{X}$. Given an appropriately specified basis, and prior over $\{\theta_j\}_{j=1}^{J}$, one can model essentially any continuous function. There are many types of basis functions that one can consider, and each one puts increased prior probability over a certain class of functions. Consequently, the choice of basis contributes to the efficiency of the estimate. With a poorly chosen basis greatly increasing the uncertainty when estimaing f.

Closely related to basis function approach is the Gaussian process (GP). The GP is a stochastic process that, when given the appropriate covariance kernel, can approximate essentially any continuous function in $\mathbb{R}^p$. GP priors define f as a realization of a stochastic process having continuous sample paths. Like the basis approximation approach a poorly chosen covariance kernel may put low probability on sample paths similar to f. Consequently the choice of the covariance kernel may impact the efficiency in estimating f. We consider the problem of estimating f from both the basis approximation perspective as well through the use of Gaussian process regression.

1.2.1 Basis Regression

Assume that one observes the vector $Y = (y_1, \ldots, y_n)'$ at $(x_1, \ldots, x_n)'$ where $x_i \in \mathcal{X}$. Here Y are observations of f, i.e, $(f_1(x_1), \ldots, f_n(x_n))'$, made with error. In the following discussion we assume that

$$y_i = f(x_i) + \epsilon_i, \tag{1.14}$$

where $\epsilon_i \sim N(0, \sigma^2)$. As in (1.13) one approximates f assuming that it is well approximated by a linear combination of basis functions. These functions are defined on the knot set $\mathcal{T} = \{\tau_1, \ldots, \tau_J\}$, defined by a specific basis. For example the natural cubic spline basis is defined to be $b_1(x) = 1$, $b_2(x) = x$, $b_3(x) = x^2$ and

$$b_j(x) = (x - \tau_j)_+^3,$$

for $j \geq 4$, where $(x)_+ = x$ for $x \geq 0$ and 0 otherwise. Other examples of basis functions include the B-spline, kernel convolutions, and wavelet bases. With a basis function chosen, one completes a Bayesian specification by pacing a prior over $\{\theta_j\}_{j=1}^J$, and possibly the number and location of the knots.

Fully nonparametric approaches (Denison et al. (1998); Biller (2000) and Dimatteo et al. (2001)) put priors over $\{\theta_j\}_{j=1}^J$ as well as the the number and location of the knots. These methods develop different reversible jump MCMC (RJMCMC) (Green 1995) algorithms for posterior computation, and are usually dependent on the type of basis chosen. For example Biller (2000) develops an algorithm for B-splines that considers only three types of moves on knots: the addition, deletion or movement of knots during any iteration. This method allows for a highly flexible framework in which f can be represented through a function whose knot locations are unknown. Though these methods put priors over essentially any continuous function, the added computational burden often does not significantly improve estimation of f for most applications.

As it is often difficult to develop efficient RJMCMC algorithms, other methods have been developed to allow a high degree of flexibility when specifying a prior over f. These methods rely on penalized smoothing splines, see for example Eilers and Marx (1996) or Brumback and Rice (1998). Smoothing splines assume that the number and location of the knots are fixed, with a prior defined over $\{\theta_j\}_{j=1}^J$ that controls the smoothness of

the curve. By appropriately defining the proper prior the model can adapt to varying amounts of curvature in f.

One example of such a smoothing approach is the Bayesian P-Spline (Lang and Brezger 2004). Here priors for the coefficients $\{\theta_j\}_{j=1}^p$ are defined using first or second order random walks, i.e.:

$$\theta_j \sim N(\theta_{j-1}, \tau^{-1})$$

or

$$\theta_j \sim N(2\theta_{j-1} - \theta_{j-2}, \tau^{-1}),$$

where $\tau^{-1} \sim Ga(\frac{r}{2}, \frac{r}{2})$ and $r > 1$. Placing such a prior over $\{\theta_j\}_{j=1}^J$ and τ^{-1} allows the model to adapt to the appropriate level of smoothing. P-splines have been shown to be only slightly inferior to that that of Biller (2000), with the computational advantage that RJMCMC algorithms need not be employed.

1.2.2 Gaussian Process

The literature on Gaussian processes (Rasmussen and C. 2006) is vast. This review focuses on the use of the GP in regression. A GP $f \sim \mathcal{GP}(0, \sigma(\cdot, \cdot))$ is a stochastic process defined on a compact domain $\mathcal{X}$. It is defined such that for any finite set of points $X = \{x_1, \ldots, x_n\} \subset \mathcal{X}$, the points $\{f(x_i)\}_{i=1}^n$ are distributed as a multivariate normal with mean 0 and $cov(f(x_i), f(x_j)) = \sigma(x_i, x_j)$. GPs are often described in terms of a zero mean process. Extensions that allow the mean to vary across the domain are straightforward.

For the regression problem defined in (1.14) the GP specifies a prior over f through the mean process and covariance kernel $\sigma(x, x')$. For f observed locations X, the prior on

$f(X)$ is specified as a finite dimensional multivariate normal distribution, i.e., $f(X) \sim N(0, \Sigma_{(X,X)})$, where

$$\Sigma_{(X,X)} = \begin{pmatrix} \sigma(x_1, x_1) & \sigma(x_1, x_2) & \cdots & \sigma(x_1, x_n) \\ \sigma(x_1, x_2) & \sigma(x_2, x_2) & \cdots & \sigma(x_2, x_n) \\ \vdots & & \ddots & \vdots \\ \sigma(x_1, x_n) & \cdots & & \sigma(x_n, x_n). \end{pmatrix}$$

This defines the equivalent prior on Y :

$$Y \sim N(0, K),$$

where, $K = \Sigma_{(X,X)} + \tau^{-1} I$, I is the $n \times n$ identity matrix, and $\tau \sim Ga(a, b)$. In a Bayesian analysis one computes the posterior for $F(X)|Y$. Then, using the conditional properties of a multivariate normal distribution, one can calculate the posterior for any set of unobserved points $X' = (x_1, \ldots, x_m)'$. That is, assuming a zero mean GP, one has:

$$F(X')|Y \sim N(\hat{F}(X'), \hat{\Sigma}_{(X',X)}),$$

where

$$\hat{F}(X') = \Sigma_{(X,X')} K^{-1} Y,$$

and

$$\hat{\Sigma}_{(X',X)} = \Sigma_{(X',X')} - \Sigma_{(X,X')} K^{-1} \Sigma_{(X',X)}.$$

Here it is seen that the covariance kernel function $\sigma(\cdot,\cdot)$ is used to form a linear combination of basis functions to predict values of f. This is seen to be related to the basis function expansion if one sets $\Theta = K^{-1}Y$, and one uses the knot set $\mathcal{T} = X$.

As in the basis function case the covariance kernel is crucial in guaranteeing large support over the class of functions of interest. Examples of two such commonly used kernel functions include the Gaussian,

$$\sigma(x, x') = \sigma_f^2 \exp(-\frac{1}{2} c \left|x - x'\right|^2), \tag{1.15}$$

, where σ_f^2 is the function variance, and c is the bandwidth parameter; and the Matern class of covariance kernels

$$\sigma(x, x') = \sigma_f^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu r}}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu r}}{l}\right) \tag{1.16}$$

with parameters ν and l, where K_ν is a modified Bessel function. Given the proper kernel, with appropriate prior support over the hyperparameters a GP can be shown to have sample paths that are dense in the space of continuous functions (Tokdar and Ghosh 2007). Again given the proper hyperparameter on the covariance kernel Ghosal and Roy (2006) showed that such a GP puts positive support within an ϵ probability on any function in the in the reproducing kernel Hilbert space ($\mathcal{RKHS}$) of the kernel covariance function $\sigma(x, x')$, which given (Tokdar and Ghosh 2007) implies a prior within an ϵ distance of all continuous functions.

Other results show that the GP is a consistent estimator for the underlying true curve. This has been shown for both continuous (Mardia and Marshall 1984), and dichotomous regression (Ghosal and Roy 2006).

Posterior computation for GP regression proceeds in a relatively straightforward

manner. Conditional on knowledge of τ and the covariance kernel the posterior distribution of f can be computed by sampling from a multivariate normal distribution as described above. The hyperparameters in the covariance kernel are often more difficult to sample from, and require a metropolis within Gibbs sampling step, and mixing is usually poor. Another caveat to posterior computation is that the computations require inversion of a n dimensional covariate matrix. As inversion of such matrices are computationally demanding, requiring an algorithm of $O(n^3)$, GP computations are often computationally intractable for moderate to large problems.

Given the computational demands a GP can be approximated using basis function regression Higdon (2002); Rasmussen and C. (2006). For example the choice of Gaussian basis function, i.e, $b(x) \propto \exp(-\frac{1}{2}\left\|x\right\|^2)$ can be shown to be related to the covariance kernel $\sigma(s,s') \propto exp(-\frac{1}{2}\left\|\frac{s-s'}{\sqrt{2}}\right\|^2)$. Such approximations are often accurate, which greatly reduces the computational burden of GP posterior estimation.

Chapter 2

Mechanistic Hierarchal Gaussian Processes

2.1 Introduction

Studies of physiologic response to muscle stress are important in developing treatment protocols to combat work-, athletic-, and age related injury. In order to investigate muscle adaptation and maladaptation following repetitive resistance-type exercises, scientists often obtain a series of functional measures (often at the beginning and end of a multi-session exercise protocol) on the force produced by the muscle as it moves through its range of motion. These force curves can be compared to determine the benefit/harm of an exercise routine to a population of interest.

In our application we investigate one such study conducted on a rodent population. In this study scientists are interested looking at physiologic response between young (3 month) and old (30 month) animals exposed to the same resistive exercise protocol. Here animals underwent 13 training sessions on the dorsiflexor (lower leg) muscle group. At the beginning and end of the training regimen the muscles underwent isometric- (muscle activation without movement) and stretch shortening-(muscle activation with joint movement) contractions. Each force tracing was recorded, as illustrated in Figure 2.1. This figure is divided into five sections where each section

is separated by a vertical line. The first and last sections represent force generation when the muscles are not contracting. The second and fourth sections represent the force generated during an isometric contraction; with the third section denoting the stretch shortening contraction. Note that in the stretch shortening contraction there is an isometric component to force generation, and modeling should estimate both the isometric and stretch shortening components.

We have 86 such force tracings, and investigators wish to model the isometric and stretch shortening force generation. The data are defined as follows: for an individual measurement, 565 evenly spaced functional observations were taken. This measurement was taken two times (pre and post excersize protocol) resulting in $2 \times 565 = 1130$ functional measurements per animal. All 43 animals (28 old and 15 young) underwent the same resistive exercise protocol resulting in $48,590$ total measurements. Our intent is to investigate possible differences in response, between groups (young/old), as well as differences in response pre- and post-training. We are interested in comparing the individual and group level force tracings for isometric as well as stretch shortening contractions.

Current methods for functional analysis are insufficient to analyze such data as they do not take into account the detailed scientific information already available on the responses. Further, without extensive modification, they are unable identify the stretch shortening and isometric components. Parametric models, based upon ordinary differential equations (ODEs) do exist but are known to be inadequate for characterizing muscle force tracings. We develop Bayesian nonparametric methods that favor shapes consistent these parametric models, but are flexible enough to account for deviations from parametric assumptions. As we are primarily interested in mean differences between groups we further extend these methods to a hierarchical setting allowing functional ANOVA style comparisons.

2.1.1 Skeletal Muscle Force

Statistical methods for functional data analysis cannot easily incorporate mechanistic information and often produce results that are challenging to interpret. There is a large literature on muscle force output based on differential equations. Such models are easily interpretable and incorporate mechanistic information but are not flexible enough to realistically characterize available data. Motivated by the need to quantify differences in physiological muscle force output as a biomarker of muscle adaptation or pathology (Erdemir et al. 2007), we develop a non-parametric Bayesian modeling approach.

The force generated by muscle activation, illustrated in Figure 2.1, is nonlinear (Maffiuletti 2010; Parsaei and Stashuk 2011) and is associated with complex physiology, such as motor systems and muscle twitch dynamics. The current lack of accurate statistical models for characterizing force tracings has made effective statistical comparisons challenging.

Models for isometric force measurements date back to Hill (1938). A popular approach uses first order differential equations relating muscle force output to a series of motor, damper, spring systems (Wexler et al. 1997; Ding et al. 1998; Phillips et al. 2004). Such models may reasonably describe areas of observed data across the force activation curve but do not represent important aspects of the response. Other modeling approaches (Geronilla et al. 2006) attempt to characterize the response curve using a time-varying combination of basis functions, leading to improvements in prediction but a lack of interpretability and accommodation of prior mechanistic knowledge.

In an effort to develop better training/rehabilitation protocols tailored to individual needs, recent studies have investigated how age affects muscle adaptation and maladaptation following specific non-injurious, repetitive, resistance-type loading protocols designed to induce increases in performance and muscle mass. Initial investigations

(Cutlip et al. 2006; Murlasits et al. 2006) and subsequent validations (Ryan et al. 2008; Baker et al. 2010; Hollander et al. 2010) have supported the use of supramaximal, electrically-evoked stretch-shortening contractions precisely prescribed for inducing adaptation (increases in performance and muscle mass) in young animals following repetitive exposures of resistive muscle contractions. We use such data to study the effects of age on resistive muscle training sessions to better understand the benefits/harm of training across age groups.

Complexities arise when modeling the force tracings of a stretch-shortening contraction. The force output is a product of the isometric force at time t and a function related to joint movement. That is, the total force $h(t)$ measured at time t is thought to be

$$h(t) = Q(t)F(t), \tag{2.1}$$

where $F : \mathbb{R}^+ \to \mathbb{R}^+$ is the isometric force at time t and $Q : \mathbb{R}^+ \to \mathbb{R}^+$ is a function representing the increase ($1 < Q(t)$) or decrease ($0 < Q(t) < 1$) in isometric muscle force generation during a stretch shortening contraction. Scientific interest focuses on differences in $Q(t)$ and $F(t)$ across experimental conditions. Interest in $F(t)$ stems from the fact it is the 'baseline' force produced by the muscle. Differences are related to the general health of the muscle. Interest in $Q(t)$ is based upon the fact that it represents the 'potential' force that is released when the muscle moves; differences here relate to the ability of the muscle to adapt. Our focus is on developing nonparametric Bayesian methods that incorporate prior information using ODEs that can estimate both $Q(t)$ and $F(t)$ using minimal assumptions.

2.1.2 Relevant Literature

From a Bayesian perspective there has been some work on estimation of parameters from ODEs. Lunn et al. (2002) develops a framework for parameter estimation in pharmacokinetic/pharmocodynamic models. Putter et al. (2002) developed methods based on partial differential equations to estimate HIV infection, and Huang et al. (2006) developed a hierarchical framework to investigate the antiviral response for HIV infection in a population of individuals. The methods assume that the differential equations are characterized through finitely-many parameters, with posterior computation relying on Metropolis-Hastings steps.

Alternatively, one can rely on a Gaussian process (GP) emulator (Kennedy and O'Hagan 2000; 2001). In the first stage, a solver is used to obtain the differential equation solution on a finite grid of points. Then, uncertainty and bias are accommodated in the second stage through centering a GP prior on the differential equation solution. Mechanistic information is not included in the Gaussian process and hence, unless one assumes a very small deviation from the differential equation solution, the resulting trajectories may be quite unrealistic, leading to poor predictive performance. Our goal is to obtain mechanistic hierarchical Gaussian processes, which favor realizations that inherit the behavior of the ODE, while also allowing variability among individual trajectories across subjects.

Recent work (Lawrence et al. 2007; Alvarez et al. 2009; Honkela et al. 2010) develops latent force models, which embed mechanistic information into a GP prior. Here the GP has a mean function and covariance kernel derived from a differential equation similar to that of a simple motor, damper, spring system. This is accomplished by specifying a GP prior with squared exponential covariance function and integrating this GP over the Greens function corresponding to the specified ODE. In our experience, this

approach cannot be applied directly to our motivating application due to extreme ill-conditioning problems in the covariance matrix. Hence, instead of directly using their methods, we develop an alternative approach that relies on accurately approximating solutions to the differential equations. This method is then extended to a hierarchical Gaussian process (Behseta et al. 2005) allowing for sharing of information among subjects in the population. By using the hierarchical Gaussian process we model individual experimental group effects as well as individual subject effects.

2.2 Mechanistic Gaussian Process

Consider modeling an unknown functional response $h : \mathcal{T} \to \mathbb{R}$, with $\mathcal{T} = [t_0, t_1] \in \mathbb{R}$ and data consisting of error-prone measurements $(y_1, \ldots, y_n)'$ of h at locations $(t_1, \ldots, t_n)'$. A common approach lets

$$y(t_l) = h(t_l) + \epsilon_l, \tag{2.2}$$

where $h \sim \mathcal{GP}(0, R(\cdot, \cdot))$, a zero mean GP with covariance kernel $R(\cdot, \cdot)$, and $\epsilon_l \overset{iid}{\sim} N(0, \tau^{-1})$, with $l = 1, \ldots, n$. The covariance kernel $R(\cdot, \cdot)$ is frequently chosen as squared exponential, exponential, Matern or some default form that leads to flexible realizations. Although prior information about h can potentially be incorporated through the mean of the Gaussian process and choice of the covariance kernel, it can be difficult to choose appropriate values in practice.

We incorporate prior information by defining a covariance kernel favoring shapes consistent with mechanistic information specified by differential equations. We assume the information is expressible in the form of a linear ordinary differential equation

$$Lh(t) = \frac{d^m h(t)}{dt^m} + a_{m-1}(t)\frac{d^{m-1}h(t)}{dt^{m-1}} + \ldots a_1(t)\frac{dh(t)}{dt} + a_0(t)h(t) = r(t). \tag{2.3}$$

Given $\{a_0(t), \ldots, a_{m-1}(t)\}$ are non-zero on $\mathcal{T}$, the solution to (2.3) exists and, given initial values, can be expressed as

$$h(t) = \int_{t_0}^{t} G(t, \xi) r(\xi) d\xi. \tag{2.4}$$

Here $G(t, \xi)$ is Green's function, and the integral operator $\int G(t, \xi) d\xi$, described in shorthand as G below, is a linear operator, and is the inverse of the differential operator L in (2.3). As G is linear, if $r(t) \sim \mathcal{GP}(0, R(\cdot, \cdot))$, then $h(t)$ is also a Gaussian process with a new covariance kernel dependent on G and $R(\cdot, \cdot)$. This defines a GP over h whose covariance kernel favors shapes consistent with (2.3).

Unfortunately, in many cases the resulting covariance matrix is extremely ill conditioned resulting in computational instability. We tried a wide variety of existing methods for addressing ill-conditioning problems in GP regression with no success. The induced covariance of $h(t)$ tends to be substantially more subject to ill-conditioning than even the squared exponential covariance. Alternatively, by relying on a Runge-Kutta approximation (Asaithambi 1995), we develop an approach that allows direct modeling of $r(t)$ for an arbitrary covariance kernel $R(\cdot, \cdot)$. In our experience this increases the numerical stability of the approximation, while bypassing the cumbersome calculations necessary to compute the covariance kernel.

2.2.1 Approximation of the Process

There is a large literature on approximate solutions to differential equations. Given a set of initial conditions corresponding to $h(t_0)$ as well as the first $m-1$ derivatives of h evaluated at the initial point t_0, Runge-Kutta (RK) methods (see chapter 9 Asaithambi (1995)) offer efficient algorithms that approximate the solution to an m^{th} order ODE. When L is linear, RK methods express the numerical solution to the ODE as a linear

combination of the forcing function $r(t)$ evaluated at a finite set of points, $\{t_l\}_{l=1}^n$, along with the initial conditions $\mathbf{h}^* = \{h_1^*, \ldots, h_m^*\}$. We illustrate the approach using the Euler-Cauchy second order approximation, though other RK approximations proceed in much the same way.

The Euler-Cauchy approximation recursively defines a solution to $h(t)$ at $\{t_l\}_{l=1}^n$, by approximating the function as a linear combination of $\mathbf{r} = (r(t_1), \ldots, r(t_n))'$ and $\mathbf{h}^*$. As an example, consider a first order differential equation (i.e., $m = 1$ in (2.3)) where points are equally spaced with $\Delta = 2(t_j - t_{l-1})$. The approximate Euler-Cauchy solution is formed recursively by:

$$\hat{h}_l = h_{l-1} + \Delta \left\{ g(t_{l-1}, h_{l-1}) \right\} \tag{2.5}$$

$$h_l = h_{l-1} + \frac{\Delta}{2} \left\{ g(t_{l-1}, h_{l-1}) + g(t_l, \hat{h}_l) \right\}. \tag{2.6}$$

Here $g(t_{l-1}, h_{l-1})$ is a function of the derivative evaluated at t_{l-1} and h_{l-1} for $l > 1$ (e.g., for (2.3) with $m = 1$ one has $g(t_{l-1}, h_{l-1}) = [r(t_{l-1}) + A_0(t_{l-1})h_{l-1}]$). As long as $g(t, f)$ is linear the approximation is a linear function of $r(t)$ and the initial conditions $\mathbf{h}^*$. Consequently the solution can alternatively be expressed as a product of a matrix G and a vector of elements $\mathbf{r}^* = (\mathbf{h}^*, \mathbf{r}')'$. We form the matrix recursively as above, with row l corresponding directly to each function evaluation described above. Continuing with the example, one defines the matrix G as follows: first set the first row to $\langle 1 \quad 0 \quad \cdots \quad 0 \rangle$, which corresponds to h_1^*. Then for $l \geq 1$ the approximation proceeds by specifying a row vector

$$\hat{G}_{\{l,:\}} = \left[1 + \Delta A_0(t_{l-1})\right] G_{\{l-1,:\}} + \hat{K},$$

where $\hat{K}$ is a row vector of zeros except at the entry l, which is set to Δ, and $G_{\{l-1,:\}}$

is the previous row. One then defines row l of G as

$$G_{\{l,:\}} = G_{\{l-1,:\}} + \frac{\Delta}{2}\left[A_0(t_{l-1})G_{\{l-1,:\}} + A_0(t_{l-1})\hat{G}_{\{l,:\}}\right] + K,$$

where K is a row vector of zeros except at entries l and $l+1$, which are set to $\frac{\Delta}{2}$. Through this alternate expression one arrives at the approximation $h(t) \approx G\mathbf{r}^*$, and $h(t)$ is seen in the context of a linear regression where $\mathbf{h}^*$ and $\mathbf{r}$ are unknown. Though we describe the method using the Euler-Cauchy approximation (a second order method), a similar G matrix can be constructed using higher order RK methods. Higher order methods do form better approximations but require more functional evaluations of $r(t)$. This may require $r(t)$ to be evaluated at points on the index set that have not been observed and may greatly increase the computational complexity when sampling from the posterior. Before implementation this trade off should be evaluated, as in many situations a lower order approximation is adequate. For example numerical experiments produced results that in most cases had a maximum difference of 10^{-3} between the actual and numerical solution, indicating higher order solutions were not needed.

2.2.2 Posterior Sampling

For the above approximation, sampling from the mechanistic GP proceeds using a series of conditionally conjugate Gibbs steps. The discussion assumes model (2.2) with $\mathbf{Y} \sim N(h, \tau)$ where $\mathbf{Y} = (y_1(t_1), \ldots, y_n(t_n))$ and $\mathbf{h} = (h(t_1), \ldots, h(t_n))'$, with $\tau \sim Ga(a_0, b_0)$. Following the above discussion, the matrix G is formed from $\mathbf{A} = (A_1(t), \ldots, A_n(t))'$, which are parameters in (2.3). Further we assume the initial conditions are specified as $\mathbf{h}^* \sim N(A_0, B_0)$, which is independent of $r(t)$.

Sampling algorithm 1

1. Sample $\mathbf{r}^* \sim N(E, W)$ where $W = (\tau G'G + \Omega^{-1})^{-1}$ and $E = W(\tau G'Y + \Omega^{-1}\rho)$. Here $\Omega = \text{block-diag}(B_0, \Sigma)$ is an $(n+m) \times (n+m)$ matrix, ρ is the the prior mean of $\mathbf{r}^*$, and Σ is the $n \times n$ covariance matrix, formed from $R(\cdot, \cdot)$.

2. Sample τ from $Ga(a_0 + n/2, b_0 + (Y - G\mathbf{r}^*)'(Y - G\mathbf{r}^*)/2)$.

3. Marginalizing out $\mathbf{r}^*$, i.e., $Y \sim N(0, G\Omega G' + \tau^{-1}I)$ where I is a $(n \times n)$ identity matrix, sample the parameters $\mathbf{A}$ using a Metropolis-Hastings or griddy Gibbs (Ritter and Tanner 1992) sampling step.

2.3 Adaptation to Muscle Force Application

The mechanistic GP is not directly applicable to the muscle force application, which has the additional complication of decomposing $h(t)$ as

$$h(t) = \begin{cases} F(t) & t \notin [t_a, t_b] \\ F(t)Q(t) & t \in [t_a, t_b] \end{cases} \tag{2.7}$$

where $F(t)$ and $Q(t)$ are describable through first and second order differential equations, respectively. Additional constraints are needed to separately identify $F(t)$ and $Q(t)$. For $F(t)$, shape constraints are needed that rule out Gaussian processes, so we use restricted splines; and $Q(t)$ is known to equal one at the beginning and end of the stretch shortening contraction, so we modify the GP to include this information. In what follows, we describe the individual ODEs governing $F(t)$ and $Q(t)$ and outline an extension of the posterior sampling algorithm of Section 2.2.

2.3.1 Prior Extended to Muscle Force Data

We define an ODE for $F(t)$ and $Q(t)$ using generalizations of models from the muscle force literature. The isometric force function $F(t)$ is historically related to the first order differential equation (Hill 1938)

$$\frac{dF(t)}{dt} - BF(t) + p(t) = 0. \tag{2.8}$$

Here B represents the damping constant of the muscle fibers and $p(t)$ corresponds to the joint action of muscle at time t. We assume that the form of the motor activation function is unknown but is linear shortly after activation.

Placing a linearity assumption on $p(t)$ only during the SS contraction, we let

$$p(t) = \sum_{s=0}^{S} \beta_s b_s(t)$$

where $b_0(t) = 1$ and $b_s(t)$, for $s \geq 1$ are defined as piecewise polynomial splines on the interval $\mathcal{T}_s = [\tau_{s-1}, \tau_{s+1}]$. For $s \neq s'$ we use cubic splines defined to be 0 prior to the interval and 1 after the interval. Here for all $s \neq s'$ these intervals are defined outside of the range of the SS contraction. For the interval including the SS contraction we let $b_{s'}$ be a linear spline on the interval, 0 prior to, and 1 after the SS contraction. In order to model a flexible curve we use a large number of splines in estimating $p(t)$.

When the joint is moved through its range of motion the force on the joint is related to the angle of the joint and other factors. Angular motion is often described using a second order differential equation, and we follow this approach. As the exact form of the differential equation is unknown (i.e., damping constant etc.) we specify this

function through the fully specified second order differential equation:

$$\frac{d^2Q(t)}{dt^2} + \lambda\frac{dQ(t)}{dt} - AQ(t) + g(t) = 0, \tag{2.9}$$

where $g(t) \sim \mathcal{GP}(0, R(\cdot,\cdot))$, $A > 0$ and the damping constant $\lambda \leq 0$. Note that when $g(t) = 0$ this defines a periodic function with a period of $\pi\sqrt{A}$.

It is further known that the multiplicative effect of $Q(t)$ should be 1 prior to and after the joint is moved through a stretch shortening contraction. We add the constraint that at the beginning t_a, and end t_b, of the stretch shortening contraction $Q(t_a) = Q(t_b) = 1$. One can easily sample from this using the conditional properties of the multivariate normal distribution.

2.3.2 Posterior Sampling Extensions

The RK approximation is used to sample both $F(t)$ and $Q(t)$. Analogous to $\mathbf{h}^*$ above, we define $\mathbf{F}^* = (F_0)'$ and $\mathbf{Q}^* = (Q_0, Q_1)'$, initial value vectors for $F(t)$ and $Q(t)$ respectively. Similarly let $\mathbf{p} = (p(t_1), \ldots, p(t_n))'$, and $\mathbf{g} = (g(t_1), \ldots, g(t_n)),'$ which, as above, are vectors of the latent forcing functions evaluated at a finite set of points for $F(t)$ and $Q(t)$ respectively. Further define $\mathbf{p}^* = (\mathbf{F}^{*'}, \mathbf{p}')'$ and $\mathbf{g}^* = (\mathbf{Q}^{*'}, \mathbf{g}').'$ For convenience we refer to G as the Euler-Cauchy approximation to either (2.8) or (2.9). For all references to $F(t)$, G is the solution to (2.8), and for all references to $Q(t)$, G is the solution to (2.9).

In sampling $F(t)$ we note $\mathbf{p} = X\boldsymbol{\beta}$, where X is the $n \times (S+1)$ matrix of spline basis functions $\{b_s(t)\}_{s=0}^{S}$ evaluated at $(t_1, \ldots, t_n)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_S)$. Letting $\boldsymbol{\beta} \sim N(0, \Sigma_\beta)$, step 1 of sampling algorithm 1 is modified as follows:

Sampling algorithm 2

1. Putting the prior $\mathbf{F}^* \sim N(A_0, B_0)$ over the initial conditions, define $V = GX$, $\rho = (A_0' \ 0)'$, and $\Omega = \text{block-diag}(B_0, \Sigma_\beta)$. Then sample $\mathbf{p}^* \sim N(E, W)$, where $W = (\tau V'V + \Omega^{-1})^{-1}$ and $E = W(V'Y + \Omega^{-1}\rho)$.

We modify algorithm 1 to sample $\mathbf{g}^*$ given $Q(t_a) = Q(t_b) = 1$. This is done using the conditional properties of the multivariate normal distribution, i.e., for

$$\begin{bmatrix} \mathcal{X}_1 \\ \mathcal{X}_2 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} C_{11} & C_{12} \\ C_{12} & C_{22} \end{bmatrix} \right),$$

one has

$$\mathcal{X}_1 | \mathcal{X}_2 \sim N\left(\mu_1 - C_{12}C_{22}^{-1}\left[\mu_2 - \mathcal{X}_2\right], C_{11} - C_{12}C_{22}^{-1}C_{21}\right) \tag{2.10}$$

In the above approximation $Q(t_a) = Q_0$ and $Q(t_b) = G_{\{n,:\}}\mathbf{g}^*$, where $G_{\{n,:\}}$ is the last row of G, we modify step one of sampling algorithm 1 as follows:

Sampling algorithm 3

1. Calculating $g^* \sim N(E, W)$ as in algorithm 1, define the following quantities

$$E^* = \begin{bmatrix} I \\ G_{\{n,:\}} \end{bmatrix} E, \quad W^* = \begin{bmatrix} I \\ G_{\{n,:\}} \end{bmatrix} W \begin{bmatrix} I & G_{\{n,:\}} \end{bmatrix}.$$

   Then sample $\mathbf{g}^* | Q(t_a)Q(t_b)$ from a normal distribution whose mean and covariance are derived from E^* and W^* as in (2.10).

On the interval $[t_a, t_b]$ sampling $Q(t)$ and $F(t)$ proceeds conditional on knowledge of the other. To sample $F(t)$ one uses algorithm 2 and multiplies each row of G by the

corresponding value of $Q(t)$ (i.e., for row l one multiplies each element in this row by $Q(t_l)$). Similarly we multiply by $F(t)$ when sampling $Q(t)$ and sample from algorithm 3.

2.4 Hierarchical Mechanistic Gaussian Process

We extend the mechanistic GP to hierarchical modeling (Behseta et al. 2005). This allows modeling of individual curves as well as population means. The extension is described in terms of our application but can be readily used in other settings.

Consider a study in which there is a single factor of interest having I levels. For subject j a functional response $h_{ijk} : \mathcal{T} \to \mathbb{R}$ is measured K times. In our application the factor is age, $I = 2$, $K = 2$ and represents measurements pre and post exercise routine, and $h_{ijk}(t)$ is the time varying force function. Here, for all i, j, k, the n functional measurements are taken at equally spaced points on the index set $\mathcal{T}$. Data are modeled as:

$$y_{ijk}(t_l) = h_{ijk}(t_l) + \epsilon_{ijkl},$$

where $\epsilon_{ijkl} \overset{iid}{\sim} N(0, \tau_j^{-1})$, and a mechanistic Gaussian process prior is defined over $h_{ijk}(t)$ as in (2.3).

For subject j, in group i, the pre and post functional measurements are modeled as

$$h_{ijk}(t) = \tilde{h}_{ij1}(t)1(k \geq 1) + \tilde{h}_{ij2}(t)1(k \geq 2), \tag{2.11}$$

where $1(\cdot)$ is an indicator function that takes the value of 1 if the argument is true, and

0 otherwise. In terms of the mechanistic GP one models the latent forcing function as

$$r_{ijk}(t) = \tilde{r}_{ij1}(t)1(k \geq 1) + \tilde{r}_{ij2}(t)1(k \geq 2). \tag{2.12}$$

Here one integrates (2.12) using (2.4) to get (2.11). For interpretability between observations and groups we use the same integral operator G across all i, j and k.

Extending (2.12) to account for variability between factors we define

$$\begin{aligned} \tilde{r}_{ijk}(t) &\sim \mathcal{GP}(r_{ik}^{(1)}, R_{ik}^{(1)}(\cdot,\cdot)) \\ r_{ik}^{(1)} &\sim \mathcal{GP}(r_{k}^{(2)}, R_{k}^{(2)}(\cdot,\cdot)), \end{aligned}$$

with $k = 1, 2$ as in (2.12) and $r_k^{(2)} \sim \mathcal{GP}(0, R^{(3)}(\cdot,\cdot))$. Sampling from this hierarchy proceeds in much the same way as algorithm 1. Analogous to the case of the single curve we define $\tilde{\mathbf{r}}_{ijk}^*$, $\mathbf{r}_{ik}^{(1)*}$, and $\mathbf{r}_{ik}^{(2)*}$ as above. Further, we define the individual vector of observations $Y_{ijk} = (y_{ijk}(t_1), \ldots, y_{ijk}(t_n))'$. Sampling from the posterior is specified in terms of $\tilde{\mathbf{r}}_{ijk}^*$, $\mathbf{r}_{ik}^{(1)*}$, and $\mathbf{r}_{ik}^{(2)*}$, and proceeds as follows:

Sampling algorithm 4

1. For each i, j, k sample $\tilde{\mathbf{r}}_{ijk}^*$ conditionally on $\tilde{\mathbf{r}}_{ijk'}^*$ where $k' = 1$ if $k = 2$ and $k' = 2$ otherwise. Here let $Y^* = (Y_{ijk} - G\tilde{\mathbf{r}}_{ijk'}^*)$ and sample $\tilde{\mathbf{r}}_{ijk}^* \sim N(E, W)$ where $W = (\tau G'G + \Omega_{ijk}^{-1})^{-1}$, $E = W(\tau G'Y^* + \Omega_{ijk}^{-1}\mathbf{r}_{ik}^{(1)*})$. Here, as in sampling algorithm 1, Ω_{ijk} is subject specific $(n+m) \times (n+m)$ covariance matrix.

2. For each i, k pair $\mathbf{r}_{ik}^{(1)*} \sim N(E, W)$ where in this case $W = (\sum_j \Omega_{ijk}^{-1} + [\Omega_{ik}^{(1)}]^{-1})$ and
$E = W\left(\sum_j \tilde{\Omega}_{ijk}^{-1}\tilde{\mathbf{r}}_{ij}^* + [\Omega_{ik}^{(1)}]^{-1}\mathbf{r}_k^{(2)*}\right)$. Here $\Omega_{ik}^{(1)}$ is an $(n+m)\times(n+m)$ covariance matrix as specified above, where $R_{ik}^{(1)}(\cdot,\cdot)$ is used to compute the finite dimensional covariance for the latent forcing function.

3. For $\mathbf{r}_k^{(2)*}$ sample as in step 2 replacing $\tilde{\mathbf{r}}_{ijk}$ with $\mathbf{r}_{ik}^{(1)*}$ etc.

4. For each i, j sample τ_j from $Ga(a_0 + n, b_0 + \sum_k (Y_{ijk} - G\mathbf{r}_{ijk}^*)'(Y_{ijk} - G\mathbf{r}_{ijk}^*)/2)$, where $\mathbf{r}_{ijk}^* = \tilde{\mathbf{r}}_{ij1}^* 1(k \geq 1) + \tilde{\mathbf{r}}_{ij2}^* 1(k \geq 2)$.

5. Sample $\mathbf{A}$ similar to algorithm 1.

Note that inference on the group average curves $h_{ik}^{(1)}(t)$ and the population average curves $h_k^{(2)}(t)$ proceed using the approximation $G\mathbf{r}_{ik}^{(1)*}$ and $G\mathbf{r}_k^{(2)*}$ respectively, and, as G is the same across all i, j, k, the population averages have the same interpretation as other curves in the hierarchy. Extending the above framework, i.e. adding more hierarchies, is straightforward. Each additional hierarchy is sampled as in step 2 noting that the previous level is used as the input vector.

2.4.1 Extensions to the Hierarchal Mechanistic Process

We extend the hierarchical mechanistic process to our application. Here

$$h_{ijk}(t) = \begin{cases} F_{ijk}(t) & t \notin [t_a, t_b] \\ F_{ijk}(t)Q_{ijk}(t) & t \in [t_a, t_b] \end{cases}$$

where $F_{ijk}(t)$ and $Q_{ijk}(t)$ are defined using (2.8) and (2.9) respectively. For $F_{ijk}(t)$ and $Q_{ijk}(t)$ we define the hierarchy over the latent forcing function, with $p_{ijk}(t)$ and $g_{ijk}(t)$ specified as in (2.12). This discussion uses the same notation as above, i.e., $\tilde{g}_{ijk}(t), \tilde{\mathbf{g}}_{ijk}^*, \tilde{p}_{ijk}(t), \mathbf{p}_{ijk}^*$ etc.

For $Q_{ijk}(t)$, the forcing functions $\tilde{g}_{ijk}(t), g_{ijk}^{(1)}(t)$, and $g_{ijk}^{(2)}(t)$, are defined such that $Q_{ijk}(t_a) = Q_{ijk}(t_b) = 1$ etc, and these constraints are implemented in exactly the same way as above. To sample $\tilde{\mathbf{g}}_{ijk}^*, \mathbf{g}_{ik}^{(1)*}, \mathbf{g}_k^{(2)*}$ one proceeds by computing E and W as in sampling algorithm 4, then sampling from the conditionally conjugate distribution specified in sampling algorithm 3.

Hierarchical extensions in modeling $F_{ijk}(t)$ are direct. Here we place multivariate normal hierarchies over the spline coefficient vector $\boldsymbol{\beta}$ vector, i.e:

$$\tilde{\boldsymbol{\beta}}_{ijk} \sim N(\boldsymbol{\beta}_{ik}^{(1)}, \Sigma_{\beta,ik}^{(1)})$$
$$\boldsymbol{\beta}_{ik}^{(1)} \sim N(\boldsymbol{\beta}_{k}^{(2)}, \Sigma_{\beta,k}^{(2)}),$$

which in turn defines $\tilde{p}_{ijk}(t), p_{ik}^{(1)}(t)$, and $p_{k}^{(2)}(t)$. Sampling each $\tilde{\mathbf{p}}_{ijk}^{*}, \mathbf{p}_{ik}^{(1)*}$, and $\mathbf{p}_{k}^{(2)*}$ proceeds by placing the modifications of sampling algorithm 2 into sampling algorithm 4.

2.5 Simulation

We conduct a simulation experiment based upon the model developed in (2.7). Here curves, similar to those expected in a muscle force application are generated, and the simulated curves are compared against posterior estimated curves. Similar to the muscle force application the hierarchy was generated assuming $I = 2$, $J = 30$ and $K = 2$. The group levels of the hierarchy, i.e., $F_{ik}^{(1)}(t)$ and $Q_{ik}^{(1)}(t)$, were generated to resemble muscle force tracings of isometric and stretch shortening contractions respectively, and were simulated based upon (2.8) and (2.9). The individual level data were generated at 565 equally spaced points, assuming $\Delta = \frac{1}{260}$. Here the first 80 observations represent the force tracing prior to muscle activation. After activation 120 observations were taken of $F_{ijk}(t)$. The next 201 observations were of $F_{ijk}(t)Q_{ijk}(t)$ with the 164 remaining observations generated from $F_{ijk}(t)$. Similar to the real data, all data was generated assuming little variability between observations; here $\tau_j = 1000$ for all observations.

We chose weakly informative priors for all hyper parameters. We place a GP prior over $g_{ijk}(t)$ where the covariance kernel is specified using the squared exponential kernel $K(t, t') = \sigma^2 exp(-\ell \|t - t'\|^2)$. We set $\sigma^{-2} \sim Ga(1, 1)$ and let $\ell \sim Ga(1000, 0.1)$, which

reflects the assumption that $g_{ijk}(t)$ is not expected to be very smooth. The same assumptions are made for all other levels of the hierarchy. For $p_{ijk}(t)$ we put normal priors over the β coefficients, with diffuse priors specified at the topmost level. The precision parameter for all other levels was assigned a $Ga(0.1, 0.1)$ prior. Finally the precision parameter τ_j was specified using a $Ga(100, 0.1)$ prior. This is a vague prior on τ_j centered approximately at the observed error found in muscle force tracings. For the parameters in (2.8) and (2.9) we defined discrete uniform priors over a range of plausible values. Here B is defined to be in $[4.1, 5.2]$, based upon analyses of isometric data with a parametric parametric model. Further the parameter A, which defines the period of $Q_{ijk}(t)$ is put in the range of $[-2.3, -0.6]$. This choice corresponds to a range representing a half to a full period. Finally the damping constant was expected to be negligible, and λ was given a plausible range of $[0.01, 1]$. Note λ can not take on values at 0 due to the identifiability constraints on the ODE.

We collected $25,000$ MCMC samples disregarding the first $5,000$ as a burn-in. Every other observation was then recorded, leaving $10,000$ samples for the analysis. Examinations of trace plots for the quantities of interest, i.e., the individual curves, as well as curves in the hierarchy, showed excellent mixing. Hyperparameters for the covariance kernel as well as the parameters specified in (2.8) and (2.9), exhibited poor mixing. This however did not affect the convergence for the quantities of interest.

For the quantities of interest (i.e., $F_{ijk}(t), Q_{ijk}(t), F_{ik}^{(1)}(t)$, and $Q_{ik}^{(1)}(t)$)), which represents 125 total curves, the true curve was within the 95% credible region at the specified level for these curves. Figure 2.2 shows the estimates of $Q_{i1}^{(1)}$ and $Q_{i2}^{(1)}$, for one of the groups. Here the true curve is given by the dashed line, the estimated curve is shown in solid black, and the 95% credible intervals on the central estimate are given by the dotted lines. One can see that the true curve, represented by the dashed line, is estimated within these regions. This figure is representative of the other estimates,

where truth is well described by the model.

2.6 Muscle Force Application

With the goal of investigating the effect of non-injurious, repetitive muscle contractions on muscle force generation, we apply our approach to data compiled from Cutlip et al. (2006), Murlasits et al. (2006), and Baker et al. (2010). In these studies, 15 young (3 months), and 28 old (30 months), rats' dorsiflexor muscles were exposed to a resistive muscle contraction protocol that included thirteen sessions. At the end of each session the dorsiflexor muscle group underwent isometric as well as stretch shortening contraction (as described in Figure 2.1). Individual observations were taken at evenly spaced intervals ($\Delta = \frac{1}{260}$ of a second). The entire measurement lasted just over 2 seconds, resulting in 565 total functional observations as in our above simulation study. Our analysis looks at possible differences between muscle force measurements pre (after the first resistive muscle contraction protocol) and post (after the last protocol) study, between young and old animals. Priors for all parameters as well as computational implementation was as specified in the simulation.

Figure 2.3 shows the individual fits of $h_{ijk}(t)$ for one animal for their pre and post observations. Here the central posterior estimated curve is shown in black, with the observed data shown using gray hash marks. The credible intervals are not shown, as they are too close to the central estimate to be visible in the figure. Figure 2.4 shows the expected mean isometric contraction for the pre (dashed line) and post (solid line) exercise protocol in the old animals (top left) and the young animals (top right). The difference (solid line) between the pre and the post training, as well as the 95% pointwise credible interval (dashed line), is shown in the bottom row for the old (bottom left) and young (bottom right) animals. Here it is seen that the young animals, as a group, displayed increased muscle performance related to stretch

shortening contractions; however, the old animals did not display a difference for much of the curve. When there were differences, they were small and not seen as biologically relevant. Likewise no difference was shown in the group average isometric contraction (i.e., $F_{ik}^{(1)}(t)$). For the group level isometric contractions, figure 2.5 shows the estimated posterior curves (top) and corresponding differences (bottom) for the young animals. Here the pre treatment (dotted line) and post treatment(solid line) estimated isometric contractions are shown in the top row, and, though the central estimates are different, the bottom row shows that there is not enough evidence to suggest differences between the two groups. Similar results (figure not shown) were observed for the older animals.

The model also allows one to look at individual estimates between curves. Here though the old animals showed no significant differences at the group level for both isometric and stretch shortening contraction force generation, individual differences were seen. Figure 2.6 shows the stretch shortening contraction difference for an individual animal. Here the pre and post treatment estimates are shown in the top graph with the estimated differences being shown in the bottom graph. Here individual differences can be seen, which is significant as it supports the idea that some some older animals vary in their physiology related to dynamic responses.

Note that there is an additional advantage of modeling the latent forcing function as it may be used to generate or support hypotheses. For example, for $F_{ijk}(t)$, the latent forcing function $p_{ijk}(t)$ represents the muscle motor action at time t. These curves (figure not shown) show a steep increase right after activation, a sharp decrease shortly thereafter, and then a stabilization to a near constant level. This is supportive of the idea that large amounts of calcium influx the cytoplasm and bind rapidly to troponin upon muscle activation (steep increase in force tracing), until finally calcium is sequestered from the cytosol upon deactivation (return to baseline in force tracing).

2.7 Discussion

This article proposes a flexible nonparametric Bayesian method that takes into account prior scientific information based upon an ODE. This method further develops an approximate sampling algorithm using a Runge-Kutta approximation to the ODE. The nature of the approximation allows for a hierarchical specification at the population level estimates by modeling the latent forcing function directly. The goal was to develop an inferential framework for muscle force tracings, and investigate the effect of non-injurious resistive exercise protocols on different muscle types (i.e., young and old muscles). It was important to accurately model both the overall functional response and the two constituent functions, which themselves have scientific interest.

Given the results it may be that the dynamic force generated through the stretch shortening contraction may be more informative and specific in showing adaptation and maladaptation following non-injurious mechanical loading. Specifically, it can be seen that the younger rats have an adaptive response in dynamic muscle force produced in that the force generated is, at the population level, greater after the exercise protocol. Whereas older rats have little if no response to the same muscle exercise protocol in terms of the dynamic force generated. Further, the maximal force for older rats appears to occur at a different joint angle, suggesting a physiologic differences in the length-tension curve dynamics as the animals age.

The proposed approach can be extended to human muscle force tracings, and may allow for in-depth study of human physiologic responses to exercise routine post training. Such an analysis, on the entire force output, has previously not been attempted. For such a study the age response, as well as other variables, can be included in the hierarchical framework. In this manner the efficacy of the current 'one size fits all' approach across the spectrum of prevention/intervention including, occupational medicine, physical therapy, strength conditioning, and wellness programs can be studied. If similar age

results are seen in human populations, this may result in different treatment protocols depending on age or other variables of interest.

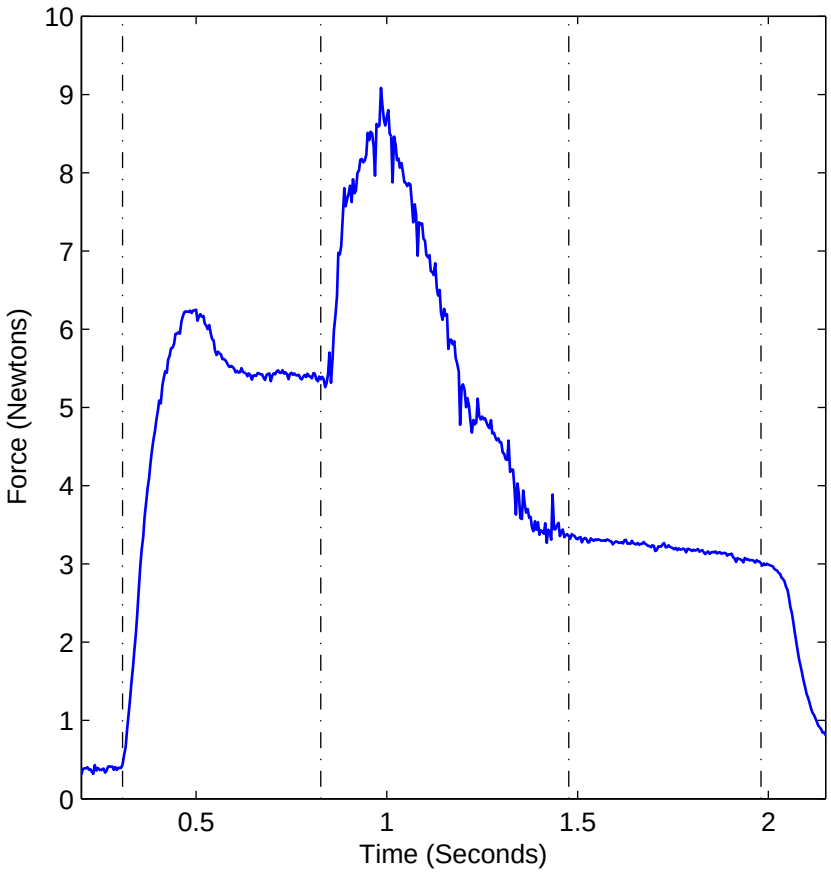

Figure 2.1: The first and second lines represent the beginning of the isometric and stretch shortening contraction, respectively. The third and fourth lines represent the end of the stretch shortening and isometric contractions, respectively.

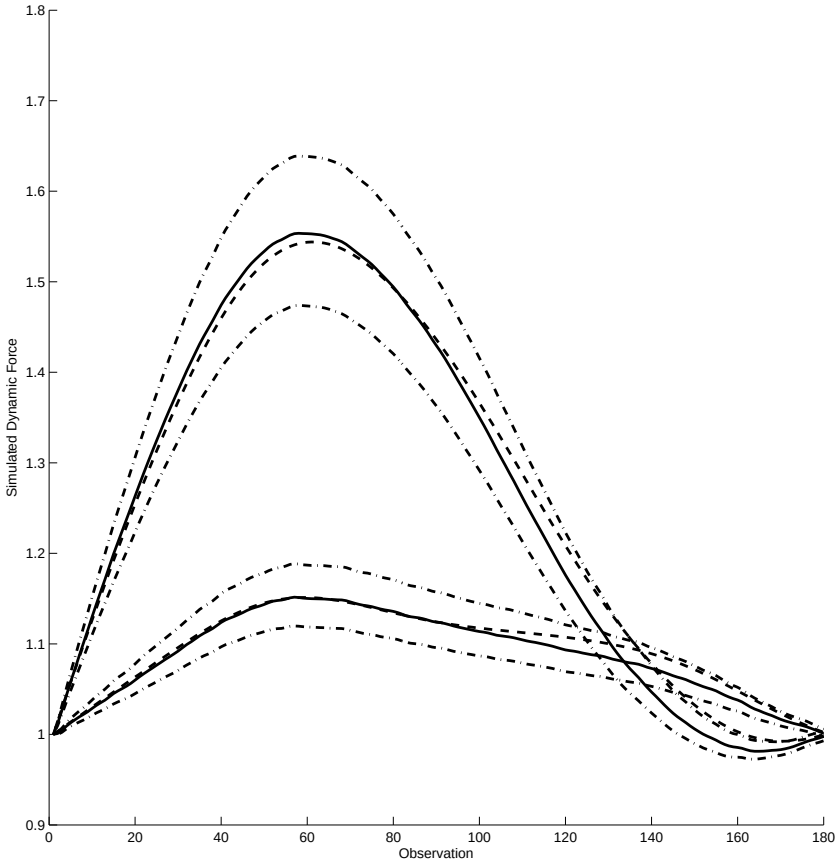

Figure 2.2: Estimated group level curves for the dynamic force in a stretch shortening contraction. Solid line, and corresponding 95% credible region (dotted line), representing the estimated curve. Here truth is represented by the dashed line.

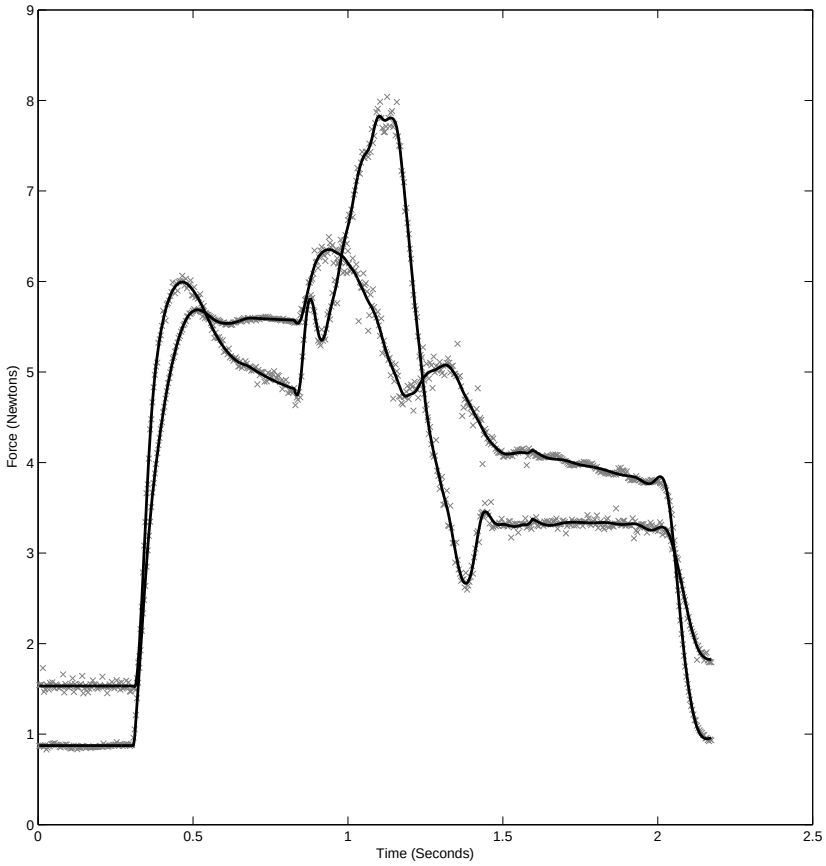

Figure 2.3: The estimated mean isometric force generated for a single animal pre and post treatment. The dark black line represents central estimates of $Q(t)F(t)$, with the dark gray hash marks representing the observed data. Here credible interval estimates are not shown as they are very narrow.

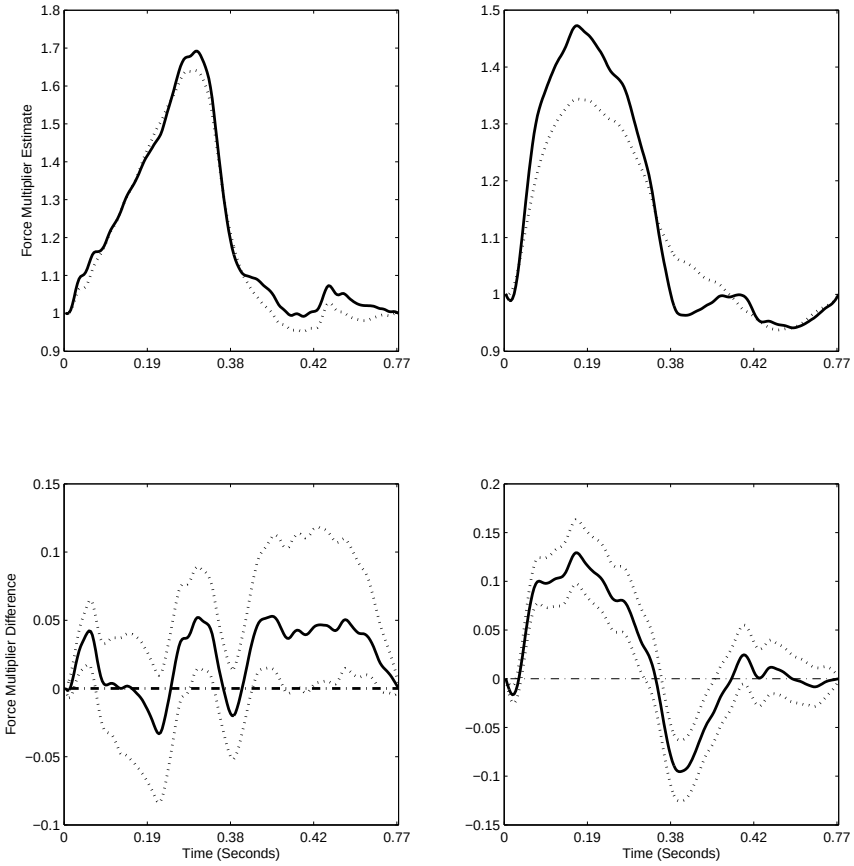

Figure 2.4: The estimated group level dynamic force multiplier generated by young (right column) and old (left column) animals. The bottom row represents the 95% pointwise credible interval for this difference.

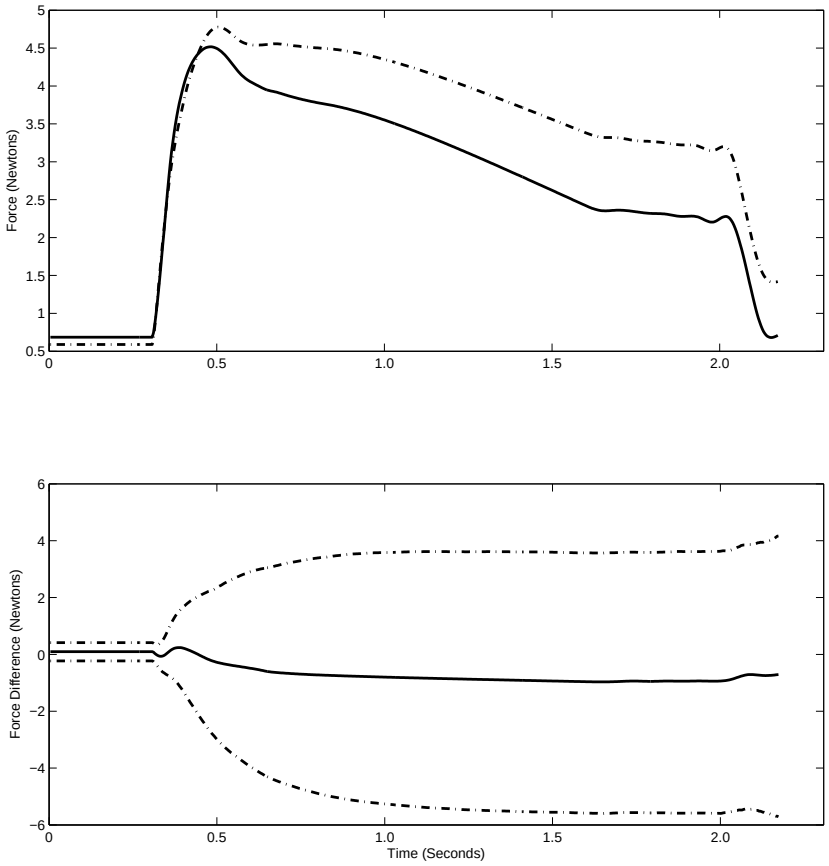

Figure 2.5: Estimated mean isometric muscle force generated for the young animals pre (dashed line) and post (solid line). The bottom row gives the estimates, and 95% pointwise credible intervals of the difference between the two.

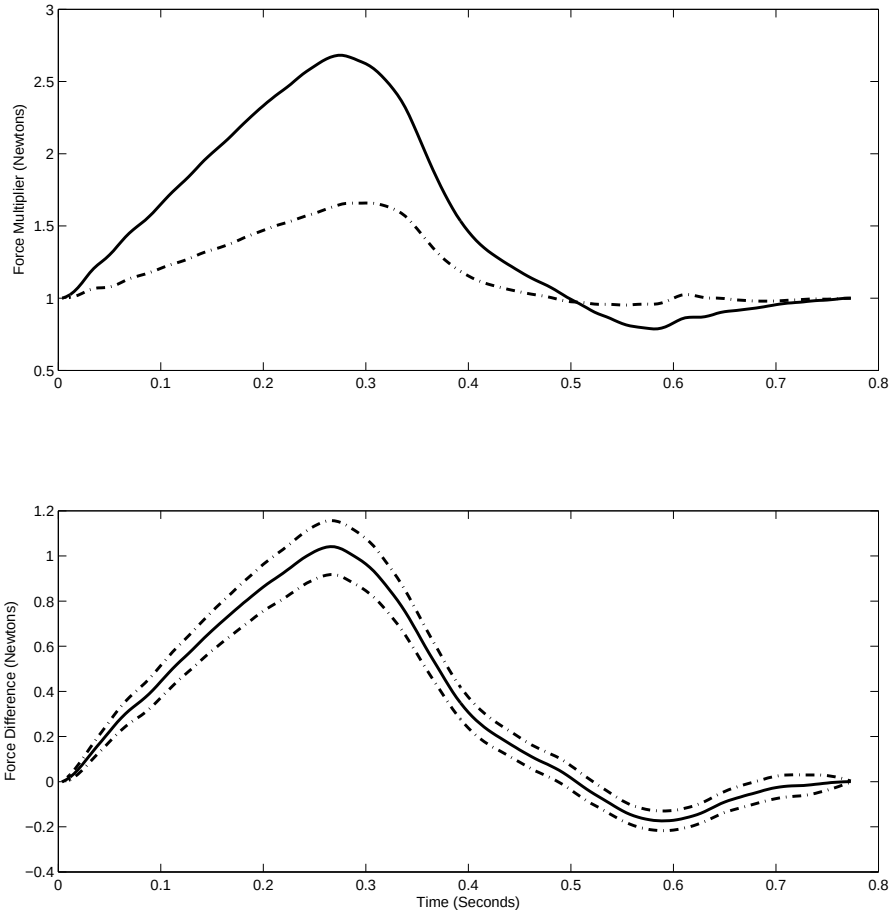

Figure 2.6: Estimated dynamic muscle force for an old animal. Here the top figure is the central estimate for the pre (dash dotted line) and post (solid line), and the bottom figure is the estimated difference between the two estimates.

Chapter 3

Local Extrema Splines

In many applications there is interest in modeling an unknown function $f : \mathcal{X} \to \mathbb{R}$. Approaches such as splines or Gaussian processes (GP) typically do not incorporate prior information on the number of extrema of the function, and often produce estimates having many bumps. It is typical in many applications to have strong prior information that the function under study has no more than a given number of local extrema. With such a restriction, one can gain efficiency while producing a more appealing estimate that agrees with prior knowledge and does not have artifactual bumps that are difficult to interpret.

U-shaped or umbrella-shaped functions are particularly common in applications. For example, in biomedical studies relating dose of an exposure to a response variable, there can be hormetic relationships in which the exposure is beneficial at low doses and is toxic at higher levels. This leads to a U-shaped function. More common are umbrella shapes. These occur due to toxicity at high dose levels causing a downturn in an initially monotone increasing dose response. Such umbrella or hill shapes are also common in financial and engineering applications. Seasonal fluctuations can also lead to multiple extrema, with such fluctuations often not well approximated by perfectly periodic basis functions such as sine curves. We develop local extrema (LX) splines that allow one to place an upper bound on the number of local extrema (changepoints), while otherwise

allowing highly flexible shapes through uncertainty in the location of the changepoints.

As a motivation, we focus on two examples through the paper. The first example relates body mass index (BMI) to mortality. Several studies have reported a U-shaped relationship between BMI and mortality, with very low and high BMI associated with higher all-cause mortality. As there are limited data available for individuals with very low and high BMIs it is useful to restrict the regression curve to have no more than one (but possibly zero) interior extrema. This restriction bypasses the well know problem of artifactual bumps in a completely unrestricted estimate, such as that produced by a GP or unrestricted splines, while also substantially reducing uncertainty in interpolating across sparse data regions. Our proposed LX spline method provides a simple and computationally efficient way to include such constraints.

Our second motivating application is daily maximum temperature data from Albany, NY. Though the daily temperatures can of course deviate substantially from this average trend, due to seasonal trends, we expect the smoothed daily temperature curve across the year to have a single nadir and single maximum. Although sin curves are widely used to model such seasonal trends, such a parametric model may be insufficiently flexible to capture the real temperature data. Hence, it is appealing to consider a nonparametric model that restricts the number of interior local extrema to be no more than two, while allowing flexibility. Our LX spline method provides a simple approach to accomplish this that conveys some practical gains.

Conceptually, when modeling an unknown curve with a finite number of local extrema, one needs to constrain the function to have monotone segments with unknown changepoints (local extrema). For strictly monotone functions, there is a rich literature on restricting functional forms from Bayesian (see for example Neelon and Dunson (2004) and Shively et al. (2009)) and frequentist (Ramsay 1988; 1998) perspectives.

Ramsay (1988) proposed I-splines for isotonic regression. Given only positivity

constraints on the basis function coefficients this approach effectively models monotone curves. Alternatively he also (1998) uses penalization methods to enforce monotonicity. Frequentist work on constraining curves to have single changepoint is sparse. Meyer (2008) discussed methods for enforcing convexity restrictions, but, as many curves do not fall into this space of functions, such a restriction is limited. More sophisticated penalization methods have been developed (Heckman and Ramsay 2000), but these do not guarantee enforcement of the shape constraint.

From a Bayesian perspective Neelon and Dunson (2004) develop a prior over piecewise linear splines for isotonic regression. This approach, given enough splines, was shown to well approximate smooth monotone functions. More recently Shively et al. (2009) developed an approach using free knot splines. For umbrella shape constraints, Hans and Dunson (2005) developed a Bayesian model that allows a single discrete-valued changepoint. As for other Bayesian changepoint models, posterior computation can be challenging in considering extensions to more than one changepoint. Shively et al. (2011) developed methods for fixed and free knot splines that model continuous monotone segments having a single unknown changepoint. This model was specifically developed for continuous U-shaped curves, and was not extended to the case of multiple changepoints.

This article proposes a fundamentally different approach to the shape constrained regression problem. Instead of only looking at the coefficients on the splines (Neelon and Dunson 2004; Shively et al. 2009; 2011), or penalizations (Ramsay 1998), we develop a novel spline construction similar in spirit to the I-spline construction of Ramsay (1988) or C-spline construction of Meyer (2008). Our construction, when paired with positivity constraints on the spline coefficients, enforces prior shape restrictions on the curve of interest by limiting the number of local extrema. In our approach the resulting model provides a flexible unified framework for inference when an upper bound on the number

of changepoints is known.

3.1 LX Splines

3.1.1 Formulation

Let Y be a vector of noisy observations of an unknown function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} = [\gamma_l, \gamma_b] \subset \mathbb{R}$. We wish to estimate f via

$$\underset{\beta \in \mathbb{R}}{argmin} \ \|\mathcal{A}(X, \boldsymbol{\beta}) - Y\|^2 + \lambda \mathbf{g}(\beta) \tag{3.1}$$

where $\|\cdot\|^2$ is the L2 norm, $\mathcal{A}(X, \boldsymbol{\beta})$ is a linear map, $\boldsymbol{\beta}$ is $(K+1) \times 1$ vector and $\lambda \mathbf{g}(\boldsymbol{\beta})$ is a norm on $\boldsymbol{\beta}$. We develop a spline based approach to the map $\mathcal{A}(X, \boldsymbol{\beta})$ that limits the curve to at most H changepoints on $\mathcal{X}$.

Consider the linear map

$$\mathcal{A}(X, \boldsymbol{\beta}) = \sum_{k=0}^{K} \beta_k b_k^*(x), \tag{3.2}$$

where β_k are basis coefficients, $b_0^*(x) = 1$, and $b_k^*(x)$, for $k \geq 1$, are basis functions defined on the knot set $\mathcal{T}^* = \{\tau_k\}_{k=1}^K$, with $\gamma_l < \tau_1 \leq \tau_2 \leq \ldots \leq \tau_K \leq \ldots \leq \tau_{K+j}$, where j is defined as the order of the spline. Given a well chosen basis, (3.2) can approximate essentially any continuous function. To limit the forms of the functions approximated, we restrict this function to have a most H changepoints by defining $b_k^*(x)$, for $k > 0$, as:

$$b_k^*(x) = CM \int_{-\infty}^{x} \frac{\prod_{h=1}^{H}(\xi - \alpha_h)}{\tau_{k+j}^H} B_{(k,j)}(\xi) \, d\xi, \tag{3.3}$$

where $B_{(k,j)}(x)$ is a B-spline basis function of order j (De Boor 2001) defined on $\mathcal{T}^*$, $C \in \{1, -1\}$, and M is a positive constant. As B-Splines are defined to be greater

than or equal to zero, if all β_k, for $k \geq 1$, are constrained to be greater than zero, then (3.2) can have at most H changepoints. These changepoints exist in $\mathcal{X}$ only if $\alpha_h \in (\gamma_l, \gamma_u)$. Note there can be less than H extrema if α_h is equal to γ_l or γ_u. Consequently to complete the LX-spline specification, we constrain $\beta_k \geq 0$ for all $k \geq 1$ and $\alpha_h \in \mathcal{X}$ for all h. In what follows we describe how to estimate $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$ and $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_H)'$ to enforce such constraints.

3.1.2 Spline Construction

In constructing the LX-spline we consider the B-spline order. Though higher order B-spline constructions may offer smoother approximations, by choosing many lower order splines with an appropriate penalty for smoothness and sparsity, one can eliminate the need for higher order splines. Consequently, we use B-splines of order 2; these splines are triangle distributions defined on $[\tau_k, \tau_{k+2}]$ having mode τ_{k+1}, and allow for closed form expressions of the LX-spline. In practice we have observed excellent performance when a large number of splines are considered.

The spline is constructed to take into account the effect of the smoothing penalty. The height of an individual spline is dependent on $\boldsymbol{\alpha}$, and when there are large differences between the height of individual splines this leads to over smoothing of certain regions and under smoothing in others. Our implementation minimizes this effect by using a two step estimation procedure fully discussed below. In the first stage $\boldsymbol{\alpha}$ is considered unknown and splines are constructed as in (3.3). Here the term τ_k^H in $\frac{\prod_{h=1}^{H}(\xi-\alpha_h)}{\tau_{(k+2)}^H}$ puts the absolute height of each spline roughly on the same scale, while keeping a polynomial representation for $b_k^*(x)$. Without this term the height of the largest and smallest spline can differ by an order of magnitude. In the second stage $\boldsymbol{\alpha}$ is known and M is used to make the height of each spline 1. Note that the τ_k^H term produces splines that may be undefined if $0 \in \mathcal{X}$. To avert this problem we map $\mathcal{X}$ to a closed set in $\mathbb{R}^+$ and

construct all splines on this set.

Figure 4.1 shows the spline construction for $H = 1$ (left) and $H = 2$ (right) on the interval $\mathcal{X} = [2, 3]$. For $H = 1$ we set $\alpha_1 = 2.5$, and for $H = 2$ we let $\alpha_1 = 2.33$ and $\alpha_2 = 2.66$. Here one can see how $\boldsymbol{\alpha}$ combined with the positivity constraint on $\boldsymbol{\beta}$ controls the shape of the curve. For the right panel the curve shapes can be: monotone increasing ($\alpha_1 = 2$), J-shaped ($\alpha_1 \in (2, 3)$), and monotone decreasing ($\alpha_1 = 3$). Note if $C = -1$ the patterns invert, and an umbrella shaped ordering is considered. When $H > 1$ more complicated shapes can be constructed. In each case if $\alpha_h = \gamma_l$ then $(x - \alpha_h)$ is positive on $\mathcal{X}$, if $\alpha_h = \gamma_u$ the quantity is negative in this range, and if $\alpha_h \in (\gamma_l, \gamma_u)$ a changepoint exists in $\mathcal{X}$. Here it is seen that the number of α_h in (γ_l, γ_u) defines the number of extrema of the curve.

3.1.3 Estimation

In our implementation we use a large number of equally spaced splines. Without the term $\lambda \mathbf{g}(\beta)$, the minimization problem is ill-posed. We regularize the problem using a variant of the L2 penalty. This penalty is defined over the differences in the $\boldsymbol{\beta}$ coefficients

$$\mathbf{g}(\beta) = (\beta_1)^2 + \sum_{k=1}^{K} (\beta_k - \beta_{k-1})^2.$$

As the number of basis functions increase many of these differences will be small. This results in minimal change in the derivative. Note that the derivative does change some as $\frac{\prod_{h=1}^{H}(\xi - \alpha_h)}{\tau_{(k+2)}^{H}}$ is non-constant. In our experience this implementation performs well across a wide array of functions.

As mentioned above we estimate $f(x)$ through a two stage process. In the first stage we estimate $\boldsymbol{\alpha}$ using constrained minimization, with the smoothing penalty λ set near

zero. Note that we allow $\alpha_h = \gamma_l$ and $\alpha_h = \gamma_u$, which allows deletion of changepoints. In the second stage $\boldsymbol{\alpha}$ is known and the spline basis functions are normalized to have a maximum absolute height of one. We then compute the optimal λ value using generalized cross-validation (GCV) (Golub et al. 1979), and estimate $\boldsymbol{\beta}$ using constrained minimization.

3.2 Spline Properties

For the LX-spline to be useful in applications it is important that it can well approximate a large class of functions. In what follows let $\mathcal{F}_H$ be the set of Lipschitz functions on $\mathcal{X}$, having no more than H changepoints. Further let $\mathcal{F}_H^*$ be the space of functions defined by the LX-spline in (3.3) having knot set $\mathcal{T}^*$. We give the following approximation theorem.

Theorem 1: Let $f \in \mathcal{F}_H$ then for some knot set $\mathcal{T}^*$ there exists a $f^* \in \mathcal{F}_H^*$ such that:

$$\|f - f^*\| < \epsilon,$$

where $\|\cdot\|$ is the sup-norm metric.

Proof:

Without loss of generality assume that the knot set $\mathcal{T}^*$ is a tight grid of evenly spaced points. Further, recalling that a B-spline of order j is defined to be nonzero only on $[\tau_k, \tau_{k+j-1}]$, let Δ_j be the width of this interval. It is enough to show for the case where there are exactly H changepoints in $\mathcal{X}$. When there are less one can always place a $\alpha_h = \gamma_l$ or $\alpha_h = \gamma_u$ (depending on the pattern) and proceed as below. Finally we prove when $C = 1$, as the construction is the same if $C = -1$.

Let $\tilde{f}$ be a taut B-spline approximation of f of order $j+1$. Note that this implies $\tilde{f}$

has only H changepoints in $\mathcal{X}$. Further define $\tilde{f}$ such that its derivative is a linear map of B-splines (all of order j) defined on the knot set $\mathcal{T}^*$. It is well known that such splines approximations exist and can approximate f within a factor of Δ_{j+1} (see chapter X, XII, and XVI in De Boor (2001)). As $\tilde{f}$ can be made arbitrarily close to f, assume that $\tilde{f}$ approximates f close enough such that its changepoints $\{\tilde{\alpha}_1, \cdots, \tilde{\alpha}_H\} \in (\gamma_l, \gamma_u)$, and let $\alpha_1 = \tilde{\alpha}, \ldots \alpha_H = \tilde{\alpha_h}$ respectively making the location of the changepoints on f^* coincide with $\tilde{f}$. Consider:

$$\begin{aligned} \|f - f^*\| &= \|f - \tilde{f} + \tilde{f} - f^*\| \\ &\leq \|f - \tilde{f}\| + \|\tilde{f} - f^*\|. \end{aligned}$$

As $\|f - \tilde{f}\|$ can be made arbitrarily close to f we investigate $\|\tilde{f} - f^*\|$. Consider:

$$\|\tilde{f} - f^*\| = \underset{x \in \mathcal{X}}{sup} |\tilde{f}(x) - f^*(x)|,$$

let $\beta_0 = \tilde{f}(\gamma_l)$, $G(x) = \frac{\prod_{h=1}^{H}(x - \alpha_h)}{\tau_{k+j}^{H}}$, and rewrite the RHS in a more convenient form using the fundamental theorem of calculus. Here we have

$$\underset{x \in \mathcal{X}}{sup} \left| \int_{-\infty}^{x} \sum_{k=1}^{K} \kappa_k B_{(k,j)}(\xi) - \beta_k G(\xi) B_{(k,j)}(\xi) \, d\xi \right|,$$

where the derivative of $\tilde{f}$ is based upon the derivative formula for B-Splines (see chapter X - property (viii)- in De Boor (2001)). As $B_{(k,j)}(x)$ is non-zero only on $[\tau_k, \tau_{k+j}]$ we have:

$$\leq \sum_{k=1}^{K} \underset{x \in \mathcal{X}}{sup} \left| \int_{\tau_k}^{x} \kappa_k B_{(k,j)}(\xi) - \beta_k G(\xi) B_{(k,j)}(\xi) \, d\xi \right|. \tag{3.4}$$

Because $\tilde{f}$ is constructed to be a taut spline we know that for all k such that $\alpha_k \notin$

$[\tau_k, \tau_{k+j}]$ one has $sgn(\kappa_k) = sgn(G(x))$, where $sgn(\cdot)$ is the signum function. Consequently on each of these intervals let:

$$\beta_k = \frac{\int_{\tau_k}^{\tau_{k+j-1}} \kappa_k B_{(k,j)}(\xi)\ d\xi}{\int_{\tau_k}^{\tau_{k+j}} G(\xi) B_{(k,j)}(\xi)\ d\xi}.$$

This is positive and implies

$$\int_{\tau_k}^{\tau_{k+j}} \kappa_k B_{(k,j)}(\xi) - \beta_k G(\xi) B_{(k,j)}(\xi)\ d\xi = 0.$$

for all k such that $x \geq \tau_{k+j}$. Note that there are at most j intervals such that $\alpha_k \notin [\tau_k, \tau_{k+j}]$ and $x \in [\tau_k, \tau_{k+j}]$, for all $x \in \mathcal{X}$. Given the construction of β_k these are bounded and are dealt with below.

There are $H \times (j-1)$ basis functions defined such that $\alpha \in [\tau_k, \tau_{k+j}]$ whose corresponding β_k has not been given a value. For these it is sufficient to let β_k be zero, as it can be shown that $\kappa_k B_{(k,j)}(\xi)$ goes to zero as Δ is made small.

Note that

$$\int_{-\infty}^{x} \kappa_k B_{(k,j)}(\xi) - \beta_k G(\xi) B_{(k,j)}(\xi)\ d\xi \tag{3.5}$$

is bounded for all k. Let M_{Δ_j} be the largest of the bound for all integrals in 3.4, for a given Δ_j. For any $x \in \mathcal{X}$ there is a finite sum of non zero integrals in (3.4). The maximum number of non-zero integrals is $H \times (j-1) + j$. This implies that (3.4) is less than or equal to

$$\left| M_{\Delta_j} (H \times (j-1) + j) \Delta \right|.$$

This construction defines a valid $f^* \in \mathcal{F}_H^*$. As f is assumed to satisfy the Lipschitz condition, one can show that for successive refinements of Δ_j the bound M_{Δ_j} goes to

zero. As $(H \times (j-1)+j)$ is unchanged as one increases the number of knots, one can make Δ_j sufficiently small with proper choice of $\mathcal{T}^*$. $\square$

The proof offers a construction that shows the LX-spline can well approximate any function in $\mathcal{F}_H$ given enough knots. In practice we have found that defining $\mathcal{T}^*$ over a finite grid of knots (typically 20 to 100) is sufficient to well approximate many continuous functions in $\mathcal{F}_H$. In these cases we have found that increasing the number of knots does not substantially alter the approximations, but does come with the price of increased computational burden.

When there are exactly H changepoints the quantity C guarantees f can be approximated by f^*. For example consider the case where f has two changepoints and is monotonic decreasing, increasing, decreasing in between the changepoints. When $H = 2$ and $C = 1$, f^* can only represent functions having two changepoints where f^* is of the form: monotonic increasing, decreasing, increasing. However, when $C = -1$ such functions can be approximated. Though one can estimate C this is typically unnecessary in practice. In many applications the shape of the function is assumed to follow a specific shape, and C can be set accordingly. When f has fewer changepoints estimating C is not necessary. Continuing the example where $H = 2$, consider the case where f has one changepoints and follows an umbrella shape. Here the estimation procedure can set one of the changepoints to the right end point, for example α_1 is set to γ_u. This removes a changepoint and allows for estimation of f.

3.3 Numerical Examples

In this section we assess the LX-spline in terms of squared error loss when analyzing different curves. As no competing method exists for constraining the number of changepoints, we compare our method to the Gaussian process as well as smoothing P-splines (Eilers and Marx 1996). The GP is fit using the software accompanying Rasmussen

and C. (2006). For all examples we use the squared exponential kernel with unknown length scale and dispersion parameters. For the smoothing spline estimates the penalty coefficient is calculated using GCV. The model fits for the LX-spline are constructed as above with knots chosen at equally spaced intervals, using the same number of knots as observations. All LX-spline confidence limits are constructed using the bootstrap.

In our experiments we simulated data on the interval $\mathcal{X} = [2, 4]$ from a large number of curves having one or two changepoints. In what follows we focus on three functional forms as they are generally illustrative of the performance gains one may expect to see in practice. For each form we varied the parameter η to investigate how the steepness of the true curve affected estimation. For the first curve we simulate data from

$$f_1(x) = -5\exp(-20(x - 2.2)) - \frac{\eta_1}{(x - 4.5)^3},$$

where $\eta_1 \in \{0, 0.1, 0.3\}$. For the second curve we chose a function having a single minimum, but varied the smoothness of the area around the changepoint. We chose this function to be

$$f_2(x) = \eta_2 \left[\frac{1}{(1 - 1.5)^3} - \frac{1}{(1 - 4.5)^3} \right]$$

where $\eta_2 \in \{0.2, 0.5, 1.0\}$. For the final simulation example we chose a curve having two changepoints:

$$\begin{aligned} f_3(x) = &-2.5\exp(-50(x - 2.5)^2) \\ &+ 4.5\exp(-50(x - 3.0)^2) - \eta_3(x - 2.75)^3, \end{aligned}$$

where $\eta_3 \in \{0, 0.5, 1.0\}$. For each simulation condition we vary the number ($n = 20, 30, 40, 60, 80,$ and 100) of sampled points.

Though most conditions showed gains when using the LX-spline, these tended to decrease as the curves became steeper relative to the noise. This is consistent with the idea that the LX-spline removes artifactual bumps from the estimation, but the number and frequency of these bumps decrease as the signal increases. In terms of bias all of the methods were similar exhibiting little to no bias, and, in what follows, we focus on squared error loss.

Figure 3.5 describes the estimated squared error loss under the first η condition for curves f_1, f_2, f_3 top to bottom respectively. Here it is seen that the LX-splines give gains of between $20\% - 45\%$ over the competing methods when there are few observations ($n = 20$). When the number of observations increase these gains decrease, but in most cases we still observe gains. When the curves become steeper the gains do persist but are much less pronounced and typically between 1% to 10%.

Figure 3.5 helps explain this phenomenon. It shows the third simulation condition where $\eta_3 = 1.0$, and compares the LX-spline (red), the smoothing spline (green), and the true curve (black). Note the GP is not shown as it is nearly indistinguishable from the smoothing spline estimate. Here one can see when $f_3(x)$ is changing the most (between 2.2 and 3.3) the two estimates are very similar. However when the signal decreases the uncertainty around the true curve increases, and the artifactual bumps shown in the unconstrained estimate increase. The LX-spline tends to average out these bumps. This is best illustrated between 3.5 and 4 in the same figure. Here the smoothing splines are seen to oscillate. Alternatively, the LX-spline is flat in this region which reflects the fact that the changepoint is not estimated to be in this region.

Figure 3.5 illustrates how this uncertainty in flat regions is reflected in the confidence limits around the curve. Here we interpolate the curve in a very flat region for the first simulation condition of f_1, and compare the LX-spline (red) to the GP (blue)(the GP was chosen as it had narrower confidence intervals as compared to the smoothing spline).

The width of the confidence intervals increases in the region of the interpolation, and they stay approximately the same width for the LX-spline.

3.4 Data Examples

3.4.1 Albany NY Temperature Data

We consider daily high temperature data in Albany NY from January 1 1997 to December 31, 1999. As a convention, we reference the data relative to October 1st in the analysis. We fit the LX-spline with $H = 2$, setting $C = -1$ which assumes the first changepoint is a minimum. Daily temperatures have a high level of autocorrelation that may impact the estimation. For example figure 3.5 shows the estimated curve when smoothing splines (red line) and p-splines (blue line) are fit using the default settings of the packages 'smooth' and 'pspline' in R. Much of this appeared due to the failure of cross-validation or GCV to provide an appropriate smoothing parameter for the data. Gaussian processes did perform better, but still had several areas with large oscillations. Note we did try some methods that attempted to model the autocorrelation (e.g. weighted least squares). These methods, though much smoother, produced unrealistic estimates at the beginning and end of the year, and often underestimated the high temperature peak by approximately 10 degrees.

Figure 3.5 shows the fit of the LX-spline (red), the Gaussian process (blue). Here one can see that the GP estimate produces several artifactual bumps that are unrealistic and caused by the high degree of autocorrelation in the data. The LX-spline's estimate is much more realistic with a much smoother estimate. This is also reflected in the variability of the estimate, which, in some areas, is reduced by as much as 50%.

3.4.2 BMI and Mortality

To illustrate our approach on data that is not continuous we take data from the third National Health and Nutrition Examination Survey (NHANES III). Here we investigate all cause mortality in $n = 8448$ adult males surveyed who had enough data to estimate the body mass index (BMI). Though this study is cross sectional, death records for each individual were obtained and linked to each observation, and time to event data are recorded. In all a total of 2443 deaths occurred after the last linkage (December 2010).

Past studies have found that there is a U-shaped relationship to BMI and mortality, but typically use step functions to estimate this functional relationship. The step functions are specified arbitrarily by grouping BMIs into bins and it is difficult to discern the relationship as the piecewise bins do not allow estimation of a smooth curve. Smoothing based methods have had some difficulty estimating the curve at the extremes. Specifically for our data set (as evident in figure 3.7) BMIs higher than 35 show oscillation in the risk. This is biologically unrealistic. We use LX-splines to study BMI in relation to all cause mortality. The analysis is conducted with offsets for age, smoking status, as well as education.

Figure 3.8 shows the LX-spline estimate of all cause mortality, as well as 95% confidence intervals of that estimate. Here the oscilation from the higher end of the BMI scale has been removed, with the same general pattern of the response estimated in figure 3.8. This estimate is smooth for most of the BMI range, but there are two flat regions at the high end of the BMI scale. These are artifacts of the LX-spline procedure removing the oscillations seen in figure 3.7.

3.5 Conclusion

We have shown, in simulated and real data, that the LX-spline provides noticeable advantages over existing smoothing approaches. The gains were both quantifiable (i.e., in terms of mean squared error), as well as qualitative (i.e., the plausibility of the estimated curve). By specifying a maximum number of changepoints in the model the LX-spline reduces the uncertainty in the estimated curve limiting changepoints to the most plausible regions of the data.

For multivariate function estimation, the LX-spline can be used for additive structures. However, the LX-spline does not generalize in the case of tensor products. We are currently investigating other ways in which variants of the LX-spline can be used in higher dimensional data.

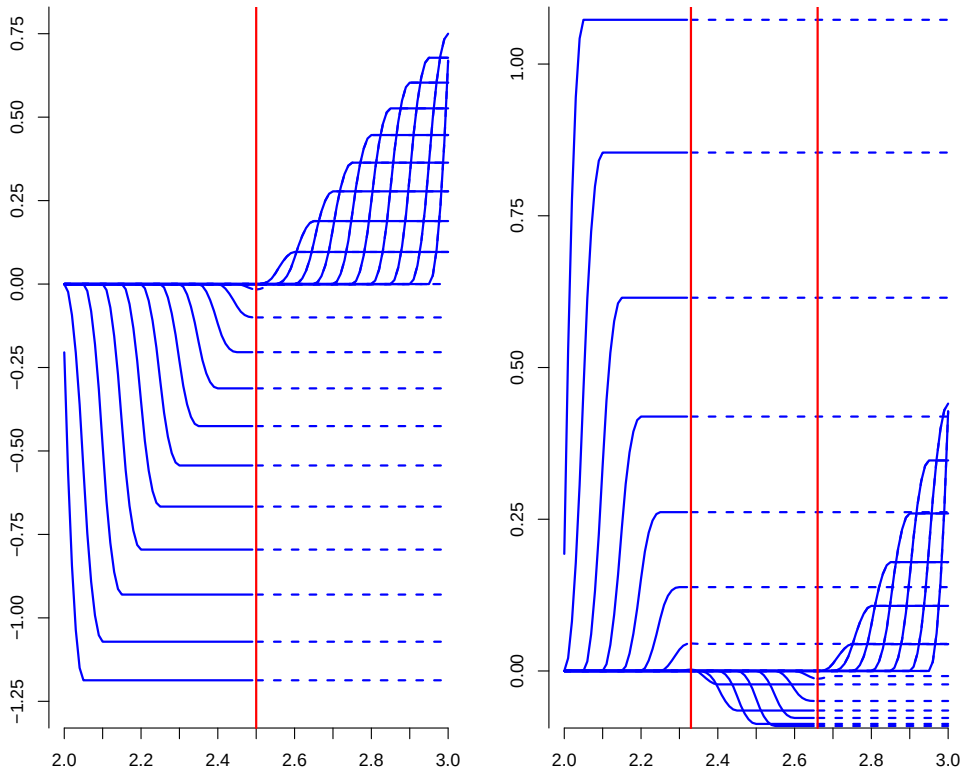

Figure 3.1: LX splines with a single changepoint at 2.5 (left), and LX splines with two changepoints at 2.33 and 2.66 (right).

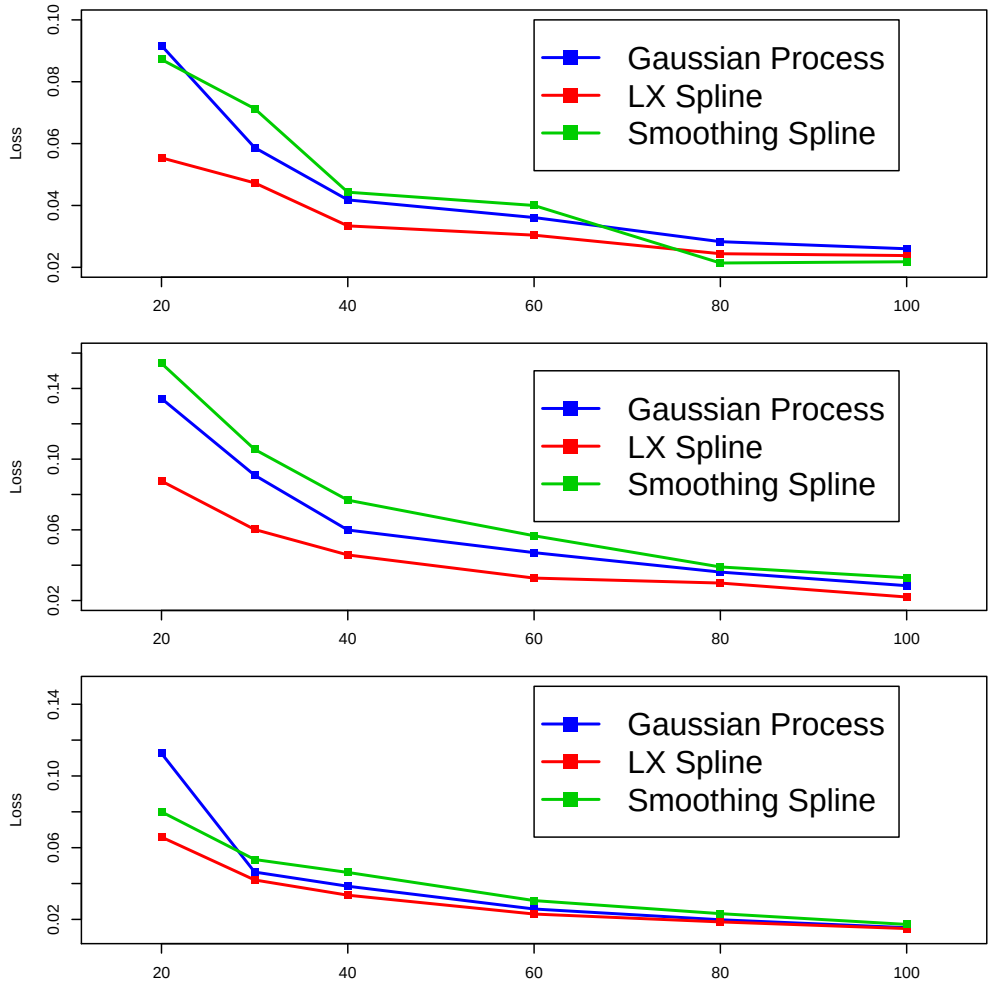

Figure 3.2: Estimated squared error loss between the GP and the LX-spline from simulation condition 1, for the first condition of all three shapes investigated.

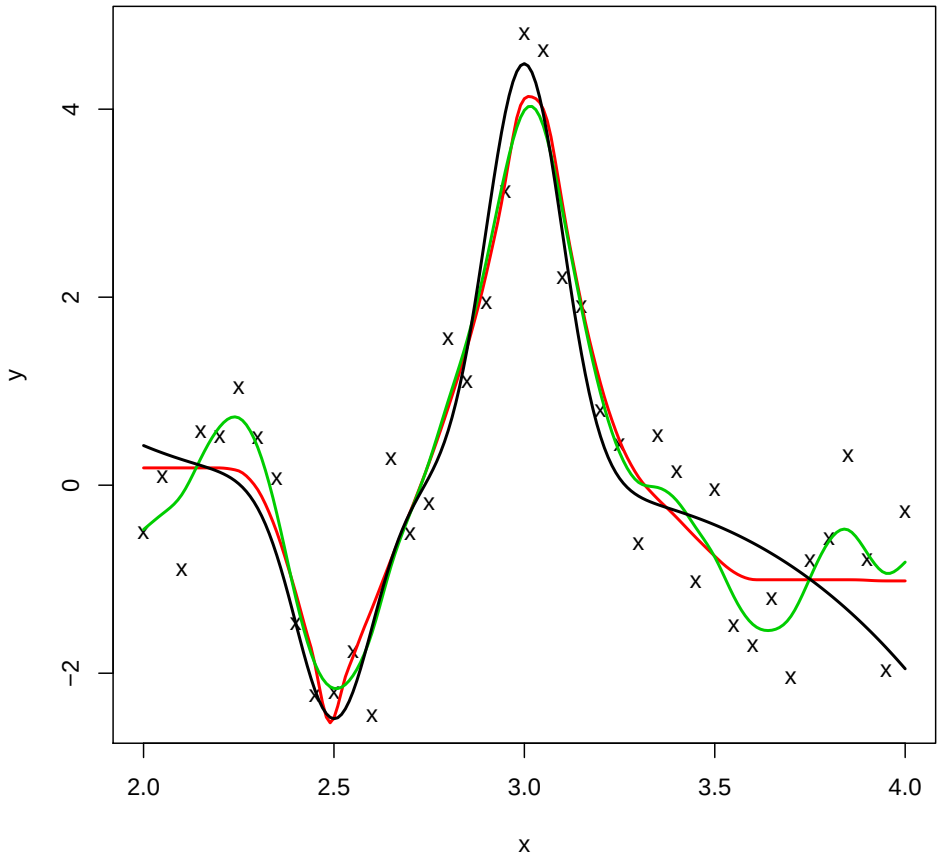

Figure 3.3: Estimation of $f_3(x)$ (black) using the LX-spline (red) and the smoothing spline (green).

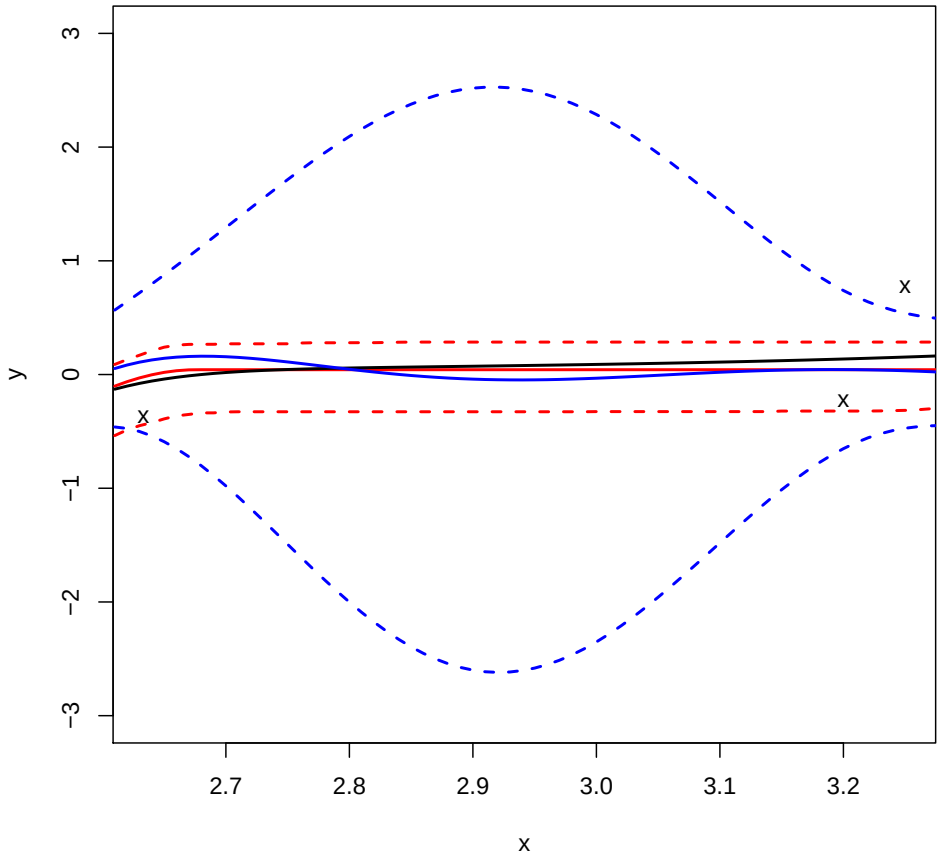

Figure 3.4: Estimated curve (solid red line) and 95% confidence intervals (dashed red line) for the LX-splines, and the Gaussian process (blue solid and dashed lines respectively) when estimating the true curve (black).

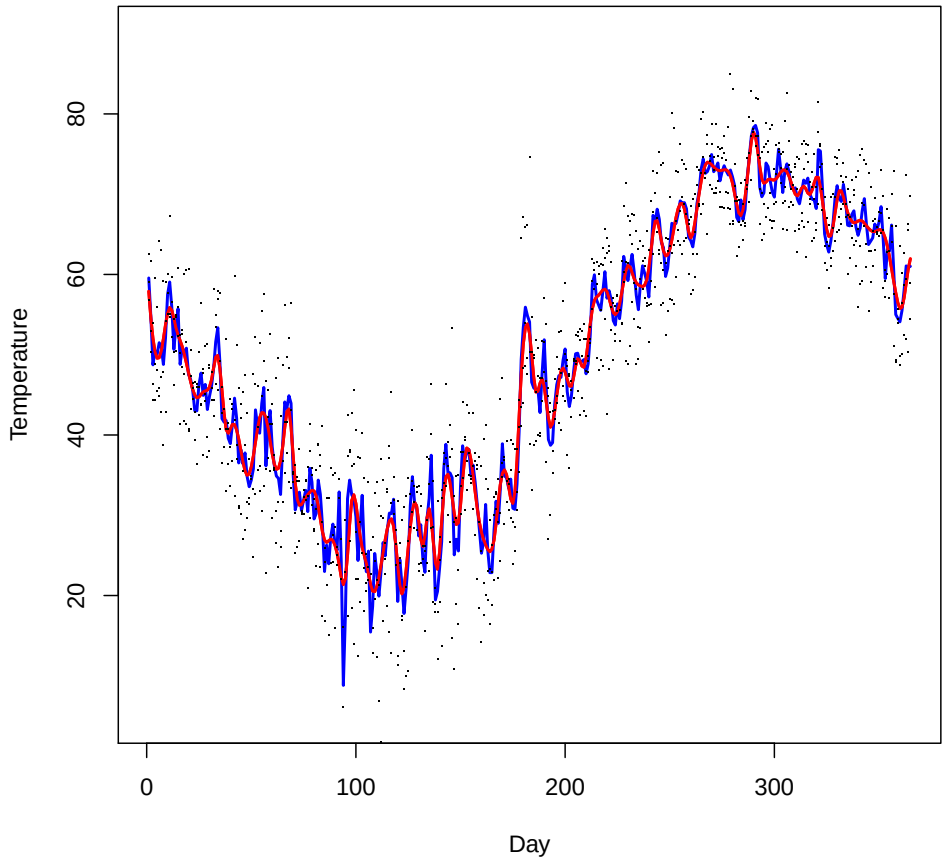

Figure 3.5: Fit of the Albany, NY temperature data when using smoothing splines (red) and P-splines (blue).

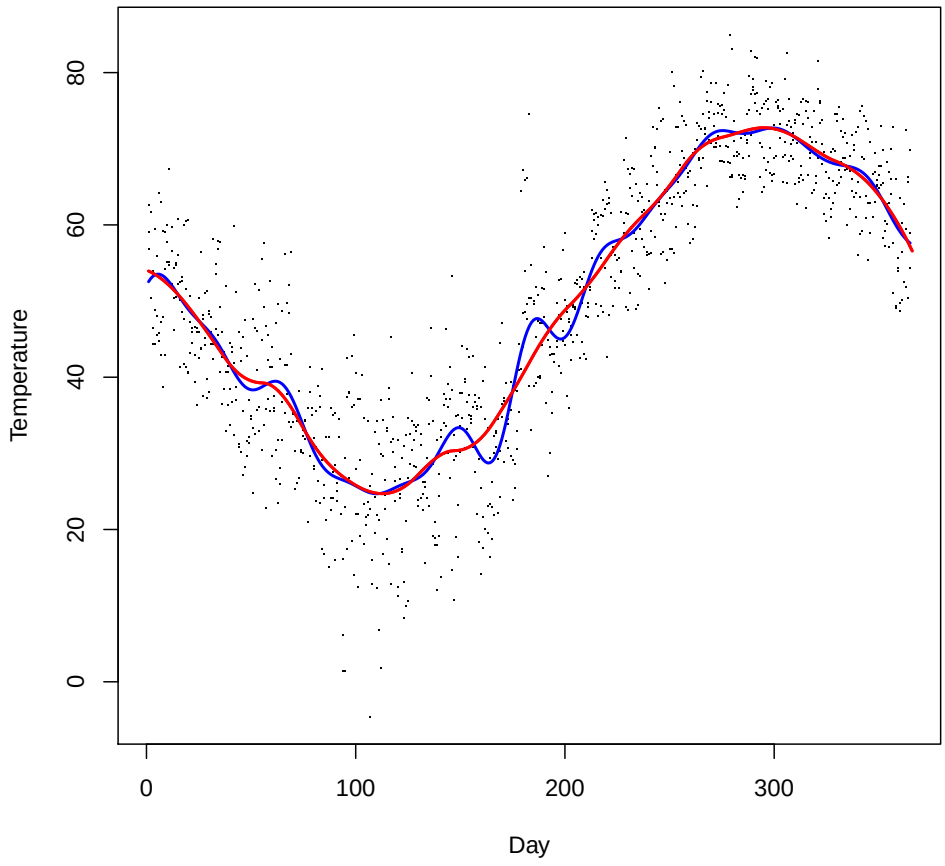

Figure 3.6: Fit of the LX spline (red) as compared to the Gaussian process (blue) based upon 3 years of daily high temperature data collected in Albany, NY.

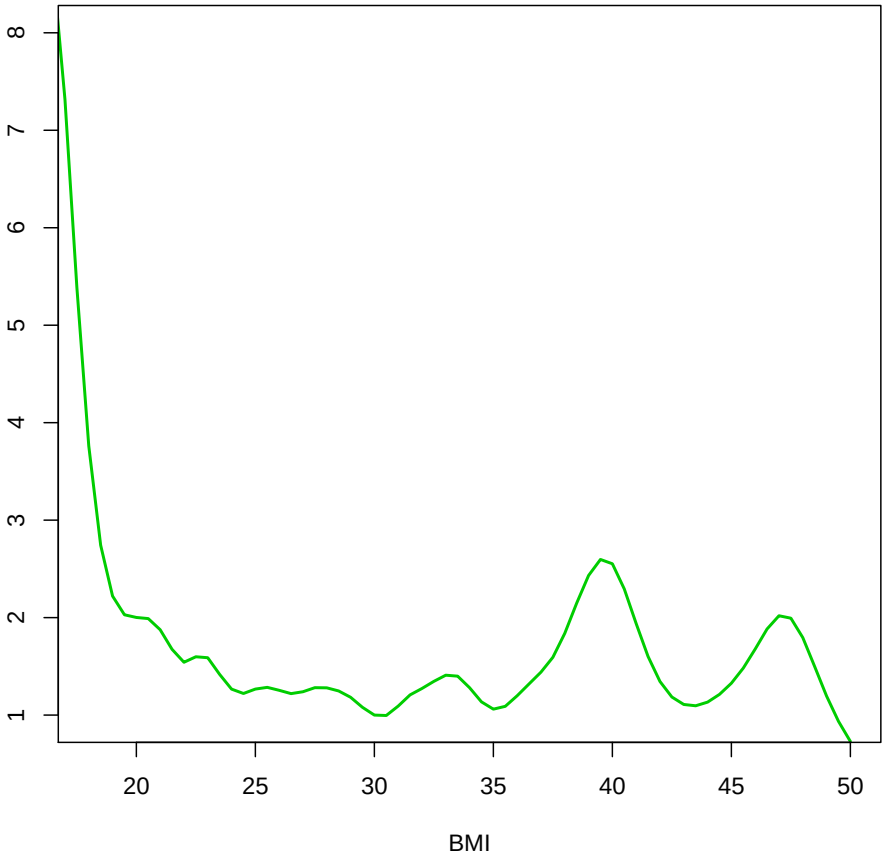

Figure 3.7: Relative risk of all cause mortality estimated using a spline based approach.

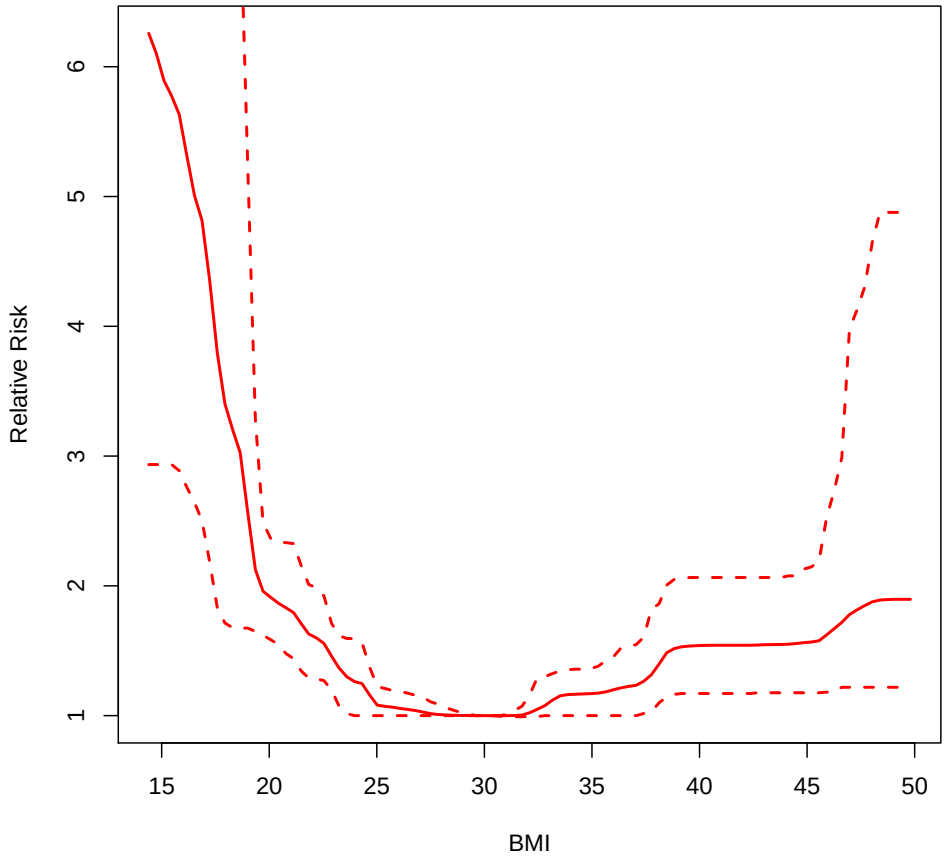

Figure 3.8: Estimated relative risk of all cause mortality, and corresponding 95% confidence intervals for different BMIs calculated using the LX spline only. Here risk is relative to the BMI associated with the minimum risk (BMI = 30.03).

Chapter 4

Bayesian Local Extrema Splines

4.1 Introduction

In many applications there is interest in modeling an unknown function $f : \mathcal{X} \to \mathbb{R}$, where it is reasonable to assume a maximum number of local extrema (changepoints in $\mathcal{X}$). Approaches to modeling f based upon splines or Gaussian processes specify priors over a large class of curves, many of which may be unrealistic for a given application. As a result, their use may produce estimates having artifactual bumps. For a given problem such estimates may be unrealistic and lead to a loss of efficiency when modeling f.

For example, in toxicology studies investigating dose-response relationships data may be hypothesized to have a J-shape. Here at low levels of exposure a chemical may have a therapeutic (often termed hormetic) effect but at higher levels of exposure the chemical produces an adverse response. More commonly an umbrella ordering is seen. Here there is a noticeable decrease in the adverse response of interest at higher doses due to acute toxicity and death. Such J-shapes and umbrella shapes are also common in the social sciences, as well as financial or engineering applications.

Applications with more than one changepoint are also prevalent in the literature. Seasonal fluctuations may lead to curves having a single nadir and single maximum

(something which is common when accounting for seasonal effects in time series data). These curves may not be well approximated by perfectly periodic basis functions. In both examples, putting strong prior knowledge on the form of the function may lead to gains in efficiency when modeling the curve. We develop the local extrema (LX) spline to put a prior over curves having at most H change points.

Further one may also wish to test if the function has a specified shape as compared to an alternative. For example a recent manuscript by Myrskylä et al. (2009) suggested that the well know decline in the fertility rate related to advances in human development (measured by the the human development index (HDI)), can be seen to reverse in the the most economically developed countries. This conclusion was subject to controversy, and was based on a smoothed regression fit that suggested a J shaped curve. Testing of the J-shaped hypothesis over a monotone decreasing alternative would allow one to formally evaluate the strength of the author's claims. We show how the LX-spline can be used to construct tests.

Conceptually, when modeling a unknown curve with a finite number of local extrema, one needs to place a flexible prior over monotone segments having unknown change points (local extrema). From a Bayesian perspective there is a rich literature on restricting functional forms to be strictly monotone (Neelon and Dunson 2004; Bornkamp and Ickstadt 2009; Shively et al. 2009). One approach (Neelon and Dunson 2004) develops a prior over piecewise linear splines for isotonic regression. This approach, given enough splines, was shown to well approximate smooth monotone functions. More recently Shively et al. (2009) developed an approach using free knot splines. Here the authors place a novel prior over the spline coefficients to enforce monotonicity.

In adding one change point Gunn and Dunson (2005) extended Neelon and Dunson (2004) to model J-shaped curves in a hierarchical setting. This approach was not fully

Bayesian as it lacked an explicitly specified prior distribution. Alternatively, Hans and Dunson (2005), developed an explicit prior over umbrella shaped orderings. Most recently Shively et al. (2011) developed methods for fixed and free knot splines that model continuous monotone segments with a single unknown changepoint. They did not extend this approach to consider multiple change points, and their methodology was not used to test the shape of the curve.

This article proposes a fundamentally different approach to the shape constrained regression problem. Instead of only looking at the coefficients on the splines and placing an appropriate prior over these coefficients (Neelon and Dunson 2004; Shively et al. 2009; 2011), we develop a novel spline construction. This is similar in spirit to the I-spline construction of Ramsay (1988) or the C-spline construction for convex splines (Meyer 2008). Our construction, when paired with positivity constraints on the spline coefficients, enforces prior shape restrictions on the curve of interest by limiting the number of changepoints. Further one can test the location of these parameters to test the shape of the response against a known alternative. In our approach we allow the number of knots to be large, and use a smoothing approach similar to Bayesian P-splines (Lang and Brezger 2004) for the spline coefficients. The resulting model provides a flexible unified framework for Bayesian inference where one can place strong prior knowledge on the maximum number of changepoints in $\mathcal{X}$.

The outline of the manuscript is as follows: section 2 proposes the model and prior structure. Section 3 proposes the MCMC sampling algorithm. Section 4 investigates the models performance through a series of numerical experiments. Section 5 applies the model to a variety of domains , and Section 6 discuses the results.

4.2 Model

4.2.1 Spline Construction

Let $Y = (y_1, y_2, \ldots, y_n)'$ be a vector of error prone observations of an unknown function $f : \mathcal{X} \to \mathbb{R}$, where $\mathcal{X} = [\gamma_l, \gamma_u] \subset \mathbb{R}$. Here we observe $f(x)$ at $\{x_i\}_{i=1}^n \in \mathcal{X}$. Assume that

$$y_i = f(x_i) + \epsilon_i, \tag{4.1}$$

with $\epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$. We develop a spline based approximation of $f(x)$ that limits the curve to at most H local extrema on $\mathcal{X}$.

Consider the approximation

$$f(x) \approx \sum_{k=0}^{K} \beta_k B_k^*(x), \tag{4.2}$$

where β_k are basis coefficients, $B_0^*(x) = 1$, and $B_k^*(x)$, for $k \geq 1$, are basis functions defined on some knot set $\mathcal{T}$. Given a well chosen basis, (4.2) can approximate essentially any continuous function. To limit the forms that $f(x)$ may have, we restrict this function to have a most H local extrema by defining $B_k^*(x)$ to be

$$B_k^*(x) = M \int_{-\infty}^{x} \frac{\prod_{h=1}^{H}(\xi - \alpha_h)}{\tau_{k+j}^H} B_{(j,k)}(\xi) d\xi, \tag{4.3}$$

where $B_{(j,k)}(x)$ is a B-spline basis function of order j (De Boor 2001) that is constructed using the knot set $\mathcal{T} = \{\tau_k\}_{k=1}^{K+j}$, $\tau_1 \leq \tau_2 \leq \ldots \leq \tau_K \leq \ldots \leq \tau_{K+j}$, and defined on $[\tau_k, \tau_{k+j}]$. Further in (4.3) the quantity M is a constant. As B-Splines are defined to be greater than or equal to zero, if all β_k are constrained to be greater than zero the derivative of (4.2) can only be zero when $x = \alpha_h$ and there can be at most H change

points in the approximation. These change points exist in $\mathcal{X}$ only if $\alpha_h \in (\gamma_l, \gamma_u)$. Consequently there can be fewer than H extrema if α_h is outside of (γ_l, γ_u). In what follows we specify flexible priors for β_k that enforce the constraints $\beta_k \geq 0$ for $k \geq 1$ and $\alpha_h \in \mathcal{X}$ for all h.

4.2.2 Prior Specification

We follow a Bayesian approach to allow: (1) curves to have at most H change points in $\mathcal{X}$, and (2) monotone regions having the possibility of flat segments. In this discussion we assume that K is large, and knots are located at data points, or are chosen to be on a tightly spaced grid. In defining an appropriate prior over $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_K)'$ we define the latent variables $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_K^*),'$ such that $\beta_k = 1_{(\beta_k^* >= \delta)}\beta_k^*$, for $k \geq 1$, with $\delta \geq 0$. We specify the prior over $\boldsymbol{\beta}^*$

$$\begin{aligned}\pi(\boldsymbol{\beta}^*) &= \pi(\beta_1^*) \prod_{k=2}^{K} \pi(\beta_k^*; \beta_{k-1}^*, \lambda_k^{-1}) \\ &= N(\beta_1^*; A_1, B_1) \prod_{k=2}^{K} N(\beta_k^*; \beta_{k-1}^*, \lambda_k^{-1}),\end{aligned}$$

where A_1, and B_1 are constants, $N(\cdot)$ is the normal distribution, and $\lambda_k \sim Ga(\frac{r}{2}, \frac{r}{2})$, for $r \geq 1$. This prior defines a random walk of T-distributed variates having r degrees of freedom.

We complete the prior specification on $\boldsymbol{\beta}$, and $\boldsymbol{\beta}^*$ through

$$\pi(\boldsymbol{\beta}, \boldsymbol{\beta}^*) = N(\beta_0; A_0, B_0)\pi(\boldsymbol{\beta}^*) \prod_{k=1}^{K} \left\{ 1_{(\beta_k = 0)} 1_{(\beta_k^* < \delta)} + 1_{(\beta_k = \beta_k^*)} 1_{(\beta_k^* \geq \delta)} \right\},$$

where $\delta > 0$. Note that β_0 is the intercept parameter and is not constrained. For a given β_k, β_{k-1}^* and δ one can marginalize out β_k^* arriving at the following mixture

distribution for $k \geq 1$:

$$\pi(\beta_k|\beta^*_{k-1}) = \Phi\left([\delta - \beta^*_{k-1}]\sqrt{\lambda_k}\right) 1_{(\beta_k=0)} + N(\beta_k; \beta^*_{k-1}, \lambda_k^{-1}) 1_{(\beta_k>\delta)}, \tag{4.4}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution, with $\beta^*_0 = A_1$ and $\lambda_0 = B_1$. This is a point-mass mixture prior for β_k, and allows the coefficient to be exactly zero or positive. The quantity δ controls the probability $\beta_k = 0$. To define a prior over $Pr(\beta_k = 0)$, we let $\delta \sim Ga(c_1, d_1)$.

Without $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_H)'$ the prior in (4.4) provides an inferential framework similar to Neelon and Dunson (2004) and Nakajima and West (2012). We allow for at most H changepoints by defining the following prior on $\boldsymbol{\alpha}$:

$$\pi(\boldsymbol{\alpha}) \propto \prod_{h=1}^{H} \left[\pi_l 1_{(\alpha_h=\gamma_l)} + \pi_u 1_{(\alpha_h=\gamma_u)} + (1 - \pi_l - \pi_u) N(\alpha_h; C_h, D_h) 1_{(\gamma_l \leq \alpha_h \leq \gamma_u)}\right],$$

which defines a mixture over γ_l, γ_u or on the interval (γ_l, γ_u). If $\alpha_h = \gamma_l$ or $\alpha_h = \gamma_u$ this is equivalent to removing a local extrema from the model. When considering J-shaped curves $H = 1$, if $M > 1$ and $\alpha_1 = \gamma_l$, the curve is monotonic increasing, and if $\alpha_1 = \gamma_u$ it is monotonic decreasing. When $\alpha_1 \in [\gamma_l, \gamma_u]$ this quantity is directly interpretable as a minimum and $Pr(\alpha_l \notin [\gamma_l, \gamma_u])$ is related to the hypothesis on the existence of the J-shape.

4.2.3 Spline Construction

When constructing the LX-spline we consider the B-spline order. Though higher order B-spline constructions may offer smoother approximations, by choosing many lower order splines with an appropriate penalty for smoothness and sparsity, one can eliminate the need for higher order splines. We use B-splines of order $j = 2$; these splines are proportional to triangle distributions defined on $[\tau_k, \tau_{k+2}]$ having mode τ_{k+1}.

In practice we have observed excellent performance when a large number of splines are considered.

We carefully construct the spline to achieve desirable smoothing properties. The height of a spline is dependent on $\boldsymbol{\alpha}$. When there are large differences between the height of individual splines, the prior over smooths certain regions and under smooths others. The term τ_{k+2}^H in $\frac{\prod_{h=1}^H(\xi-\alpha_h)}{\tau_{k+2}^H}$ makes the absolute height of each spline roughly on the same scale, while keeping a polynomial representation for $B_k^*(x)$. Without this term the height of the largest and smallest spline can differ by an order of magnitude, and problems with smoothing may occur. Note that the τ_{k+2}^H term produces splines that may be undefined if $0 \in \mathcal{X}$. This problem is easily averted by mapping the domain to a closed set that does not contain zero and construct all splines on this set.

Figure 4.1 shows the spline construction when $H = 1$ (left) and $H = 2$ (right) on the interval $\mathcal{X} = [2, 3]$. For $H = 1$ we set $\alpha_1 = 2.5$, and for $H = 2$ we let $\alpha_1 = 2.33$ and $\alpha_2 = 2.66$. Here one can see how $\boldsymbol{\alpha}$ combined with the positivity constraint on $\boldsymbol{\beta}$ controls the shape of the curve. For the right panel the curve shapes can be: monotone increasing ($\alpha_1 = 2$), J-shaped ($\alpha_1 \in (2, 3)$), and monotone decreasing ($\alpha_1 = 3$). Note if $M < 0$ the patterns invert, and an umbrella shaped ordering is considered. When $H > 1$ more complicated shapes can be constructed. In each case if $\alpha_h = \gamma_l$ then $(x - \alpha_h)$ is positive on $\mathcal{X}$, if $\alpha_h = \gamma_u$ the quantity is negative in this range, and if $\alpha_h \in (\gamma_l, \gamma_u)$ a changepoint exists in $\mathcal{X}$.

4.2.4 Inference on the change point parameters

As discussed above the change point parameters determine the shape of the curve. When performing a Bayesian test on the shape of the curve one needs to monitor $\boldsymbol{\alpha}$.

For example, in testing between

$$H_0 : f(x) \text{ is monotone increasing}$$

$$H_1 : f(x) \text{ is J shaped,}$$

if $\alpha_1 = \gamma_l$ then the the change point is removed from the model, and the resulting model is monotonic increasing. Though this is a sufficient condition to remove the change point from $\mathcal{X}$ it is not necessary. There is positive probability $\sum_{k:\tau_k \leq \alpha_1} \beta_k = 0$ and this implies no changepoint in the interval. Continuing the J-shaped example above the MCMC samples can be monitored for the above two conditions, and the posterior probability of a monotone shape over a J-shape can be assessed.

Hypothesis testing when $H \geq 2$ is similar to the case when $H = 1$. Note that the labels on $\boldsymbol{\alpha}$ are not identifiable and label switching may occur. For the LX-splines this is not an issue as the sorted $\alpha' s$ maintain the same interpretation regardless of the label switching. For example when $H = 3$, and $M > 0$ the smallest α is always interpretable as the minimum. Consequently all one needs to do is sort the posterior sample each iteration to guarantee the correct interpretation.

4.2.5 Extensions to Multiple Predictors

The LX-spline can be adapted to models where there are multiple predictors using an additive structure. In particular let $(t_1, \ldots, t_S)'$ be vector of predictors with corresponding covariates $(\theta_1, \ldots, \theta_S)$, model (4.1) becomes:

$$y_i = \sum_{s=1}^{S} \theta_s t_s + \sum_{k=0}^{K} \beta_k B_k^*(x_i) + \epsilon_i.$$

Some of these parameters may be unconstrained or may represent another LX-spline construction. Posterior sampling these parameters is discussed below and requires

minimal modification from the LX spline formulation.

4.2.6 Extensions to Dichotomous Outcomes

The above model can be adapted to categorical responses by following the approach of Albert and Chib (1993). Here one observes $Z = (z_1, \ldots, z_n)$ be independently distributed Bernoulli random variables with

$$Pr(z_i = 1|\boldsymbol{\beta}, \boldsymbol{\beta}^*, \boldsymbol{\alpha}, \delta, y_i) = \Phi(X_i(\boldsymbol{\alpha})\boldsymbol{\beta}),$$

here $X_i(\boldsymbol{\alpha})$ is row i of $X(\boldsymbol{\alpha})$ a $n \times k$ design matrix that is a function of $\boldsymbol{\alpha}$ (described below), and y_i is an random variable such that $y_i \sim N(X_i(\boldsymbol{\alpha})\boldsymbol{\beta}, 1)$. This model is equivalent to assuming $z_i = 1_{(y_i>0)}$. Posterior computation proceeds by alternating between two steps. In the first step one samples from the algorithm below where τ is set to 1; in the second step one samples each y_i from its conditional density given $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and z_i. Here $y_i \sim \pi(y_i; \boldsymbol{\beta}, \boldsymbol{\alpha}, z_i) \stackrel{d}{=} N(X_i(\boldsymbol{\alpha})\boldsymbol{\beta}, 1)$ truncated above by zero if $z_i = 0$ and truncated below by 0 if $z_i = 1$.

4.3 Posterior Computation

Posterior computation proceeds through a series of conditionally conjugate Gibbs sampling steps. Here the joint posterior density is proportional to

$$L(y|\boldsymbol{\beta}, \tau, \boldsymbol{\alpha}, X(\boldsymbol{\alpha}))\pi(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau),$$

where $\pi(\boldsymbol{\beta}, \boldsymbol{\beta}^*, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \tau) = \pi(\boldsymbol{\beta}, \boldsymbol{\beta}^* | \boldsymbol{\lambda}) \pi(\boldsymbol{\lambda}) \pi(\boldsymbol{\alpha}) \pi(\tau)$, and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_k)'$. In outlining the sampling algorithm we decompose $X(\boldsymbol{\alpha})$ as

$$X(\boldsymbol{\alpha}) = z_r(\boldsymbol{\alpha}) X^r + \ldots z_0(\boldsymbol{\alpha}) X^0.$$

Here X^r, $0 \leq r \leq H$, is a $n \times k$ matrix where each element $X^r_{(i,k)}$ from row i and column k of the matrix X^r is computed using

$$X^r_{(i,k)} = \int_{\tau_k}^{x_i} \frac{\xi^r}{\tau_{k+1}^H} B_{(k,j)}(\xi) \, d\xi.$$

Further $z_r(\boldsymbol{\alpha})$ is a function of $\boldsymbol{\alpha}$ corresponding to the coefficient of x^r in the polynomial $\prod_{i=1}^{H}(x - \alpha_h)$. For example when $H = 2$ one has $z_0(\boldsymbol{\alpha}) = \alpha_2 \alpha_1$, $z_1(\boldsymbol{\alpha}) = -(\alpha_1 + \alpha_2)$ and $z_2(\boldsymbol{\alpha}) = 1$. Note $X(\boldsymbol{\alpha})$ is used when sampling $\boldsymbol{\beta}$ and $\{X^r\}_{r=0}^H$ is used when sampling $\boldsymbol{\alpha}$. Finally $TN(\mu, \sigma^2, a, b)$ specifies truncated normal distribution having parameters μ and σ^2 that is truncated below at a and above at b.

Sampling Algorithm

1. For $1 \leq k \leq K$ when sampling (β_k^*, β_k) let $Y^* = Y - X(\boldsymbol{\alpha})_{-k} \boldsymbol{\beta}_{-k}$, where $\boldsymbol{\beta}_{-k}$ is $\boldsymbol{\beta}$ without entry k, and $X(\boldsymbol{\alpha})_{-k}$ is the design matrix without column k. Letting $w = X(\boldsymbol{\alpha})_k$, a $n \times 1$ column vector representing column k in $X(\boldsymbol{\alpha})$. One can show that

$$\begin{aligned} \pi(\beta_k^*, \beta_k) \propto & 1_{(\beta_k = 0)} 1_{(\beta_k^* < \delta)} \left[\frac{N(\beta_k^*, E_{(k,0)}, V_{(k,0)})}{N(0, E_{(k,0)}, V_{(k,0)})} \right] \\ & + 1_{(\beta_k = \beta^*)} 1_{(\beta_{k^*} \geq \delta)} \left[\frac{N(\beta_k^*, \hat{E}_k, \hat{V}_k)}{N(0, \hat{E}_k, \hat{V}_k)} \right], \end{aligned}$$

where $V_{(k,0)} = \left[\lambda_{k-1} + \lambda_{k+1}\right]^{-1}$, $E_{(k,0)} = V_{(k,0)} \left[\lambda_{k-1} \beta_{k-1}^* + \lambda_{k+1} \beta_{k+1}^*\right]$, $\hat{V}_k = \left[\tau(w'w) + \lambda_{k-1} + \lambda_{k+1}\right]^{-1}$, and $\hat{E}_k = \hat{V}_k \left[\tau w' Y^* + \lambda_{k-1} + \lambda_{k+1}\right]^{-1}$.

This expression has the following normalizing constant

$$C = \frac{F(\delta; E_{(k,0)}, V_{(k,0)})}{N(0; E_{(k,0)}, V_{(k,0)})} + \frac{1 - F(\delta; \hat{E}_k, \hat{V}_k)}{N(0; \hat{E}_k, \hat{V}_k)} = A + B.$$

Where $F(\theta; \mu, \sigma^2)$ is the CDF of a normal distribution with mean μ and variance σ^2 evaluated at θ. Here one samples $\beta_k = 0$, and $\beta_k^* \sim TN(E_{(k,0)}, V_{(k,0)}, -\infty, \delta)$ with probability A/C, and samples $\beta_k = \beta_k^*$ and $\beta_k^* \sim TN(\hat{E}_k, \hat{V}_k, \delta, \infty)$ with probability B/C. Note that $\lambda_{K+1} = 0$ and $\beta_{K+1}^* = 0$, when sampling from (β_K, β_K^*).

2. For the intercept, where $k = 0$, let $Y^* = Y - X(\boldsymbol{\alpha})_{-0}\boldsymbol{\beta}_{-0}$ and sample $\beta_0 \sim N(E, V)$ where $V = (\tau n + B_0^{-1})^{-1}$ and $E = V(\tau Y^* + B_0^{-1} A_0)$.

3. For each α_h in $\boldsymbol{\alpha}$ define $Y^* = Y - \left\{\sum_{r=0}^{H} [z_r^-(\boldsymbol{\alpha}, \alpha_h) X^r]\right\} \boldsymbol{\beta}$. Here $z_r^-(\boldsymbol{\alpha}, \alpha_h)$ is a function representing the terms in $z_r(\boldsymbol{\alpha})$ that do not involve α_h. For example when $H = 2$ then $z_2(\boldsymbol{\alpha}) = 1$ and $z_2^-(\boldsymbol{\alpha}, \alpha_1) = 1$, $z_1(\boldsymbol{\alpha}) = -(\alpha_1 + \alpha_2)$ and $z_1^-(\boldsymbol{\alpha}, \alpha_1) = -\alpha_2$, and $z_0^-(\boldsymbol{\alpha}, \alpha_1) = 0$ as all terms in $z_0(\boldsymbol{\alpha})$ contain α_1. Let $w = \left\{\sum_{r=0}^{H} [z_r^*(\boldsymbol{\alpha}, \alpha_h) X^r]\right\} \boldsymbol{\beta}$, where $z_r^*(\boldsymbol{\alpha}, \alpha_h)$ is a function that contains only the terms in $z_r(\boldsymbol{\alpha})$ having α_h factored out. Again when $H = 2$ for α_1 one has $z_0^*(\boldsymbol{\alpha}, \alpha_1) = \alpha_2$, $z_1^*(\boldsymbol{\alpha}, \alpha_1) = -1$, and $z_2^*(\boldsymbol{\alpha}, \alpha_1) = 0$ as no term in $z_2^*(\boldsymbol{\alpha})$ contains α_1. Given these quantities sample α_h from a distribution proportional to

$$\pi_l N(Y^*; \gamma_l w, \tau^{-1}) 1_{(\alpha_h = \gamma_l)} + \pi_u N(Y^*; \gamma_u w, \tau^{-1}) 1_{(\alpha_h = \gamma_u)} +$$
$$(1 - \pi_l - \pi_u) \frac{\sqrt{D_h^{-1}} \left[N(\alpha_h; E, V)/N(0; E, V)\right] 1_{(\gamma_l < \alpha_h < \gamma_u)}}{\Phi\left[\sqrt{D_h^{-1}}(\gamma_u - C_h)\right] - \Phi\left[(\sqrt{D_h^{-1}}(\gamma_l - C_h)\right]},$$

where $V = \left[\tau(w'w) + D_h^{-1}\right]^{-1}$ and $E = V\left[\tau w' Y^* + D_h^{-1} C_h\right]$, and $\Phi[\cdot]$ is the CDF of the standard normal distribution.

4. Sample τ from a $Ga(A, B)$ where $A = \frac{n}{2} + A_0$ and $B = \frac{[Y - X(\boldsymbol{\alpha})\boldsymbol{\beta}]'[Y - X(\boldsymbol{\alpha})\boldsymbol{\beta}]}{2} + B_0$.

5. For $k \geq 2$ sample λ_k from a $Ga(A, B)$ where $A = \frac{1}{2} + \frac{r}{2}$ and $B = \frac{(\beta_k - \beta_{k-1})^2}{2} + \frac{r}{2}$.

6. Sample δ from a distribution proportional to $Ga(c_1, d_1)1_{LB \leq \delta \leq UB}$ where $LB = max(0, \{\beta_k^* : \beta_k = 0\}_{k=0}^K)$ and $UB = min(\{\beta_k^* : \beta_k = \beta_k^*\}_{k=0}^K)$.

In estimating the underlying curve we have observed adequate mixing with the above algorithm. For $20,000$ samples the median effective sample size across 101 equally spaced points along $f(x)$ was $1,875$, and 80% of these points had effective sample sizes greater than 500. We noticed a decrease in the effective sample size at the beginning of the curve. Here there is a large amount of correlation with β_0 and this increases autocorrelation in the estimate of $f(x)$. If one is performing inference on the shape of the curve inference on $\boldsymbol{\alpha}$ must be performed. In our experience $\boldsymbol{\alpha}$ can mix much slower with approximately $50,000$ samples frequently being adequate. When $H = 1$ we have observed effective sample sizes of approximately 500 per $50,000$ samples. For both curve estimation and shape testing the mixing greatly depended on the choice of the prior for δ, and the number of knots. Mixing became markedly worse when δ was made diffuse, or when there were a large number of knots relative to the number of observations. In these cases the algorithm is relatively computationally inexpensive. Our testing was done using unoptimized MATLAB code. Here, for most of our simulation examples, it took approximately 13 seconds per $1,000$ samples on a $2.6ghz$ Intel processor.

Modifications to this algorithm to add multiple predictors are straightforward. Here one adds a step to sample from the distribution of these predictors given the other parameters. For example if a normal conjugate prior is used for $\boldsymbol{\theta}$, a block Gibbs step can be used.

4.4 Numerical Experiments

In evaluating the performance of this approach we conduct a series of simulation studies designed to investigate certain aspects of the model. We compare the method with the P-spline approach (Lang and Brezger 2004) in curve estimation. We also conduct a small simulation study that investigates sample size for hypothesis testing on the existence of a J-shaped curve. In these studies we estimate all curves on the interval $\mathcal{X} = [2, 3]$. The prior was specified letting $\beta_0 \sim N(0, 100)$, $\beta_1^* \sim N(0, 100)$, and by specifying $r = 5$ for $\lambda_k \sim Ga(\frac{r}{2}, \frac{r}{2})$, $k \geq 2$. Also we let $\delta \sim Ga(1, 20)$. For $H \geq 1$ we let $\alpha_1 \sim N(2, 10)$, and $\alpha_2 \sim N(3, 10)$. Further we let $\pi_l = \pi_u = 1/3$. Finally we specify $\tau^{-1} \sim Ga(0.01, 0.01)$. Though we chose this specification based the data examples the priors on $\delta, \boldsymbol{\alpha}$, and $\boldsymbol{\lambda}$ were varied and the results were robust to prior specification. When sampling from the posterior note that when inference on the curve was of primary interest $25,000$ MCMC draws were taken from the posterior distribution with the first $5,000$ discarded as burn in samples. When the location/hypothesis testing of the change point was of interest $55,000$ MCMC draws were taken with the first $5,000$ disregarded as burn in. Trace plots showed convergence occurred quickly.

4.4.1 Curve estimation

We compare our method to the Bayesian P-spline approach of Lang and Brezger (2004) using a number of simulated data sets having one or more change points. A number of examples were considered. These examples included: cases where the true number of change points were varied (0, 1, or 2 changepoints), H was set at or above the true number of changepoints, the derivative of the curves was varied, and the sampling variance was modified. In relation to the P-spline these results showed removal of artifactual bumps, narrower credible regions, and little to no bias in estimating the true underlying curve. The following examples are indicative of the increases in efficiency

seen when using the LX-spline.

In the first example we set the true curve to be $f(x) = 2.5exp(-100(x-2.2)^2) - 4.5exp(-100(x-2.85)^2)$. For the simulation 101 observations were taken at equally spaced intervals on $\mathcal{X}$. We show results when $\tau^{-1} = 2$ (which we consider a high variance condition) and $\tau^{-1} = 0.1$ (which we consider a low variance condition). Figure 4.2 shows the fit of our model (black line) with corresponding 95% credible intervals (black dotted lines), as well as the P-spline approach (gray line), for the low variability condition (top plot) and high variability conditions (bottom plot). Both methods performed similarly in curve estimation. However, as expected, there were no artifactual bumps using the LX-spline. For the P-spline the bumps are most noticeable for flat regions of the curve in the higher variability condition. When there was a large signal relative to variance the LX-spline and P-spline performed similarly often giving nearly identical estimates. When the signal relative to variance decreased large gains in the LX-spline were observed in terms of squared error loss.

To investigate this behavior a simulation study was performed that compared the squared error loss for the LX-spline and the P-spline approaches. Here all data are simulated from the line $y = Ax + \epsilon$, with $\epsilon \sim N(0, \tau^{-1})$. For the simulation the slope parameter A was given values $0, 5, 10$, and 20, and the variance τ^{-1} was assigned values of $0.5, 0.1$, and 0.01. We report the ratio of the estimated loss for the LX-spline when compared to the P-spline methodologies. Here values higher than one indicate the P-spline has a higher squared error loss than the LX-spline method. Table 4.1 shows the results for all conditions. The LX-spline's estimated loss is lower for every simulation condition. In the worst case - when truth is a flat curve- the estimated squared error loss is approximately six and a half times greater when estimating with P-splines. The table also shows a consistent pattern where the ratio decreases as the derivative/slope increases and/or the variability decreases. This suggests that the efficiency gains in the

LX-spline are greatest when there is a large amount of noise in relation to the slope.

We also note that in these simulations $H = 1$ even though the true curve had no change points in the interval. With $H = 1$ there was little increase in noise when adding an extraneous changepoint to the model. This observation was repeated in other examples where there too was little difference in estimation when H was greater than the number of changepoints in the model. In these cases one (or more) of the $\boldsymbol{\alpha}$ was estimated to be on the boundary of $\mathcal{X}$ effectively removing it from the model.

4.4.2 Power Simulation

To test the performance of the method when conducting a test on the shape of the function we perform a simulation study when testing a J-shaped curve against a monotone alternative. Here data was simulated assuming the model $y = -exp(10(x - a)^2) + \epsilon$ on the interval $\mathcal{X} = [2, 3]$, with $\epsilon \sim N(0, 0.01)$. We investigated the behavior of the test in relation to the number of points sampled. Initially we sampled 101 points evenly spaced across the interval $\mathcal{X}$. We then added $50, 100$, and 150 evenly spaced points across [22.3] which contained the minimum. In this study we looked at three curves: one where the true shape was monotonic increasing $a = 2$, one where there was a shallow minimum $a = 2.15$ and one where there was a well defined minimum $a = 2.3$. Finally all data was simulated from independent draws of a $N(0, 0.01)$ distribution. This gave a moderate amount of noise given the signal to investigate the behavior of the test in relation to the sample size. We simulate data from each condition 100 times and look at the mean of the posterior probability of H_1 : J-shaped curve.

Table 4.2 reports the average posterior probability of H_1. When the curve was truly monotone there was very little evidence that suggested the curve was J-shaped. Further as n increased this evidence decreased. When the curve had a well defined minimum, the opposite result was seen. Here there was consistently a large amount of evidence

suggesting that there was a true minimum. Further, this minimum was in most cases estimated to be close to the true minimum of 2.3. In the middle condition a relatively large number of observations are required to increase the evidence in favor of H_1. This suggests that when the minimum is shallow a large number of data are required to effectively test if the minimum exists.

4.5 Data Examples

4.5.1 HDI and Fertility

A recent article by Myrskylä et al. (2009) suggested the negative association between human development and fertility reverses in countries that are the most economically developed. The authors hypothesized that this upturn occurs after the human development index (HDI) reaches 0.86. They concluded this based upon a LOESS regression where formal hypothesis testing was not conducted. Supporting this hypothesis Furuoka (2009), used threshold models to test the location of the changepoint. They estimate the critical HDI was 0.77. They used a linear threshold model. As only two linear segments were used, the model can not effectively model smooth changes that may occur. Other modeling approaches rely on other parametric assumptions, and do not answer the question if there is enough evidence to suggest that the data support a J-shape over a monotonic decreasing curve.

We use the LX-spline approach to test the J-shaped hypothesis against a monotonic decreasing alternative. In this analysis we use $K = 50$ equally spaced knots and sample the posterior using $105,000$ samples with the first $5,000$ samples disregarded as burn-in. In this analysis we specify the prior as in the simulation conditions. We note that we varied the number of the knots as well as the prior over $\boldsymbol{\lambda}$ and δ, and the results are consistent with those reported here. This analysis investigates the 2005 fertility data

as studied by Myrskylä et al. (2009). In all, 141 data points were analyzed. In this dataset the minimum HDI score was 0.30, and the maximum HDI score was 0.97.

Figure 4.3 shows the posterior estimate of the LX-spline (solid black line) model and corresponding 95% credible intervals (black dotted line) against the same LOESS spline model fit in Myrskylä et al. (2009) for the HDI region of interest. This fit shows that the LX and LOESS estimates are very similar across the HDI region of interest. Here the minimum occurs at approximately the same location with the minimum calculated to be at an HDI of 0.90 with a 95% CI of $(0.86, 0.97)$. This is reflected in the observed J shape of the posterior curve estimate. When compared against the probability the curve is monotonic decreasing one finds $Pr(\alpha_1 < \gamma_u) = 0.84$. This suggest there is relatively weak evidence in favor of the J-shaped hypothesis.

We caution that these results do not contradict the findings of Myrskylä et al. (2009). Given the simulation study, there simply may not be enough observations to show that there is a true minimum. Our analysis did produce a similar, though shallower, J-shaped relationship as found in that work. This analysis does argue that there are not yet enough data to discount the idea that the decrease in fertility may merely be stabilizing to a constant, and not increasing for high HDI levels.

4.5.2 Seasonal Adjustments

Time series data often are assumed to be the sum of two deterministic components: the trend T_i and the seasonal component S_i, i.e.,:

$$y_i = T_i + S_i + \epsilon_i, \tag{4.5}$$

where $\epsilon_i \sim N(0, \tau^{-1})$. In many cases it is reasonable to assume that the seasonal component has a single maximum and single nadir. Perfectly periodic basis functions

may not appropriately describe the seasonal trend. We use LX-splines to adjust for the seasonal component on data taken from monthly ambient air CO_2 measurements taken between 1980 and 1995. Due to the northern hemisphere having greater landmass (and consequently more vegetation), ambient CO_2 decreases between the summer months of May and September, while increasing in the intervening months. There is expected to be an overall increasing trend in CO_2 concentrations due to economic development, and we wish to model this trend without the seasonal component.

In modeling S_i we use an LX-spline with $H = 2$, $K = 12$, and $r = 5$. Here the the same prior specification was used as in the simulation. In estimating the trend, T_i, we use 30 equally spaced B-splines where a diffuse prior is placed upon the basis coefficients. We compare this approach in estimating the seasonally adjusted time series to an ARIMA aproach. Here the seasonal component is removed using the X-12 ARIMA seasonal adjustment (Findley et al. 1998). The seasonally adjusted trend as well as seasonal component is estimated using 'PROC X12,' in the SAS system.

Figure 4.4 shows the estimated seasonally adjusted trend (dashed line, top plot) and the corresponding unadjusted estimate (solid line, top plot) fit to the observed data. The bottom plot of this figure compares the seasonally adjusted estimate using the LX-spline as an adjustment (dashed line) to a seasonally adjusted estimate using the X-12 ARIMA method (solid line). As seen in the top portion of the plot the method effectively describes the given data, and recovers the observed trend. The bottom plot shows that the method's seasonal adjusted trend estimate is nearly identical when compared to the X-12 ARIMA method. Note the plot shows the years between 1980 and 1983, and not the entire range of data, as the lines are essentially indistinguishable when the complete trend is shown. Figure 4.5 shows the estimated seasonal adjustment (black line) and 95% credible intervals. Here it is seen that periodic basis functions would not adequately describe this trend and that something more flexible, such as the

LX-spline is necessary.

4.5.3 Benchmark Dose Risk Assessment

Human health risk assessment often utilizes toxicology data from studies having relatively few dose groups. These studies estimate the adverse risk of disease (often cancer) given exposure to some chemical. In many cases, where mode of action is unknown, fitted dose response curves are used to interpolate the dose response between dose groups. These interpolations allow estimation of risk for doses that are not observed. Alternatively, given a specific risk level, these models allow estimation of a dose associated with this risk. This dose, known as the benchmark dose (BMD), is often used in regulatory decisions to determine an acceptable level of human exposure.

Dose response models often assume monotonic increasing responses; however, in some cases, high doses produce acute toxicity and death, which occurs prior to tumorigenesis. This high dose effect can result in downturns at the higher tested doses. Due to the monotonic assumptions the higher dose data are often removed from the analysis as standard models fail to adequately describe this phenomenon. Removal of such information may lead to increased uncertainty in the BMD estimate. We investigate data that come from a National Toxicology Program (NTP) long term bioassay studying the health effects of exposure to tumeric oleoresin (NTP 1993). These data were also studied by Peddada et al. (2004) and Hans and Dunson (2005) to test for possible downturns at each dose group. Their methodologies only estimated the probabilities of adverse response for an observed dose groups, and could not be extended to BMD estimation as the models were not continuous. In this analysis we investigate the dose response data using the 2 year female mice data. Here hepatocellular adenomas were the response of interest. These data are described in Table 4.3. We estimate the BMD using the LX-spline, and compare the results to another method based upon monotonic

splines (Wheeler and Bailer 2012). The later monotonic method is fit using only the first three dose groups. Further, as is often standard in regulatory contexts, we study the dose response relationship without adjustment for the age of the animal at death. Following Hans and Dunson (2005), we use three similar NTP studies control groups in developing a prior on the background response rate.

For the LX-spline data were fit with $K = 5$ equally spaced splines. Again we set $r = 5$. Further we set $\pi_l = 0$ to restrict our analysis to only positive dose response relationships. An informative prior was placed on β_0 (i.e., the background rate), based upon the mean and variance control data found in table 4.3. Here the prior mean was set to the mean of the control data and the prior variance was made to equal 10 times the variance of the data. The posterior was sampled $55,000$ times with the first $5,000$ removed as burn-in.

Figure 4.6 describes the fit of the LX-spline model to the data. The other model, which assume only monotonicity, is not shown because it is indistinguishable in the figure. This model estimates a slight downturn and is consistent with estimates from other analyses (Peddada et al. 2004; Hans and Dunson 2005). For the LX-spline, $Pr(\text{Downturn}) = 0.935$, which is lower than those reported in other analyses. These analyses adjusted for the time to tumor, which might increase this probability. In terms of BMD estimation, the LX-spline is similar to the other method when estimating the BMD. Here the LX spline estimated the BMD to be 4350, while the monotonic method estimated the BMD to be 4670 ppm. However, the 95% lower bound on the BMD, which is the quantity often used in regulatory settings, was 1950 ppm for the LX-spline and 1640 ppm for the nonparametric method. This has important regulatory implications. We believe that the narrower interval was a result of estimating the downturn. Here the downturn reduced the probability of having very steep curves and resulted in narrower confidence interval widths on the BMD.

4.6 Conclusion

The LX-spline gives is a novel spline construction allowing for constrained functional estimation where a maximum number of changepoints can be assumed. Given a proper prior we show, both through simulation experiment and data example, that the method can be used in a wide number of contexts, and that there are gains in efficiency that can be expected when the LX spline is used. These gains efficiency increase as derivative, in relation to the overall variability, increases.

As mentioned above the LX-spline can be used in additive structures for estimating multidimensional surfaces. Here the total number of change points becomes a multiple of all of the change points in the model. Note however that their properties do not transfer in the case of tensor products. We are currently investigating methods that are similar to the LX-spline and can be used for higher dimensional data.

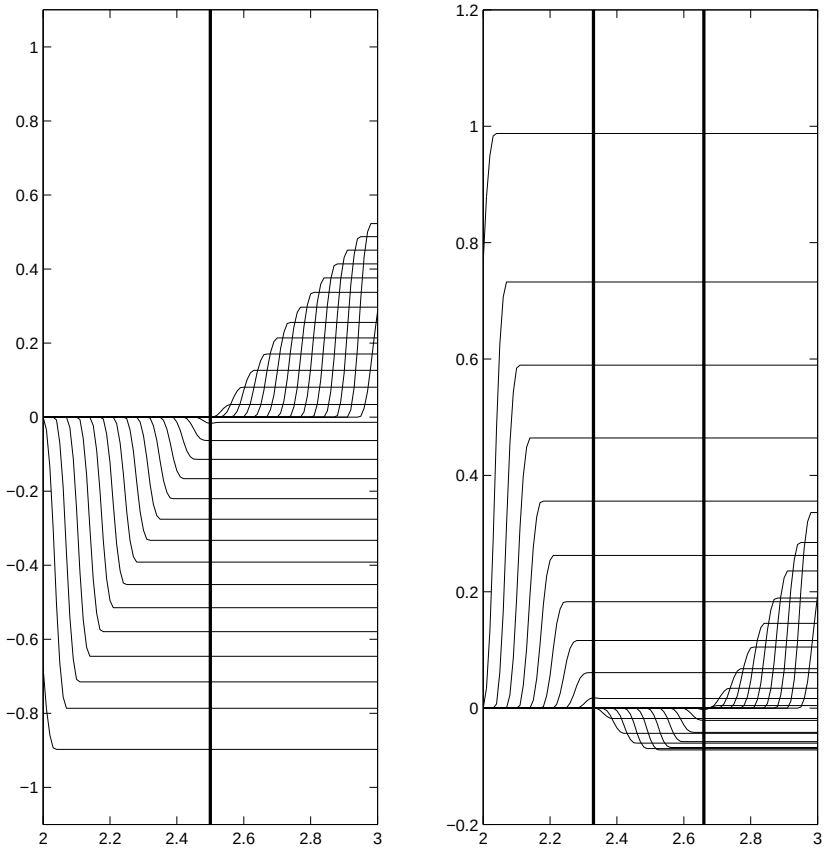

Figure 4.1: Order restricted splines with a single change point at 2.5 (left), and order restricted splines with two change points at 2.33 and 2.66 (right).

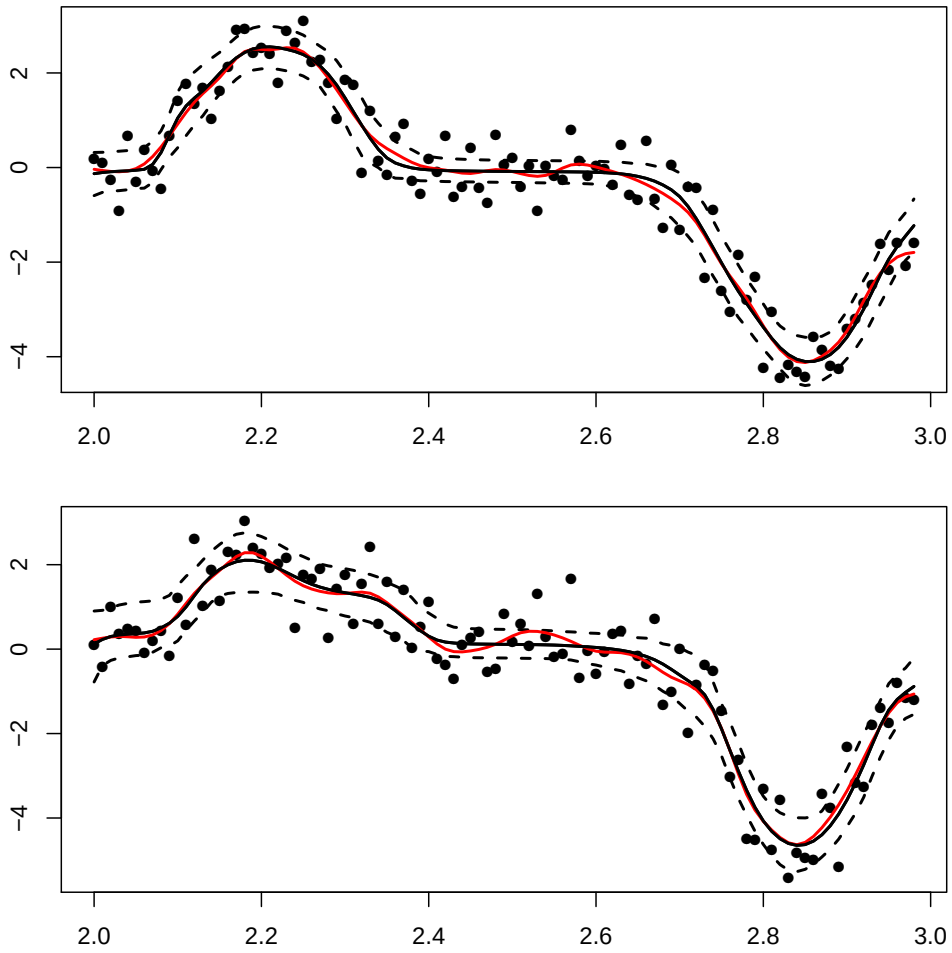

Figure 4.2: Fit of the LX-spline (black line) with corresponding 95% credible intervals (dotted line) and Bayesian P-spline (red) for the top and bottom plots. The top plot represents a simulation with lower variance and the bottom plot represents a higher variance condition.

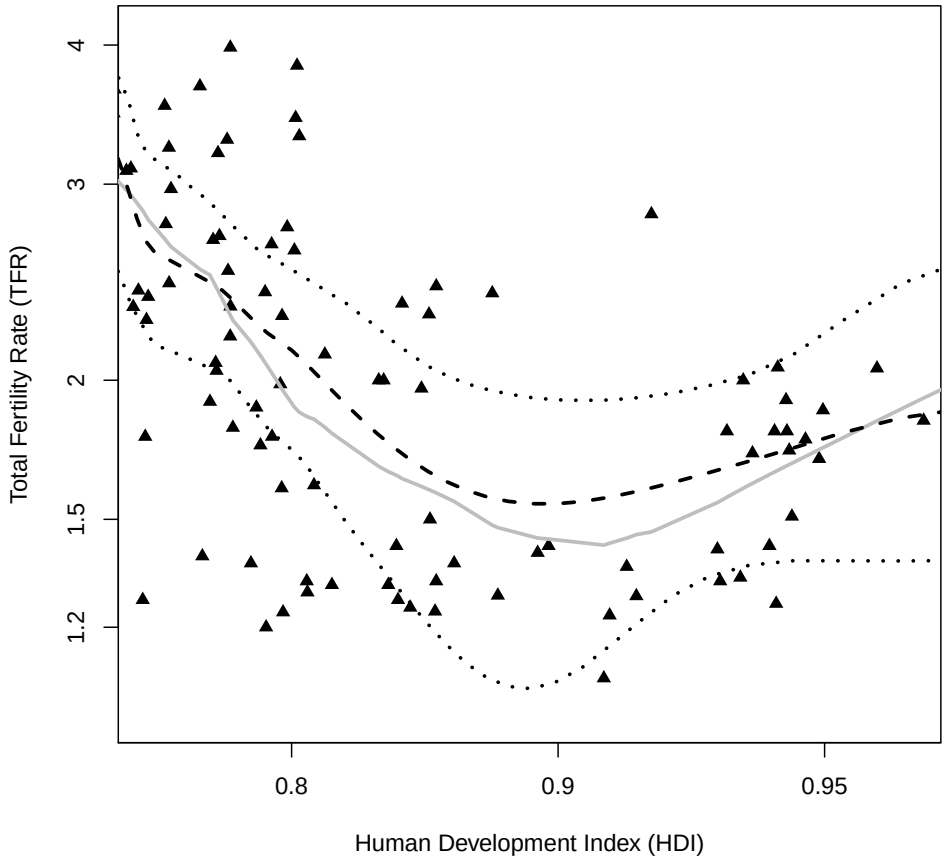

Figure 4.3: Plot of the observed total fertility rate against the HDI. The gray solid line represents the results reported (Myrskylä et al. 2009). The LX-spline is show using the solid black line, with corresponding 95% credible intervals of the LX-spline fit to the same data.

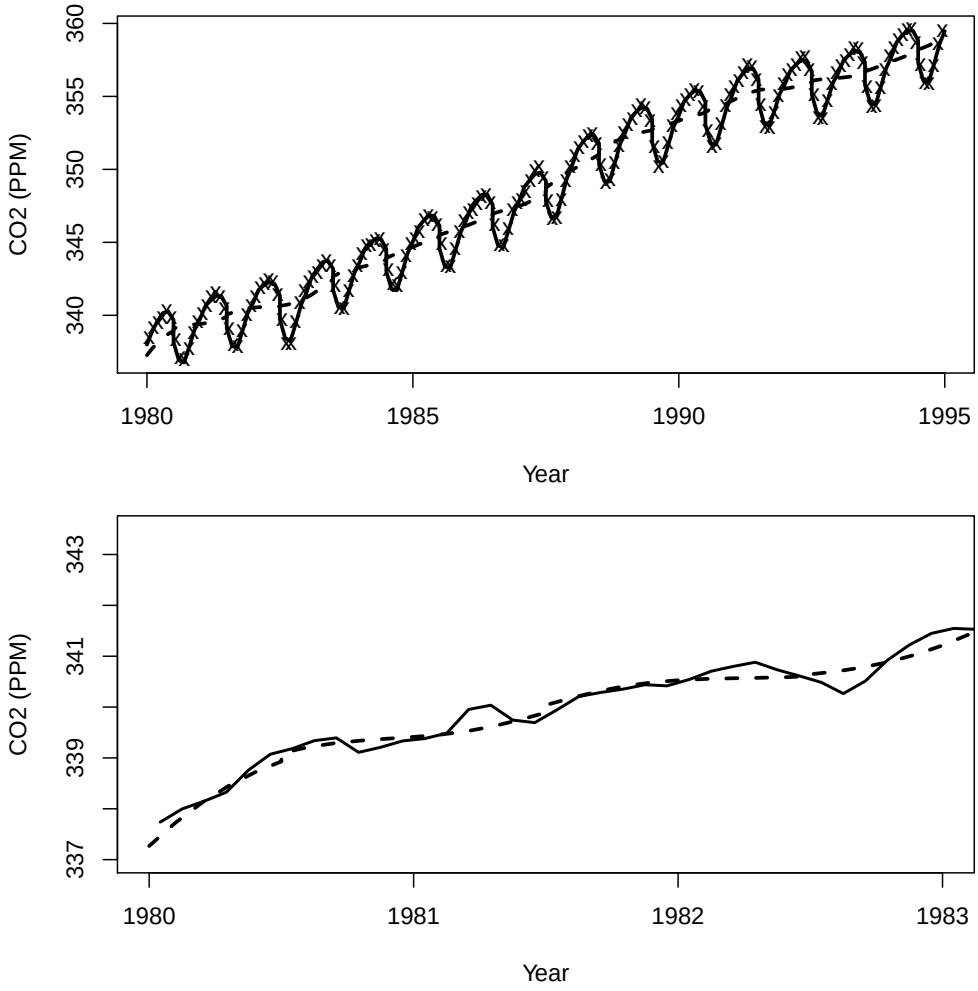

Figure 4.4: The top plot shows the seasonally adjusted world CO_2 trend line (dashed line) as compared to the seasonally unadjusted estimate (dark black line) fit to the observed data. The bottom plot compares the seasonally adjusted world CO_2 estimate where the adjustment was based upon the LX-spline (dashed line) and the $X-12$ ARIMA adjustment (solid line).

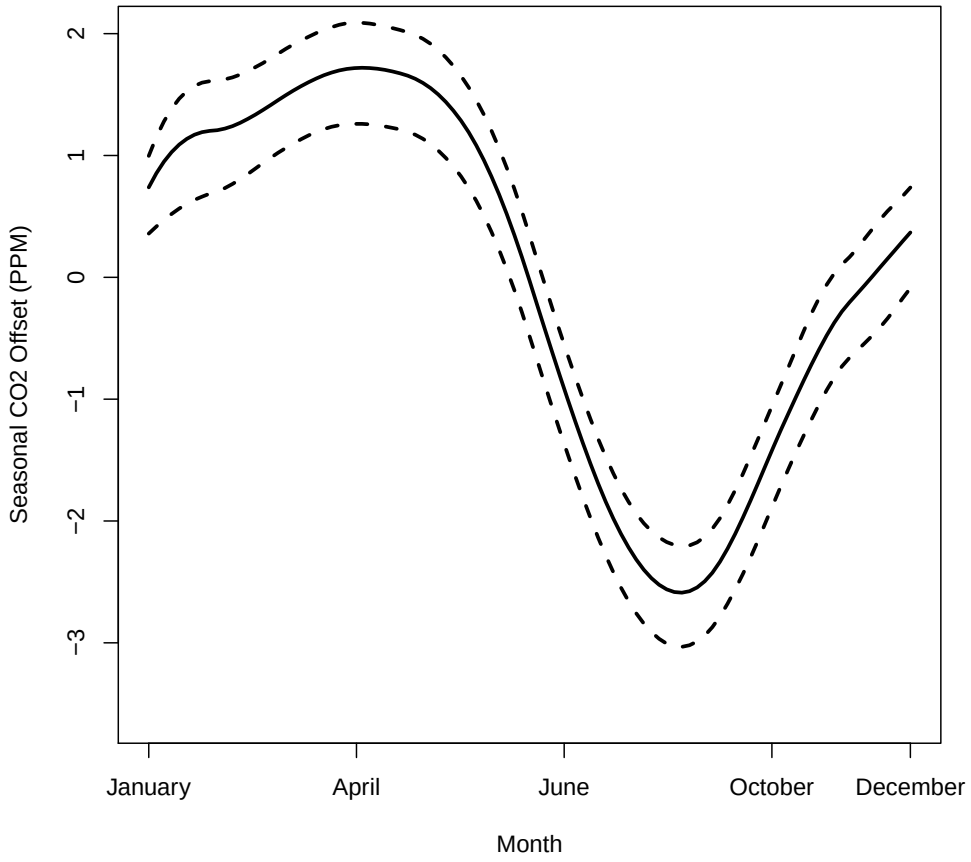

Figure 4.5: Estimated seasonal adjustment for the observed yearly CO_2 concentration data (black line) and its 95% estimated credible interval (dotted line).

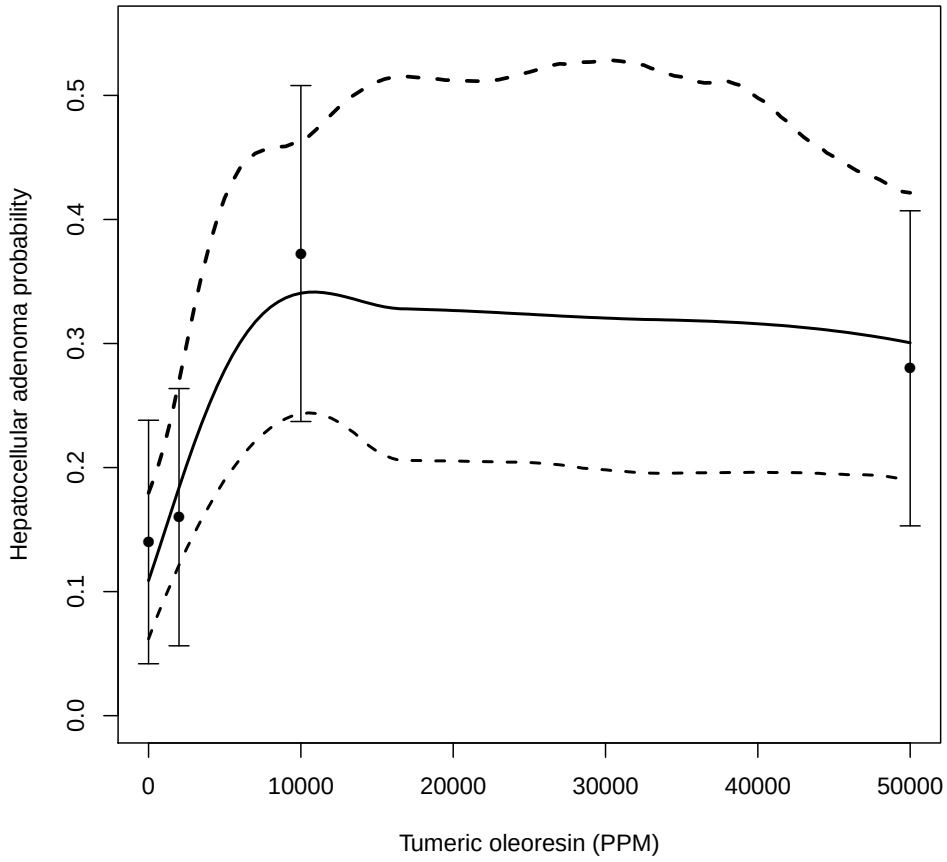

Figure 4.6: Estimated dose-response curve for tumeric oleoresin in a two year bioassay of B6CF1 female mice. The curve represents the probability of observing hepatocellular adenomas given increasing levels of tumeric oleoresin (ppm).

	Derivative			
Variance	**0**	**5**	**10**	**20**
0.50	6.52	6.72	3.66	2.22
0.10	5.10	5.18	2.55	1.77
0.01	2.98	3.45	1.63	1.23

Table 4.1: Ratio of squared error loss between the LX-spline and the P-spline for line segments on $\mathcal{X} = [2, 3]$ given a specified derivative and variance condition.

Sample Size	**Monotone Increasing**	**Shallow Minimum**	**Well Defined Minimum**
n=100	0.12 (0.00,0.27)	0.24 (0.00,0.49)	0.99 (0.99,1.00)
n=150	0.09 (0.00,0.24)	0.38 (0.04,0.72)	0.99 (0.99,1.00)
n=200	0.09 (0.00,0.30)	0.66 (0.24,1.00)	0.99 (0.99,1.00)
n=250	0.07 (0.00,0.26)	0.89 (0.59,1.00)	0.99 (0.99,1.00)

Table 4.2: Results of a simulation study looking at the posterior probability (with corresponding 95% confidence intervals) the estimated curve contains a single minimum when compared to a monotone increasing curve for three simulation conditions. The three conditions considered for the true curve were: monotone increasing (i.e., no minimum), shallow minimum near the boundary of $\mathcal{X}$, and a well defined minimum.

NTP Study #	dose	obs	n
394	control	3	49
419	control	5	50
439	control	6	49
427	control	7	50
427	2000	8	50
427	10000	19	51
427	50000	14	50

Table 4.3: Summary of hepatocellular adenomas data of female B6CF1 mice exposed to tumeric oleoresin. The top three lines show control data for NTP studies that were used to develop priors for the analysis.

Chapter 5

Conclusion

This dissertation considered novel priors on functions $f : \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} \subset \mathbb{R}$. Here, rather than placing a prior over f directly, priors were developed through the use of a differential operator applied to f. In the case of the mechanistic GP the prior was applied using a linear differential operator L. This operator can be based upon parametric scientific models, as was our case, or can be used to put a prior over curves that one might expect in an application. The LX-spline was defined by placing a prior over the maximum number of zeros in the first derivative, which in turn limited the maximum number of changepoints in f. This prior produced flexible models that removed artifactual bumps that are commonly seen in many applications.

Though the mechanistic GP develops a prior over functions having shapes consistent with scientific information, more research is needed to understand how such a prior may improve the efficiency of estimation. Specifically research on the rate of posterior contraction similar to van der Vaart and van Zanten (2008), may be helpful in understanding the possible gains in efficiency when using such a prior. Further, though the approximation used does allow for posterior computation, the problem of matrix inversion in GP regression still holds. Consequently tractability issues may arise for large data sets. Further efficient computational methods such as Banerjee et al. (2012) may not be directly applicable due to the nature of the Runge Kutta approximation.

A further area of research would be to develop more efficient methods for large data sets when using the mechanistic hierarchical GP

In terms of the LX-spline various extensions can be envisioned. The prior currently supports single functions, and thus can not be used for longitudinal data. Development of a prior that could be used for longitudinal data may aid in estimation of subject specific curves, especially when there are few longitudinal observations. Further, the prior is defined over a fixed knot set. Such a prior does not place prior probability over an arbitrary function having at most H changepoints in $\mathcal{X}$, and research on alternative priors that allow the location and number of knots to be unknown may be preferred. Such a prior would guarantee the prior positivity within an ϵ distance for all functions having at most H changepoints. Finally the LX-spline was only defined for one dimensional surfaces. It would be useful to extend it to multiple dimensions as most covariates cannot be assumed to have an additive structure.

Bibliography

Albert, J. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association* **88,** 669–679.

Alvarez, M., Luengo-Garcia, D., and Lawrence, N. (2009). Latent force models. *Proceedings of the Twelfth International Workshop on Artificial Intelligence and Statistics* **5,** 9–16.

Asaithambi, N. (1995). *Numerical Analysis: Theory and Practice*. Saunders College Pub.

Baker, B., Hollander, M., Kashon, M., and Cutlip, R. (2010). Effects of glutathione depletion and age on skeletal muscle performance and morphology following chronic stretch-shortening contraction exposure. *European Journal of Applied Physiology* **108,** 619–630.

Banerjee, A., Dunson, D., and Tokdar, S. (2012). Efficient Gaussian process regression for large data sets. *pre-print* .

Behseta, S., Kass, R., and Wallstrom, G. (2005). Hierarchical models for assessing variability among functions. *Biometrika* **92,** 419–434.

Biller, C. (2000). Adaptive Bayesian regression splines in semiparametric generalized linear models. *Journal of Computational and Graphical Statistics* **9,** 122–140.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via Pólya urn schemes. *The Annals of Statistics* **1,** 353–355.

Bornkamp, B. and Ickstadt, K. (2009). Bayesian nonparametric estimation of continuous monotone functions with applications to dose–response analysis. *Biometrics* **65,** 198–205.

Brumback, B. and Rice, J. (1998). Smoothing spline models for the analysis of nested and crossed samples of curves. *Journal of the American Statistical Association* **93,** 961–976.

Cutlip, R., Baker, B., Geronilla, K., Mercer, R., Kashon, M., Miller, G., Murlasits, Z., and Alway, S. (2006). Chronic exposure to stretch-shortening contractions results in skeletal muscle adaptation in young rats and maladaptation in old rats. *Applied Physiology, Nutrition, and Metabolism* **31,** 573–587.

De Boor, C. (2001). *A Practical Guide to Splines*, volume 27. Springer Verlag.

Denison, D., Mallick, B., and Smith, A. (1998). Automatic Bayesian curve fitting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **60,** 333–350.

Dimatteo, I., Genovese, C., and Kass, R. (2001). Bayesian curve-fitting with free-knot splines. *Biometrika* **88,** 1055–1071.

Ding, J., Binder-Macleod, S., and Wexler, A. (1998). Two-step, predictive, isometric force model tested on data from human and rat muscles. *Journal of Applied Physiology* **85,** 2176–2189.

Duan, J., Guindani, M., and Gelfand, A. (2007). Generalized spatial Dirichlet process models. *Biometrika* **94,** 809.

Eilers, P. and Marx, B. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science* **11,** 89–102.

Erdemir, A., McLean, S., Herzog, W., and van den Bogert, A. (2007). Model-based estimation of muscle forces exerted during movements. *Clinical Biomechanics* **22,** 131–154.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89,** 268–277.

Escobar, M. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association* **90,** 577–588.

Ferguson, T. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics* **1,** 209–230.

Ferguson, T. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics* **2,** 615–629.

Findley, D., Monsell, B., Bell, W., Otto, M., and Chen, B. (1998). New capabilities and methods of the x-12-arima seasonal-adjustment program. *Journal of Business & Economic Statistics* **16,** 127–152.

Furuoka, F. (2009). Looking for a j-shaped development-fertility relationship: Do advances in development really reverse fertility declines? *Economics Bulletin* **29,** 3067–3074.

Gelfand, A. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85,** 398–409.

Gelfand, A., Kottas, A., and MacEachern, S. (2005). Bayesian nonparametric spatial modeling with Dirichlet process mixing. *Journal of the American Statistical Association* **100,** 1021–1035.

Geman, S. and Geman, D. (1993). Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *Journal of Applied Statistics* **20,** 25–62.

Geronilla, K., Wu, J., Baker, B., and Cutlip, R. (2006). Characterization of isometric contractions of rat skeletal muscle in vivo: Duty cycle effects. *Bio-Medical Materials and Engineering* **16,** 369–380.

Ghosal, S. and Roy, A. (2006). Posterior consistency of Gaussian process prior for nonparametric binary regression. *The Annals of Statistics* **34,** 2413–2429.

Golub, G., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21,** 215–223.

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

Gunn, L. and Dunson, D. (2005). A transformation approach for incorporating monotone or unimodal constraints. *Biostatistics* **6,** 434–449.

Hans, C. and Dunson, D. (2005). Bayesian inferences on umbrella orderings. *Biometrics* **61,** 1018–1026.

Heckman, N. and Ramsay, J. (2000). Penalized regression with model-based penalties. *Canadian Journal of Statistics* **28,** 241–258.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative Methods for Current Environmental Issues*, pages 37–56. Springer.

Hill, A. (1938). The heat of shortening and the dynamic constants of muscle. *Proceedings of the Royal Society of London: Series B (Biological Sciences)* **126,** 136–195.

Hollander, M., Baker, B., Ensey, J., Kashon, M., and Cutlip, R. (2010). Effects of age and glutathione levels on oxidative stress in rats after chronic exposure to stretch-shortening contractions. *European Journal of Applied Physiology* **108,** 589–597.

Honkela, A., Girardot, C., Gustafson, E., Liu, Y., Furlong, E., Lawrence, N., and Rattray, M. (2010). Model-based method for transcription factor target identification with limited data. *Proceedings of the National Academy of Sciences* **107,** 7793–7798.

Huang, Y., Liu, D., and Wu, H. (2006). Hierarchical bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics* **62,** 413–423.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96,** 161–173.

Ishwaran, H. and Zarepour, M. (2000). Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models. *Biometrika* **87,** 371–390.

Kennedy, M. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87,** 1–13.

Kennedy, M. and O'Hagan, A. (2001). Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B, Statistical Methodology* **63,** 425–464.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics* **13,** 183–212.

Lawrence, N., Sanguinetti, G., and Rattray, M. (2007). Transcriptional regulation using Gaussian processes. *Advances in Neural Information Processing Systems* **19,** 785.

Lo, A. (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *The Annals of Statistics* **12,** 351–357.

Lunn, D., Best, N., Thomas, A., Wakefield, J., and Spiegelhalter, D. (2002). Bayesian analysis of population pk/pd models: General concepts and software. *Journal of Pharmacokinetics and Pharmacodynamics* **29,** 271–307.

MacEachern, S. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics-Simulation and Computation* **23,** 727–741.

MacEachern, S. (1999). Dependent nonparametric processes. In *ASA Proceedings of the Section on Bayesian Statistical Science*, pages 50–55.

MacEachern, S. and Müller, P. (1998). Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7,** 223–238.

Maffiuletti, N. (2010). Physiological and methodological considerations for the use of neuromuscular electrical stimulation. *European Journal of Applied Physiology* **110,** 223–234.

Mardia, K. and Marshall, R. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71,** 135–146.

Meyer, M. (2008). Inference using shape-restricted regression splines. *The Annals of Applied Statistics* **2,** 1013–1033.

Murlasits, Z., Cutlip, R., Geronilla, K., Rao, K., Wonderlin, W., and Alway, S. (2006). Resistance training increases heat shock protein levels in skeletal muscle of young and old rats. *Experimental Gerontology* **41,** 398–406.

Myrskylä, M., Kohler, H., and Billari, F. (2009). Advances in development reverse fertility declines. *Nature* **460,** 741–743.

Nakajima, J. and West, M. (2012). Bayesian analysis of latent threshold dynamic models. *Journal of Business and Economic Statistics,In Press* .

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9,** 249–265.

Neelon, B. and Dunson, D. (2004). Bayesian isotonic regression and trend analysis. *Biometrics* **60,** 398–406.

Nguyen, X. and Gelfand, A. (2011). The Dirichlet labeling process for clustering functional data. *Statistica Sinica* **21,** 639–642.

NTP (1993). Carcinogenesis studies of turmeric oleoresin (cas no. 8024-37-1) in f344/n rats and b6c3f1 mice (feed studies). *National Toxicoligy Program, Research Triangle Park, North Carolina* **427,** 1–275.

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika* **95,** 169–186.

Parsaei, H. and Stashuk, D. (2011). Adaptive motor unit potential train validation using mup shape information. *Medical Engineering and Physics* **33,** 581–589.

Peddada, S., Prescott, K., and Conaway, M. (2004). Tests for order restrictions in binary data. *Biometrics* **57,** 1219–1227.

Petrone, S., Guindani, M., and Gelfand, A. (2009). Hybrid Dirichlet mixture models for functional data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71,** 755–782.

Phillips, C., Repperger, D., Neidhard-Doll, A., and Reynolds, D. (2004). Biomimetic model of skeletal muscle isometric contraction: I. An energetic-viscoelastic model for the skeletal muscle isometric force twitch. *Computers in Biology and Medicine* **34,** 307–322.

Pitman, J. (1995). Exchangeable and partially exchangeable random partitions. *Probability Theory and Related Fields* **102,** 145–158.

Pitman, J. (1996). Some developments of the Blackwell-MacQueen urn scheme. *Lecture Notes-Monograph Series* **30,** 245–267.

Pitman, J. and Yor, M. (1997). The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability* **25,** 855–900.

Putter, H., Heisterkamp, S., Lange, J., and De Wolf, F. (2002). A Bayesian approach to parameter estimation in HIV dynamical models. *Statistics in Medicine* **21,** 2199–2214.

Ramsay, J. (1988). Monotone regression splines in action. *Statistical Science* **3,** 425–441.

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

Rasmussen, C. and C., W. (2006). *Gaussian Processes in Machine Learning.* MIT Press.

Ritter, C. and Tanner, M. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association* **89,** 861–868.

Rodriguez, A. and Dunson, D. (2011). Nonparametric bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6,** 145–177.

Ryan, M., Dudash, H., Docherty, M., Geronilla, K., Baker, B., Haff, G., Cutlip, R., and Alway, S. (2008). Aging-dependent regulation of antioxidant enzymes and redox status in chronically loaded rat dorsiflexor muscles. *The Journals of Gerontology Series A: Biological Sciences and Medical Sciences* **63,** 1015–1026.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4,** 639–650.

Shively, T., Sager, T., and Walker, S. (2009). A bayesian approach to non-parametric monotone function estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **71,** 159–175.

Shively, T., Walker, S., and Damien, P. (2011). Nonparametric function estimation subject to monotonicity, convexity and other shape constraints. *Journal of Econometrics* **161,** 166–181.

Tokdar, S. and Ghosh, J. (2007). Posterior consistency of logistic Gaussian process priors in density estimation. *Journal of Statistical Planning and Inference* **137,** 34–42.

van der Vaart, A. and van Zanten, J. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections* **3,** 200–222.

Walker, S. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics-Simulation and Computation* **36,** 45–54.

Wexler, A., Ding, J., and Binder-Macleod, S. (1997). A mathematical model that predicts skeletal muscle force. *IEEE Transactions on Biomedical Engineering* **44,** 337–348.

Wheeler, M. and Bailer, A. (2012). Monotonic Bayesian semiparametric benchmark dose analysis. *Risk Analysis* **32,** 1207–1218.