Bayesian methods in single and multiple curve fitting

Carsten Holm Botts

Iowa State University
Bayesian methods in single and multiple curve fitting

by

Carsten Holm Botts

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Statistics

Program of Study Committee:
Michael Daniels, Major Professor
Mark Kaiser
Taps Maiti
Jean Opsomer
Rohan Fernando

Iowa State University
Ames, Iowa
2005
Copyright © Carsten Holm Botts, 2005. All rights reserved.
INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.
This is to certify that the doctoral dissertation of

Carsten Holm Botts

has met the dissertation requirements of Iowa State University

Signature was redacted for privacy.

Major Professor

Signature was redacted for privacy.

For the Major Program
TABLE OF CONTENTS

1 INTRODUCTION ....................................................... 1
  1.1 Statistical inference ........................................... 1
    1.1.1 Frequentist (Classical) methods ....................... 1
    1.1.2 Bayesian inference ....................................... 2
  1.2 Estimating one curve ........................................... 4
    1.2.1 The general problem of function estimation ............. 4
      1.2.1.1 Parametric ........................................... 4
      1.2.1.2 Nonparametric ....................................... 5
    1.2.2 Spectral density estimation .............................. 6
  1.3 Estimating multiple curves .................................... 7
  1.4 Detailed outline of dissertation ............................ 9
  1.5 References .................................................... 10

2 A SHRINKAGE ESTIMATOR FOR SPECTRAL DENSITIES ................. 15
  2.1 Abstract ...................................................... 15
  2.2 Introduction .................................................. 15
  2.3 Estimator ..................................................... 17
    2.3.1 The DC estimator ......................................... 17
    2.3.2 New estimator ............................................ 19
      2.3.2.1 Choice of $f_p(\omega; \theta)$ ......................... 21
      2.3.2.2 Priors on $\psi, \sigma^2$, and $\tau^2$ .............. 22
      2.3.2.3 Consistency and evaluation of the estimator at any set of frequencies 23
  2.4 Simulations and computations ................................ 24
  2.5 Results ....................................................... 26
  2.6 Example ....................................................... 27
  2.7 Discussion .................................................... 27
  2.8 References ................................................... 28
3 LIKELIHOOD APPROXIMATIONS IN BAYESIAN MULTIPLE CURVE FITTING

3.1 Abstract

3.2 Introduction

3.3 The model and its corresponding methods

3.3.1 A general strategy for sampling from \( p(k, t_k, k', t_{k'} | y) \)

3.3.2 Calculating \( p(y | k, t_k, k', t_{k'}) \)

3.3.2.1 The 'plugged in' approximation

3.3.2.2 The Laplace approximation

3.4 The penalty of the approximations

3.4.1 Which approximation to use? A question of model selection

3.4.2 Quantifying the penalty of the Laplace approximation

3.5 Simulation studies

3.5.1 Likelihood ratios

3.5.2 Gibbs sampler

3.6 Discussion

3.7 References

3.8 Appendix

3.8.1 RJMCMC details

3.8.2 Proof of detailed balance

4 PRINCIPAL COMPONENT REDUCTION IN LINEAR MIXED B-SPLINES

4.1 Abstract

4.2 Introduction

4.3 The model

4.3.1 The mixed model

4.3.2 Priors

4.4 Sampling \( p(k, t_k, r | y) \)

4.5 Simulation study

4.6 Application

4.7 Discussion

4.8 References

4.9 Appendix
4.9.1 RJMCMC details ............................................. 82
4.9.2 Proof of detailed balance ..................................... 83
4.9.3 Proof of Theorem 4 ............................................ 84
4.9.4 Derivatives for $S$ ............................................... 86

5 CONCLUSIONS ......................................................... 92
  5.1 General contributions of thesis .................................. 92
  5.2 Open issues ..................................................... 92
  5.3 References ...................................................... 93

6 ACKNOWLEDGEMENTS .............................................. 95
1 INTRODUCTION

This chapter briefly reviews statistical inference and function estimation. Section 1.1 discusses the general concepts and techniques of classical and Bayesian inference. Sections 1.2 and 1.3 briefly review the literature and problems associated with single and multiple curve estimation, and Section 1.4 provides a detailed outline of this dissertation. This last section discusses how Bayesian methods, in this thesis, are used to estimate single and multiple curves.

1.1 Statistical inference

Statisticians often explain variability in a data set using a model, $M$, that is indexed by a fixed, unknown parameter vector $\theta$. Accounting for the variability observed in a data set becomes a problem of estimating, or learning about, $\theta$. Two paradigms are typically used to make inference about the parameter $\theta$: frequentist (classical) methods and the Bayesian methods. These two methods differ in how they define probability, and are reviewed in the sections below.

1.1.1 Frequentist (Classical) methods

For a frequentist, the probability of an event is its long run frequency in repeated trials. No meaningful probabilistic statements can thus be made about the parameter $\theta$. The probability that the parameter $\theta$ exists in a particular region, for example, is either 1 or 0; $\theta$ is a fixed quantity and either is in the specified region, or is not. In the frequentist literature, one of the most commonly used point estimators for $\theta$ is the maximum likelihood estimator. This estimator, denoted as $\hat{\theta}_{ml}$, is selected to maximize the probability of the observed data (the likelihood), and is defined mathematically as

$$\hat{\theta}_{ml} = \text{argmax}_{\theta} \{ p(y|\theta) \},$$

where $y$ is the observed data, and $p(y|\theta)$ is the likelihood of the data as a function of $\theta$. The quality of this maximum likelihood estimator is usually measured in terms of how close, on average, it is to the true value of $\theta$. This is measured by variability and bias. The variance and bias are combined to calculate the estimator's mean squared error (MSE), where
\[ \text{MSE} = \int \left( \hat{\theta}_{\text{ml}} - \theta \right)' \left( \hat{\theta}_{\text{ml}} - \theta \right) p(y|\theta) \, dy \]
\[ = \mathbb{E} \left[ \left( \hat{\theta}_{\text{ml}} - \theta \right)' \left( \hat{\theta}_{\text{ml}} - \theta \right) \right] \]
\[ = \text{Var} \left( \hat{\theta}_{\text{ml}} \right) + \text{Bias}^2 \left( \hat{\theta}_{\text{ml}} \right). \]

If the estimator is unbiased, the MSE is equal to the variance of the estimator. This variance is estimated using

\[ \text{Var} \left( \hat{\theta}_{\text{ml}} \right) = -\left( \frac{\partial^2 \log p(y|\theta)}{\partial \theta \partial \theta'} \right)^{-1} \bigg|_{\theta = \hat{\theta}_{\text{ml}}}. \quad (1.1) \]

Note that equation (1.1) inversely relates the peakedness of the likelihood at \( \hat{\theta}_{\text{ml}} \) to the estimated variance of \( \hat{\theta}_{\text{ml}} \).

### 1.1.2 Bayesian inference

The principles of Bayesian inference are fundamentally different than those associated with frequentist methods. For Bayesians, probability measures one’s belief that an event is true. Meaningful probabilistic statements about \( \theta \) can be made. This belief is formulated as a posterior distribution of \( \theta \). This distribution, \( p(\theta|y) \), is a function of one’s prior belief and the data collected. It is mathematically defined as

\[ p(\theta|y) = \frac{p(y|\theta) p(\theta)}{\int p(y|\theta) p(\theta) \, d\theta} \propto p(y|\theta) p(\theta), \quad (1.2) \]

where \( p(\theta) \) is a prior distribution placed on \( \theta \). Inference on \( \theta \) can be made from its posterior distribution.

Point estimators of \( \theta \) derived from its posterior distribution include the posterior’s mean, mode, and median. These points are defined, analytically, as

\[ \hat{\theta}_{\text{mean}} = \int \theta p(\theta|y) \, d\theta \]
\[ \hat{\theta}_{\text{mode}} = \left\{ \theta^* : p(\theta^*|y) = \max_{\theta} p(\theta|y) \right\} \]
\[ \hat{\theta}_{\text{median}} = \left\{ \theta^* : \int_{-\infty}^{\theta^*} p(\theta|y) \, d\theta = .5 \right\} \]

It is common that the posterior can not be evaluated analytically since the denominator in expression (1.2) can not written in closed form. If this were true, point estimators such as the ones discussed above
would be difficult to calculate, and inference on $\theta$ may involve sampling values of $\theta$ from $p(\theta|y)$. Several algorithms allow one to sample from a posterior distribution without having to evaluate the denominator in (1.2). Some of these algorithms are discussed next.

In this section, the Gibbs sampler, the Metropolis-Hastings algorithm, and the reversible jump MCMC algorithm are reviewed. The need for any one of these sampling algorithms varies with the kind of posterior being sampled from. If $\theta$ is $j-$dimensional with $j > 1$, and all of the conditional posteriors have a recognizable form, the Gibbs sampler can be used. This algorithm, given below, simply involves iteratively sampling from all $j$ conditional posterior distributions.

Gibbs Sampler:
Call the current value of $\theta$ $\theta^{\text{old}}$, where $\theta^{\text{old}} = (\theta_1^{\text{old}}, \theta_2^{\text{old}}, \ldots, \theta_j^{\text{old}})$, and repeat the following steps:
1. Sample $\theta_1^{\text{new}}$ from $p(\theta_1|\theta_2^{\text{old}}, \ldots, \theta_j^{\text{old}}, y)$
2. Sample $\theta_2^{\text{new}}$ from $p(\theta_2|\theta_1^{\text{new}}, \theta_3^{\text{old}}, \ldots, \theta_j^{\text{old}}, y)$
   
   ... 
   
   j. Sample $\theta_j^{\text{new}}$ from $p(\theta_j|\theta_1^{\text{new}}, \ldots, \theta_{j-1}^{\text{new}}, y)$
   
   j+1. Rename $\theta^{\text{new}}$ $\theta^{\text{old}}$.

If the full conditionals are not standard distributions, however, other methods need to be considered. The Metropolis-Hastings algorithm, given below, does not require the full conditionals to be recognizable distributions, but it does require the specification of a jumping distribution $J(\cdot|\cdot)$.

Metropolis Hastings:
Call the current value of $\theta$ $\theta^{\text{old}}$, and repeat the following steps

1. Draw a new value of $\theta$, $\theta^{\text{new}}$, from the jumping distribution $J(\cdot|\theta^{\text{old}})$.
2. Accept $\theta^{\text{new}}$ with probability $\frac{p(\theta^{\text{new}}|y)J(\theta^{\text{old}}|\theta^{\text{new}})}{p(\theta^{\text{old}}|y)J(\theta^{\text{new}}|\theta^{\text{old}})}$
3. If accepted, rename $\theta^{\text{new}}$ $\theta^{\text{old}}$.

Now consider the problem of drawing from the posterior $p(j, \theta|y)$, where, again, $j = \dim(\theta)$. In this case, one of the parameters being sampled indexes the dimension of another parameter being sampled, and the support for the posterior becomes $\bigoplus_{j=1}^{\infty} \Theta_j$, where $\Theta_j$ is the parameter space of $\theta$ when $\dim(\theta) = j$. Reversible jump MCMC methods are required in this case. The reversible jump MCMC algorithm is
nearly identical to the Metropolis-Hastings algorithm given above. The only difference between the two is that reversible jump methods require a jumping distribution which proposes a new value of \( \theta, \theta_{\text{new}} \), and a new value of \( j, j_{\text{new}} \).

This dissertation introduces novel Bayesian methods in both single and multiple curve estimation. The following sections briefly review both frequentist and Bayesian methods associated with these two estimation problems.

1.2 Estimating one curve

This section reviews the motivations and techniques associated with single curve estimation. The first part of the section discusses the general problem of function estimation. The second part of the section focuses on spectral density estimation, which is the subject of Chapter 2.

1.2.1 The general problem of function estimation

Social and physical scientists are often interested in how certain variables depend on one another. An economist, for example, may want to learn about the relationship between the average value of an individual’s savings given his annual income, or a chemist may wish to see how the pressure of a closed system changes with rising temperature. Scientists often assume that a functional relationship exists between these variables. These functional relationships can be written as

\[
Y = f(x)
\]  

(1.3)

where \( Y \) is the dependent variable and \( x \) the independent, or controlled, variable. To learn about this function, \( f(\cdot) \), scientists conduct experiments or collect samples. They are aware, however, that the collected data does not follow the deterministic equation given in (1.3), but rather follows the stochastic equation

\[
Y = f(x) + \epsilon
\]  

(1.4)

where \( \epsilon \) is random error with mean 0. Estimating the function \( f(x) \) under model (1.4) is one of the most common problems in statistics. This section reviews parametric and nonparametric methods of function estimation.

1.2.1.1 Parametric

While estimating \( f \) at all values of \( x \) initially appears intimidating, the problem can be simplified if \( f \) is assumed to take some parametric form. One of the most common parametric forms assumed
is the simple linear regression model \( f(x) = \beta_0 + \beta_1 x \), where \( \beta_0 \) and \( \beta_1 \) are the two parameters that uniquely identify the function. Estimating \( f(\cdot) \) now becomes a problem of estimating \( \beta_0 \) and \( \beta_1 \). Other parametric forms may also be assumed. If the relationship between \( Y \) and \( x \) is thought to be quadratic, for example, the model fit to the data could be \( f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 \). These ordinary regression models can be written as \( Y = X\beta + \epsilon \), where \( X \) is a design matrix for all \( n \) subjects in the study, \( Y \) is the response vector for all subjects, and typically \( \epsilon \sim N(0, \sigma^2 I) \). Classical inference returns the maximum likelihood estimators \( \hat{\beta}_{ml} = (X^T X)^{-1} X^T Y \) and \( \hat{\sigma}^2_{ml} = \frac{(Y - X\hat{\beta})^T (Y - X\hat{\beta})}{(n - \text{dim}(\beta))} \). Bayesian analysis, however, bases parametric inference on the posterior \( p(\beta, \sigma^2 | y) \) (Gelman, et al., 1995).

### 1.2.1.2 Nonparametric

There are a variety of alternatives to the parametric models introduced in Section 1.2.1.1. For example, the function \( f(\cdot) \) can be defined as the sum of \( k \) basis functions. This allows \( f(\cdot) \) to be written as

\[
f(x) = \sum_{j=1}^{k} \alpha_j \xi_j (x),
\]

where \( \xi_1 (\cdot), \ldots, \xi_k (\cdot) \) are the basis functions and \( \alpha_1, \ldots, \alpha_k \) are \( k \) coefficients. This model is attractive because it provides its user with great flexibility (Ramsey and Silverman, 1997). This flexibility comes with the wide selection of basis functions.

One of the most common choices is the regression polynomial basis. A regression polynomial basis of order \( u \) sets

\[
\xi_j (x) = (x - t_j)^u
\]

where \( z_u^+ = z_0^+ (z > 0) \), and \( t_j \) is a knot located at some point on the domain of \( f(x) \), the regression spline. Knots are points on the \( x \)-axis where the function \( f(x) \) is not continuous. For a regression polynomial of order \( u \), the \( u^{\text{th}} \) derivative of \( f(\cdot) \) is discontinuous at all points \( (t_1, \ldots, t_k) \). Figure 1.1 shows four piecewise cubic splines, each with 10 knots. Note how, on average, the knots are located at points where the function is not smooth. Identifying the function \( f(\cdot) \), in cases where polynomial splines are assumed, thus involves correctly identifying the number and location of knots associated with the function. Classical (frequentist) methods of knot selection typically involve fitting multiple sets of knots and identifying that set which maximizes the likelihood and predictive quality of the corresponding model (Friedman, 1991; Stone et al., 1997; Lou and Wahba, 1997; Zhou and Shen, 2001; Lindstrom, 1999). Bayesian methods often require sampling from a posterior distribution on some functional of the knots. In cases where the posterior is \( p(k, t_k = (t_1, \ldots, t_k) | y) \), reversible jump MCMC methods are required (Denison et al., 1998; DiMatteo et al., 2001).
Smoothing splines provide yet another way to estimate a function. In such cases, an estimator for the function \( f(\cdot) \), \( \hat{f}(\cdot) \), is derived in minimizing the expression

\[
\sum_i \left( Y(x_i) - \hat{f}(x_i) \right)^2 + \lambda \int \left( \hat{f}''(x) \right)^2 dx.
\]

The value of \( \hat{f}(\cdot) \) which minimizes (1.5) can be written as

\[
\hat{f}(x_j) = \sum_i W_\lambda(|x_i - x_j|) Y(x_i),
\]

where \( W_\lambda(|x_i - x_j|) \) is the weight applied to a response which is \(|x_i - x_j|\) units away from \( x_j \), and \( \lambda \) is referred to as the roughness penalty, or bandwidth, of the estimator. This parameter, \( \lambda \), governs how smooth, or “rough,” the eventual estimator becomes. If the bandwidth is small, the estimator becomes less biased but more variable and thus more “wiggly” (since less weight is placed on \( \int \left( \hat{f}''(x) \right)^2 dx \)). The opposite occurs with large bandwidths. Much research has been devoted on how to select an optimal bandwidth (Bowman, 1985; Terrell and Scott, 1985; Stone, 1984; Hall and Marron, 1987).

Penalized regression splines provide yet another way to model a function. A penalized spline of order \( u \) is written as

\[
f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \cdots + \beta_u x^u + \sum_{j=1}^k \alpha_j (x - t_j)^u,
\]

but it is fit under constraints on the regression coefficients. This constraint is typically written such that the estimator of \( f(\cdot) \) will minimize the quantity

\[
\sum_i (Y(x_i) - f(x_i))^2 + \lambda \sum_j \alpha_j^2.
\]

The parameter \( \lambda \) is the “penalty” introduced into this penalized regression spline. Rather than penalize the function for being bumpy (as was the case in smoothing splines), penalized regression splines penalize for excessively estimated coefficients \( \alpha \).

While the methods outlined in this section can be used to estimate a variety of single curves, Chapter 2 of this dissertation specifically focuses on spectral density estimation. Spectral densities are reviewed in the following section.

### 1.2.2 Spectral density estimation

The spectral density \( f(\omega) \) of a stationary temporal process \( \{X_1, X_2, \ldots, X_t, \ldots\} \) is a function defined on the interval \((0, \pi)\) which completely preserves the second order properties of the process. It is defined in terms of the covariance function as

\[
f(\omega) = \frac{1}{2\pi} \sum_{h=-\infty}^{\infty} e^{-i\omega h} \text{Cov}(X_t, X_{t+h})
\]

(1.7)
A model-based estimator of the spectral density can be constructed if the observed data in time is assumed to be the realization of a process that can be parametrically specified. For example, if the process is assumed to be ARMA(p, q) (have p autoregressive and q moving average parameters), an estimator for the spectral density would be

\[ \hat{f}(\omega) = \frac{1}{2\pi} \left| \frac{\hat{\eta}(\exp(-i\omega))}{\hat{\phi}(\exp(-i\omega))} \right|^2, \]

where \( \hat{\eta}(\cdot) \) is the moving average polynomial evaluated at estimates of the moving average parameters, \( \hat{\eta} \), and \( \hat{\phi}(\cdot) \) is the autoregressive polynomial evaluated at estimates of the autoregressive parameters, \( \hat{\phi} \).

An alternative estimator to the spectral density is the periodogram. The periodogram (defined on the set of Fourier frequencies \( \left\{ \frac{2\pi j}{n} \right\}_{j=1,...,F} \), where \( [a] \) is the largest integer less than or equal to \( a \), and \( n \) is the sample size) is not model based, is inconsistent for the truth, and is calculated as

\[ I(\omega_j) = \frac{1}{2\pi n} \left[ \left( \sum_{t=1}^{n} X_t \sin(\omega_j t) \right)^2 + \left( \sum_{t=1}^{n} X_t \cos(\omega_j t) \right)^2 \right], \]

where \( \omega_j \) is the \( j^{th} \) Fourier frequency. Estimates of the spectral density are often calculated in smoothing the periodogram as in (1.6) (Daniell, 1946; Parzen, 1961; Blackman and Tukey, 1958; Beltrao, 1987). Several Bayesian methods have also been developed to estimate spectral densities. These methods either involve specifying a parametric time series model and then placing priors on the parameters (Huerta and West, 1999), or placing prior distributions on the smoothness and shape of the spectral density (Shaman, 1977; Dension et al., 2002). Estimators for the spectral density are then calculated from the corresponding posterior distributions. Chapter 2 of this dissertation derives an estimator for the spectral density, a single curve, using Bayesian methods. Chapters 3 and 4, however, focus on estimating multiple curves. Methods associated with multiple curves are briefly reviewed in the section below.

### 1.3 Estimating multiple curves

The problem of function estimation can also arise in more complicated settings. For example, when each individual in a sample is assumed to have a unique mean response curve, model (1.4) is inadequate. A model which allows for the variability of curves between subjects is

\[ Y_i(x) = f(x) + G_i(x) + \epsilon_i \]

where \( Y_i(x) \) is the response of individual \( i \) at all values of \( x \), \( G_i(x) \) is a random function (with mean 0) specific to subject \( i \) with realization \( g_i(x) \), and \( f(x) + g_i(x) \) is interpreted as the mean response curve for individual \( i \). Parametric and nonparametric methods can also be used to learn about these functions.
The most common parametric specification sets \( f(x) = \mathbf{x}_{F,i}^{T} \beta \) and \( G_i(x) = \mathbf{x}_{R,i}^{T} b_i \) where \( \mathbf{x}_{F,i} \) is a vector of fixed effect (often polynomial) covariates, \( \beta \) is a vector of coefficients, \( \mathbf{x}_{R,i} \) is a vector of random effect (often polynomial) covariates, and \( b_i \) is a vector of random effects specific to subject \( i \) typically modeled as \( b_i \sim N(0, \Sigma) \), and \( \epsilon \sim N(0, \sigma^2) \). The theoretical properties and methods associated with this linear mixed model have been thoroughly studied and reviewed in the literature (Harville, 1976, 1977; Bealle and Little, 1975; Dempster et al., 1977; Hayes, 1973; Patterson and Thompson, 1971). This linear mixed model has also been studied using Bayesian methods. These methods typically involve sampling from the posterior distribution of \( \beta \), \( \Sigma \), and \( \sigma^2 \) (Zeger and Kari, 1991; Lange et al., 1992; Gilks et al., 1993; McNeil and Gore, 1996).

If these two functions, \( f(x) \) and \( G_i(x) \), are assumed to be splines that can be identified with a set of knots, then learning about these functions becomes a problem of estimating the number and location of knots associated with each curve. A variety of models and techniques exist for problems such as these, and they allow for more flexibility than the methods associated with the parametric models discussed above (Crainiceanu et al., 2004; Brumback and Rice, 1996; Shi et al., 1996).

Some methods even go beyond trying to identify the form of \( G_i(\cdot) \); people often attempt to reduce the number of principal component curves associated with these random effects. The principal component curves associated with a \( p \times 1 \) random effect vector are \( p \) orthogonal curves

\[
\left( h_1(\cdot), h_2(\cdot), \ldots, h_p(\cdot) : \int h_k(x) h_l(x) \, dx = \mathbb{I}(k = l) \right)
\]

which, from \( h_1(\cdot) \) to \( h_p(\cdot) \), decreasingly account for the variability observed in the random effects. Equation (1.9) re-expresses equation (1.8) as a function of principal component curves.

\[
Y_{ij} = f(x_j) + \sum_{l=1}^{p} \sqrt{\lambda_l} \delta_{ij} h_l(x_j) + \epsilon_{ij}.
\] (1.9)

In (1.9) \( Y_{ij} \) is the observed value of subject \( i \) at \( x_j \), \( \lambda_l \) is the \( l^{th} \) largest eigenvalue of the random effects covariance matrix, \( h_l(\cdot) \) is the \( l^{th} \) principal component curve, and \( \delta_{ij} \sim N(0,1) \). Observe that the largest eigenvalue, \( \lambda_1 \), is paired with the first principal component curve, \( h_1(x) \). It thus follows that this principal component curve, \( h_1(\cdot) \), accounts for most of the variability observed in the random effects. The second principal component curve, \( h_2(\cdot) \), is paired with the second largest eigenvalue, \( \lambda_2 \). This curve thus accounts for the second largest amount of variability observed in the random effects. Figure 1.2 illustrates this idea. While the plot on top shows three orthogonal principal component curves, the plot on bottom displays five random effect curves, \( G_i(x) = (G_i(x_1), G_i(x_2), \ldots, G_i(x_{100})) \), which were generated using the function

\[
G_i(x_j) = 100 \times \delta_{ij} \times h_1(x_j) + \delta_{ij} \times h_2(x_j) + .1 \times \delta_{ij} \times h_3(x_j)
\] (1.10)
where $\delta_{ij} \sim N(0,1) \forall i,j$. Since $h_1(\cdot)$ is paired with an eigenvalue of 100 in equation (1.10), while $h_2(\cdot)$ and $h_3(\cdot)$ are paired with eigenvalues 1 and .1, $h_1(\cdot)$ should account for most of the variability in the random effect curves. Figure 1.2 makes this clear (the trends across the x-axes for $h_1(\cdot)$ and the realizations of $G_1(\cdot)$ are nearly identical). So in this case, if one wanted to minimize the number of principal component curves which explain the variance in the random effects, $h_2(\cdot)$ and $h_3(\cdot)$ would be eliminated. It is often not this easy, however, to identify those principal component curves which should be eliminated or retained. James et al. (2000), considering the model in (1.9), measured the amount of variability explained by the first $r$ principal components curves, and retained the first $r$ principal component curves which accounted for at least 90% of the variability observed in the random effects. Shi et al. (1996), however, retained the first $r$ principal component curves which minimized a cross-validation score. The third and fourth chapters of this dissertation study the model given in (1.8). $f(\cdot)$ and $G(\cdot)$ are modeled as regression splines in these chapters, and Bayesian methods are used to both identify the knots associated with these curves and to reduce the number of principal components associated with the random effects.

1.4 Detailed outline of dissertation

This dissertation focuses on Bayesian methods in function estimation. Chapter 2 focuses on estimating a particular function, the spectral density. The semiparametric estimator derived in this chapter combines a smoothed version of the periodogram (as in (1.6) with a parametric estimator of the spectral density (the parametric estimator assumes the observed data comes from an ARMA($p,q$) process). This semi-parametric estimator, which shrinks towards the parametric form provided it is correct, is derived from a hierarchical model. This estimator improves upon the Daniels and Cressie (2001) shrinkage estimator in that it is consistent, it is competitive with other estimators (as seen through simulation studies), and it does not require the specification of a parametric form.

The third and fourth chapters of this dissertation focus on the model given in (1.8). In both chapters, the functions $f(\cdot)$ and $G(\cdot)$ are modeled as splines with $k$ and $k'$ knots at locations $t_k$ and $t_{k'}$, respectively. The problem of knot selection is solved using Bayesian methods. In Chapter 3, reversible jump MCMC methods are used to sample from the posterior $p(k, t_k, k', t_{k'} | y)$. Sampling from such a posterior using reversible jump methods, however, requires evaluation of the marginal likelihood $p(y|k, t_k, k', t_{k'})$, which can not be calculated. Two sampling methods are thus considered in this chapter; these two methods correspond to two different approximations of $p(y|k, t_k, k', t_{k'})$ and are compared based on how effectively they penalize models with unnecessarily large values of $k'$. Regardless of which approximation is better, the methods introduced in this chapter are innovative in that several sets of knots, which fit the two curves well, are identified in an automated fashion.
In the fourth chapter, a similar posterior is considered. This posterior, however, relies on the decomposition of the random effect curve, $G_i(\cdot)$, into $p$ orthogonal principal component curves as in (1.9) and restricts the random effects curve to have the same knots as the fixed effect curve. Decomposing the random effect curve in such a fashion allows model (1.8) to be approximated with

$$Y_i(x_j) \approx f(x_j) + \sum_{l=1}^{r} \sqrt{\lambda_l} \delta_{ij} h_l(x_j) + \epsilon_{ij}$$

where $r < p$ and where $\{h_1(), \ldots, h_r()\}$ are the $r$ principal component curves which explain a majority of the variability observed in the random effects. In Chapter 4, the number of significant principal component curves, $r$, is identified by sampling from the posterior $p(k, t_k, r|y)$. The methods introduced in this chapter are especially appealing since the knots associated with the two curves, along with the number of effective principal component curves, are all sampled at once.

1.5 References


Figure 1.1 Four piecewise cubic splines. Knots are shown as *.
Figure 1.2 The top graph plots three principal component curves. (---): $h_1$, (— —): $h_2$, (· · · ·): $h_3$. The bottom graph displays five realized random effect curves.
2 A SHRINKAGE ESTIMATOR FOR SPECTRAL DENSITIES

Carsten H. Botts and Michael Daniels

A paper submitted to *Biometrika*.

2.1 Abstract

We propose a shrinkage estimator for spectral densities based on a multilevel normal hierarchical model. The first level captures the sampling variability via a likelihood constructed using the asymptotic properties of the periodogram. At the second level, the spectral density is shrunk towards a parametric time series model. To avoid selecting a particular parametric model for the second level, a third level is added which induces an estimator that averages over a class of parsimonious time series models. The estimator derived from this model, the model averaged shrinkage estimator (MASE), is consistent, is shown to be highly competitive with other spectral density estimators via simulations, and is computationally inexpensive.

2.2 Introduction

Data collected over time can be of interest for a variety of reasons. For example, one may wish to predict future values, or understand the cyclic behavior of a particular process. Regardless of what the final objective is, to answer almost any question related to a time series process, one must estimate its dependence structure. The spectral density of a time series completely captures this dependence structure. In fact, the covariance for any stationary time series \( \{X_1, X_2, \ldots, X_t, \ldots\} \) with innovation variance \( \sigma^2 \) can be calculated from its spectral density, \( f_X(\omega) \), using the formula

\[
\text{Cov}(X_t, X_{t+h}) = 2\sigma^2 \int_0^\pi f_X(\omega) \cos(\omega h) d\omega; \quad h = 0, 1, 2, 3, \ldots
\]  

(2.1)

The dependence structure of a time series can also be assessed by just looking at its spectral density. Spectral densities which peak at low frequencies, for example, reveal that at short lags, observations in
the time series are positively correlated. The opposite would be true for spectral densities which peak at high frequencies. Since the dependence structure in a time series can be so easily deduced from the spectral density itself, spectral density estimation has been an active area of research.

Statistical research in regression and nonparametric function estimation can easily be applied to the estimation of spectral densities. However, estimating the spectral density of a time series has attracted research all of its own. The primary motivation of this research has been the statistical properties of the periodogram, which is an inconsistent estimator of the spectral density. The periodogram is defined on a set of \( F \) frequencies (the Fourier frequencies = \( \frac{2\pi j}{n} \) \( j=1,\ldots,F=[\frac{n}{2}] \)) and is calculated as

\[
I(\omega_j) = \frac{1}{2\pi} \frac{1}{n} \left[ \left( \sum_{t=1}^{n} X_t^* \sin(\omega_j t) \right)^2 + \left( \sum_{t=1}^{n} X_t^* \cos(\omega_j t) \right)^2 \right],
\]

where \( X_t^* = X_t - \bar{X} \), \( n \) is the sample size, \( \omega_j = \frac{2\pi j}{n} \), and \( [a] \) is the largest integer less than or equal to \( a \). A typical correction for the periodogram's lack of consistency is to smooth the periodogram across frequencies. A large percentage of the research devoted to spectral density estimation focuses on how the periodogram should be smoothed (Daniell 1946; Blackman and Tukey, 1958; Hall et al., 1994; Parzen, 1961; Beltrao and Bloomfield, 1987; Ombao et al., 2001).

A variety of other estimation methods besides smoothing have also been explored. Pawitan and Gangopadhyay (1991) used linear model theory to derive an estimator of the spectral density. The linear model considered was \( \hat{f}_{\text{smooth}}(\omega) = f(\omega) + f''(\omega)x + e(f(\omega)) \) where \( \hat{f}_{\text{smooth}}(\omega) \) is a smoothed version of the periodogram, \( f(\omega) \) is the true spectral density, \( f''(\omega) \) its second derivative, and \( e(f(\omega)) \) is random error with a distribution depending on \( f(\omega) \). Estimating the parameters of this linear model gives an estimator of the spectral density. Taniguchi (1987) derived a minimum contrast estimator of the spectral density. His estimator of the spectral density minimizes the distance \( \int_{0}^{\pi} K(f_0(\omega)/\hat{f}(\omega))d\omega \) between \( f_0(\omega) \), a parametric approximation to the true spectral density, and the estimator, \( \hat{f}(\omega) \).

Others have tried to estimate the spectral density using Bayesian methods. Shaman (1977) constructed a Bayesian estimator which allows one to incorporate prior information regarding the smoothness and shape of the spectral density. Huerta and West (1999) considered spectral densities of autoregressive (AR) processes; they placed priors on the roots of AR polynomials and then sampled from the corresponding posterior distributions to get estimates of the spectral density. Denison et al. (2002) fit piecewise polynomials to the log periodogram. The piecewise polynomials they fit were defined with a given set of knots at specific locations. They placed priors on the knots, the locations of the knots, and additional parameters in the piecewise polynomial. They then used reversible jump MCMC methods to sample from their posterior distributions.

A shrinkage estimator of the spectral density, based on a two-level hierarchy that requires one to
specify a parametric time series model, was proposed by Daniels and Cressie (2001). Their estimator, which we will call the DC estimator, shrinks toward the parametric estimate of the spectral density. Although the ideas behind this shrinkage estimator are appealing, the asymptotic properties of this estimator have not been explored, it has not been studied via simulations, and it requires the selection of a parametric time series model.

The estimator studied in this paper builds on the ideas in Daniels and Cressie (2001). We derive a consistent shrinkage estimator of the spectral density similar to that of the DC estimator. We also avoid selection of a particular parametric time series model by averaging over a class of parametric models.

Section 2.3 of this paper describes the DC estimator and the model averaged shrinkage estimator (MASE) in detail. In Section 2.4, the simulations and computations used to assess the estimator are described. Section 2.5 presents the results of the simulations, and in Section 2.6, our methods are applied to a real data set.

2.3 Estimator

In this section, we present the model from which the shrinkage estimator is derived. The original model and methods of Daniels and Cressie are reviewed in Section 2.3.1. This section includes how the estimator was derived and why the asymptotic properties of the DC estimator might be of concern. In Section 2.3.2, we present a new shrinkage estimator and propose a strategy to avoid choosing a particular member of a parametric family. We also examine the feasibility of a fully Bayesian analysis. Finally, we show that this shrinkage estimator can be evaluated at any set of frequencies on the interval $(0, \pi)$, and that at those frequencies, the shrinkage estimator is consistent for the true spectral density.

2.3.1 The DC estimator

Daniels and Cressie constructed a two-level hierarchical model from which their shrinkage estimator was derived. The first level models the stochastic behavior of the periodogram, while the second requires the specification of a parametric time series model. The asymptotic properties of this estimator have not been extensively studied.

The first level of the hierarchical model describes the random behavior of the periodogram (2.2). Asymptotically, the periodogram of any time series has a $\chi^2_2$ distribution (the last periodogram ordinate has a $\chi^2_1$ distribution if $n$ is even). Since the fourth root of a $\chi^2_2$ random variable is approximately normally distributed (Hawkins and Wixley, 1986),

$$M_j I(\omega_j)^{\frac{1}{4}} \sim N(f(\omega_j)^{\frac{1}{2}}, V_j f(\omega_j)^{\frac{1}{2}}),$$ (2.3)
where for $n$ odd, $M_j = 2^{\frac{1}{2}} \Gamma(1.25)$ and $V_j = 2^{\frac{1}{2}} \Gamma(1.5) - M_j^2$ for $j = 1, 2, \ldots, F$ with $F$ = the number of Fourier frequencies, and for $n$ even, $M_j = 2^{\frac{1}{2}} \Gamma(1.25)$ and $V_j = 2^{\frac{1}{2}} \Gamma(1.5) - M_j^2$ for $j = 1, 2, \ldots, F - 1$, and $M_F = 2^{\frac{1}{2}} \frac{\Gamma(0.75)}{\Gamma(0.5)}$ and $V_F = 2^{\frac{1}{2}} \Gamma(0.5) - M_F^2$. Diggle and al Wasel (1997) used the asymptotic distribution of the periodogram to construct an approximate likelihood, but without the fourth root transformation.

In the second level of the hierarchical model, a prior was placed on the fourth root of the true spectral density at all of the Fourier frequencies. This prior can be written as

$$f(\omega)^{\frac{1}{4}} = \begin{pmatrix} f(\omega_1)^{\frac{1}{4}} \\ f(\omega_2)^{\frac{1}{4}} \\ \vdots \\ f(\omega_F)^{\frac{1}{4}} \end{pmatrix} \sim N \left( \begin{pmatrix} f_p(\omega_1; \theta)^{\frac{1}{4}} \\ f_p(\omega_2; \theta)^{\frac{1}{4}} \\ \vdots \\ f_p(\omega_F; \theta)^{\frac{1}{4}} \end{pmatrix}, \tau^2 \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \right)$$

The mean of the prior given in (2.4) is the spectral density of an ARMA(p,q) time series with unknown parameters $\theta$. Note that $\theta$ is a vector which contains all of the parameters that specify a spectral density. For an ARMA(p,q) time series, $\theta = (\psi, \sigma^2)$ where $\sigma^2$ is the innovation variance, and $\psi = (\phi, \eta)$ where $\phi$ are the autoregressive parameters and $\eta$ are the moving average parameters. Throughout this paper, $f_p(\omega; \theta)$ will be denoted as $f_p(\omega; \psi, \sigma^2)$ when appropriate. $\tau^2$ measures the uncertainty of the true spectral density around the ARMA model. The motivation behind this hierarchy was to obtain improved estimation in small samples by shrinking towards a parametric form (see, e.g., Chen, 1979; Daniels and Kass, 1999).

With this model, the posterior distribution of the true spectral density at the Fourier frequencies, conditional on $\tau^2$ and $\theta$, is easy to calculate. The mean of the conditional posterior, $p \left( f(\omega)^{\frac{1}{4}} | \theta, \tau^2, I(\omega) \right)$, is used as the estimator of the spectral density's fourth root at each of the Fourier frequencies. This mean is given below in equation (2.5),

$$E \left( f(\omega_j)^{\frac{1}{4}} | \hat{\theta}, \hat{\tau^2}, I(\omega) \right) = \frac{V_j \left( M_j I(\omega_j)^{\frac{1}{4}} \right)^2}{V_j \left( M_j I(\omega_j)^{\frac{1}{4}} \right)^2 + \hat{\tau^2}} f_p(\omega_j; \hat{\theta})^{\frac{1}{4}} + \frac{\hat{\tau^2}}{V_j \left( M_j I(\omega_j)^{\frac{1}{4}} \right)^2 + \hat{\tau^2}} M_j I(\omega_j)^{\frac{1}{4}} \quad (2.5)$$

where $\hat{\tau^2}$ was estimated using a method-of-moments estimator, and $\hat{\theta}$ was calculated using maximum likelihood estimation. The estimator given in (2.5) is a weighted sum of the periodogram and the parametric estimate both on the fourth root scale. Daniels and Cressie used the fact that the posterior distribution of $f(\omega)^{\frac{1}{4}} | \theta, \tau^2, I(\omega)$ was normal when deriving their estimator for the true spectral density on the original scale, $f(\omega)$. Their estimator for the true spectral density is the fourth moment of a normal random variable with approximate mean $m = E \left( f(\omega_j)^{\frac{1}{4}} | \hat{\theta}, \hat{\tau^2}, I(\omega) \right)$ and variance $\nu^2 = \left( \frac{1}{V_j} \left( M_j I(\omega_j)^{\frac{1}{4}} \right)^2 + \frac{1}{\hat{\tau^2}} \right)^{-1}$. 

The fourth moment of this random variable is

\[ \hat{f}_{\text{DC}}(\omega_j) = \mathbb{E} \left( \left( f(\omega_j)^{\frac{1}{4}} \right)^4 \right) = 3\nu^4 + 6\nu^2m^2 + m^4. \tag{2.6} \]

Although the asymptotic properties of this estimator have not been carefully studied, this estimator is clearly not consistent when the true model, \( f_p(\omega; \theta) \), is incorrectly specified. To understand the behavior of the estimator when \( f_p(\omega; \theta) \) is correctly specified, we first examine their estimator of \( \tau^2 \). DC calculated \( \tau^2 \) as

\[ \hat{\tau}^2 = \max \left( 0, \sum w_j \left( M_j I(\omega_j)^{\frac{1}{4}} - f_p(\omega_j; \theta)^{\frac{1}{4}} \right)^2 \left( \sum w_j - \left( \sum w_j^2 \right) \left( \sum w_j \right)^{-1} \right)^{-1} \right), \]

where \( w_j \), the weight at frequency \( \omega_j \), should ideally be \( \left( \text{Var}(M_j I(\omega_j)^{\frac{1}{4}}) \right)^{-1} \). Daniels and Cressie estimated the weights at Fourier frequency \( \omega_j \) using \( V_j \left( M_j I(\omega_j)^{\frac{1}{4}} \right)^2 \), which is inconsistent for \( V_j f(\omega_j)^{\frac{1}{4}} \).

It should be noted, however, that if \( w_j \xrightarrow{\text{prob}} \left( \text{Var}(M_j I(\omega_j)^{\frac{1}{4}}) \right)^{-1} \), and if \( f_p(\omega; \theta) \) is correctly specified, the DC estimator is consistent.

**Theorem 1.** Provided \( w_j \xrightarrow{\text{prob}} \left( \text{Var}(M_j I(\omega_j)^{\frac{1}{4}}) \right)^{-1} \), \( \hat{f}_{\text{DC}} \) is consistent for \( f \) if the true model, \( f_p(\omega; \theta) \), is correctly specified. (See Appendix for Proof)

In the next section, we propose a modification to this estimator so that it is consistent regardless of the choice of \( f_p(\omega; \theta) \).

### 2.3.2 New estimator

We modify the methods of Daniels and Cressie in several respects. In the first level of the hierarchy, we minimally smooth the fourth root of the periodogram and adjust the likelihood in (2.3) accordingly. Second, we add an additional level to the hierarchy by placing a prior distribution on \( \tau^2 \). Section 2.3.2.2 discusses the prior in detail. Third, we avoid selection of a particular parametric model for the estimator by model averaging. This is discussed in Section 2.3.2.1.

We begin by replacing the periodogram with a minimally smoothed nonparametric estimate of the spectral density’s fourth root at each of the Fourier frequencies,

\[ \hat{f}_s(\omega_j)^{\frac{1}{4}} = \sum_{k=1}^{F} W_h(\omega_k - \omega_j) M_k I(\omega_k)^{\frac{1}{4}} \tag{2.7} \]

where \( W_h(\omega_k - \omega_j) = K_h(\omega_k - \omega_j) \left( \sum_{k=1}^{F} K_h(\omega_k - \omega_j) \right)^{-1} \), \( \omega_k \) is the \( k \)th Fourier frequency, and \( K_h(\cdot) \) is a normal kernel with bandwidth proportional to \( n^{-\frac{1}{2}} \). Notice that this is nearly the smallest bandwidth we can choose while maintaining the consistency of our estimator (see Theorem 3). Our motivation for
doing this should be clear: smooth the periodogram as little as possible so as not to lose sudden and high peaks. With this smoothing, the likelihood in (3) is replaced with

$$
\hat{f}_s(\omega) \frac{1}{2} = \left( \begin{array}{c}
\hat{f}_s(\omega_1) \\
\hat{f}_s(\omega_2) \\
\vdots \\
\hat{f}_s(\omega_F) 
\end{array} \right) \sim N \left( \left( \begin{array}{c}
f(\omega_1) \\
f(\omega_2) \\
\vdots \\
f(\omega_F) 
\end{array} \right), WVW' \right),
$$

where $W = [W_h(\omega_k - \omega_j)]_{k=1,\ldots,F}^{j=1,\ldots,F}$, $V = \text{diag} \left( \text{Var}(\hat{f}_s(\omega_1) \frac{1}{2}), \text{Var}(\hat{f}_s(\omega_2) \frac{1}{2}), \ldots, \text{Var}(\hat{f}_s(\omega_F) \frac{1}{2}) \right)$, 
\[\text{Var} \left( \hat{f}_s(\omega_j) \frac{1}{2} \right) = V_j \hat{f}_s(\omega_j) \frac{1}{2},\text{ and } \hat{f}_s(\omega_j) \frac{1}{2} = \left( \sum_{k=1}^{F} W_h(\omega_k - \omega_j) M_k I(\omega_k) \frac{1}{2} \right)^2.\] We treat the likelihood above as an actual likelihood. This is similar in spirit to the modified likelihoods constructed by Sun et al. (2000) on pre-whitened residuals. We also smooth the periodogram in our estimate of the variance. This is done to ensure that our estimator of the covariance matrix, $WVW'$, is consistent (see Theorem 3). As an additional benefit, we expect this approximate likelihood to be more accurate than (3) due to the averaging (smoothing).

The prior distribution of $f(\omega) \frac{1}{2}$ evaluated at all the Fourier frequencies is given in (2.4). This prior should actually be a truncated multivariate normal distribution (with truncation at 0). However, we replace this truncated distribution with a non-truncated multivariate normal distribution to derive the estimator. We also place a prior distribution on $\tau^2, p(\tau^2)$, to exert additional control over the amount of shrinkage towards $f_p(\omega; \theta)$.

With these distributions, the conditional posterior mean of $f(\omega) \frac{1}{2}|\theta, \tau^2, \hat{f}_s(\omega)$ becomes

$$
E \left( f(\omega) \frac{1}{2}|\theta, \tau^2, \hat{f}_s(\omega) \right) = \Sigma + \tau^2 I)^{-1} f_p(\omega; \theta) \frac{1}{2} + \tau^2 (\Sigma + \tau^2 I)^{-1} \hat{f}_s(\omega) \frac{1}{2},
$$

where $\Sigma = WVW'$. Our estimator for the fourth root under squared error loss is then

$$
\hat{f}(\omega) \frac{1}{2} = \text{E}(f(\omega) \frac{1}{2}|\theta = \hat{\theta}, \tau^2 = \hat{\tau}^2, \hat{f}_s(\omega)),\text{ where } \hat{\tau}^2 \text{ is the mode of } \log(\text{Lik}(\tau^2|\theta = \hat{\theta}, I(\omega))p(\tau^2)) \text{ and } \hat{\theta} \text{ is the mle of the ARMA(p,q) process fit on original data.}\text{ The variance of this estimator is given by}
$$

$$
\text{Var} \left( f(\omega) \frac{1}{2}|\sigma^2, \tau^2 \right) = \Sigma (\Sigma + \tau^2 I)^{-1} \text{Var}(f_p(\omega; \hat{\psi}, \sigma^2) \frac{1}{2}) (\Sigma (\Sigma + \tau^2 I)^{-1})^T + \tau^2 I (\Sigma + \tau^2 I)^{-1} \text{Var}(\hat{f}_s(\omega) \frac{1}{2}) (\tau^2 I (\Sigma + \tau^2 I)^{-1})^T,
$$

where $\hat{\psi} = \left( \hat{\phi}, \hat{\eta} \right)$, 
\[\text{Var}(f_p(\omega; \hat{\psi}, \sigma^2) \frac{1}{2}) = R \sigma^2 \left[ \begin{array}{cc}
\text{EU}_tU_t' & \text{EU}_tV_t' \\
\text{EV}_tU_t' & \text{EV}_tV_t'
\end{array} \right]^{-1}RT,
\]
\[R = \left[ \frac{\partial f_p(\omega_1; \theta) \frac{1}{2}}{\partial \hat{\psi}}, \ldots, \frac{\partial f_p(\omega_F; \theta) \frac{1}{2}}{\partial \hat{\psi}} \right]'.\]
\textbf{U}_t = (U_t, U_{t-1}, \ldots, U_{t+1-p})', \quad \textbf{V}_t = (V_t, V_{t-1}, \ldots, V_{t+1-q})', \quad \text{and} \quad \{U_t\}, \{V_t\} \text{ are the processes } U_t = \sum_{i=1}^{p} \phi_i U_{t-i} + Z_t \quad \text{and} \quad V_t = \sum_{i=1}^{q} \eta_i V_{t-i} + Z_t, \quad \text{with } Z_t \sim N(0, \sigma^2) \quad \text{(Brockwell and Davis, 1990).}

Our estimator of \( f(\omega) \), \( \hat{f}(\omega) \), takes a similar form to that of Daniels and Cressie. Rather than use the fourth moment of a univariate normal distribution as our estimator for \( f \), we use the fourth moment of the multivariate normal random vector \( f(\omega)^\frac{1}{2} \theta, \tau^2, \hat{f}_s(\omega) \), which is

\[
\hat{f}(\omega) = 3\bar{\nu}^2 + 6\bar{\nu}^2 \bar{m}^2 + \bar{m}^4,
\]

where \( \bar{\nu} \) and \( \bar{m} \) are \( F \times 1 \) vectors such that \( \bar{\nu} = \text{diag}(\Sigma^{-1} + \frac{1}{2}I_{F \times F}) \) and \( \bar{m} = \mathbb{E}(f(\omega)^\frac{1}{2} \theta, \tau^2, \hat{f}_s(\omega)). \)

The variance of this fourth moment can then be computed as

\[
\text{Var}(\hat{f}(\omega)) = 144 \times \text{diag}(\bar{\nu}_1^2 \bar{m}_1^2, \ldots, \bar{\nu}_F^2 \bar{m}_F^2) \text{Var}(f(\omega)^\frac{1}{2}) + 16 \times \text{diag}(\bar{m}_1^6, \ldots, \bar{m}_F^6) \text{Var}(\hat{f}(\omega)^\frac{1}{2}).
\]

\textbf{2.3.2.1} Choice of \( f_p(u; \theta) \)

The shrinkage estimator derived in the previous section shrinks the nonparametric spectral density estimate (the smoothed periodogram) to some parametric ARMA\( (p,q) \) model. But which values of \( p \) and \( q \) should be selected? This has not been addressed in other work with these types of shrinkage priors for dependence (Chen 1979; Daniels and Kass, 1999 and 2001; and Daniels and Pourahmadi 2002). To avoid having to answer this question, we propose an estimator which averages over a class of parametric models.

The estimator we propose averages over a class of ARMA\( (p,q) \) models where \( p \in \{0,1,2,3\} \) and \( q \in \{0,1,2,3\} \). We limit the number of autoregressive and moving average parameters to three because we would like to study our estimator in small samples. However, any class of parsimonious models would do. This model averaged shrinkage estimator (MASE) can be written as

\[
\hat{f}_{\text{avg}}(\omega) = \sum_{\text{all } p,q} p(p,q|\hat{f}_s(\omega)) \hat{f}_{p,q}(\omega),
\]

where \( \hat{f}_{p,q}(\omega) \) denotes the shrinkage estimator derived from a model with \( p \) autoregressive and \( q \) moving average parameters, and \( p(p,q|\hat{f}_s(\omega)) \) denotes the corresponding posterior probability. With the MASE written as it is in (2.10), it is clear that averaging in this way adds an additional level of smoothing to the estimator given in (2.9). For the terms in (2.10), it should be noted that

\[
p(p,q|\hat{f}_s(\omega)) = p(\hat{f}_s(\omega)|p,q)p(p,q) \left( \sum_{p,q} p(\hat{f}_s(\omega)|p,q)p(p,q) \right)^{-1}
\]

where
\begin{equation}
p(\hat{f}_s(\omega) | p, q) = \int_{\mathbb{R}^p} \int_{0}^{\infty} \int_{0}^{\infty} \text{Lik}(\psi, \sigma^2, \tau^2 | I(\omega)) p(\psi) p(\sigma^2) p(\tau^2) d\sigma^2 d\tau^2 d\psi,
\end{equation}

$p(p, q)$ is a prior placed on $p$ and $q$, and

\begin{equation}
\text{Lik}(\psi, \sigma^2, \tau^2 | I(\omega)) = \left( \frac{1}{(2\pi)^{\frac{p+q}{2}}} \right)^{\frac{1}{2}} \times \exp \left( -\frac{1}{2} \left( \tilde{f}_s(\omega) - f_p(\omega; \psi, \sigma^2) \right)^T (\tau^2 I + \Sigma)^{-1} \left( \tilde{f}_s(\omega) - f_p(\omega; \psi, \sigma^2) \right) \right) \cdot 
\end{equation}

The marginal likelihood of $p$ and $q$ is calculated using a Laplace approximation. Specifically, we let

\begin{equation}
p \left( \hat{f}_s(\omega) | p, q \right) = \text{Lik}(\psi, \sigma^2, \tau^2 | I(\omega)) p(\psi) p(\tau^2) p(\sigma^2) (2\pi)^{\frac{p+q}{2}} \times \left| \left[ \frac{\partial^2 \log \text{Lik}(\psi, \sigma^2, \tau^2)}{\partial (\psi, \sigma^2, \tau^2)} \right]^{-1} \right|^{\frac{1}{2}}_{\psi = \hat{\psi}, \sigma^2 = \hat{\sigma}^2, \tau^2 = \hat{\tau}^2},
\end{equation}

where $p(\psi, \sigma^2, \tau^2) = p(\psi) p(\sigma^2) p(\tau^2)$, $\hat{\psi}$ is the maximum likelihood estimates of the ARMA parameters, $\hat{\sigma}^2$ is the maximum likelihood estimator of the innovation variance, and $\hat{\tau}^2$ is the mode of the posterior distribution $p(\tau^2 | \psi = \hat{\psi}, \sigma^2 = \hat{\sigma}^2, I(\omega))$.

A variety of priors can be placed on the number of autoregressive and moving average parameters, $p$ and $q$. The prior we consider places a majority of its weight on the lower order models. In fact, we let $p(p, q) = p(p)p(q)$ where $p(p)$ and $p(q)$ are both truncated Poisson distributions with $\lambda = 1.5$. Other priors that can be placed may be flat, or may place little weight on small values of $p$ and $q$ (favoring more complex models).

### 2.3.2.2 Priors on $\psi$, $\sigma^2$, and $\tau^2$

In the model above, flat priors are implicit on the ARMA parameters, $\psi$ and innovation variance $\sigma^2$. That is,

\begin{equation}
p(p) \propto \mathbb{I} \{ p \in \mathbb{R}^{p+q} \} \quad \text{and} \quad p(\sigma^2) \propto \mathbb{I} \{ \sigma^2 \in (0, \infty) \}.
\end{equation}

We consider two priors for $\tau^2$; $p(\tau^2) \propto (\tau^2)^{-1}$ which, of course, has more mass towards 0, and $p(\tau^2) \propto 1$ which, compared to the prior given above, should shrink less towards the selected parametric form.

Although we use an empirical Bayesian approach to compute our estimator, a fully Bayesian analysis can be performed. We show that when shrinking towards an ARMA$(p,0)$ model, if $p(\psi)$ is proper and $p(\sigma^2)$ and $p(\tau^2)$ are flat on the positive real line, $p(\psi, \sigma^2, \tau^2 | I(\omega))$ is proper, and a fully Bayesian analysis can be done. If $p(\tau^2) \propto (\tau^2)^{-1}$, however, $p(\psi, \sigma^2, \tau^2 | I(\omega))$ is improper.
Theorem 2. Consider an ARMA(p,0) model where the AR parameters are denoted \( \phi = \{\phi_1, \ldots, \phi_p\} \). If both \( p(\sigma^2) \) and \( p(\tau^2) \) are flat on the positive real line, and the two conditions below are met, \( p(\phi, \sigma^2, \tau^2|I(\omega)) \) is proper. (see Appendix for Proof)

1. \( \int_{\mathbb{R}^p} p(\phi) d\phi < \infty \)
2. \( \int_{\mathbb{R}^p} |\phi_j \phi_k| p(\phi) d\phi < \infty \quad \forall \text{ pairs } j, k. \)

This provides justification for doing a fully Bayesian analysis with these priors. It should be noted that if independent normal priors are placed on the autoregressive parameters, the two conditions listed above are satisfied. For an ARMA(p,q) model with \( q \geq 1 \), propriety of \( p(\psi, \sigma^2, \tau^2|I(\omega)) \) remains unresolved.

2.3.2.3 Consistency and evaluation of the estimator at any set of frequencies

As developed so far, we obtain an estimator for the spectral density at only the Fourier frequencies. However, it is straightforward to compute an estimate of the spectral density at any set of frequencies. Suppose we choose a set of \( K \) frequencies on the interval \((0, \pi)\). Calling the new set of frequencies \( \omega_l = \{\omega_{l1}, \omega_{l2}, \ldots, \omega_{lK}\} \), one can write an estimator for \( f(\omega) \frac{1}{i} \) at \( \omega_l \) as

\[
\tilde{E}(f(\omega_l) \frac{1}{i} | \hat{\theta}, \hat{\tau}^2, \hat{f}_s(\omega)) = \tilde{\Sigma}_l (\tilde{\Sigma}_l + \hat{\tau}^2 I)^{-1} f_p(\omega_l; \hat{\theta}) \frac{1}{i} + \hat{\tau}^2 I (\tilde{\Sigma}_l + \hat{\tau}^2 I)^{-1} \hat{f}_s(\omega_l) \frac{1}{i},
\]

where \( \tilde{\Sigma}_l \) is a \( K \times K \) symmetric matrix with \( \sum_{k=1}^P W_h(\omega_k - \omega_l_i) W_h(\omega_k - \omega_l_j) V_k \hat{f}_s(\omega_k) \frac{1}{i} \) as its \((t, l)\)th entry, \( \hat{f}_s(\omega_l) \frac{1}{i} \) is a \( K \times 1 \) vector with \( \sum_{k=1}^P W_h(\omega_k - \omega_l_i) M_k I(\omega_k) \frac{1}{i} \) as its \( l \)th entry, \( \hat{\tau}^2 \) is that value estimated from the Fourier frequencies, and recall that \( W_h(\omega_k - \omega_l_i) = K_h(\omega_k - \omega_l_i) \left( \sum_j K_h(\omega_j - \omega_l_i) \right)^{-1} \). The estimator for the true spectral density then becomes \( \tilde{f}(\omega_l) \) where

\[
\tilde{f}(\omega_l) = 3\tilde{\nu}_l + 6\hat{\tau}^2 \tilde{m}_l + \tilde{m}_l^4,
\]

where \( \tilde{\nu}_l = \text{diag} \left( \tilde{\Sigma}_l^{-1} + \frac{\hat{\tau}^2}{2} I_{K \times K} \right)^{-1} \), \( \tilde{m}_l = E(f(\omega_l) \frac{1}{i} | \hat{\theta}, \hat{\tau}^2, \hat{f}_s(\omega)) \), and \( \hat{\tau}^2 \) and \( \hat{\theta} \) are, again, the maximum likelihood estimates calculated from the original data (i.e. the 'smoothed' periodogram at the Fourier frequencies). \( \tilde{f}(\omega_l) \) is consistent for the true spectral density at any set of frequencies, \( \omega_l \).

Theorem 3. Let \( \omega_l \) be a set of \( K \) frequencies (not necessarily Fourier frequencies) on the interval \((0, \pi)\). This set can be written as \( \omega_l = \{\omega_{l1}, \ldots, \omega_{lK}\} \). Assume that

(i) \( f(\omega) \) is bounded, and \( f''(\omega) \) is continuous on the interval \([0, \pi]\),

(ii) The kernel, \( K_h \), is symmetric about \( \theta \) and supported on \([-\pi, \pi]\),
(iii) The bandwidth, \( h = h_n \), is a sequence satisfying \( h_n \rightarrow 0 \) and \( nh_n \rightarrow \infty \), and

(iv) \( h \pi < \omega_1 < \omega_{1K} < \pi - \pi h \) for all \( n \) after a particular value of \( n_0 \).

Then regardless of the parametric structure specified, \( \hat{f}(\omega) \), given in (2.12), is consistent for \( f(\omega) \).

(see Appendix for Proof)

In the next section, we study the non-large sample properties of the estimator.

2.4 Simulations and computations

To explore the behavior of the MASE in small to medium sample sizes, we simulated 500 realizations from eight different time series models at three different sample sizes (\( n = 32, 64, \) and \( 128 \)). For each realization, the MASE and seven other spectral density estimators were calculated. The mean integrated squared error, \( \text{MISE} \), and the mean maximum squared deviation, \( \text{MMSD} \), of each of these estimators are calculated. These are calculated using

\[
\text{MISE} = \frac{1}{n_s} \sum_{k=1}^{n_s} \left[ \frac{1}{n_K} \sum_{i=1}^{n_K} \left( \hat{f}(\omega_i) - f(\omega_i) \right)^2 \right] \quad \text{and} \quad \text{MMSD} = \frac{1}{n_s} \sum_{k=1}^{n_s} \max_{\omega_i \in \Omega} \left\{ \left( \hat{f}(\omega_i) - f(\omega_i) \right)^2 \right\},
\]

where \( n_s \) is the number of simulations performed, \( n_K \) is the number of frequencies at which the estimator is evaluated, and \( \Omega \) is the set of frequencies at which the estimator is evaluated. The \( \text{MISE} \) is calculated to measure how each estimator performs in estimating an entire spectral density, while the \( \text{MMSD} \) measures the largest deviation and is introduced to see how each estimator captures peaks.

Of the eight time series simulated from, five were true stationary ARMA processes, and three were not. The time series were all \( n \times 1 \) zero-mean multivariate normal random vectors with covariances derived from a given spectral density. For the ARMA time series processes, the data were simulated using the fracdiff.sim procedure in R. Figure 2.1 shows the spectral densities of the eight time series simulated from. The first five of these time series are true ARMA models. The general model is listed below, and the parameters we considered are listed in Table 2.1.

\[
X_t - \sum_{j=1}^{p} \phi_j X_{t-j} = Z_t + \sum_{i=1}^{q} \eta_i Z_{t-i} \quad \text{where} \quad Z_t \sim N(0,1)
\]

The last three time series simulated from were not ARMA processes. Their spectral densities are given in Table 2.2 (note: \( \text{dnorm}(\mu_0, \sigma_0^2) = \left( \frac{1}{2\pi\sigma_0^2} \right)^{\frac{3}{2}} \exp \left( -\frac{1}{2} \frac{(\omega-\mu_0)^2}{\sigma_0^2} \right) \)).

In calculating the MASE, a total of 16 parametric models were fit to the data (ARMA(p,q) for \( p \in \{0,1,2,3\} \) and \( q \in \{0,1,2,3\} \)). The maximum likelihood estimators for all models were calculated
using the arima0 function in the tseries package of R. We compared the MASE to several other spectral density estimators including: (1) the periodogram, (2) an adaptive bandwidth smoother which was calculated using the glkerns function in R (Gasser et al., 1986, 1991; and Hermann, 1997), (3) the maximum likelihood estimator under each of the ARMA models, (4) a wavelet estimator, (5) the DC estimator (given in equation (2.6)), (6) the shrinkage estimator (given in equation (2.12)) which had fit the model with the lowest AIC returned by the arima0 function in R, and (7), the shrinkage estimator (given in equation (2.12)) which had the highest posterior probability, given in (2.11). The adaptive bandwidth smoother, which is calculated through the 'glkerns' function, is a kernel regression estimate of the spectral density of the form

$$\hat{f}_{\text{Adpt Bnd}}(\omega; b) = \frac{1}{2\pi} \sum_{i=1}^{F} w_i \left( \frac{\omega_i - \omega}{b} \right) I(\omega_i),$$

where $b$ is the bandwidth and $w(\cdot)$ is a polynomial kernel. In this case, the bandwidth is selected to minimize the asymptotic mean integrated squared error $E \left[ \int_0^\pi \left( f(\omega) - \hat{f}_{\text{Adpt Bnd}}(\omega; b) \right)^2 d\omega \right]$, with the heteroskedasticity of the periodogram taken into account. The wavelet estimator is calculated using the methods outlined in Percival and Walden (2000), and is computed using the functions available in the R package 'wavethresh' (For $F = 16, 32, 64$, the values of $J_\alpha$ were set to 4, 5, and 6, respectively. Alternative values of $J_\alpha$ either could not be calculated using the functions available, or returned estimates with higher $MISE$s and $MMSSD$s.) The wavelet estimator was then evaluated at the frequencies of interest using the 'approx' function in R.
2.5 Results

The mean integrated squared error (MISE) and mean maximum squared deviation (MMSD) of all the estimators considered at all the sample sizes are shown in Tables 2.4 - 2.9. These tables give the MISE and MMSD of the periodogram (Prdg), the DC estimator which shrinks to the correct parametric model (DC), the Wavelet estimator (Wvlt), the Adaptive estimator (Adapt), the Maximum Likelihood (ML) and Shrinkage (Shrnk) estimators which correspond to the correct parametric model, and the model averaged shrinkage estimator (MASE). In Figure 2.2, true spectral densities 3, 4, and 8 are plotted with the MASE, Wavelet, and Adaptive estimators averaged over five randomly selected samples at $n = 128$. Figure 2.2 intends to reveal, visually, how the MASE compares in capturing the peaks of a spectral density. Throughout this discussion, the ML, Shrinkage, and DC estimators will be referred to as model dependent estimators; the others, model independent estimators. On average, the MASE that we are proposing outperforms the model independent estimators. Although it loses in only a few of the simulated cases shown, the MASE never consistently loses to the same estimator.

At $n = 32$, the MASE appears to be the best at estimating an entire spectral density and capturing its peaks. With the exception of spectral density 4 and 8, the MISE and MMSD of the MASE are smaller than those of all the other model independent estimators (note that the Adaptive estimator beats the MASE in these cases). The Maximum Likelihood and the Shrinkage estimators occasionally outperform the MASE, but this happens because the model being fit is the correct structure (which, in practice, is unknown).

At sample sizes 64 and 128, the results are similar. The MASE outperforms all of the other model independent estimators, except at spectral densities 4 and 8. For spectral density 8 the MASE has an MISE and MMSD higher than the Adaptive estimator, and for spectral density 4 it loses to the Wavelet estimator. The Maximum Likelihood and Shrinkage estimators exhibit similar behavior at $n = 64$ and $n = 128$ as they did at $n = 32$. They tend to do well when the correct model is fit to the data. On spectral density 3 at sample size 64, for example, the Maximum Likelihood and Shrinkage estimators significantly outperform any of the other estimators. It is unfair, however, to compare these estimators to the MASE since they depend on knowing the true parametric structure of the time series.

Figure 2.2 shows that the MASE performs just as well (if not better) as the Wavelet and Adaptive estimator at capturing peaks. Spectral densities 3, 4, and 8 are the “spikiest” of the eight spectral densities considered and are the only spectral densities where the Adaptive and Wavelet estimators significantly compete with the MASE (notice that the Wavelet estimator performs poorly on the smoother spectral densities). At $n = 128$, the MASE locates and estimates the peaks in each spectral density at least as well as the others. For spectral density 4, all three estimators behave equivalently. For spectral densities 3 and 8, the Wavelet estimator tends to overestimate the height of each peak, while the
Adaptive estimator underestimates the height of each peak. The MASE, in both of these cases, captures the magnitude of each peak correctly.

2.6 Example

We used the MASE, Adaptive, and Wavelet estimators to estimate the spectral density of a few randomly selected patients involved in a medical experiment. The experiment tried to relate clinical depression to the frequency at which luteinizing hormone (LH) is released in the blood stream. The concentration of LH in the blood fluctuates with time, and Grambsch et al. (2002) hypothesized that the frequency at which this concentration fluctuates differs between those women who are clinically depressed and those who are not. An experiment to test this hypothesis was conducted. The sample taken in this study includes 26 women who were diagnosed as “clinically depressed” and 24 women who were not. Blood was taken every 10 minutes for 8 hours from each woman, and the concentration of LH was recorded at each time (this gives 49 observations in time for each subject).

In our analysis, we estimated the spectral density of nine randomly selected subjects using the first 32 observations in time (this is the largest sample size which allows the Wavelet estimator to be calculated). In these nine cases, we used each spectral density estimator to predict the next five observations. The estimators corresponding to one of these subjects is shown in Figure 2.3. Recall that in this figure, the estimates can not be evaluated on how well they estimate the true spectral density, since this is not known. The mean predicted squared errors of these three estimators (the MASE, Adaptive, and Wavelet estimators) and the periodogram are shown below in Table 2.3. On average, the MASE and the Wavelet estimator do the best at predicting future observations.

<table>
<thead>
<tr>
<th>Prdg</th>
<th>Wvlt</th>
<th>Adapt</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.05</td>
<td>.74</td>
<td>1.97</td>
<td>.70</td>
</tr>
</tbody>
</table>

Table 2.3  PSE Averaged Across 9 Randomly Selected Subjects

2.7 Discussion

The MASE is very competitive with all of the other spectral density estimators considered in this paper. In addition to its excellent performance in small samples, the MASE is appealing because it is easy to compute, and although it uses parametric time series models to improve its stability, it does not require the selection of a particular parametric time series model. The first author hopes to make a function in R publicly available to compute the MASE.
The concepts and models behind this shrinkage estimator may be suited to a variety of other problems involving spectral density estimation. In longitudinal studies, for example, time series are observed for multiple subjects, and a common periodic behavior within each series may be of interest as in the example in Section 2.6. (Grambsch et al. (2002)). Hierarchical models, similar to the one proposed in this paper, might be adapted to such a setting by shrinking the subject specific spectral densities to a population spectral density. The authors are currently working on such models. Similar models have previously been explored in Diggle and al Wasel (1997).

2.8 References


2.9 Appendix

2.9.1 Proof of theorems

Proof of Theorem 1: Let \( w_j = \sqrt{\text{Var} \left( M_j I(\omega_j)^{1/2} \right)} = V_j \left( f_s(\omega_j)^{1/4} \right)^2 \xrightarrow{\text{prob}} V_j f(\omega_j)^{1/2} = \text{Var} \left( M_j I(\omega_j)^{1/4} \right). \)

We begin by showing that \( \tau^2 \xrightarrow{\text{prob}} 0 \) when the model is correctly specified. To prove this, we consider the function

\[
h(I(\omega_1), \ldots, I(\omega_F)) = \sum_{j=1}^{F} \left( V_j \left( f_s(\omega_j)^{1/4} \right)^2 \right)^{-1} \left( M_j I(\omega_j)^{1/4} - f_p(\omega_j; \theta)^{1/4} \right)^2 - F
\]

and show that \( h(I(\omega_1), I(\omega_2), \ldots, I(\omega_F)) \xrightarrow{\text{prob}} 0 \) when the model is correctly specified. To do this, we consider the numerator in the function \( h(I(\omega_1), \ldots, I(\omega_F)) = \frac{1}{F} h(I(\omega_1), \ldots, I(\omega_F)) \). As long as \( \frac{1}{p} \sum_j \left\{ \left( M_j I(\omega_j)^{1/4} - f_p(\omega_j; \theta)^{1/4} \right)^2 / V_j f_s(\omega_j)^{1/4} \right\} \) goes to 1 in probability, \( h \) should converge to 0 in probability. To study the limiting behavior of this sum, consider the array \( \left\{ \frac{M_j I(\omega_j)^{1/4} - f_p(\omega_j; \theta)^{1/4}}{V_j f_s(\omega_j)^{1/4}} \right\} \) for \( 1 \leq j \leq F = \left[ \frac{n}{\lambda} \right] \). Since \( V_j \left( f_s(\omega_j)^{1/4} \right)^2 \xrightarrow{\text{prob}} V_j f(\omega_j)^{1/2} \), it is clear that

\[
\mathbb{E} \left( \frac{\left( M_j I(\omega_j)^{1/4} - f_p(\omega_j; \theta)^{1/4} \right)^2}{V_j \left( f_s(\omega_j)^{1/4} \right)^2} \right) \rightarrow \frac{\text{Var}(M_j I(\omega_j)^{1/4})}{V_j f(\omega_j)^{1/2}} = 1.
\]

Now consider the variance of the terms in the sequence.

\[
\text{Var} \left( \frac{\left( M_j I(\omega_j)^{1/4} - f_p(\omega_j; \theta)^{1/4} \right)^2}{V_j \left( f_s(\omega_j)^{1/4} \right)^2} \right) \approx \frac{1}{\left( V_j f(\omega_j)^{1/2} \right)^2} \text{Var} \left\{ \left( M_j I(\omega_j)^{1/4} - f_p(\omega_j; \theta) \right)^2 \right\}
\]

\[
\leq \frac{1}{\min_j \left( V_j \right)^2} \left( \max_j (2M_j^4) + \max_j \left( V_j \right) + 1 \right).
\]
This inequality follows after evaluating \( \text{Var} \left\{ \left( M_j (\omega_j)^{\frac{1}{4}} - f_p(\omega_j; \theta)^{\frac{1}{4}} \right)^2 \right\} \), and substituting \( f(\omega_j) \) for \( f_p(\omega_j; \theta) \) (remember that the correct parametric form has been specified). From this inequality, one gets

\[
\frac{1}{F^2} \sum_{j=1}^{F} \left( \text{Var} \left( \frac{\left( M_j I(\omega_j)^{\frac{1}{4}} - f_p(\omega_j; \theta)^{\frac{1}{4}} \right)^2}{V_j \left( f_{\theta}(\omega_j)^{\frac{1}{4}} \right)^2} \right) \right) < \frac{1}{F} \left( \frac{1}{\min_j (V_j)^2} \left( \max_j \left( 2M_j^4 \right) + \max_j (V_j + 1) \right) \right) \to 0.
\]

From Theorem 5.4 in Durrett (1996), it is clear that \( \frac{1}{F} \sum_{j=1}^{F} \left( \frac{\left( M_j I(\omega_j)^{\frac{1}{4}} - f_p(\omega_j; \theta)^{\frac{1}{4}} \right)^2}{V_j \left( f_{\theta}(\omega_j)^{\frac{1}{4}} \right)^2} \right) \xrightarrow{\text{prob}} 1 \). This implies that the numerator in the function \( \hat{h}(\cdot) \) goes to 0 in probability. With this result, it follows that \( \hat{\tau}^2 \xrightarrow{\text{prob}} 0 \) when the correct parametric form is specified.

Now that it has been established that \( \hat{\tau}^2 \xrightarrow{\text{prob}} 0 \) when the model is correctly specified, the limiting behavior of the DC estimator must be examined under these conditions. The estimate of the spectral density’s fourth root is given by \( m \), and it converges to the truth in this case. The estimated fourth moment of the spectral density’s fourth root also converges to the truth, making the DC Estimator consistent. \( \Box \)

**Proof of Theorem 2** One should note that this proof is developed under the assumption that the prior placed on \( f(\omega)^{\frac{1}{4}} \) is the multivariate normal distribution in (2.4) truncated at 0. As a result, this proof uses Results 1 and 2 given below.

**Result 1.** If \( Y \sim TN_0 (\mu_Y, \sigma_Y^2) \) with \( \mu_Y > 0 \), then \( \mathbb{E} (Y^3) \leq 6\mu_Y \sigma_Y^2 + 2\mu_Y^3 + 8\sigma_Y^3 + 14\mu_Y^2 \sigma_Y \).

**Proof of Result 1:** Let \( Y \sim TN_0 (\mu_Y, \sigma_Y^2) \). \( \mathbb{E} (Y^3) = \left( \frac{d^3 M_Y(t)}{dt^3} \right)_{t=0} \) where

\[
M_Y(t) = \mathbb{E} (e^{ty}) = \int_0^{\infty} e^{ty} \frac{1}{\sigma_Y \sqrt{2\pi}} \left( 1 - \Phi \left( \frac{\mu_Y}{\sigma_Y} \right) \right)^{-1} \exp \left( \frac{(y - \mu_Y)^2}{2\sigma_Y^2} \right) dy =
\]

\[
\left( 1 - \Phi \left( \frac{\mu_Y}{\sigma_Y} \right) \right)^{-1} \exp \left( \left( 2\mu_Y t\sigma_Y^2 + t^2 \sigma_Y^2 / 2\sigma_Y^2 \right) \left( 1 - \Phi \left( \frac{\mu_Y + t\sigma_Y^2}{\sigma_Y} \right) \right) \right).
\]

Simple Calculus and basic algebra show that

\[
\frac{d^3}{dt^3} \left( 1 - \Phi \left( \frac{\mu_Y}{\sigma_Y} \right) \right)^{-1} \exp \left( \left( 2\mu_Y t\sigma_Y^2 + t^2 \sigma_Y^2 / 2\sigma_Y^2 \right) \left( 1 - \Phi \left( \frac{\mu_Y + t\sigma_Y^2}{\sigma_Y} \right) \right) \right) \bigg|_{t=0} \leq
\]

\[
2 \left( 3\mu_Y \sigma_Y^2 + \mu_Y^3 + 4\sigma_Y^3 + 7\mu_Y^2 \sigma_Y \right).
\]

**Result 2.** Let \( Y \) be a \( K \times 1 \) random vector such that \( Y \sim MVN (\mu, \Lambda) \) and let \( \mu \sim TMVN (\beta, \Gamma) \) (a multivariate normal distribution with mean \( \beta \) and covariance \( \Gamma \) truncated at 0). Then \( Y|\Lambda, \beta, \Gamma \sim p_Y (y) \) where
\[ p_Y(y) = \frac{c(\beta, \Gamma)}{(2\pi)^{\frac{|\Lambda + \Gamma|}{2}}} \exp \left( -\frac{1}{2} (y - \beta)' (\Lambda + \Gamma)^{-1} (y - \beta) \right) P(W \geq 0), \]

\[ c(\beta, \Gamma) = (P(\bar{\mu} \geq 0))^{-1}, \bar{\mu} \sim MVN(\beta, \Lambda), \text{ and} \]

\[ W \sim MVN\left( (\Lambda + \Gamma)^{-1} \beta + \Gamma (\Lambda + \Gamma)^{-1} Y, (\Lambda^{-1} + \Gamma^{-1})^{-1} \right). \]

**Proof of Result 2:** First observe that since \( \mu \sim TMVN(\beta, \Gamma) \), \( \mu \sim p_\mu(\mu) \) where

\[ p_\mu(\mu) = \begin{cases} \frac{c(\beta, \Gamma)}{(2\pi)^{\frac{|\Lambda|}{2}}} \exp \left( -\frac{1}{2} (\mu - \beta)' \Gamma^{-1} (\mu - \beta) \right) & \mu_i \geq 0 \forall i \\ 0 & \text{o.w.} \end{cases} \]

From this, it is clear that

\[ p_Y(y) = \int_{(\mathbb{R}^+)^K} \frac{1}{(2\pi)^{\frac{|\Lambda|}{2}}} \exp \left( -\frac{1}{2} (y - \mu)' \Lambda^{-1} (y - \mu) \right) \times \frac{c(\beta, \Gamma)}{(2\pi)^{\frac{|\Gamma|}{2}}} \exp \left( -\frac{1}{2} (\mu - \beta)' \Gamma^{-1} (\mu - \beta) \right) \prod_{i=1}^{K} d\mu_i \]

\[ = \frac{c(\beta, \Gamma)}{(2\pi)^{\frac{|\Gamma|}{2}}} \exp \left( -\frac{1}{2} (\mu' (\Lambda^{-1} + \Gamma^{-1}) \mu - 2\mu (\Lambda^{-1} y + \Gamma^{-1} \beta)) \right) \prod_{i=1}^{K} d\mu_i. \]

Letting \( \gamma = \Lambda (\Lambda + \Gamma)^{-1} \beta + \Gamma (\Lambda + \Gamma)^{-1} y \), the above expression can be set equal to

\[ = \frac{c(\beta, \Gamma)}{(2\pi)^{\frac{|\Gamma|}{2}}} \exp \left( -\frac{1}{2} (\mu' (\Lambda^{-1} + \Gamma^{-1}) \beta + \gamma' (\Lambda^{-1} + \Gamma^{-1}) \gamma) \right) \]

\[ \times \int_{(\mathbb{R}^+)^K} (2\pi)^{\frac{|\Lambda + \Gamma|}{2}} \exp \left( -\frac{1}{2} (\mu - \gamma)' (\Lambda^{-1} + \Gamma^{-1}) (\mu - \gamma) \right) \prod_{i=1}^{K} d\mu_i \]

\[ = \frac{c(\beta, \Gamma)}{(2\pi)^{\frac{|\Lambda + \Gamma|}{2}}} \exp \left( -\frac{1}{2} (y - \beta)' (\Lambda + \Gamma)^{-1} (y - \beta) \right) P(W \geq 0) \]

**Proof of Theorem 2** To begin the proof of Theorem 2, first observe that using Result 2

\[ p\left( I(\omega) \frac{1}{2} | \phi, \sigma^2, \tau^2 \right) = \frac{c(\mu, \tau^2)}{(2\pi)^{\frac{1}{2}}} \exp \left( -\frac{1}{2} \left( I(\omega) \frac{1}{2} (\sigma^2)^{-1} \frac{1}{2} g(\omega; \phi) \right)' (\tau^2 I + \Sigma)^{-1} \left( I(\omega) \frac{1}{2} (\sigma^2)^{-1} g(\omega; \phi) \right) \right) P(T \geq 0) \]
where $I(\omega)^{\frac{1}{2}} = \left(I(\omega_1)^{\frac{1}{2}}, \ldots, I(\omega_F)^{\frac{1}{2}}\right)^T$, $g(\omega; \phi) = \frac{1}{\sigma^2} \left(f(\omega_1; \phi)^{\frac{1}{2}}, \ldots, f(\omega_F; \phi)^{\frac{1}{2}}\right)^T$, $\omega$ is the $j^{th}$ Fourier frequency, and $T \sim MVN\left(\Sigma + \tau^2 I, (\Sigma + \tau^2 I)^{-1} I(\omega)^{\frac{1}{2}}(\Sigma^{-1} + \tau_1^2 I)^{-1}\right)$. Knowing that $c(f_p, \tau^2) \leq 0.5^{-F}$, it is clear that

$$p\left(I(\omega)^{\frac{1}{2}} \mid \phi, \sigma^2, \tau^2\right) \leq \frac{\cdot 5^{-F}}{(2\pi)^{\frac{F}{2}}} \left|\tau^2 I + \Sigma\right|^{\frac{1}{2}} \times \exp\left[-\frac{1}{2} \left(I(\omega)^{\frac{1}{2}} - (\sigma^2)^{\frac{1}{2}} g(\omega; \phi)\right)^T (\tau^2 I + \Sigma)^{-1} \left(I(\omega)^{\frac{1}{2}} - (\sigma^2)^{\frac{1}{2}} g(\omega; \phi)\right)\right]$$

$$= \cdot 5^{-F} \bar{p}(I(\omega) \mid \phi, \sigma^2, \tau^2)$$

$$\rightarrow \int_{\mathbb{R}^F} \int_{0}^{\infty} p\left(I(\omega)^{\frac{1}{2}} \mid \phi, \sigma^2, \tau^2\right) p(\phi) d\sigma^2 d\tau^2 d\phi$$

$$\leq \int_{\mathbb{R}^F} \int_{0}^{\infty} \int_{0}^{\infty} \cdot 5^{-F} \bar{p}(I(\omega) \mid \phi, \sigma^2, \tau^2) p(\phi) d\sigma^2 d\tau^2 d\phi$$

$$= \int_{\mathbb{R}^F} \int_{0}^{\infty} \int_{0}^{\infty} \frac{\cdot 5^{-F}}{(2\pi)^{\frac{F}{2}}} \left(\Pi_{i=1}^{F}(\tau^2 + \lambda_i)^{\frac{1}{2}}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2} \left(b - \sigma^2 e\right)^T \Lambda^{-1} \left(b - \sigma^2 e\right)\right] p(\phi) d\sigma^2 d\tau^2 d\phi$$

where $\lambda_i$ is the $i^{th}$ eigenvalue of $\Sigma$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_F)$, $c = P^\prime g(\omega; \phi)$ where $P = [e_1, \ldots, e_F]$ and $e_j$ is the $j^{th}$ eigenvector of $\Sigma$, and $b = P^\prime I(\omega)^{\frac{1}{2}}$. Since $\Sigma$ is a nonsingular p.d. matrix, $\lambda_i > 0 \forall i$. Call the smallest eigenvalue of $\Sigma \lambda_F$. The following inequality then follows for the above integral.

$$\leq \int_{\mathbb{R}^F} \int_{0}^{\infty} \int_{0}^{\infty} \frac{\cdot 5^{-F}}{(2\pi)^{\frac{F}{2}}} \left(\Pi_{i=1}^{F}(\tau^2 + \lambda_i)^{\frac{1}{2}}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2} \left(c^2 \frac{\lambda_i}{\tau^2 + \lambda_i}\right) \left(b_i - \sigma^2 e_i\right)^2\right] p(\phi) d\sigma^2 d\tau^2 d\phi$$

$$\leq \int_{\mathbb{R}^F} \int_{0}^{\infty} \frac{\cdot 5^{-F}}{(2\pi)^{\frac{F}{2}}} \left(\Pi_{i=1}^{F}(\tau^2 + \lambda_i)^{\frac{1}{2}}\right)^{\frac{1}{2}} \frac{\cdot 4 (\tau^2 + \lambda_i)^{\frac{1}{2}}}{(2\pi)^{\frac{1}{2}}} \sqrt{c^2}$$

$$\times \left\{ \int_{0}^{\infty} \frac{\cdot 5^{-F}}{(2\pi)^{\frac{1}{2}}} \left(\frac{\lambda_i}{\tau^2 + \lambda_i}\right)^{\frac{1}{2}} \exp\left[-\frac{1}{2} \left(c^2 \frac{\lambda_i}{\tau^2 + \lambda_i}\right) \left(b_i - \sigma^2 e_i\right)^2\right] d\sigma^2 \right\} p(\phi) d\tau^2 d\phi$$

The equality under the underbrace follows from Result 1. In the above integral, the only quantity depending on $\phi$ is $c_1$. To remind the reader of this dependence, we will denote $c_1$ as $c_1(\phi)$. The above integral is finite provided $F \geq 5$, $\int_{\mathbb{R}^F} \frac{1}{c_1(\phi)} P(\phi) d\phi < \infty$, $\int_{\mathbb{R}^F} \frac{1}{c_1(\phi)} \cdot 5^{-F} P(\phi) d\phi < \infty$, and $\int_{\mathbb{R}^F} \frac{1}{c_1(\phi)^2} P(\phi) d\phi < \infty$. The last three inequalities hold when an ARMA($p, 0$) model is specified. The following is a proof of the first case.
The last integral written above is finite given that $p(\phi)$ is proper and $\int_{\mathbb{R}^p} |\phi_i\phi_k| p(\phi) d\phi < \infty$ for all pairs of autoregressive parameters. □

**Proof of Theorem 3** In proving this theorem, we use six different results (Results 2-7). Results 3-7 (and their proofs or references to them) are given below. Throughout these proofs, it should be clear that $\omega_j$ refers to the $j^{th}$ Fourier frequency and $\omega_k$ refers to the $l^{th}$ frequency in the set $\omega_l$, and that we are letting $p\left(f^2(\omega)\right) f_p^2(\omega), \tau^2 \right)$ be the multivariate normal distribution in (2.4) truncated at 0.

**Result 3.** $f_p(\omega_1; \hat{\theta}) \overset{\text{prob}}{\rightarrow} f(\omega_1; \theta_0)$ for some value $\theta_0$ (when the model is correctly specified, $\theta_0 = \theta$). (Dalhaus and Wefelmeyer (1996))

**Result 4.** $\hat{f}_s(\omega_l)^{\frac{1}{2}} = \sum_{j=1}^{P} W_h(\omega_j - \omega_l) M_j I(\omega_j)^{\frac{1}{2}} \overset{\text{prob}}{\rightarrow} f(\omega_l)^{\frac{1}{2}}$ where $\omega_l$ is the $l^{th}$ frequency in the set $\omega_j$. (Wand and Jones (1995))

**Result 5.** $\hat{\Sigma}_l = \text{Var} \left(\hat{f}_s(\omega_l)^{\frac{1}{2}}\right) \overset{\text{prob}}{\rightarrow} [0]_{K \times K}$

**Proof of Result 5:** Just consider the $l^{th}$ diagonal element of $\hat{\Sigma}_l$, $\text{Var} \left(\hat{f}_s(\omega_l)^{\frac{1}{2}}\right)$. Remember that $\text{Var} \left(\hat{f}_s(\omega_l)^{\frac{1}{2}}\right) = \sum_{j=1}^{P} W_h^2(\omega_j - \omega_l) V_j \left(\hat{f}_s(\omega_j)^{\frac{1}{2}}\right)^2$. Result 4 shows that $\hat{f}_s(\omega_j)^{\frac{1}{2}} \overset{\text{prob}}{\rightarrow} f(\omega_j)^{\frac{1}{2}}$. This implies that for any arbitrarily small numbers $\delta_1 > 0$ and $\delta_2 > 0$, there exists a sample size $n^{**}$ such that

$$P \left(\sum_{j=1}^{P} W_h^2(\omega_j - \omega_l) V_j \left(\hat{f}_s(\omega_j)^{\frac{1}{2}}\right)^2 < V_j \left(\sup_{\omega \in [0, \pi]} f(\omega)^{\frac{1}{2}} + \delta_1\right) \sum_{j=1}^{P} W_h^2(\omega_j - \omega_l)\right) > 1 - \delta_2$$

for all $n > n^{**}$. Since $\sum W_h^2(\omega_j - \omega_l)$ can be made arbitrarily small with a sufficiently large sample size $\hat{n}$, it is clear that for a given $\delta_1$ and $\delta_2$ and for any $n > \max(n^{**}, \hat{n})$, $P \left(\text{Var} \left(\hat{f}_s(\omega_l)^{\frac{1}{2}}\right) < \epsilon\right) < 1 - \delta_2$.

**Result 6.** $\hat{\tau}^2 \overset{\text{prob}}{\rightarrow} 0$ when the correct parametric model is specified.

**Proof of Result 6:** From Result 2, it is clear that $p\left(\hat{f}_s(\omega_j)^{\frac{1}{2}} \mid f_p(\omega_j)^{\frac{1}{2}}, \Sigma_l, \tau^2\right) =$
\[\frac{c \left(f_p(\omega)\frac{1}{2}, \tau^2\right)}{(2\pi)^{\frac{K}{2}} |\Sigma_I + \tau^2 I|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} \left( f_s(\omega)\frac{1}{2} - f_p(\omega)\frac{1}{2} \right) \left( \Sigma_I + \tau^2 I \right)^{-1} \left( f_s(\omega)\frac{1}{2} - f_p(\omega)\frac{1}{2} \right) \right) P(U \geq 0),\]

where \(U \sim MVN \left( \Sigma_I (\Sigma_I + \tau^2 I)^{-1} f_p(\omega)\frac{1}{2} + \tau^2 \left( \Sigma_I + \tau^2 I \right)^{-1} f_s(\omega)\frac{1}{2}, \left( \Sigma_I^{-1} + \frac{1}{2} \tau^2 I \right)^{-1} \right) \). Define \(z(\omega) = f_s(\omega)\frac{1}{2} - f_p(\omega)\frac{1}{2} \). We will begin by assuming that \(z(\omega) \not\rightarrow 0\) is not true. If \(z(\omega) \not\rightarrow 0\), it is implied that \(z(\omega) > 0\) without loss of generality, let \(l = 1\). Then

\[
P(z(\omega_1) \geq \epsilon) = \int_{\epsilon}^{\infty} \int_{R^{K-1}} \frac{c \left(f_p(\omega)\frac{1}{2}, \tau^2\right)}{(2\pi)^{\frac{K}{2}} |\Sigma_I + \tau^2 I|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} z' \left( \Sigma_I + \tau^2 I \right)^{-1} z \right) P(U \geq 0) \left( \prod_{j=2}^{K} dz(\omega_j) \right) dz(\omega_1)
\]

\[
\geq (0.5)^K \int_{\epsilon}^{\infty} \int_{R^{K-1}} \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma_I + \tau^2 I|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} z' \left( \Sigma_I + \tau^2 I \right)^{-1} z \right) \left( \prod_{j=2}^{K} dz(\omega_j) \right) dz(\omega_1)
\]

\[
= (0.5)^K \left( 1 - \Phi \left( \frac{\epsilon}{(\Sigma[I, I] + \tau^2)^{\frac{1}{2}}} \right) \right) > 0 \forall \epsilon
\]

which contradicts \(P(z(\omega_1) \geq \epsilon) \rightarrow 0 \forall \epsilon\). (Note: the first inequality results since \(c \left(f_p(\omega)\frac{1}{2}, \tau^2\right) \geq 1\) and \(P(U \geq 0) \geq 0.5^K\).)

**Result 7.** \(\tau^2\) is bounded away from 0 when the model is incorrectly specified.

**Proof of Result 7:** This will be another proof by contradiction. Using the same notation given in Result 6, assume that \(f_s(\omega)\frac{1}{2} - f_p(\omega)\frac{1}{2} = z(\omega) \not\rightarrow 0\) is not true, yet assume that \(\tau^2 \not\rightarrow 0\). It is known that \(\epsilon \not\rightarrow 0\) is not true, then there exists some number \(j \in \{1, 2, \ldots, K\}\) such that \(P(z(\omega_j) \geq \epsilon) \geq \delta > 0\) for some \(\epsilon\) and \(\delta\) and for infinitely many \(n\). Without loss of generality, assume that \(j = 1\). Then

\[
P(z(\omega_1) \geq \epsilon) = \int_{\epsilon}^{\infty} \int_{R^{K-1}} \frac{c \left(f_p(\omega)\frac{1}{2}, \tau^2\right)}{(2\pi)^{\frac{K}{2}} |\Sigma_I + \tau^2 I|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} z' \left( \Sigma_I + \tau^2 I \right)^{-1} z \right) P(U \geq 0) \left( \prod_{j=2}^{K} dz(\omega_j) \right) dz(\omega_1)
\]

\[
< (0.5)^{-K} \left\{ \int_{\epsilon}^{\infty} \int_{R^{K-1}} \frac{1}{(2\pi)^{\frac{K}{2}} |\Sigma_I + \tau^2 I|^{\frac{1}{2}}} \exp \left( -\frac{1}{2} z' \left( \Sigma_I + \tau^2 I \right)^{-1} z \right) \left( \prod_{j=2}^{K} dz(\omega_j) \right) dz(\omega_1) \right\}
\]

since \(c \left(f_p(\omega)\frac{1}{2}, \tau^2\right) \leq 0.5^{-K}\) and \(P(U \geq 0) \leq 1\). Since it is known that \(\Sigma_I \rightarrow [0]_{K \times K}\) and it is assumed that \(\tau^2 \not\rightarrow 0\), it is clear that the quantity in curly brackets goes to 0. This contradicts the assumption that \(z(\omega) \not\rightarrow 0\) is not true, however. As a result, we can conclude that \(\tau^2 \not\rightarrow 0\) is not true.
Proof of Theorem 3: To prove that \( \tilde{f}(\omega_I) = 3\tilde{v}_I + 6\tilde{v}_I^2 \tilde{m}_I^0 + \tilde{m}_I^4 \xrightarrow{\text{prob}} f(\omega_I) \), the limiting behavior of both \( \tilde{v}_I^2 \) and \( \tilde{m}_I \) have to be studied. First observe that

\[
\tilde{v}_I = \text{diag} \left( \frac{1}{\tilde{v}_I} + \frac{1}{\tilde{v}_I^2} I_{K \times K} \right)^{-1} \xrightarrow{\text{prob}} 0_{K \times 1}
\]

since \( \tilde{\Sigma}_I \xrightarrow{\text{prob}} [0]_{K \times K} \). Secondly, note that regardless of whether the parametric model is specified correctly or incorrectly,

\[
\tilde{m}_I = \tilde{\Sigma}_I \left( \tilde{v}_I^2 I_{K \times K} + \tilde{\Sigma}_I \right)^{-1} f_p \left( \omega; \theta \right) + \tilde{v}_I^2 \left( \tilde{v}_I^2 I_{K \times K} + \tilde{\Sigma}_I \right)^{-1} \tilde{f}_s (\omega) \xrightarrow{\text{prob}} f(\omega).
\]

The limiting behavior of \( \tilde{\Sigma}_I \xrightarrow{\text{prob}} [0]_{K \times K} \) and \( \tilde{v}_I \) ensures this. When the model is incorrectly specified, for instance, the first term written above disappears as \( \tilde{\Sigma}_I \xrightarrow{\text{prob}} [0]_{K \times K} \) and the second dominates. Since \( \tilde{f}_s (\omega) \xrightarrow{\text{prob}} f(\omega) \), \( \tilde{m}_I \xrightarrow{\text{prob}} f(\omega) \) when the model is incorrectly specified. When the model is correctly specified, \( \tilde{m}_I \) is just a linear combination of two estimators that are consistent for \( f(\omega) \). Given these observations, it is clear that \( \tilde{f}(\omega_I) \xrightarrow{\text{prob}} f(\omega_I) \). \( \square \)
Table 2.4  *MISE* of Estimators at *n* = 32. Monte Carlo standard errors are given in parentheses.

<table>
<thead>
<tr>
<th>SD</th>
<th>Prd</th>
<th>DC</th>
<th>Wvlt</th>
<th>Adpt</th>
<th>ML</th>
<th>Shrnk</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.030</td>
<td>.017</td>
<td>.061</td>
<td>.015</td>
<td>.007</td>
<td>.007</td>
<td>.008</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.001)</td>
<td>(.004)</td>
<td>(.001)</td>
<td>(.0004)</td>
<td>(.0004)</td>
<td>(.0004)</td>
</tr>
<tr>
<td>2</td>
<td>.50</td>
<td>5</td>
<td>1.3</td>
<td>.27</td>
<td>3</td>
<td>.24</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(5)</td>
<td>(.1)</td>
<td>(.02)</td>
<td>(3)</td>
<td>(.04)</td>
<td>(.01)</td>
</tr>
<tr>
<td>3</td>
<td>.069</td>
<td>.046</td>
<td>.124</td>
<td>.045</td>
<td>.038</td>
<td>.035</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>(.003)</td>
<td>(.003)</td>
<td>(.008)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.001)</td>
</tr>
<tr>
<td>4</td>
<td>50</td>
<td>56.2</td>
<td>54.2</td>
<td>47</td>
<td>96</td>
<td>42</td>
<td>51</td>
</tr>
<tr>
<td></td>
<td>(1)</td>
<td>(.8)</td>
<td>(.9)</td>
<td>(2)</td>
<td>(25)</td>
<td>(2)</td>
<td>(2)</td>
</tr>
<tr>
<td>5</td>
<td>.040</td>
<td>.092</td>
<td>.025</td>
<td>.065</td>
<td>.067</td>
<td>.067</td>
<td>.020</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.007)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>6</td>
<td>.32</td>
<td>.68</td>
<td>.137</td>
<td>.169</td>
<td>.4</td>
<td>1.3</td>
<td>.144</td>
</tr>
<tr>
<td></td>
<td>(.01)</td>
<td>(.04)</td>
<td>(.006)</td>
<td>(.165)</td>
<td>(.165)</td>
<td>(.165)</td>
<td>(.165)</td>
</tr>
<tr>
<td>7</td>
<td>.83</td>
<td>1.6</td>
<td>.52</td>
<td>.45</td>
<td>.55</td>
<td>.41</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>(.06)</td>
<td>(.1)</td>
<td>(.05)</td>
<td>(.03)</td>
<td>(.05)</td>
<td>(.02)</td>
<td>(.01)</td>
</tr>
<tr>
<td>8</td>
<td>1.27</td>
<td>2.1</td>
<td>.83</td>
<td>.55</td>
<td>.41</td>
<td>.34</td>
<td>.30</td>
</tr>
<tr>
<td></td>
<td>(.08)</td>
<td>(.2)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.03)</td>
</tr>
</tbody>
</table>

Table 2.5  *MMSD* of Estimators at *n* = 32. Monte Carlo standard errors are given in parentheses.

<table>
<thead>
<tr>
<th>SD</th>
<th>Prd</th>
<th>DC</th>
<th>Wvlt</th>
<th>Adpt</th>
<th>ML</th>
<th>Shrnk</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.381</td>
<td>.172</td>
<td>.576</td>
<td>.142</td>
<td>.064</td>
<td>.065</td>
<td>.067</td>
</tr>
<tr>
<td></td>
<td>(.03)</td>
<td>(.02)</td>
<td>(.04)</td>
<td>(.01)</td>
<td>(.007)</td>
<td>(.007)</td>
<td>(.007)</td>
</tr>
<tr>
<td>2</td>
<td>6.8</td>
<td>170</td>
<td>19</td>
<td>2.9</td>
<td>169</td>
<td>4</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>(.5)</td>
<td>(165)</td>
<td>(2)</td>
<td>(.3)</td>
<td>(165)</td>
<td>(1)</td>
<td>(.1)</td>
</tr>
<tr>
<td>3</td>
<td>.93</td>
<td>.53</td>
<td>1.4</td>
<td>.45</td>
<td>.55</td>
<td>.41</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>(.06)</td>
<td>(.04)</td>
<td>(.1)</td>
<td>(.03)</td>
<td>(.05)</td>
<td>(.02)</td>
<td>(.01)</td>
</tr>
<tr>
<td>4</td>
<td>1353</td>
<td>1583</td>
<td>1382</td>
<td>1205</td>
<td>4588</td>
<td>1173</td>
<td>1404</td>
</tr>
<tr>
<td></td>
<td>(26)</td>
<td>(15)</td>
<td>(20)</td>
<td>(30)</td>
<td>(1500)</td>
<td>(52)</td>
<td>(43)</td>
</tr>
<tr>
<td>5</td>
<td>.62</td>
<td>1.6</td>
<td>.34</td>
<td>.34</td>
<td>.34</td>
<td>.34</td>
<td>.34</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.2)</td>
<td>(.04)</td>
<td>(.04)</td>
<td>(.04)</td>
<td>(.04)</td>
<td>(.04)</td>
</tr>
<tr>
<td>6</td>
<td>3.4</td>
<td>7.3</td>
<td>.82</td>
<td>.82</td>
<td>.82</td>
<td>.82</td>
<td>.82</td>
</tr>
<tr>
<td></td>
<td>(.2)</td>
<td>(.7)</td>
<td>(.07)</td>
<td>(.07)</td>
<td>(.07)</td>
<td>(.07)</td>
<td>(.07)</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>26</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(2)</td>
<td>(1)</td>
<td>(1)</td>
<td>(1)</td>
<td>(1)</td>
<td>(1)</td>
</tr>
<tr>
<td>8</td>
<td>21</td>
<td>28</td>
<td>9.9</td>
<td>11.3</td>
<td>11.3</td>
<td>11.3</td>
<td>11.3</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(3)</td>
<td>(.6)</td>
<td>(.3)</td>
<td>(.3)</td>
<td>(.3)</td>
<td>(.3)</td>
</tr>
</tbody>
</table>
Table 2.6 $MISE$ of Estimators at $n = 64$. Monte Carlo standard errors are given in parentheses.

<table>
<thead>
<tr>
<th>SD</th>
<th>Prd</th>
<th>DC</th>
<th>Wvlt</th>
<th>Adpt</th>
<th>ML</th>
<th>Shrnk</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.035</td>
<td>.015</td>
<td>.029</td>
<td>.011</td>
<td>.004</td>
<td>.004</td>
<td>.005</td>
</tr>
<tr>
<td></td>
<td>(.001)</td>
<td>(.001)</td>
<td>(.002)</td>
<td>(.001)</td>
<td>(.0002)</td>
<td>(.0002)</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>.52</td>
<td>.29</td>
<td>.48</td>
<td>.18</td>
<td>.09</td>
<td>.09</td>
<td>.098</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.05)</td>
<td>(.02)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.004)</td>
</tr>
<tr>
<td>3</td>
<td>.088</td>
<td>.040</td>
<td>.066</td>
<td>.043</td>
<td>.029</td>
<td>.029</td>
<td>.037</td>
</tr>
<tr>
<td></td>
<td>(.003)</td>
<td>(.002)</td>
<td>(.003)</td>
<td>(.002)</td>
<td>(.001)</td>
<td>(.001)</td>
<td>(.001)</td>
</tr>
<tr>
<td>4</td>
<td>53</td>
<td>29.3</td>
<td>34.3</td>
<td>46</td>
<td>113</td>
<td>72</td>
<td>42</td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td>(.8)</td>
<td>(.8)</td>
<td>(3)</td>
<td>(31)</td>
<td>(14)</td>
<td>(2)</td>
</tr>
<tr>
<td>5</td>
<td>.042</td>
<td>.039</td>
<td>.020</td>
<td>.037</td>
<td>.037</td>
<td>.041</td>
<td>.013</td>
</tr>
<tr>
<td></td>
<td>(.003)</td>
<td>(.003)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>.318</td>
<td>.33</td>
<td>.38</td>
<td>.45</td>
<td>.45</td>
<td>.60</td>
<td>.986</td>
</tr>
<tr>
<td></td>
<td>(.008)</td>
<td>(.02)</td>
<td>(.004)</td>
<td>(.04)</td>
<td>(.04)</td>
<td>(.05)</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>.84</td>
<td>.78</td>
<td>.38</td>
<td>.54</td>
<td>.51</td>
<td>.53</td>
<td>.22</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td>(.05)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>1.46</td>
<td>1.7</td>
<td>.66</td>
<td>.53</td>
<td>.53</td>
<td>.53</td>
<td>.94</td>
</tr>
<tr>
<td></td>
<td>(.08)</td>
<td>(.1)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.02)</td>
<td>(.03)</td>
<td></td>
</tr>
</tbody>
</table>

Table 2.7 $MMSD$ of Estimators at $n = 64$. Monte Carlo standard errors are given in parentheses.

<table>
<thead>
<tr>
<th>SD</th>
<th>Prd</th>
<th>DC</th>
<th>Wvlt</th>
<th>Adpt</th>
<th>ML</th>
<th>Shrnk</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.69</td>
<td>.17</td>
<td>.34</td>
<td>.11</td>
<td>.037</td>
<td>.037</td>
<td>.041</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td>(.01)</td>
<td>(.03)</td>
<td>(.01)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>2</td>
<td>9.8</td>
<td>4</td>
<td>6.5</td>
<td>2.1</td>
<td>.45</td>
<td>.45</td>
<td>.60</td>
</tr>
<tr>
<td></td>
<td>(.6)</td>
<td>(2)</td>
<td>(.5)</td>
<td>(.2)</td>
<td>(.04)</td>
<td>(.04)</td>
<td>(.05)</td>
</tr>
<tr>
<td>3</td>
<td>1.8</td>
<td>.63</td>
<td>.89</td>
<td>.66</td>
<td>.54</td>
<td>.51</td>
<td>.53</td>
</tr>
<tr>
<td></td>
<td>(.1)</td>
<td>(.04)</td>
<td>(.06)</td>
<td>(.06)</td>
<td>(.03)</td>
<td>(.02)</td>
<td>(.02)</td>
</tr>
<tr>
<td>4</td>
<td>1397</td>
<td>889</td>
<td>961</td>
<td>1139</td>
<td>6198</td>
<td>3431</td>
<td>1291</td>
</tr>
<tr>
<td></td>
<td>(73)</td>
<td>(21)</td>
<td>(22)</td>
<td>(75)</td>
<td>(2000)</td>
<td>(840)</td>
<td>(90)</td>
</tr>
<tr>
<td>5</td>
<td>1.0</td>
<td>.76</td>
<td>.36</td>
<td>.21</td>
<td>.21</td>
<td>.21</td>
<td>.21</td>
</tr>
<tr>
<td></td>
<td>(.1)</td>
<td>(.08)</td>
<td>(.07)</td>
<td>(.07)</td>
<td>(.02)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>4.6</td>
<td>3.9</td>
<td>.64</td>
<td>.51</td>
<td>.51</td>
<td>.51</td>
<td>.51</td>
</tr>
<tr>
<td></td>
<td>(.2)</td>
<td>(.2)</td>
<td>(.05)</td>
<td>(.05)</td>
<td>(.03)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>20</td>
<td>14</td>
<td>6.4</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
<td>3.2</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(1)</td>
<td>(.7)</td>
<td>(.8)</td>
<td>(.8)</td>
<td>(.8)</td>
<td>(.8)</td>
</tr>
<tr>
<td>8</td>
<td>36</td>
<td>37</td>
<td>9.2</td>
<td>17</td>
<td>17</td>
<td>17</td>
<td>17</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(4)</td>
<td>(.5)</td>
<td>(1)</td>
<td>(1)</td>
<td>(1)</td>
<td>(1)</td>
</tr>
</tbody>
</table>
Table 2.8  \textit{MISE} of Estimators at n = 128. Monte Carlo standard errors are given in parentheses.

<table>
<thead>
<tr>
<th>SD</th>
<th>Prd</th>
<th>DC</th>
<th>Wvlt</th>
<th>Adpt</th>
<th>ML</th>
<th>Shrnk</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>.052</td>
<td>.018</td>
<td>.015</td>
<td>.007</td>
<td>.003</td>
<td>.003</td>
<td>.003</td>
</tr>
<tr>
<td></td>
<td>(.001)</td>
<td>(.001)</td>
<td>(.001)</td>
<td>(.0003)</td>
<td>(.0001)</td>
<td>(.0001)</td>
<td>(.0002)</td>
</tr>
<tr>
<td>2</td>
<td>.74</td>
<td>.26</td>
<td>.22</td>
<td>.106</td>
<td>.042</td>
<td>.042</td>
<td>.050</td>
</tr>
<tr>
<td></td>
<td>(.02)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.005)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>3</td>
<td>.150</td>
<td>.053</td>
<td>.046</td>
<td>.035</td>
<td>.022</td>
<td>.022</td>
<td>.025</td>
</tr>
<tr>
<td></td>
<td>(.004)</td>
<td>(.005)</td>
<td>(.001)</td>
<td>(.002)</td>
<td>(.001)</td>
<td>(.001)</td>
<td>(.001)</td>
</tr>
<tr>
<td>4</td>
<td>.74</td>
<td>19.3</td>
<td>22</td>
<td>38</td>
<td>56</td>
<td>54</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(5)</td>
<td>(.9)</td>
<td>(.9)</td>
<td>(3)</td>
<td>(6)</td>
<td>(5)</td>
<td>(1)</td>
</tr>
<tr>
<td>5</td>
<td>.057</td>
<td>.20</td>
<td>.11</td>
<td>.079</td>
<td>.021</td>
<td>.022</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td>(.002)</td>
<td>(.001)</td>
<td>(.001)</td>
<td>(.0005)</td>
<td>(.001)</td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>6</td>
<td>.486</td>
<td>.157</td>
<td>.056</td>
<td>.21</td>
<td>.21</td>
<td>.29</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>(.009)</td>
<td>(.006)</td>
<td>(.002)</td>
<td>(.001)</td>
<td>(.002)</td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>7</td>
<td>.69</td>
<td>.36</td>
<td>.23</td>
<td>.83</td>
<td>.67</td>
<td>.48</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>(.04)</td>
<td>(.02)</td>
<td>(.01)</td>
<td>(.04)</td>
<td>(.04)</td>
<td>(.03)</td>
<td>(.03)</td>
</tr>
<tr>
<td>8</td>
<td>2.66</td>
<td>1.21</td>
<td>.46</td>
<td>.48</td>
<td>.48</td>
<td>.50</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>(.08)</td>
<td>(.07)</td>
<td>(.02)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.03)</td>
</tr>
</tbody>
</table>

Table 2.9  \textit{MMSD} of Estimators at n = 128. Monte Carlo standard errors are given in parentheses.

<table>
<thead>
<tr>
<th>SD</th>
<th>Prd</th>
<th>DC</th>
<th>Wvlt</th>
<th>Adpt</th>
<th>ML</th>
<th>Shrnk</th>
<th>MASE</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.19</td>
<td>20</td>
<td>22</td>
<td>.079</td>
<td>.021</td>
<td>.022</td>
<td>.026</td>
</tr>
<tr>
<td></td>
<td>(.05)</td>
<td>(.01)</td>
<td>(.02)</td>
<td>(.005)</td>
<td>(.001)</td>
<td>(.002)</td>
<td>(.002)</td>
</tr>
<tr>
<td>2</td>
<td>16.8</td>
<td>2.8</td>
<td>3.0</td>
<td>1.3</td>
<td>21</td>
<td>21</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>(.8)</td>
<td>(.2)</td>
<td>(.2)</td>
<td>(.1)</td>
<td>(.01)</td>
<td>(.01)</td>
<td>(.01)</td>
</tr>
<tr>
<td>3</td>
<td>4.0</td>
<td>1.1</td>
<td>.83</td>
<td>.67</td>
<td>.48</td>
<td>.48</td>
<td>.50</td>
</tr>
<tr>
<td></td>
<td>(.2)</td>
<td>(.2)</td>
<td>(.04)</td>
<td>(.04)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.02)</td>
</tr>
<tr>
<td>4</td>
<td>2374</td>
<td>583</td>
<td>635</td>
<td>1124</td>
<td>2922</td>
<td>2829</td>
<td>1046</td>
</tr>
<tr>
<td></td>
<td>(200)</td>
<td>(30)</td>
<td>(26)</td>
<td>(100)</td>
<td>(350)</td>
<td>(320)</td>
<td>(100)</td>
</tr>
<tr>
<td>5</td>
<td>1.40</td>
<td>.40</td>
<td>.18</td>
<td>.079</td>
<td>.021</td>
<td>.022</td>
<td>.084</td>
</tr>
<tr>
<td></td>
<td>(.09)</td>
<td>(.04)</td>
<td>(.02)</td>
<td>(.005)</td>
<td>(.001)</td>
<td>(.002)</td>
<td>(.007)</td>
</tr>
<tr>
<td>6</td>
<td>9.2</td>
<td>2.3</td>
<td>.46</td>
<td>.48</td>
<td>.48</td>
<td>.29</td>
<td>.29</td>
</tr>
<tr>
<td></td>
<td>(.3)</td>
<td>(.1)</td>
<td>(.04)</td>
<td>(.03)</td>
<td>(.03)</td>
<td>(.02)</td>
<td>(.02)</td>
</tr>
<tr>
<td>7</td>
<td>27</td>
<td>5.6</td>
<td>4.0</td>
<td>1.22</td>
<td>1.22</td>
<td>1.22</td>
<td>1.22</td>
</tr>
<tr>
<td></td>
<td>(2)</td>
<td>(5)</td>
<td>(.4)</td>
<td>(.5)</td>
<td>(.4)</td>
<td>(.5)</td>
<td>(.5)</td>
</tr>
<tr>
<td>8</td>
<td>68</td>
<td>30</td>
<td>7</td>
<td>10.7</td>
<td>10.7</td>
<td>10.7</td>
<td>10.7</td>
</tr>
<tr>
<td></td>
<td>(4)</td>
<td>(2)</td>
<td>(.4)</td>
<td>(.4)</td>
<td>(.4)</td>
<td>(.4)</td>
<td>(.4)</td>
</tr>
</tbody>
</table>
Figure 2.1 Spectral Densities
Figure 2.2  Estimators Averaged Over 5 Randomly Selected Samples. (······) Wavelet Estimator, (--- - - - - -) Adaptive Estimator, (-- --) MASE, (--- - - - - -) True Spectral Density
Figure 2.3 Spectral Density Estimates of Two Grambsch Subjects. (-----) Wavelet Estimator, (---) Adaptive Estimator, (---) MASE, (-----) Periodogram
3 LIKELIHOOD APPROXIMATIONS IN BAYESIAN MULTIPLE CURVE FITTING

Carsten H. Botts and Michael Daniels

A paper to be submitted.

3.1 Abstract

We model functional data from many subjects with a regression spline linear mixed model. In this model, the expected values for any subject (conditioned on the random effects) can be written as the sum of a population curve and a subject specific deviate from this population curve. The population curve and the subject specific deviates are both modeled as b-splines with \(k\) and \(k'\) knots located at \(t_{fc}\) and \(t_{fc}'\), respectively. We sample from the posterior \(p(k, t_{fc}, k', t_{fc}'|y)\), where \(y\) is the observed data, using reversible jump MCMC methods. Sampling from this posterior distribution is complicated by the fact that no analytical form for \(p(y|k, t_{fc}, k', t_{fc}')\) exists. We explore a variety of approximations to this likelihood and study how each approximation penalizes linear mixed models with too many knots.

3.2 Introduction

Estimating the functional relationship between two variables \(Y\) and \(x\) has always been an active area of research in statistics. One of the simplest models relating a response, \(Y\), to a single predictor, \(x\), is the linear regression model

\[
Y(x) = f(x) + \epsilon = \beta_0 + \beta_1 x + \epsilon
\]

where the \(\epsilon\) is random error with mean 0. Needless to say, more complicated models have since been developed. Some of the most complicated models assume that the function \(f(x)\) can not be described by a finite number of parameters. These models are said to be nonparametric. In cases such as this, the model relating \(Y\) and \(x\) is simply given as

\[
Y(x) = f(x) + \epsilon. \tag{3.1}
\]
Because of this model’s popularity, a variety of nonparametric estimates of $f(x)$ have been developed.

A common technique in estimating this function nonparametrically is to smooth the responses. To be more specific, an estimator of the function at a particular value, $x_j$, can be written as

$$\hat{f}(x_j) = \sum_i W_\lambda(|x_i - x_j|)Y(x_i).$$

This estimator smooths the data. The degree to which this data is smoothed is governed by the weight, $W_\lambda(\cdot)$, assigned to the response which is $|x_i - x_j|$ units away from the response of interest. The weights, in turn, depend on the bandwidth selected, $\lambda$. The theoretical features and practical value of this smoother have been extensively studied (Nadaraya 1964; Priestley and Chao, 1972; Gasser and Muller, 1979, 1984; Rice, 1984; and Cleveland, 1979).

Another technique in estimating the function $f(x)$ is to assume that it is a smooth function which can be described with splines. Associated with each spline is a set of $k$ knots, $t_k$. Although the definition of a knot varies with the kind of spline it is associated with, a knot can generally be described as a point on the $x$ axis at which the function, $f(x)$, is not smooth. A piecewise cubic spline is written as

$$f(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3 + \sum_{i=1}^{k} (x - t_i)^3 a_{3+i},$$

where $(t_1, t_2, \ldots, t_k)$ are the knots, and $\alpha = (a_0, a_1, \ldots, a_{3+k})$ is a fixed set of parameters. In this case, the knots are points on the $x$ axis at which $f'''(x)$ is not continuous. The piecewise cubic spline written above illustrates the concept behind splines, but it is unstable and thus rarely used. B-splines, which are computationally more stable, are preferred. (Hastie and Tibshirani, 1990; Zhou and Shen, 2001). The functional form of a B-spline is more complex than that of the piecewise cubic spline, but is readily available (deBoor, 1978). If $f(x)$ is modeled as a B-spline, it is simply written as

$$f(x) = b(x, t_k)\alpha,$$

where $b(x, t_k)$ is the design vector associated with a b-spline evaluated at $x$ with $k$ knots at $t_k = (t_1, t_2, \ldots, t_k)$. If $f(x)$ were modeled as such, (3.1) could then be re-written as

$$Y(x) = b(x, t_k)\alpha + \epsilon,$$

and for observations collected at $(x_1, x_2, \ldots, x_n)$, it would be written as

$$Y(x) = \begin{pmatrix} Y(x_1) \\ Y(x_2) \\ \vdots \\ Y(x_n) \end{pmatrix} = \begin{bmatrix} b(x_1, t_k) \\ b(x_2, t_k) \\ \vdots \\ b(x_n, t_k) \end{bmatrix} \alpha + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}.$$
With this model, estimating $f(x)$ becomes a problem of estimating the number of knots, $k$, the locations of those knots, $t_k$, and $\alpha$.

A variety of methods have been developed in estimating $k$ and $t_k$. Halpern (1973) approached this problem using Bayesian methods. Allowing knots to only be placed at the design points in the experiment, he considered all of the subsets of the design points. Halpern assigned prior probabilities to all of these subsets, and calculated the corresponding posterior probabilities of these subsets. His estimator of the function was based on these posterior probabilities. Dension, Mallick, and Smith (1998) placed priors on the number of knots, $k$, and their locations, $t_k$. With these priors, they calculated a joint posterior distribution which included the variables $k$ and $t_k$ and then sampled from this posterior distribution using reversible jump MCMC methods. They too restricted the knots to be located only at the design points of the experiment. DiMatteo, Genovese, and Kass (2001) proposed a method similar to that of Denison, et al. They did not restrict the knots to be located only at the design points of the experiment, however, and they correctly penalize models with unnecessarily large values of $k$ by integrating out $\alpha$.

Friedman (1991), Stone et al. (1997), and Luo and Wahba (1997) try to solve the problem of knot selection using backward and/or forward knot selection. In the case of backward knot selection, one begins with a large set of knots, deletes a knot, and then compares the fit of the two models using a model selection criterion. This process continues until the "best" model has been identified. Zhou and Shen (2001) used an alternative method in identifying the locations of the knots. They constructed an algorithm which favored adding knots in locations where several knots had already been added. Lindstrom (1999) used similar methods when selecting knot locations.

Splines have also been used to model curves that vary within and between subjects (Crainiceanu et al., 2004; Brumback and Rice, 1996; Shi et al., 1996). These mixed models take the general form

$$Y_i(x_j) = f(x_j) + G_i(x_j) + \epsilon_{ij}$$

where $Y_i(x_j)$ is the observation of the $i^{th}$ individual at $x_j$, $f(x)$ can be thought of as a population curve, and $G_i(x)$ is a random curve specific to subject $i$. We model these two functions as b-splines with $k$ and $k'$ knots located at $t_k$ and $t_{k'}$, respectively. We then hope to learn about the fixed curve, $f(x)$, and random effect curve, $G_i(x)$, by sampling from the posterior $p(k, t_k, t_{k'}, y)$. This procedure, however, is complicated by the fact that the corresponding likelihood has no analytical form. As a result, we explore different ways to approximate this likelihood and study how these approximations penalize models with large values of $k'$. Section 3.3 of this paper discusses, in detail, the mixed linear model that we consider and the approximations to the intractable likelihood needed for inference. Section 3.4 then addresses how the approximations we use may penalize mixed linear models with too many random
effect knots. Section 3.5 describes two simulation studies performed to explore the penalty issue.

3.3 The model and its corresponding methods

Let $Y_i(x_j)$ be the observed value of the $i^{th}$ curve at $x_j$. We specify the following mixed model for functional data in this paper.

$$Y_i(x_j) = b(x_j, t_k)\alpha + b(x_j, t_{k'})\gamma_i + \epsilon_{ij},$$

where $\gamma_i \sim N(0, \Sigma_\gamma)$, $\epsilon_{ij} \sim N(0, \sigma^2)$, and $k' \leq k$. This model is a special case of the model given in (3.2). While the population curve, $f(x)$, takes the form $b(x, t_{fc})\alpha$, the more interesting part of model (3.3) is in the random effect curves. In letting $G_i(x) = b(x, t_{fc'})\gamma_i$ and $\gamma_i \sim N(0, \Sigma_\gamma)$, we assume that all individuals have the same random effect knots. However, we do not restrict $t_{fc'} \subset t_k$. This restriction is often imposed to ease computations, but it limits the flexibility of the model.

We place prior distributions on the number of fixed and random effect knots ($k$ and $k'$), the locations of the fixed and random effect knots ($t_k$ and $t_{k'}$), and $\alpha$. The prior distributions for these parameters and $\sigma^2$ are given below.

$$k \sim \text{Poi}(\mu_k), \quad \quad \quad p(t_k | k) \propto 1 \left( a < t_{k(1)} < t_{k(2)} < \cdots < t_{k(k)} < b \right),$$

$$k' \sim \text{Poi}(\mu_{k'}), \quad \quad \quad p(t_{k'} | k') \propto 1 \left( a < t_{k'(1)} < t_{k'(2)} < \cdots < t_{k'(k')} < b \right),$$

$$\alpha \sim \text{MVN}(0, \Sigma_\alpha = c \times I_{k+2})$$

$$p(\sigma^2) \propto \sigma^{-2}$$

where $t_{k(j)}$ ($t_{k'(j)}$) is the $j^{th}$ smallest knot in the vector $t_k$ ($t_{k'}$), $(a, b)$ is the domain of $f$ and $G$, and $c$ is a constant significantly greater than 0. The prior on $\Sigma_\gamma$ will, at this point, simply be referred to as $p(\Sigma_\gamma | k', t_{k'})$. Its functional form will be provided in Section 3.3.2.2.

Learning about the fixed and random effect functions now becomes a problem in estimating $k$, $t_k$, $k'$, and $t_{k'}$. We do this by sampling from the posterior distribution $p(k, t_k, k', t_{k'} | y)$. While this is intuitively appealing, a problem arises since the likelihood $p(y | k, t_k, k', t_{k'})$ is intractable. We thus consider a variety of ways to sample from this posterior, and we study how these methods penalize models with large values of $k'$.

Section 3.3.1 gives a basic strategy of how we plan to sample from $p(k, t_k, k', t_{k'} | y)$ and shows why the marginal likelihood $p(y | k, t_k, k', t_{k'})$ is necessary in this scheme. In Section 3.3.2 we show why this likelihood cannot be calculated, and how we chose to approximate it.
3.3.1 A general strategy for sampling from \( p(k, t_k, k', t_{k'})|y) \)

Reversible jump MCMC methods (Green, 1995) can be used when trying to sample from a distribution of a random variable \( \theta \) and \( \text{dim}(\theta) \). In this particular case, the dimensions of the vectors \( t_k \) and \( t_{k'} \) vary with the values of \( k \) and \( k' \), so reversible jump MCMC methods are necessary when sampling from the posterior distribution \( p(k, t_k, k', t_{k'}|y) \).

Rather than sampling the fixed and random effect knots together in one MCMC iteration, we sample from the posterior \( p(k, t_k, k', t_{k'}|y) \) using Gibbs Sampling. This was done to optimize the acceptance rate of the Markov chain. The steps of this Gibbs Sampler are given below.

Gibbs Sampler:

1. Call the current set of knots \((k_{\text{old}}, t_{k_{\text{old}}}, k'_{\text{old}}, t_{k'_{\text{old}}})\)

2. Sample from \( p(k, t_k|k'_{\text{old}}, t_{k'_{\text{old}}}, y) \)

   (a) Propose a move for the fixed effect knots. Accept with probability

   \[ P_{\text{Fixed}} = \min \left( 1, \frac{P_{\text{Jump}}(k_{\text{new}}, t_{k_{\text{new}}} \rightarrow k_{\text{old}}, t_{k_{\text{old}}}|k'_{\text{old}}, t_{k'_{\text{old}}}, y) p(k_{\text{new}}, t_{k_{\text{new}}}|k'_{\text{old}}, t_{k'_{\text{old}}}, y)}{P_{\text{Jump}}(k_{\text{old}}, t_{k_{\text{old}}} \rightarrow k_{\text{new}}, t_{k_{\text{new}}}|k'_{\text{old}}, t_{k'_{\text{old}}}) p(k_{\text{new}}, t_{k_{\text{new}}}|k'_{\text{old}}, t_{k'_{\text{old}}}, y)} \right), \]

   where \( P_{\text{Jump}}(k_{\text{old}}, t_{k_{\text{old}}} \rightarrow k_{\text{new}}, t_{k_{\text{new}}}|k'_{\text{old}}, t_{k'_{\text{old}}}, y) \) is the probability of jumping from \((k_{\text{old}}, t_{k_{\text{old}}})\) to \((k_{\text{new}}, t_{k_{\text{new}}})\) conditioned on the current value of \((k', t_{k'})\).

   (b) If accepted, change \((k_{\text{old}}, t_{k_{\text{old}}})\) to the accepted values \((k_{\text{new}}, t_{k_{\text{new}}})\).

3. Sample from \( p(k', t_{k'}|k_{\text{old}}, t_{k_{\text{old}}}, y) \)

   (a) Propose a move for the random effect knots. Accept with probability

   \[ P_{\text{Random}} = \min \left( 1, \frac{P_{\text{Jump}}(k'_{\text{new}}, t'_{k'_{\text{new}}} \rightarrow k'_{\text{old}}, t'_{k'_{\text{old}}}|k_{\text{old}}, t_{k_{\text{old}}}) p(k'_{\text{new}}, t'_{k'_{\text{new}}}|k_{\text{old}}, t_{k_{\text{old}}, y})}{P_{\text{Jump}}(k'_{\text{old}}, t'_{k'_{\text{old}}} \rightarrow k'_{\text{new}}, t'_{k'_{\text{new}}}|k_{\text{old}}, t_{k_{\text{old}}}) p(k'_{\text{new}}, t'_{k'_{\text{new}}}|k_{\text{old}}, t_{k_{\text{old}}, y})} \right), \]

   where \( P_{\text{Jump}}(k'_{\text{old}}, t'_{k'_{\text{old}}} \rightarrow k'_{\text{new}}, t'_{k'_{\text{new}}}|k_{\text{old}}, t_{k_{\text{old}}}) \) is the probability of jumping from \((k'_{\text{old}}, t_{k'_{\text{old}}})\) to \((k'_{\text{new}}, t_{k'_{\text{new}}})\) conditioned on the current value of \((k, t_k)\).
(b) If accepted, change the value of \((k_{old}', t_{old}')\) to the newly accepted pair \((k_{new}', t_{new}')\).

We now provide details on how the posterior ratios are calculated. The ratio in the first step of the Gibbs Sampler can be written as

\[
\frac{p(k_{new}, t_{new} | k_{old}', t_{old}')}{p(k_{old}, t_{old} | k_{old}', t_{old}')}
\]

Given the priors in Section 3.3, it is clear that (3.4) reduces to

\[
\frac{p(y | k_{new}, t_{new}, k_{old}', t_{old}')}{p(y | k_{old}, t_{old}, k_{old}', t_{old}')} = \frac{p(y | k_{new}, t_{new}, k_{old}', t_{old}')}{p(y | k_{old}, t_{old}, k_{old}', t_{old}')} \times \frac{p(t_{new} | k_{new})}{p(t_{old} | k_{old})}.
\]

The details of this algorithm and how the jump probabilities are calculated are given in the Appendix. The proof of detailed balance is given in the Appendix.

### 3.3.2 Calculating \(p(y | k, t, k', t')\)

In order to correctly implement the Gibbs sampler shown in Section 3.3.1, the marginal likelihood \(p(y | k, t, k', t')\) must be calculated. Unfortunately, no analytical form exists for this likelihood. The likelihood of interest can be expressed as
\[ p(y|k, k', k', t_{k'}) = \int_{\Sigma_{\gamma}} \int_{\alpha, \gamma} \prod_{i=1}^{n} \left[ \int_{\gamma_i} \int_{\sigma^2} \int_{\alpha, \gamma_i, \sigma^2} p(y_i|k, k', k', \alpha, \gamma_i, \sigma^2) p(\gamma_i|\Sigma_{\gamma}) d\gamma_i \right] p(\alpha) p(\Sigma_{\gamma}|k', t_{k'}) \, p(\sigma^2) \, d\alpha \, d\gamma_i \, d\sigma^2 \] 

\[ = \int_{\Sigma_{\gamma}} \int_{\alpha, \gamma} \prod_{i=1}^{n} \int_{\gamma_i} \int_{\sigma^2} \left[ \int_{\gamma_i} \int_{\alpha, \gamma_i, \sigma^2} p(y_i|k, k', k', \gamma_i, \sigma^2) p(\gamma_i|\Sigma_{\gamma}, k', t_{k'}) \, p(\alpha) \, p(\Sigma_{\gamma}|k', t_{k'}) \, p(\sigma^2) \, d\gamma_i \right. 
\int_{\alpha, \gamma_i, \sigma^2} \left. \int_{\alpha, \gamma_i, \sigma^2} \right] \right] \, d\alpha \, d\gamma_i \, d\sigma^2, \]

where

\[ p(y|k, k', t_{k'}, \Sigma_{\gamma}, \sigma^2) = \left| \mathbf{A}^{-1} \right|^\frac{1}{2} \prod_{i=1}^{n} (2\pi)^{-\frac{n}{2}} \left[ \mathbf{B}_{R_i} \mathbf{B}_{R_i}' + \sigma^2 \mathbf{I}_{m_i} \right]^{-\frac{1}{2}} \left| \Sigma_{\alpha} \right|^{-\frac{1}{2}} \]

\[ \times \exp \left( -\frac{1}{2} \left( \sum_{i=1}^{n} \mathbf{y}_i^T \left( \mathbf{B}_{R_i} \mathbf{B}_{R_i}' + \sigma^2 \mathbf{I}_{m_i} \right)^{-1} \mathbf{y}_i - \mathbf{d}^T \mathbf{A} \mathbf{d} \right) \right), \]

\[ \mathbf{d} = \mathbf{A}^{-1} \sum_{i=1}^{n} \mathbf{B}_{F_i}' \left( \mathbf{B}_{R_i} \mathbf{B}_{R_i}' + \sigma^2 \mathbf{I}_{m_i} \right)^{-1} \mathbf{y}_i \]

\[ \mathbf{A} = \sum_{i=1}^{n} \mathbf{B}_{F_i}' \left( \mathbf{B}_{R_i} \mathbf{B}_{R_i}' + \sigma^2 \mathbf{I}_{m_i} \right)^{-1} \]

\[ \mathbf{B}_{F_i} = \left( \mathbf{b} (x_{1_i}; t_k)^T, \mathbf{b} (x_{2_i}; t_k)^T, \ldots, \mathbf{b} (x_{m_i}; t_k)^T \right)^T, \]

\[ \mathbf{B}_{R_i} = \left( \mathbf{b} (x_{1_i}; t_{k'})^T, \mathbf{b} (x_{2_i}; t_{k'})^T, \ldots, \mathbf{b} (x_{m_i}; t_{k'})^T \right)^T, \]

\[ \mathbf{x}_i = (x_{1_i}, x_{2_i}, \ldots, x_{m_i}) \] is the vector containing the \( x \)-values corresponding to subject \( i \), \( n \) is the total number of subjects (curves) in the study, and \( m_i \) is the total number of observations for subject \( i \).

The integral in (3.6) can not be calculated analytically. Two different approximations to the likelihood \( p(y|k, k', t_{k'}) \) are explored in this paper. The first simply plugs in the maximum likelihood estimate of \( \Sigma_{\gamma} \) and \( \sigma^2 \). The second approximates (3.6) using a Laplace approximation. These two approximations are discussed in Sections 3.3.2.1 and 3.3.2.2, respectively.

### 3.3.2.1 The 'plugged in' approximation

The first approximation that we consider estimates \( p(y|k, k', t_{k'}) \) with

\[ p \left( y|k, k', t_{k'}, \hat{\Sigma}_{\gamma} (t_k, t_{k'}), \hat{\sigma}^2 (t_k, t_{k'}) \right) \]

where \( \hat{\Sigma}_{\gamma} (t_k, t_{k'}) \) and \( \hat{\sigma}^2 (t_k, t_{k'}) \) are the maximum likelihood estimators of \( \Sigma_{\gamma} \) and \( \sigma^2 \) corresponding to the model with fixed effect knots at \( t_k \) and random effect knots at \( t_{k'} \). The dependence of these maximum likelihood estimators on \( (t_k, t_{k'}) \) will be suppressed to simply \( \hat{\Sigma}_{\gamma} \) and \( \hat{\sigma}^2 \). With this approximation, the two likelihood ratios necessary in the reversible jump MCMC steps,
\[ p \left( y \mid k_{\text{new}}, t_{k_{\text{new}}}, k'_{\text{old}}, t_{k'_{\text{old}}} \right) \quad \text{and} \quad p \left( y \mid k_{\text{old}}, t_{k_{\text{old}}}, k'_{\text{new}}, t_{k'_{\text{new}}} \right) \]

are replaced with

\[ p \left( y \mid k_{\text{new}}, t_{k_{\text{new}}}, k'_{\text{old}}, t_{k'_{\text{old}}}, \Sigma_{\gamma}, \sigma^2 \right) \quad \text{and} \quad p \left( y \mid k_{\text{old}}, t_{k_{\text{old}}}, k'_{\text{new}}, t_{k'_{\text{new}}}, \Sigma_{\gamma}, \sigma^2 \right) \]

This is similar in spirit to the RJMCMC method employed by Dominici et al. (2002).

This approach ignores the penalty from increased dimension of \( \Sigma_\gamma \) but accounts for some penalty by integrating out the random effects themselves.

### 3.3.2.2 The Laplace approximation

An alternative, yet computationally more complex approach, is to integrate out \( \Sigma_\gamma \) and \( \sigma^2 \) using an approximation method. To estimate \( p(y \mid k, t_k, k', t'_{k'}) \) using Laplace methods (Tierney and Kadane, 1986), \( \Sigma_\gamma \) and \( \sigma^2 \) should be parametrized in such a way that the log-likelihood of the corresponding parameters is as close to normal as possible. We thus use the parametrization proposed by Pourhamadi (1999). The parameters he uses to specify a covariance matrix do not have to be constrained to guarantee the positive definite restriction of \( \Sigma_\gamma \). A normal approximation to the likelihood of these parameters will thus likely be more accurate due to removal of constraints on the original elements of \( \Sigma_\gamma \). Pourhamadi decomposes the inverse of the covariance matrix \( \Sigma_\gamma \) as \( \Sigma^{-1}_\gamma = TDT^T \), where

\[
T = \begin{bmatrix}
1 & 0 & 0 & \cdots & 0 \\
-\phi_{2,1} & 1 & 0 & \cdots & 0 \\
-\phi_{3,1} & -\phi_{3,2} & 1 & \cdots & 0 \\
\vdots & \ddots & \ddots & \ddots & \vdots \\
-\phi_{p,1} & -\phi_{p,2} & \cdots & -\phi_{p,p-1} & 1
\end{bmatrix}, \quad \text{and} \quad D = \begin{bmatrix}
\sigma^2_1 & 0 & \cdots & 0 \\
0 & \sigma^2_2 & \cdots & 0 \\
\vdots & \ddots & \ddots & \vdots \\
0 & \cdots & 0 & \sigma^2_p
\end{bmatrix}.
\]

These parameters can be calculated directly from the elements of the covariance matrix using the functions \( g(\cdot; \cdot) \) and \( h(\cdot; \cdot) \) where \( \sigma^2_j = g(\Sigma_\gamma; j) = \Sigma_j[j, j] - \sigma^2_j \Sigma_j^{-1} \sigma_j \) for \( 2 \leq j \leq p \) (\( \sigma^2_1 = \Sigma_1 [1, 1] \)) \( \phi_j = (\phi_{j,1}, \phi_{j,2}, \ldots, \phi_{j,j-1}) = h(\Sigma_\gamma; j) = \Sigma^{-1}_{\gamma,j} \sigma_j \), \( \Sigma^{-1}_{\gamma,j} \) is the minor matrix within \( \Sigma_\gamma \) composed of its first \( j - 1 \) columns and rows, and \( \sigma_j \) is the vector within \( \Sigma_\gamma \) composed of the first \( j - 1 \) elements of the \( j^{th} \) column. The values of \( \phi_{j,k} \) (\( 2 \leq j \leq p, 1 \leq k \leq p - 1 \)) and \( \log(\sigma^2_i) \) (\( 1 \leq i \leq p \)) are unconstrained and "integrated out" in the Laplace approximation.\(^1\) The Laplace approximation to the likelihood \( p(y \mid k, t_k, k', t'_{k'}) \) can then be calculated as

\(^1\)Given the Pourhamadi decomposition described in this section, the functional form of \( p(\Sigma_\gamma | k', t'_{k'}) \) can be written as

\[ p(\Sigma_\gamma | k', t'_{k'}) \propto \left( \prod_{i=1}^{p} \sigma^{-2}_i \right) \left( \phi_j, k \in (-\infty, +\infty), 2 \leq j \leq p, 1 \leq k \leq p - 1 \right). \]
The difference in how these two methods perform (with respect to penalizing models with large values of $k'$) is discussed in Section 3.4.

### 3.4 The penalty of the approximations

Both methods penalize models with large values of $k'$ (the plugged-in estimator averages over the random effects, and the Laplace estimator averages over the random effects and their covariance matrix), yet the magnitude of the difference between these penalties is unclear. The penalty associated with the Laplace estimator should be greater since it averages over $\Sigma_\gamma$ in addition to $\gamma$, but this difference may be trivial given the computational burden of the Laplace estimator. Section 3.4.1 shows how this problem can be thought of as a problem in model selection, and Section 3.4.2 discusses how the additional penalty incurred by integrating out $\Sigma_\gamma$ (the penalty of the Laplace approximation) can be quantified.

#### 3.4.1 Which approximation to use? A question of model selection

When selecting among $q$ models, $\{M_1, M_2, \ldots, M_q\}$, that model with the largest predictive density is typically selected (Wasserman, 2000). The predictive density for model $i$ is denoted $p(y|M_i)$ and is calculated as

$$p(y|M_i) = \int_{\Theta_i} \int_{\Theta_j} p(y|\theta_1, \theta_2, M_i) p(\theta_1, \theta_2 | M_i) \, d\theta_1, d\theta_2, \quad (3.7)$$

where $\theta_1$ and $\theta_2$ are two parameters corresponding to model $i$, and $p(\theta_1, \theta_2 | M_i)$ is the conditional prior distribution on these parameters. Assume that the first integral in (3.7) can be calculated, but the second can only be approximated using computationally expensive techniques. In such a case,
one would be left to select a model based on \( p\left(y|\hat{\theta}_{2i}, M_i\right) \) or \( \hat{p}(y|M_i) \), where \( \hat{p}(y|M_i) \) approximates the full integral in (3.7). If \( p\left(y|\hat{\theta}_{2i}, M_i\right) \) were used as a model selection criterion instead of \( \hat{p}(y|M_i) \), however, would the same model be selected? Putting this in the context of the problem presented in this paper, the question written above could be rephrased as 'Would the same model be selected under 

\[
p\left(y|k, \ell, k', \ell', \Sigma, \sigma^2\right)
\]

and \( \hat{p}_{\text{Laplace}}\left(y|k, \ell, k', \ell'\right) \)? One would expect that \( \hat{p}_{\text{Laplace}}\left(y|k, \ell, k', \ell'\right) \) would favor more parsimonious models since the parameters \( \Sigma \) and \( \sigma^2 \) have been averaged over. Some research has been devoted to a generalized version of this problem. This research tries to quantify how likelihoods at various stages in a hierarchy penalize non-parsimonious models. This is the subject of Section 3.4.2.

### 3.4.2 Quantifying the penalty of the Laplace approximation

We are interested in how the likelihood, evaluated at different stages of a hierarchical model, varies in its support for a model. Trevisani and Gelfand (TG) (2002) address a related issue. They begin by considering two models \( M_1 \) and \( M_2 \). \( M_1 \) is a general hierarchical model and will be written as

\[
M_1: \quad y \sim p(y|\theta) \quad \theta \sim p(\theta|\phi) \quad \text{and} \quad \phi \sim p(\phi|\psi),
\]

while \( M_2 \) will simply be written as \( y \sim p(y|\xi) \). TG study how the support for Model 2 changes when Model 1 is evaluated at different stages of its likelihood. They are specifically interested in how the quantities

\[
\frac{p(y|\xi)}{p(y|\theta)}, \quad \frac{p(y|\xi)}{\int p(y|\theta) p(\theta|\phi) d\theta}, \quad \text{and} \quad \frac{p(y|\xi)}{\int \int p(y|\theta) p(\theta|\phi) p(\phi|\psi) d\theta d\phi}
\]

compare. TG prove that \( E_{Y|\phi,\theta}(\Delta) > 0 \), where

\[
\Delta = \log\left(\frac{p(y|\xi)}{\int p(y|\theta) p(\theta|\phi) d\theta}\right) - \log\left(\frac{p(y|\xi)}{p(y|\theta)}\right) = \log\left(\frac{p(y|\theta)}{p(y|\phi)}\right).
\]

This implies that if Model 1 and 2 are being compared in terms of likelihood, support for Model 2, \( p(y|\xi) \), increases as more terms are integrated out of the hierarchy of Model 1.

The penalty considered in this paper is related to the quantity \( \Delta \). Consider two hierarchical models \( M_1^* \) and \( M_2^* \), where \( M_1^* \) is nested in \( M_2^* \) (\( M_1^* \subset M_2^* \)). These models can be written more explicitly as

\[
M_1^* : \quad y \sim p(y|\theta_1) \quad \theta_1 \sim p(\theta_1|\phi_1) \quad \phi_1 \sim p(\phi_1|\psi_1), \quad \text{(3.8)}
\]

and
\[ M^*_2 : \quad y \sim p(y|\theta_2) \quad \theta_2 \sim p(\theta_2|\phi_2) \quad \phi_2 \sim p(\phi_2|\psi_2). \quad (3.9) \]

The penalty in the support of Model 2 (the larger model) when evaluating both of these models at
the first stage rather than the second stage of the hierarchy can be calculated as

\[ \Delta_1 = \log \left( \frac{p(y|\theta_2)}{p(y|\theta_1)} \right) - \log \left( \frac{\int p(y|\theta_2) p(\theta_2|\phi_2) d\theta_2}{\int p(y|\theta_1) p(\theta_1|\phi_1) d\theta_1} \right), \]

and the penalty in the support for Model 2 when evaluating the models at the second rather than third
stage of the models can be written as

\[ \Delta_2 = \log \left( \frac{\int \int p(y|\theta_2) p(\theta_2|\phi_2) p(\phi_2|\psi_2) d\theta_2 d\phi_2}{\int \int p(y|\theta_1) p(\theta_1|\phi_1) p(\phi_1|\psi_1) d\theta_1 d\phi_1} \right) \]

The problem addressed in this paper equates \( \theta \) (as seen in (3.8) and (3.9)) with the fixed effect coefficients
\( \alpha \), and \( \phi \) (as seen in (3.8) with (3.9)) as \( \Sigma_\gamma \) and \( \sigma^2 \). The ratios of integrated likelihoods we are thus
interested in are \( LR_1, LR_2, \) and \( LR_3 \), are

\[ LR_1 = \log \left( \frac{p(y|k_2, t_{k_2}, k'_2, t_{k'_2}, \hat{\alpha}, \Sigma_\gamma, \sigma^2)}{p(y|k_2, t_{k_2}, k'_1, t_{k'_1}, \hat{\alpha}, \Sigma_\gamma, \sigma^2)} \right) \quad (3.10) \]

\[ LR_2 = \log \left( \frac{p(y|k_2, t_{k_2}, k'_2, t_{k'_2}, \Sigma_\gamma, \sigma^2)}{p(y|k_2, t_{k_2}, k'_1, t_{k'_1}, \Sigma_\gamma, \sigma^2)} \right) \quad (3.11) \]

\[ LR_3 = \log \left( \frac{p(y|k_2, t_{k_2}, k'_2, t_{k'_2})}{p(y|k_2, t_{k_2}, k'_1, t_{k'_1})} \right) \quad (3.12) \]

where \( \hat{\alpha} \) is the maximum likelihood estimator for \( \alpha \), \( k'_2 > k'_1 \), and \( t_{k'_1} < t_{k'_2} \). The value of \( LR_1 \) should
indicate how integrating out the random effects penalizes models with large values of \( k' \). \( LR_2 \) should
indicate how integrating out the fixed effects further modifies the support for a model with fixed and
random effect knots at \( t_{k_2} \) and \( t_{k'_2} \), and \( LR_3 \) should suggest whether integrating out the random effects
covariance matrix carries a penalty (in addition to the one incurred by averaging over the random effects)
for models with larger values of \( k' \).

To explore how much integrating out the random effects covariance matrix penalizes linear mixed
models with large values of \( k' \) (in addition to the penalty gained by averaging over the random effects),
we conducted two simulation studies. The first focuses on how the log of the ratios listed in (3.10) -
(3.12) vary, and the second runs reversible jump MCMC algorithms using the plugged in method.
and the Laplace approximation on various data sets. The results of the Gibbs sampler should reveal the additional penalty and the additional computational expense associated with the Laplace method. These simulation studies are described in Section 3.5.

3.5 Simulation studies

3.5.1 Likelihood ratios

Using 4 different “truths”, we simulated 500 data sets and calculated the average log of the ratios given in (3.10) - (3.12). Each data set simulated had \( n \) subjects, and for each subject there were 20 observations. For each individual, 5 of the 20 simulated observations were selected at random. As a result, \( m_i \), the number of observations recorded for subject \( i \), was 5 for all \( i \). Recall that the true models that we simulated from can be written as

\[
Y_i(x) = \mathbf{B}_F \alpha + \mathbf{B}_R \gamma_i + \epsilon_i,
\]

where \( Y_i(x) = (Y_i(x_{i1}), \ldots, Y_i(x_{i20}))' \), \( \gamma_i \sim N(0, \Sigma_\gamma) \), \( \epsilon_i \sim N(0, \sigma^2 I_5) \), \( \mathbf{B}_F \) and \( \mathbf{B}_R \) are b-spline design matrices (calculated using the 'ns' function in R) with \( t_k \) and \( t_k' \) knots, respectively, and \( (a, b) \), the domain of the response, is the interval \( (0, \pi) \). Table 3.1 gives the values of the parameters for each true model simulated from, and Figure 3.1 shows the function \( f(x) \) and 5 realized values of \( f(x) + G_i(x) \).

<table>
<thead>
<tr>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>( t_k )</td>
<td>(.35, 2.3, 2.4, 3)</td>
<td>(.2, .58, 1, 1.7, 2.8)</td>
<td>(1, 1.3, 1.4, 2.8, 3)</td>
</tr>
<tr>
<td>( t_k' )</td>
<td>(2.4)</td>
<td>(.2, .58)</td>
<td>(2.8, 3.0)</td>
</tr>
<tr>
<td>( \Sigma_\gamma )</td>
<td>( 7 I_{3 \times 3} )</td>
<td>( 2.5 \times I_{4 \times 4} )</td>
<td>( 8 \times I_{4 \times 4} )</td>
</tr>
<tr>
<td>( \sigma^2 )</td>
<td>.05</td>
<td>.05</td>
<td>.05</td>
</tr>
</tbody>
</table>

Table 3.1 True Models Simulated From

Table 3.2 shows the results of the simulations. For each data set simulated, we let \( t_{k'} = (t_{k'}, t_{\text{random}}) \rangle \) where \( t_{\text{random}} \sim \text{Uniform}(0, \pi) \). The values of \( t_{k3} \), and \( t_{k3}' \) were set to \( t_k \), and \( t_{k'} \), respectively (these values are given in Table 3.1). Note that \( \overline{LR}_j \) is the value of \( LR_j \) averaged over these 500 samples, and \( C\left(\{LR_j\}\right) \) is the number of samples in which \( LR_j \) is less than 0.

These results indicate that the Laplace approximation penalizes models with large values of \( k' \) more severely than the plugged-in approximation. In almost every simulated case, \( \overline{LR}_3 < \overline{LR}_2 < \overline{LR}_1 \). For Model 1 at \( n = 25 \) and \( n = 40 \), and Model 2 at \( n = 25 \), \( \overline{LR}_3 \) is negative while \( \overline{LR}_2 \) is positive. This suggests that, in these settings, the plugged in approximation favors adding an extra random effect knot while the Laplace approximation discourages it.
Table 3.2 Average of Likelihood Ratios

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>n = 25</td>
<td>n = 40</td>
<td>n = 25</td>
<td>n = 40</td>
</tr>
<tr>
<td>( \overline{LR}_1 )</td>
<td>1.52</td>
<td>1.57</td>
<td>1.81</td>
<td>2.12</td>
</tr>
<tr>
<td>( C({LR_1}) )</td>
<td>14</td>
<td>10</td>
<td>55</td>
<td>15</td>
</tr>
<tr>
<td>( \overline{LR}_2 )</td>
<td>1.46</td>
<td>1.54</td>
<td>1.59</td>
<td>1.98</td>
</tr>
<tr>
<td>( C({LR_2}) )</td>
<td>12</td>
<td>10</td>
<td>54</td>
<td>15</td>
</tr>
<tr>
<td>( \overline{LR}_3 )</td>
<td>-1.32</td>
<td>-1.43</td>
<td>-2.71</td>
<td>1.68</td>
</tr>
<tr>
<td>( C({LR_3}) )</td>
<td>255</td>
<td>304</td>
<td>268</td>
<td>144</td>
</tr>
</tbody>
</table>

55

\( C(\{LR_j\}) \), the cardinality of the sets \{LR_j : LR_j < 0\}, also suggest that the penalty associated with the Laplace approximation is significantly greater than that of the plugged-in approximation. Every time the log of the likelihood ratio considered is less than 0, the likelihood with the additional random effect knot is less than the likelihood without the unnecessary knot. In all simulated cases, \( C(\{LR_3\}) > C(\{LR_2\}) \) and \( C(\{LR_3\}) > C(\{LR_1\}) \).

The one anomaly in the results presented in Table 3.2 is seen in Model 3 at \( n = 40 \). In this case, \( \overline{LR}_3 \) is greater than both \( \overline{LR}_2 \) and \( \overline{LR}_1 \). This is a result of outliers, since \( \overline{LR}_3 \) is negative significantly more times than \( LR_2 \) or \( LR_1 \).

The results in Table 3.2 may be skewed by the large \("signal to noise"\) ratio \( \frac{d}{\sigma^2} \), where \( \Sigma_\gamma = d \times I \). To see how the results would be affected at smaller values of \( \delta = \frac{d}{\sigma^2} \), another simulation study was conducted. 250 data sets were simulated at \( n = 25 \) for all four models at various values of \( \delta \). \( C(\{LR_j\}) \), the cardinality of the set \{LR_j : LR_j < 0\}, and the fifth percentile of the likelihood ratios are recorded in Table 3.3-3.4. The fifth percentile is of interest because it ignores how the distribution of the likelihood ratios behaves when \( LR_j > 0 \). Recall that the eventual goal is to run a Gibbs sampler, and all cases where the likelihood ratio is greater than 1 (\( LR_j > 0 \) are equivalent (in that case, the move is accepted), yet all cases where the likelihood ratio is less than 1 are not equivalent. The results in these are clear: as \( \delta \) decreases, the penalty of the Laplace method decreases. However, the fifth percentiles of the distributions of \( LR_3 \) and the values of \( C(\{LR_3\}) \) still suggest that the Laplace method penalizes the complex model more so than the plugged in method.

3.5.2 Gibbs sampler

Five data sets were simulated from the first two models listed in Table 3.1, and the posterior distribution \( p(k, t_k, k', t_k' | y) \) was sampled from using the Gibbs Samplers discussed in Section 3.3 (one Gibbs Sampler approximates \( p(y | k, t_k, k', t_k') \) using the plugged in method, and the other approximates
Table 3.3 Average of Likelihood Ratios

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>Model 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th percentile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( C({LR_1}) )</td>
<td>0.09</td>
<td>0.23</td>
</tr>
<tr>
<td>( C({LR_2}) )</td>
<td>0.10</td>
<td>0.22</td>
</tr>
<tr>
<td>5th percentile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( C({LR_3}) )</td>
<td>1.8</td>
<td>2.3</td>
</tr>
</tbody>
</table>

Table 3.4 Average of Likelihood Ratios

<table>
<thead>
<tr>
<th></th>
<th>Model 3</th>
<th>Model 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>5th percentile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( C({LR_1}) )</td>
<td>21</td>
<td>24</td>
</tr>
<tr>
<td>( C({LR_2}) )</td>
<td>6.2</td>
<td>4.35</td>
</tr>
<tr>
<td>5th percentile</td>
<td></td>
<td></td>
</tr>
<tr>
<td>( C({LR_3}) )</td>
<td>84</td>
<td>57</td>
</tr>
</tbody>
</table>

\( p(y|k, t_k, k', t_{k'}) \) using the Laplace method). Each data set contained \( n \) subjects, and each subject had 20 observations, 5 of which were randomly selected (\( m_i = 5 \forall i \)). The average number of iterations for each chain was between 9000 and 10,000. For one chain, the Laplace method took approximately 10 times as long to run as the plugged-in method. The acceptance rate of the Laplace (on average, .15), was also lower than that of the plugged in (on average, .2). The marginal distributions of the posterior \( p(k, t_k, k', t_{k'}|y) \) corresponding to each Gibbs sampler are shown in Figures 3.2-3.9.

The marginal distributions of the resulting chains show, again, that the Laplace method, compared to the plugged-in method, favors models with smaller values of \( k' \). In each case, the mode of the marginal distributions of \( k' \) do not match; the peaks for the Laplace estimator occur at smaller values of \( k' \) than those of the plugged-in estimator. There also appears to be a difference in how well these two estimators locate the positions of the knots. The distributions of \( t_k \) and \( t_{k'} \) for the Laplace estimator are more peaked at the true locations than the plugged-in estimator. While both methods seem to accurately locate the true number and locations of the fixed effect knots, they both over estimate the number of random effect knots. This might be a result of the small sample size (of subjects, and per subject).
3.6 Discussion

Although the Laplace method is computationally more burdensome, it appears to penalize models with unnecessarily large values of $k'$ more appropriately than the plugged in method. The Laplace method is thus preferred. Other methods of sampling from $p(k, t_k, k', t_{k'} | y)$ certainly exist. An alternative approach would be to sample from $p(k, t_k, k', t_{k'}, \Sigma, \sigma^2 | y)$ and then extract the marginal distributions of $k$, $t_k$, $k'$, and $t_{k'}$ from this sample. This would require us to specify, however, a distribution that would generate candidate values of $\Sigma$ and $\sigma^2$. One would have to be very careful in specifying such a distribution since a poor selection can lead to a drastically low acceptance rate in the Markov chain. This sampling technique was thus not employed in this project.

Another issue that deserves more attention in future research is how the penalty of the Laplace approximation changes with the “signal-to-noise” ratio $\delta$. It is possible that the Laplace approximation, in cases where $\delta$ is small, is not accurate. Perhaps alternative parametrizations of $\Sigma$ could lead to greater accuracy of the Laplace approximation in such cases.

The authors are currently investigating other methods of reducing the number of random effect parameters. Rather than sampling from the posterior $p(k, t_k, k', t_{k'} | y)$, we set $t_k = t_{k'}$ and try to simultaneously reduce the number of principal component curves associated with the random effects by sampling from the posterior $p(k, t_k, r | y)$ where $r$ = the number of principal components.

3.7 References


### 3.8 Appendix

#### 3.8.1 RJMCMC details

This section gives the details on the Gibbs sampler used in this paper. The details show how a move is made from a set of old knots to a new set of knots. The old set of knots will be denoted as \( (k_{\text{old}}, t_{\text{old}}, k'_{\text{old}}, t'_{\text{old}}) \).

The first step of the Gibbs Sampler involves a step in the space of the fixed effect knots, \( \Omega^k \). Note that \( \Omega^k \) is the sum of spaces \( \Omega_1, \Omega_2, \ldots \) where \( \Omega_j = [a, b]^j \). Therefore \( \Omega^k \) can be written as \( \Omega^k = \bigoplus_{j=1}^\infty \Omega_j \).

The first decision to be made is whether to give birth, relocate, or kill a fixed knot. These probabilities are labeled as \( b_F, r_F, \) and \( d_F \), respectively, and are calculated as

\[
\begin{align*}
  b_F (k_{\text{old}}) &= 0.4 \times \min \left( 1, \frac{p_k (k_{\text{old}} + 1)}{p_k (k_{\text{old}})} \right), \\
  d_F (k_{\text{old}}) &= \begin{cases} 
  0 & k_{\text{old}} = k'_{\text{old}} \\
  0.4 \times \min \left( 1, \frac{p_k (k_{\text{old}} - 1)}{p_k (k_{\text{old}})} \right) & \text{o.w.}
  \end{cases}
\end{align*}
\]

and \( r_F (k_{\text{old}}) = 1 - b_F (k_{\text{old}}) - d_F (k_{\text{old}}) \), where \( p_k (k_{\text{old}}) \) is the prior distribution assigned to \( k \) evaluated at \( k_{\text{old}} \).

If birth is chosen, a fixed effect knot, \( t_{k_{\text{selected}}} \), is selected at random and a new knot is given birth to. The new knot, \( t_{k_{\text{new}}} \), is sampled from the distribution \( t_{k_{\text{new}}} \sim \text{TN}_a^b \left( t_{k_{\text{selected}}}, 0.2 \right) \) where \( \text{TN}_a^b \) is the normal distribution truncated at \( a \) and \( b \). If death is selected, then one of the fixed effect knots is uniformly selected and killed. When relocation is chosen, a fixed effect knot, \( t_{k_{\text{relocated}}} \), is uniformly selected and then relocated to a position \( t_{k_{\text{new}}}^{\text{loc}} \) where \( t_{k_{\text{new}}}^{\text{loc}} \sim \text{TN}_a^b \left( t_{k_{\text{relocated}}}, 0.2 \right) \). The jump probabilities
corresponding to each move is given in Table 3.5. Note: \( h(y|x) = \phi(y; x, 2) (\Phi(b; x, 2) - \Phi(a; x, 2))^{-1} \) where \( \phi(\cdot; \mu, \sigma^2) \) is the normal density with mean \( \mu \) and variance \( \sigma^2 \), and \( \Phi(\cdot; \mu, \sigma^2) \) is the cumulative distribution function of a normal density with mean \( \mu \) and variance \( \sigma^2 \).

<table>
<thead>
<tr>
<th>Move in ( \Omega^k )</th>
<th>( P_{\text{Jump}} \left( k_{\text{old}}, t_{k_{\text{old}}} \rightarrow k_{\text{new}}, t_{k_{\text{new}}} \mid k'<em>{\text{old}}, t</em>{k'_{\text{old}}} \right) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth</td>
<td>( b_F(k_{\text{old}}) \frac{1}{k_M} \sum_{x \in k_{\text{old}}} h(t_{k_{\text{new}}}^x</td>
</tr>
<tr>
<td>Relocation</td>
<td>( b_R(k_{\text{old}}') = \begin{cases} 0 &amp; k_{\text{old}}' = k_{\text{old}} \ A \times \min \left( 1, \frac{p_{k'}(k_{\text{old}}')}{p_{k'}(k_{\text{old}})} \right) &amp; \text{o.w.} \end{cases} ) ( d_R(k_{\text{old}}') = \begin{cases} 0 &amp; k_{\text{old}}' = 1 \ A \times \min \left( 1, \frac{p_{k'}(k_{\text{old}}')}{p_{k'}(k_{\text{old}})} \right) &amp; \text{o.w.} \end{cases} )</td>
</tr>
<tr>
<td>Death</td>
<td>( r_R(k_{\text{old}}') = 1 - b_R(k_{\text{old}}') - d_R(k_{\text{old}}') ), where ( p_{k'}(k') ) is the prior distribution of ( k' ) evaluated at ( k_{\text{old}}' ). If death is selected, then one of the random effect knots is uniformly selected and then killed. If birth is selected, then one of the random effect knots is uniformly selected (call this ( t_{k_{\text{old}}}^{\text{selected}} ), and a knot is added to the set of random effect knots (call this knot ( t_{k_{\text{old}}}^{\text{new}} )). If relocation is selected, then one of the random effect knots is uniformly selected, ( t_{k_{\text{old}}}^{\text{selected}} ), and moved to another location (call this ( t_{k_{\text{old}}}^{\text{new, loc}} )). Again, ( t_{k_{\text{old}}}^{\text{new, loc}} \sim T N_{a}(t_{k_{\text{old}}}^{\text{selected}}, 2) ). The jump probabilities associated with these moves are given in Table 3.6.</td>
</tr>
</tbody>
</table>
These jump probabilities do guarantee detailed balance, as shown in the Appendix.

### 3.8.2 Proof of detailed balance

In this case, it needs to be shown that

\[
P_{\text{Trans}}(k_\text{old}, t_{k_\text{old}}, k'_\text{old}, t_{k'_\text{old}} \rightarrow k_\text{new}, t_{k_\text{new}}, k'_\text{new}, t_{k'_\text{new}}) p(k_\text{old}, t_{k_\text{old}}, k'_\text{old}, t_{k'_\text{old}} | y) = P_{\text{Trans}}(k_\text{new}, t_{k_\text{new}}, k'_\text{new}, t_{k'_\text{new}} \rightarrow k_\text{old}, t_{k_\text{old}}, k'_\text{old}, t_{k'_\text{old}}) p(k_\text{new}, t_{k_\text{new}}, k'_\text{new}, t_{k'_\text{new}} | y)
\]

(Note: throughout the rest of this proof, \(t_{k_\text{new}}\) will be denoted as \(t_{k_\text{new}}\), \(k_\text{old}\) will be denoted as \(k_\text{old}\), and so on.) Assume that both a fixed and random knot are being added. In this case,

\[
P_{\text{Trans}}(k_0, t_{k_0}, k'_0, t_{k'_0} \rightarrow k_n, t_{k_n}, k'_n, t_{k'_n}) p(k_0, t_{k_0}, k'_0, t_{k'_0} | y) = b_F(k_0) \frac{1}{k_0} \sum_{z \in t_{k_0}} h(t_k^\text{new} | x) \min \left( \frac{P_{\text{Jump}}(k_n, t_{k_n} \rightarrow k_0, t_{k_0}, k'_0, t_{k'_0}) p(k_n, t_{k_n}, k'_n, t_{k'_n} | y)}{P_{\text{Jump}}(k_0, t_{k_0} \rightarrow k_n, t_{k_n}, k'_n, t_{k'_n}) p(k_0, t_{k_0}, k'_0, t_{k'_0} | y)} \right)
\]

\[
\times b_R(k'_0) \frac{1}{k'_0} \sum_{z' \in t'_{k'_0}} h(t_{k'}^\text{new} | x) \min \left( \frac{P_{\text{Jump}}(k'_n, t_{k'_n} \rightarrow k'_0, t_{k'_0}, k'_n, t_{k'_n}) p(k'_n, t_{k'_n}, k'_n, t_{k'_n} | y)}{P_{\text{Jump}}(k'_0, t_{k'_0} \rightarrow k'_n, t_{k'_n}, k'_n, t_{k'_n}) p(k'_0, t_{k'_0}, k'_n, t_{k'_n} | y)} \right)
\]

\[
\times p(k_0, t_{k_0}, k'_0, t_{k'_0} | y).
\]

Assume that the ratios within each min argument is < 1 (if this is not the case, the reciprocal of the ratios will be less than 1, and the proof can be done in the other direction), and recall that

1. \(b_F(k_0) \frac{1}{k_0} \sum_{z \in t_{k_0}} h(t_k^\text{new} | x) = P_{\text{Jump}}(k_0, t_{k_0} \rightarrow k_n, t_{k_n}, k'_0, t_{k'_0})\)

2. \(b_R(k'_0) \frac{1}{k'_0} \sum_{z' \in t'_{k'_0}} h(t_{k'}^\text{new} | x) = P_{\text{Jump}}(k'_0, t_{k'_0} \rightarrow k'_n, t_{k'_n}, k'_0, t_{k'_0})\)

3. \(p(k_n, t_{k_n} | k'_0, t_{k'_0}, y) / p(k_0, t_{k_0} | k'_0, t_{k'_0}, y) = \frac{p(y|k_n, t_{k_n}, k'_0, t_{k'_0}) p(k_n)}{p(y|k_0, t_{k_0}, k'_0, t_{k'_0}) p(k_0)}\)

4. \(p(k'_n, t_{k'_n} | k_n, t_{k_n}, y) / p(k'_0, t_{k'_0} | k_n, t_{k_n}, y) = \frac{p(y|k'_n, t_{k'_n}, k_n, t_{k_n}) p(k'_n)}{p(y|k'_0, t_{k'_0}, k_n, t_{k_n}) p(k'_0)}\)

With these four observations it is clear that the above expression reduces to
This argument only shows detailed balance when a fixed and random effect knot are added. Similar arguments can be made to prove detailed balance when a fixed and/or random effect knot is deleted or relocated.
Figure 3.1 Plots of Models 1-4. (—): $f(x)$, (---): $G_t(x)$, (——): $f(x)$. 
Figure 4.2 Marginal Distributions of $k$ and $\hat{k}$ for Truth 1 at $n = 25$.
Figure 3.3 Marginal Distributions of $t_k$ and $t_{k'}$ for Truth 1 at $n = 25$
Figure 3.4 Marginal Distributions of $k$ and $k'$ for Truth 2 at $n = 25$
Figure 3.5 Marginal Distributions of $t_k$ and $t_{k'}$ for Truth 2 at $n = 25$
Figure 3.6: Marginal Distributions of $k$ and $\lambda$ for Truth $I = u = 40$

Laplace

Plugged In
Figure 3.7  Marginal Distributions of $t_k$ and $t_{k'}$ for Truth 1 at $n = 40$
Figure 3.8: Marginal Distributions of $k$ and $k'$ for Truth 2 at $n = 40$.
Figure 3.9  Marginal Distributions of $t_k$ and $t_{k'}$ for Truth 2 at $n = 40$
4 PRINCIPAL COMPONENT REDUCTION IN LINEAR MIXED B-SPLINES

Carsten H. Botts and Michael Daniels

A paper to be submitted.

4.1 Abstract

We consider a linear mixed effects model for longitudinal trajectories. In this mixed model, the population curve and the subject specific deviates are both modeled as a b-spline with \( k \) knots located at \( t_k \). We propose a novel Bayesian method of identifying possible values of \( k \) and \( t_k \) while also reducing the number of principal component curves associated with the random effects. Specifically, we use reversible jump MCMC methods to sample from a posterior distribution of \( k, t_k, \) and \( r \), where \( r \) is the number of principal component curves retained. This method of knot selection and principal component reduction is studied through simulations and then applied to a data set.

4.2 Introduction

The variability observed in longitudinal data results from the variation between subjects and the variation within subjects. A common model which allows for this is the linear mixed effects model

\[
\mathbf{Y}_i = \mathbf{W}_i \beta + \mathbf{Z}_i b_i + \epsilon_i
\]

where \( \mathbf{Y}_i \) is the observed values of subject \( i \), \( \mathbf{W}_i \) and \( \mathbf{Z}_i \) are the fixed and random design matrices, respectively, \( b_i \) is the \( p \times 1 \) 'random effect' of subject \( i \), and is typically modeled as \( b_i \sim N(0, \Sigma) \), and \( \epsilon_i \sim N(0, \sigma^2 I) \) is random error. Estimating the magnitude of these two sources of variation now becomes a problem of estimating the within subject variance, \( \sigma^2 \), and the between subject covariance matrix, \( \Sigma \). The random effects covariance matrix, \( \Sigma \), can have a potentially large number of parameters. An unnecessarily large number of parameters can result in "overfitting" the data and lead to estimates with large variability. For this reason, statisticians have focused much attention on principal component
analysis; it allows one to estimate between subject variability without overparametrizing the random effects covariance matrix. This section briefly reviews principal component curves and shows how, in recent literature, the number of principal component curves associated with the random effects can be reduced.

Principal component analysis can be understood by writing the random effects in model (4.1) as the sum of \( p \) principal component curves,

\[
Y_i(x_j) = w_{ij}\beta + \sum_{l=1}^{p} \sqrt{\lambda_l}\delta_{lj}h_l(x_j) + \epsilon_{ij}
\]

where \( Y_i(x_j) \) is the observation of subject \( i \) at \( x_j \), \( w_{ij} \) is the \( j^{th} \) row of \( W_i \), \( \delta_{lj} \sim N(0,1) \), \( \lambda_l \) is the \( l^{th} \) largest eigenvalue of \( \Sigma \), and \( h_l(\cdot) \) is the \( l^{th} \) principal component curve. Recall that the principal component curves \( \{h_1(\cdot), h_2(\cdot), \ldots, h_p(\cdot)\} \) are orthogonal \( \int h_k(x)h_l(x)dx = 1(k = l) \) and, from \( h_1(\cdot), \ldots, h_p(\cdot) \), decreasingly account for the variability in the random effects. This follows since the sum in (4.2) pairs the \( l^{th} \) principal component curve with the \( l^{th} \) largest eigenvalue, \( \lambda_l \).

Equation (4.2) appeals to those who wish to reduce the number of parameters in \( \Sigma \); if only the first few principal component functions explain a majority of the variability in the random effects, then only those few principal component curves have to be retained and estimated. Reducing the dimension of \( \Sigma \) in this way raises an obvious question: how many principal component curves should be retained? James et al. (2000) suggested two methods for selecting the number of principal components to keep. Their first suggestion was to simply calculate the amount of variability that is accounted for in the first \( r \) principal components. The smallest value of \( r \) which accounted for at least 90% of the variability observed in the data was the number of principal components retained. The second method they suggested involves examining the log likelihood of the reduced rank model. This log likelihood, graphed against increasing values of \( r \), will obviously be an increasing function. If the log likelihood peaks at an early value of \( r \) and then plateaus, however, that value of \( r \) at which it peaks could be selected. Shi et al. (1996) considered the same problem, yet their methods for selecting the number of principal components was slightly different. They fit the full rank model and then selected a value of \( r \) which minimizes a cross-validation score. Research related to this technique of principal component reduction can also be found in Rice and Silverman (1991), and Rice and Wu (2001).

In this study, a Bayesian model is specified which identifies the fixed and random effects as b-splines with \( k \) knots at \( t_k \). We try to locate these knots while also trying to reduce the number of principal component curves associated with the random effects by sampling from the posterior distribution of \( k, t_k, \) and \( r \), where \( r \) is the number of principal component curves retained. Section 4.3 provides the details of the model considered in this paper, and Section 4.4 outlines how the posterior distribution is sampled. Section 4.5 presents results of a simulation study which evaluates this method of dimension...
reduction, and Section 4.6 describes a data set to which our methods were applied and then presents
the corresponding results.

4.3 The model

In this paper, we consider a mixed model similar to that given in (4.1). The design matrices, \( W_i \)
and \( Z_i \), however, correspond to b-spline design matrices, both with \( k \) knots located at \( t_k \). To identify
knots which describe the set well while reducing the number of principal component curves associated
with the random effects, we sample from the marginal posterior \( p(k, t_k, r | y) \), where \( r \) is the number of
principal components retained in the model.

The details of the model studied in this paper are given in this section. Section 4.3.1 describes the
linear mixed model we work with in more detail and discusses how we plan to sample from the posterior
\( p(k, t_k, r | y) \). Section 4.3.2 then specifies the prior distributions used in this study.

4.3.1 The mixed model

This section identifies the mixed model studied in this paper. The section also discusses, generally,
how we plan to reduce the number of principal components associated with the random effects. The
model considered in this paper is

\[
Y_i (x) = B (x; k, t_k) \alpha + \tilde{B} (x; k, t_k) \gamma_i + \epsilon_i, \tag{4.3}
\]

where \( B (x; k, t_k) \) is a b-spline design matrix with \( k \) knots at \( t_k \), \( \alpha \) is a vector of regression coefficients,
\( \tilde{B} (x; k, t_k) \) is an orthogonalized version of \( B (x; k, t_k) \) (\( \tilde{B} (x; k, t_k)' \tilde{B} (x; k, t_k) = I \); the reasons for this
orthogonality constraint are made clear in the Appendix), and \( \gamma_i \) are subject-specific random effect
coefficients such that \( \gamma_i \sim N (0, \Sigma_\gamma) \). This is similar in form to the mixed model seen in (4.1). Re-
expressing model (4.3) just as equation (4.2) re-expressed (4.1), we get

\[
Y_i (x_j) = B (x_j; k, t_k) \alpha + \sum_{l=1}^{p} \sqrt{\lambda_l} \delta_{ij} h_l (x_j) + \epsilon_{ij}, \tag{4.4}
\]

where \( \lambda_l \) is the \( l^{th} \) largest eigenvalue of \( \Sigma_\gamma \), \( \delta_{ij} \sim N (0, 1) \), \( \epsilon_{ij} \sim N (0, \sigma^2) \), and \( h_1 (\cdot), \ldots, h_p (\cdot) \) are the
\( p \) principal component curves associated with the random effects.

To reduce the number of parameters associated with the random effects, we hope to isolate only
those principal component curves which account for a majority of the variability observed in the random
effects. In other words, we wish to approximate (4.4) with

\[
Y_i (x_j) \approx B (x_j; k, t_k) \alpha + \sum_{l=1}^{r} \sqrt{\lambda_l} \delta_{ij} h_l (x_j) + \epsilon_{ij}
\]
where \( r \ll p \). We hope to achieve this principal component reduction by sampling from the posterior \( p(k, t_k, r|y) \), where \( r \) is the number of principal component curves retained. To sample from this posterior distribution, we use the fact that

\[
p(k, t_k, r|y) = p(r|k, t_k, y) p(k, t_k|y).
\]  

(4.5)

Decomposing the posterior distribution as in (4.5) reveals an obvious strategy for sampling from the posterior distribution \( p(k, t_k, r|y) \):

1. sample from \( p(y|k, t_k) \) and, (2) conditioned on the sampled value, sample from \( p(r|k, t_k, y) \). These two conditional posterior distributions are decomposed below.

\[
p(k, t_k|y) \propto p(y|k, t_k) p(t_k|k) p(k)
\]

\[
p(r|k, t_k, y)
\]

\[
\propto p(y, r|k, t_k) p(t_k|k) p(k)
\]

\[
\propto \left\{ \int_{\Sigma_k} \int_{\Sigma_{\gamma}^r} \int_{\sigma^2} p(y|k, t_k, r, \Sigma_{\gamma}^r, \sigma^2) p(\Sigma_{\gamma}^r|r, \Sigma_\gamma) p(r|\Sigma_\gamma) p(\sigma^2|k, t_k) p(\Sigma_\gamma|k, t_k) d\sigma^2 d\Sigma_{\gamma} d\Sigma_\gamma \right\}
\]

\[
\times p(t_k|k) p(k)
\]

where

\[
p(y|k, t_k, r, \Sigma_{\gamma}^r, \sigma^2) = \int_{\alpha} \left( \prod_{i=1}^{n} \int_{\gamma_i} p(y_i|k, t_k, \alpha, \gamma_i, \sigma^2) p(\gamma_i|r, \Sigma_{\gamma}^r) d\gamma_i \right) p(\alpha) d\alpha
\]

\[
y_i|k, t_k, \alpha, \gamma_i, \sigma^2 \sim N \left( B(x; k, t_k) \alpha + \tilde{B}(x; k, t_k) \gamma_i, \sigma^2 \mathbf{I} \right)
\]

\[
\gamma_i|r, \Sigma_{\gamma}^r \sim N(0, \Sigma_{\gamma}^r) \text{ with rank } (\Sigma_{\gamma}^r) = r,
\]

\( n \) is the number of subjects, and \( p(r|\Sigma_\gamma) \), \( p(\alpha), p(\Sigma_{\gamma}^r|r, \Sigma_\gamma), p(\sigma^2|k, t_k), p(\Sigma_\gamma|k, t_k), p(t_k|k) \), and \( p(k) \) are the priors discussed in the following section.

### 4.3.2 Priors

The prior on the number of principal components given \( \Sigma_\gamma \) is a key to the model and will guide how many principal components we keep. In this paper, we choose \( p(r|\Sigma_\gamma) \) as a prior which places a majority
of its weight on the largest value of \( r \) such that the eigenvalues \( \lambda_{(1)}, \ldots, \lambda_{(r)} \) account for no more than 90\% of the variability in \( \Sigma_\gamma \) (Note: \( \lambda_{(i)} \) is the \( i^{th} \) largest eigenvalue in \( \Sigma_\gamma \)). The prior that we selected was

\[
p(r|\Sigma_\gamma) \propto \exp \left( \frac{(d - r)^2}{\lambda_{(2)} / \lambda_{(1)}} \right)
\]

where \( d = \max (s) \) and \( s = \left\{ j : \left( \sum_{i=1}^{j} \lambda_i \right) / \left( \sum_{all \; \lambda} \lambda_i \right) \leq \min \left( 0.9, \lambda_{(1)} / \sum_{all \; \lambda} \lambda_i \right) \right\} \). The variance of this prior distribution, \( \lambda_{(2)}/\lambda_{(1)} \), allows the peakedness of the prior at \( d \) to increase as the ratio of the largest eigenvalue, \( \lambda_{(1)} \), to the second largest eigenvalue, \( \lambda_{(2)} \), increases. A plot of this prior given six different covariance structures is shown in Figure 4.1.

The other priors specified in this model are listed below.

\[
\begin{align*}
\alpha &\sim N(0, \Sigma_\alpha = 10000 \times I) \\
p\left( \Sigma_r^\gamma | \Sigma_\gamma \right) &\propto I(\text{rank} (\Sigma_r^\gamma) = r) \\
p(\sigma^2 | k, t_k) &\propto \sigma^{-2} \\
\sigma_2^2 \mid k, t_k &\propto I(|\Sigma_\gamma| > 0) \\
p(t_k | k) &\propto I(a < t_1 < t_2 < \ldots < t_k < b) \\
k &\sim \text{Poi} (\mu)
\end{align*}
\]

where \( \lambda_j \) is the \( j^{th} \) eigenvalue of \( \Sigma_\gamma \), \( t_i \) is the \( i^{th} \) smallest knot, and \( (a, b) \) is a known domain for the fixed and random effect curves. With the model completely specified, the posteriors in (4.5) can be sampled. This is the subject of Section 4.4.

### 4.4 Sampling \( p(k, t_k, r|y) \)

As indicated in Section 4.3 above, we plan to sample from the posterior distribution \( p(k, t_k, r|y) \) in two steps: (1) sample from \( p(k, t_k|y) \) and (2) sample from \( p(r|k, t_k, y) \). These two steps are described next.

Sampling from the posterior \( p(k, t_k|y) \) is the more difficult of the two steps. Complications arise since the likelihood \( p(y|k, t_k) \), can not be calculated analytically. Recall that this likelihood is formulated as

\[
p(y|k, t_k) = \int_{\Sigma_\gamma} \int_{\Sigma_r^\gamma} \int_{\sigma^2} \int_{\sigma_2^2} p(y|k, t_k, r, \Sigma_\gamma, \Sigma_r^\gamma) p(\Sigma_r^\gamma | \Sigma_\gamma) p(r | \Sigma_\gamma) p(\sigma^2 | k, t_k) p(\sigma_2^2 | k, t_k) d\sigma^2 d\Sigma_r^\gamma d\Sigma_\gamma
\]

(4.6)
The first two integrals are intractable. The likelihood $p(y|k, t, r, \Sigma_\gamma)$ is thus approximated using Laplace methods (Tierney and Kadane, 1986). Since we expect much of the penalty to be captured by the eigenvalues of $\Sigma_\gamma$, where $\Sigma_\gamma = [\Delta_1, \ldots, \Delta_r] \text{diag} (\lambda_1, \ldots, \lambda_r) [\Delta_1, \ldots, \Delta_r]^T$, we only consider these terms in the Laplace approximation to $\int_{\Sigma_\gamma} p(y, \Sigma_\gamma|k, t, r, \Sigma_\gamma) \, d\Sigma_\gamma$. We thus calculate the quantity $p_{\text{Laplace}}(y|k, t, r, \Sigma_\gamma)$, where

$$p(y|k, t, r, \Sigma_\gamma) \approx p_{\text{Laplace}}(y|k, t, r, \Sigma_\gamma) = \frac{1}{(2\pi)^{\frac{n+m}{2}}} \left| S \left( \hat{\lambda}_1, \ldots, \hat{\lambda}_r, \hat{\sigma}^2 \right) \right|^{-\frac{1}{2}} \times p(y, \hat{\Sigma}_\gamma, \hat{\sigma}^2|k, t, r, \Sigma_\gamma),$$

$$S (\lambda_1, \ldots, \lambda_r, \sigma^2) = -\left( \begin{array}{c} D_{rr} \\ D_{tr} \end{array} \right),$$

$$D_{rr} = \left( \begin{array}{c} D_{rr} \\ D_{tr} \end{array} \right) \left( \begin{array}{c} \partial^2 \log(p(y, \Sigma_\gamma^r, \sigma^2|k, t, r, \Sigma_\gamma)) \\ \partial \log(\lambda_j) \partial \log(\sigma^2) \end{array} \right),$$

$$D_{tr} = \left( \begin{array}{c} \partial^2 \log(p(y, \Sigma_\gamma^r, \sigma^2|k, t, r, \Sigma_\gamma)) \\ \partial \log(\lambda_j) \partial \log(\sigma^2) \end{array} \right),$$

$$\hat{\Sigma}_\gamma = [\hat{\Delta}_1, \hat{\Delta}_2, \ldots, \hat{\Delta}_r] \text{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_r) [\hat{\Delta}_1, \hat{\Delta}_2, \ldots, \hat{\Delta}_r]^T,$$

and the eigenvector, eigenvalue pairs $(\hat{\Delta}_j, \hat{\lambda}_j)$ are discussed below. (Note: expressions for the derivatives in $S$ are provided in the Appendix). To avoid fitting the reduced rank model $Y_i \sim N \left( B\alpha, B\Sigma_\gamma B' + \sigma^2 I \right)$ with $\alpha$ integrated out, we calculated $\hat{\Delta}_1, \ldots, \hat{\Delta}_r, \hat{\lambda}_1, \ldots, \hat{\lambda}_r$, and $\hat{\sigma}^2$ which are maximum likelihood estimates of $\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r$, and $\sigma^2$ under the model $Y_i \sim N \left( B\hat{\alpha}, B\Sigma_\gamma B' + \sigma^2 I \right)$, where $\hat{\alpha}$ is the value of $\alpha$ that maximizes the full-rank model and is consistent for $\alpha$. A closed form solution to these maximum likelihood estimators are given in the theorem below.

**Theorem 4.** The values of $\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r$, and $\sigma^2$ which maximize the likelihood

$$\text{Lik}(\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2|y, \alpha = \hat{\alpha}) = \text{Lik}(\Sigma_\gamma^r|y, \alpha = \hat{\alpha})$$

$$= \left| B\Sigma_\gamma B' + \sigma^2 I \right|^{-\frac{n}{2}} (2\pi)^{-\frac{n(m+2)}{2}} \times \prod_{i=1}^{n} \exp \left( -\frac{1}{2} \left( y_i - B\hat{\alpha} \right)' (B\Sigma_\gamma B' + \sigma^2 I)^{-1} (y_i - B\hat{\alpha}) \right)$$

where $n$ is the number of subjects, $m$ is the number of observations per subject, $\Sigma_\gamma^r = (\Delta_1, \ldots, \Delta_r) \text{diag}(\lambda_1, \ldots, \lambda_r) (\Delta_1, \ldots, \Delta_r)'$, and $\hat{\alpha}$ is the value of $\alpha$ that maximizes the full-rank model are.

\footnote{To ease the notational burden in this paper, $B(x; k, t_k)$ and $B^T(x; k, t_k)$ will be written as $B$ and $B^T$, respectively.}
1. \( \hat{\Delta}_j = \Delta_j \) where \( \hat{\Delta}_j \) is the maximum likelihood estimator of the full-rank model.

2. \( \hat{\sigma}^2 = \frac{1}{n} \left( \sum_{i=1}^{n} r_i (r_i - \sum_{j=1}^{r} (r_i \Delta_j)) \right) / (m - r) \)

3. \( \hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \left( r_i \Delta_j \right)^2 - \hat{\sigma}^2 \)

where \( r_i = y_i - B \delta \).

The proof of this theorem is given in the Appendix.

With this approximation to \( p(y | k, t_k, r, \Sigma_\gamma) \), we are still need to calculate \( p(y | k, t_k) \) where

\[
p(y | k, t_k) = \int_{\Sigma_\gamma} \sum_{r=1}^{p} p(y | k, t_k, r, \Sigma_\gamma) p(r | \Sigma_\gamma) p(\Sigma_\gamma | k, t_k).
\]

This is approximated with

\[
\hat{p}(y | k, t_k) = \sum_{r=1}^{p} \hat{p}_{\text{laplace}}(y | k, t_k, r, \hat{\Sigma}_\gamma) p(r | \hat{\Sigma}_\gamma) p(\hat{\Sigma}_\gamma | k, t_k).
\]

(Recall that \( p(\Sigma_\gamma | k, t_k) \) is a flat prior over all positive definite matrices. The influence of \( \Sigma_\gamma \) thus comes entirely from \( p(r | \Sigma_\gamma) \).) \( \hat{p}_w(y | k, t_k) \) can now be used to sample from the posterior \( p(k, t_k | y) \).

When sampling from the posterior \( p(k, t_k | y) \), it is important to remember that as \( k \) changes, the dimension of \( t_k \) changes. Reversible jump MCMC (Green, 1995) methods are thus required to sample from this posterior distribution. The details of this algorithm are given below:

---

**Algorithm I: Sampling from \( p(k, t_k | y) \)**

1. Call the current set of knots and their locations \( (k_{\text{old}}, t_{k_{\text{old}}}) \)

2. Propose a move to a new set of knots, \( (k_{\text{new}}, t_{k_{\text{new}}}) \).

3. Accept this move with probability \( \hat{p}_{\text{accept}} = \min (1, r_{\text{accept}}) \), where

\[
\hat{r}_{\text{accept}} = \frac{P_{\text{jump}}(k_{\text{new}}, t_{k_{\text{new}}} \rightarrow k_{\text{old}}, t_{k_{\text{old}}}) \hat{p}(y | k_{\text{new}}, t_{k_{\text{new}}}) p(k_{\text{new}}) p(t_{k_{\text{new}}})}{P_{\text{jump}}(k_{\text{old}}, t_{k_{\text{old}}} \rightarrow k_{\text{new}}, t_{k_{\text{new}}}) \hat{p}(y | k_{\text{old}}, t_{k_{\text{old}}}) p(k_{\text{old}}) p(t_{k_{\text{old}}})}
\]

where \( P_{\text{jump}}(k_{\text{old}}, t_{k_{\text{old}}} \rightarrow k_{\text{new}}, t_{k_{\text{new}}}) \) is the probability of proposing a move to \( k_{\text{new}}, t_{k_{\text{new}}} \) given you are at state \( k_{\text{old}}, t_{k_{\text{old}}} \).

4. If the move is accepted, replace \( k_{\text{old}}, t_{k_{\text{old}}} \) with \( k_{\text{new}}, t_{k_{\text{new}}} \)

---

More details of this RJMCMC algorithm are given in the Appendix. Proof of detailed balance is also given in the Appendix.
Sampling from $p(r|k, t_k, y)$ just involves sampling from a multinomial distribution with probabilities

$$p(r|k, t_k, y) \propto \hat{p}\text{Laplace}\left(y|k, t_k, r, \Sigma_r\right) p\left(r|\Sigma_r\right) p\left(\Sigma_r\right)$$

With our model and methods clearly defined, we now evaluate how effectively this technique of principal component reduction works. We thus performed a simulation study and applied our method to a real data set. The details are given in the following two sections.

### 4.5 Simulation study

To determine how well this technique reduces the number of principal component curves, we conducted a simulation study. From eight different settings, we simulated 50 data sets. Each data set contained 25 subjects with 20 observations each. These settings are given in Table 4.1, and differ with respect to the value of $\sigma^2$ and the magnitude of the eigenvalues of the random effects covariance matrix. For each data set, we sampled 100 values of $r$ from the conditional posterior $p(r|k = k_{\text{true}}, t_k = t_{k_{\text{true}}}, y)$. The prior $p(r|\Sigma_r)$ and conditional posterior $p(r|k = k_{\text{true}}, t_k = t_{k_{\text{true}}}, y)$ corresponding to each case are provided in Figure 4.2.

Based on the results given in Figure 4.2, it is clear that the methods proposed in this paper sensibly reduce the number of principal components. In cases 1 and 5, the posterior distribution of $r$ has a mode at large values, which should be the case since the diagonal elements in the random effects covariance matrix are all identical. Several principal components should be retained in these two cases. In cases 3, 4, 7, and 8, the mode of $p(r|k = k_{\text{true}}, t_k = t_{k_{\text{true}}}, y)$ is at smaller values of $r$. We would expect this to be the case since only the first few principal components (the first few diagonal elements in $\Sigma_r$) account for a majority of the variability in the random effects. The unusually high mode in case 2 is a result of the erratic and unusual behavior of the penalty introduced by the Laplace approximation.

These simulations also display the effect of $\sigma^2$ on the posterior $p(r|k = k_{\text{true}}, t_k = t_{k_{\text{true}}}, y)$. Clearly, the smaller the estimated value of $\sigma^2$, the larger the mode in $p(r|k = k_{\text{true}}, t_k = t_{k_{\text{true}}}, y)$ (compare cases 1 and 5, 2 and 6, 3 and 7, and 4 and 8). While the reason for these differences still remains unclear, it is likely that with larger values of $\sigma^2$ the variability produced by $\sigma^2$ swamps the variability produced by the smaller eigenvalues.

### 4.6 Application

We apply this technique of dimension reduction to a longitudinal data set in Grambsch et al. (2002). The data set was collected to learn of potential differences between those women who are clinically depressed, and those who are not. They hypothesized that the frequency at which Leutenizing Hormone
Table 4.1 Varying values of $\Sigma_\gamma$ with $k^{true} = 9$ and $t_k^{true} = (.1,.2,.35,1,1.5,1.7,2.5)$

<table>
<thead>
<tr>
<th>Case</th>
<th>$\Sigma_\gamma$</th>
<th>$\sigma^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>$5 \times I_9$</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>diag(5,4,3,2,1,.1,.01,.01,.01)</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>diag(10,4,1,.1,.01,.01,.01,.01,.01)</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>diag(100,4,1,.1,.01,.01,.01,.01,.01)</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>$5 \times I_9$</td>
<td>.05</td>
</tr>
<tr>
<td>6</td>
<td>diag(5,4,3,2,1,.1,.01,.01,.01)</td>
<td>.05</td>
</tr>
<tr>
<td>7</td>
<td>diag(10,4,1,.1,.01,.01,.01,.01,.01)</td>
<td>.05</td>
</tr>
<tr>
<td>8</td>
<td>diag(100,4,1,.1,.01,.01,.01,.01,.01)</td>
<td>.05</td>
</tr>
</tbody>
</table>

(LH) is released in the blood stream is different between these two groups. 26 women in the study were diagnosed as clinically depressed, and 24 were not. Blood was drawn every 8 minutes for 10 hours, giving a total of 49 observations in time for each subject. Only those subjects with a complete data set (with no missing observations) were considered in this analysis.

We used the methods outlined in this paper to estimate the population curves across time for each group. The estimator for each population curve was calculated using Bayesian model averaging. More specifically, estimates of each population curve were averaged over the values of $k$, $t_k$, and $r$ sampled in the reversible jump MCMC chain. A majority of the weight in these averages thus corresponds to those values of $k$, $t_k$, and $r$ which had a high posterior probability. The estimator for the population curve can thus be written as

$$\hat{f}_{BMA}(x) = \mathbb{E}(f(x)|y) = \frac{1}{n_j} \sum_{i=1}^{n_j} \mathbb{E}(f(x)|k_i,t_k,r_i,y)$$

where $n_j$ is the total number of triplets $(k, t_k, r)$ sampled in the MCMC chain ($n_j = 10000$ for each group (the depressed and not depressed)), and $(k_j, t_{k_j}, r_j)$ is the $j^{th}$ triplet of $(k, t_k, r)$ sampled in the MCMC chain. The results are shown in Figures 4.3 and 4.4.

Looking at the time series of each group, it initially appears that the concentration of LH varies more rapidly in those women who are not depressed. Such differences should be noticed in the distributions of $k$, $t_k$, $r$, and the shape of any principal component curves retained. The depressed group has more fixed effect knots than the not depressed, and only one principal component curve was retained in each group. The first two principal component curves are also shown in Figures 4.3 and 4.4. The principal component curves retained suggest, as the time series do, that the concentration of LH changes more quickly in those who are not depressed. For those women, the one principal component curve kept is
more discontinuous ("bumpy") than the one principal component curve kept for the group of depressed women. Despite these results, additional simulations show that for longer chains (with larger values of $n_j$), the posterior mode for $k$ moves to the right for each group, while the mode for the posterior of $r$ stays at 1.

### 4.7 Discussion

The methods proposed in this paper seem to effectively reduce the dimension of the random effects covariance matrix. What is especially appealing about the method proposed in this paper is that it allows the knots, their locations, and the number of principal components retained to be simultaneously identified in an automated fashion. Moreover, these parameters are found by sampling from the posterior distribution $p(k, t_k, r|y)$. A single set of parameters is thus not selected; a variety of likely parameters are identified. Estimates of model parameters can thus be calculated using Bayesian model averaging (as was seen in the example).

The most notable drawback of the methods proposed in this paper regard the accuracy of the Laplace approximation. We initially hypothesized that failure to integrate out the eigenvectors of $\Sigma^r$ resulted in the Laplace approximation’s lack of accuracy. Additional simulation studies performed made it clear that this was not the case. Poor estimates of the eigenvalues (especially when $\lambda_{(\text{max})} \approx \lambda_{(\text{min})}$) can also make the Laplace approximation less credible. Correcting for poor estimates of may involve placing shrinkage priors on $\Sigma^r$ and $\Sigma^r$ (Yang and Berger, 1994; Daniels and Kass, 2001).

### 4.8 References


4.9 Appendix

4.9.1 RJMCMC details

In the reversible jump MCMC algorithm, a decision must initially be made to either kill, give birth to, or relocate a knot. The probability that any one of these moves will be made are calculated through the functions \( d(k_{\text{old}}) \), \( b(k_{\text{old}}) \), and \( r(k_{\text{old}}) \), respectively, where

\[
d(k_{\text{old}}) = \begin{cases} 
0 & k_{\text{old}} = 1 \\
0.4 \times \min \left( 1, \frac{p_k(k_{\text{old}} - 1)}{p_k(k_{\text{old}})} \right) & \text{o.w.}
\end{cases}
\]

\[b(k_{\text{old}}) = 0.4 \times \min \left( 1, \frac{p_k(k_{\text{old}} + 1)}{p_k(k_{\text{old}})} \right), \quad r(k_{\text{old}}) = 1 - b(k_{\text{old}}) - d(k_{\text{old}}), \quad \text{where } p_k(k_{\text{old}}) \text{ is the prior distribution of } k \text{ evaluated at } k_{\text{old}}.
\]

If birth is chosen, a fixed effect knot is selected at random and a new knot is given birth to. The new knot, \( t_{\text{new}} \), is sampled from the distribution \( t_{\text{new}} \sim TN^b_\alpha(t_{\text{selected}}, 2) \) where \( TN^b_\alpha \) is the normal distribution truncated at \( a \) and \( b \) and \( t_{\text{selected}} \) is the fixed effect knot initially selected. If death is selected, then one of the fixed effect knots, \( t_{\text{deleted}} \), is uniformly selected and killed. When relocation is chosen, a fixed effect knot, \( t_{\text{relocated}} \), is uniformly selected and then relocated to a position \( t_{\text{new loc}} \) where \( t_{\text{new loc}} \sim TN^b_\alpha(t_{\text{relocated}}, 2) \). The jump probabilities corresponding to each move is given in Table 4.2.

Note: \( h(y|x) = \phi(y; x, 2) (\Phi(c; x, 2) - \Phi(0; x, 2))^{-1} \) where \( \phi(\cdot; \mu, \sigma^2) \) is the normal density with mean \( \mu \) and variance \( \sigma^2 \) and \( \Phi(\cdot; \mu, \sigma^2) \) is the cumulative distribution function of a normal density with mean \( \mu \) and variance \( \sigma^2 \).
Table 4.2 Jump Probabilities

These jump probabilities do guarantee detailed balance, as proven in the Appendix.

4.9.2 Proof of detailed balance

To prove detailed balance, we must show that

$$P_{\text{Trans}} (k_{\text{old}}, t_{\text{old}} \rightarrow k_{\text{new}}, t_{\text{new}}) p(k_{\text{old}}, t_{\text{old}} | y) = P_{\text{Trans}} (k_{\text{new}}, t_{\text{new}} \rightarrow k_{\text{old}}, t_{\text{old}}) p(k_{\text{new}}, t_{\text{new}} | y)$$

(4.8)

holds, where $P_{\text{Trans}} (k_{\text{old}}, t_{\text{old}} \rightarrow k_{\text{new}}, t_{\text{new}})$ is the transition probability of moving from a model with $k_{\text{old}}, t_{\text{old}}$ knots to a model with $k_{\text{new}}, t_{\text{new}}$ knots. Assume, for the moment, that death of a knot is proposed. In this case, $k_{\text{new}} = k_{\text{old}} - 1$. Then

$$P_{\text{Trans}} (k_{\text{old}}, t_{\text{kold}} \rightarrow k_{\text{new}}, t_{\text{knew}}) p(k_{\text{old}}, t_{\text{kold}} | y) = d(k_{\text{old}}) \frac{1}{k_{\text{old}}} \min \left\{ 1, \frac{P_{\text{Jump}} (k_{\text{new}}, t_{\text{knew}} \rightarrow k_{\text{old}}, t_{\text{kold}}) p(y | k_{\text{new}}, t_{\text{knew}}) p(k_{\text{new}})}{P_{\text{Jump}} (k_{\text{old}}, t_{\text{kold}} \rightarrow k_{\text{new}}, t_{\text{knew}}) p(y | k_{\text{old}}, t_{\text{kold}}) p(k_{\text{old}})} \right\}$$

$$\times p(k_{\text{old}}, t_{\text{kold}} | y)$$

where the term in curly brackets is $P_{\text{Trans}} (k_{\text{old}}, t_{\text{kold}} \rightarrow k_{\text{new}}, t_{\text{knew}})$. Assume $r_{\text{accept}} < 1$ in this case. (If this is not the case, then $r_{\text{accept}}$ for the move from the new to old knots is < 1, and the proof can be done in the opposite direction). Then we can continue with the above as
4.9.3 Proof of Theorem 4

\[ \text{Lik} (\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2 | y, \alpha = \hat{\alpha}) = \text{Lik} (\sigma^2, \Sigma_\gamma | y, \alpha = \hat{\alpha}) = \prod_{i=1}^{n} \int_{\gamma_i} \frac{1}{(2\pi)^{\frac{r}{2}} (\sigma^2)^{\frac{r}{2}}} \exp \left( -\frac{1}{2} (y_i - B\hat{\alpha} - \hat{B}\gamma_i) \right)^t \left( y_i - B\hat{\alpha} - \hat{B}\gamma_i \right) p (\gamma_i | \Sigma_\gamma) \, d\gamma_i. \]

where \( \Sigma_\gamma = [\Delta_1, \ldots, \Delta_r] \text{ diag} (\lambda_1, \ldots, \lambda_r) [\Delta_1, \ldots, \Delta_r]^t \). Setting \( \gamma_i = \sum_{j=1}^{r} \Delta_j \sqrt{\lambda_j} z_{ji} \) where \( z_{ji} \sim N (0, 1) \) \( \forall i, j \), the above becomes

\[ = \prod_{i=1}^{n} \int_{\mathbb{R}^r} \frac{1}{(2\pi)^{\frac{r}{2}} (\sigma^2)^{\frac{r}{2}}} \exp \left( -\frac{1}{2} \left( y_i - B\hat{\alpha} - \hat{B} \sum_{j=1}^{r} \Delta_j \sqrt{\lambda_j} z_{ji} \right) \right)^t \left( y_i - B\hat{\alpha} - \hat{B} \sum_{j=1}^{r} \Delta_j \sqrt{\lambda_j} z_{ji} \right) \times \prod_{j=1}^{r} \phi (z_{ji}) \, dz_{ji} \]

\[ = c (\sigma^2)^{-\frac{n}{2}} \left( \prod_{j=1}^{r} \left( \frac{\sigma^2}{\sigma^2 + \lambda_j} \right)^{\frac{r}{2}} \right) \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^{n} r_i \frac{r_i}{2} \sum_{j=1}^{r} \Delta_j^2 \lambda_j \right) \]

where \( r_i = y_i - B\hat{\alpha} \) and \( c \) is a constant independent of \( \sigma^2 \), any \( \lambda_j \), and any \( \Delta_j \). The log-likelihood can then be expressed as
\[
\Rightarrow \log \text{Lik} (\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2 | y, \alpha = \hat{\alpha})
\]
\[
= \log(c) - \frac{nm}{2} \log(\sigma^2) + \sum_{j=1}^{r} \frac{n}{2} \log \left( \frac{\sigma^2}{\sigma^2 + \lambda_j} \right) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} r_i^T r_i + \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{r} \left( r_i^T \tilde{B} \Delta_j \right)^2 \lambda_j
\]
Maximizing this log likelihood with the obvious orthogonality restrictions on the eigenvectors, we take the derivative of
\[
g (\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2)
\]
\[
= \log \text{Lik} (\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2 | y, \alpha = \hat{\alpha}) - \sum_{j=1}^{r} \delta_j (\Delta_j' \Delta_j - 1) - 2 \sum_{k<j}^{r} \delta_{kj} \Delta_k^T \Delta_j
\]
Methods equivalent to those in Flury (1988) show that
\[
\delta_{kj} = 0 \quad \forall \ k, j
\]
\[
\delta_j = \frac{\lambda_j}{2\sigma^2 (\sigma^2 + \lambda_j)} \sum_{i=1}^{n} \left( \text{tr} \left( r_i^T \tilde{B} \Delta_j \right) \right) \Delta_j' \tilde{B}' r_i
\]
The maximum likelihood estimate of \( \Delta_j \) can thus be found by solving the equation
\[
\frac{\partial g (\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2)}{2\partial \Delta_j} = \lambda_j \sum_{i=1}^{n} \frac{1}{2} \left( \text{tr} \left( r_i^T \tilde{B} \Delta_j \right) \right) \tilde{B}' r_i
\]
\[
- \frac{\lambda_j}{2\sigma^2 (\sigma^2 + \lambda_j)} \sum_{i=1}^{n} \left( \text{tr} \left( r_i^T \tilde{B} \Delta_j \right) \right) \Delta_j' \tilde{B}' r_i \Delta_j = 0
\]
which is independent of \( r \) and, as can be seen when multiplying through by \( \frac{\sigma^2 (\sigma^2 + \lambda_j)}{\lambda_j} \), independent of the eigenvalues and \( \sigma^2 \). We can conclude that

1. \( \hat{\Delta}_j = \Delta_j \) where \( \hat{\Delta}_j \) maximizes the full-rank model.

The maximum likelihood estimates of \( \lambda_j \) and \( \sigma^2 \) can then be found by solving the equations
\[
\frac{\partial g (\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2)}{\partial \lambda_j} = -\sum_{i=1}^{n} \frac{1}{\sigma^2 + \lambda_j} + \sum_{i=1}^{n} \left( r_i^T \tilde{B} \Delta_j \right)^2 = 0
\]
and
\[
\frac{\partial g(\Delta_1, \ldots, \Delta_r, \lambda_1, \ldots, \lambda_r, \sigma^2)}{\partial \sigma^2} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{r} \left( r'_i \tilde{B} \Delta_j \right)^2 \lambda_j \left( \frac{2\sigma^2 + \lambda_j}{(\sigma^2 + \lambda_j)^{\frac{3}{2}}} \right) - \frac{\sigma^2 \lambda_j}{\lambda_j + \sigma^2} \\
+ n \sigma^2 - \sum_{i=1}^{n} r'_i r_i = 0
\]

Using these two equations to solve for the maximum likelihood estimators for \( \lambda_j \) and \( \sigma^2 \), we get

2. \( \hat{\lambda}_j = \frac{1}{n} \sum_{i=1}^{n} \left( r'_i \tilde{B} \Delta_j \right)^2 - \hat{\sigma}^2 \), and

3. \( \hat{\sigma}^2 = \frac{1}{n} \left( \sum_{i=1}^{n} r'_i r_i - \sum_{i=1}^{n} \sum_{j=1}^{r} \left( r'_i \tilde{B} \Delta_j \right)^2 \right) / (m - r) \)

4.9.4 Derivatives for \( S \)

This section gives the equations necessary to calculate the derivatives in the \( S \) matrix in (4.7).

\[
p(y, \Sigma, \sigma^2 | k, t, r, \Sigma_y)
= \left| nB \left( B \Sigma_y \tilde{B}' + \sigma^2 I \right)^{-1} \right|^{-\frac{1}{2}} \left( 2\pi \right)^{\frac{n}{2}} \left| B \Sigma_y \tilde{B}' + \sigma^2 I \right|^{-\frac{n}{2}} |\Sigma_\alpha|^{-\frac{1}{2}}
\times \exp \left( -\frac{1}{2} \left( \sum_{i} y'_i \left( \tilde{B} \Sigma_y \tilde{B}' + \sigma^2 I \right)^{-1} y_i \right) - \left( \sum_{i} y'_i \tilde{B}' \right) \left( nB \left( B \Sigma_y \tilde{B}' + \sigma^2 I \right)^{-1} \right) \left( \sum_{i} \tilde{B} y_i \right) \right)
\]

The second derivatives used to calculate \( S \) can be derived in recognizing that for any matrix \( A \) and for any parameters \( \theta_1 \) and \( \theta_2 \)

\[
\frac{\partial^2 \log (|A|)}{\partial \theta_1 \partial \theta_2} = \text{tr} \left[ - \left( A^{-1} \frac{\partial \Lambda}{\partial \theta_1} A^{-1} \right)' + A^{-1} \frac{\partial^2 \Lambda}{\partial \theta_1 \partial \theta_2} \right]
\]

\[
\frac{\partial^2}{\partial \theta_1 \partial \theta_2} (y'_i A^{-1} y_i) = -y'_i \left( -A^{-1} \left( \frac{\partial \Lambda}{\partial \theta_1} A^{-1} \frac{\partial \Lambda}{\partial \theta_2} - \frac{\partial^2 \Lambda}{\partial \theta_1 \theta_2} \right) + \frac{\partial \Lambda}{\partial \theta_1} A^{-1} \frac{\partial \Lambda}{\partial \theta_2} \right) A^{-1} \right) y_i
\]

The derivatives specific to this problem include:
\[
\frac{\partial \left( \tilde{B} \Sigma \tilde{B}' + \sigma^2 I \right)}{\partial \lambda_j} = \tilde{B} [\Delta_1, \ldots, \Delta_r] \text{diag} \left( 0, \ldots, 0, 1, 0, \ldots, 0 \right) [\Delta_1, \ldots, \Delta_r]' \tilde{B}'
\]

\[
\frac{\partial^2 \left( \tilde{B} \Sigma \tilde{B}' + \sigma^2 I \right)}{\partial \lambda_j \partial \lambda_k} = 0
\]

\[
\frac{\partial^2 \left( \tilde{B} \Sigma \tilde{B}' + \sigma^2 I \right)}{\partial \lambda_j \partial \sigma^2} = 0
\]

\[
\frac{\partial \left( \tilde{B} \Sigma \tilde{B}' + \sigma^2 I \right)}{\partial \sigma^2} = I
\]

Note: Since \( I (\Sigma, \sigma^2, \alpha) = \left( \begin{array}{cc} I (\Sigma, \sigma^2) & 0 \\ 0 & I (\alpha) \end{array} \right) \), the Laplace approximation can just as easily be calculated by taking the derivative of \( p(y, \Sigma, \sigma^2, \alpha | k, t, r, \Sigma) \). This was the posterior differentiated in our analysis.
Figure 4.1: Plots of $p(x|x^{-1})$ for various values of $\beta$. 

(\begin{align*} 
\beta &= \text{diag}(10.1, 10.4, 4.3, 2.2) \\
\beta &= \text{diag}(10.1, 14.2, 9.2, 2.1) \\
\beta &= \text{diag}(5.5, 5.5, 5.5, 5.5) \\
\beta &= \text{diag}(40.3, 4.3, 2.2, 1.1) \\
\beta &= \text{diag}(50.4, 4.3, 2.2, 1.1) \\
\beta &= \text{diag}(50.4, 4.3, 2.2, 1.1) \\
\end{align*})
Figure 4.2 Simulation Results for Cases 1-8.
Figure 4.3 (a): (——) Avg. Time Series for Depressed Subjects, (······) Estimated $f(x)$. (b) - (d): Distributions of $k$, $t_k$, $r$. (e): (——) First Principal Component Curve Retained, (······): Second Principal Component Curve Retained
Figure 4.4  (a): (----): Avg. Time Series for Non-Depressed Subjects, 
(· · · · · · ) Estimated \( f(x) \).  (b)-(d): Distributions of \( k, \ t_k, \ r \).  (e): 
(---): First Principal Component Curve Retained.  (· · · · · ·): Sec­ond Principal Component Curve Retained
CONCLUSIONS

5.1 General contributions of thesis

The techniques developed in this thesis make a positive contribution to the literature of Bayesian methods in function estimation. The shrinkage estimator proposed in Chapter 2 estimates the spectral density of a stationary time series well. It outperformed a variety of other estimators in both capturing a spectral density’s peaks and smooth regions. This shrinkage estimator is especially appealing because it is consistent, easy to compute, and does not require the specification of a particular parametric time series model.

Chapters 3 and 4 developed methods for mixed linear regression splines. In these two chapters, the trajectories observed in longitudinal studies were assumed to be the sum of a population curve with \( k \) knots at \( t_k \), a subject specific curve with \( k' \) knots at \( t_{k'} \), and random error. Chapter 3 specifically discussed two methods of sampling from the posterior \( p(k, t_k, k', t_{k'}, y) \). The two techniques explored in this chapter varied in how they approximated the likelihood corresponding to this posterior, \( p(y|k, t_k, k', t_{k'}) \). In the first approach, the likelihood was estimated using Laplace methods. In the second approach, maximum likelihood estimates of variance parameters were plugged in (these parameters could not analytically be averaged over). Of these two methods, the Laplace approximation is preferred. Despite its additional computational expense, it more appropriately penalizes mixed models with a large number of random effect knots.

Chapter 4 identified the fixed and random effect knots in a similar fashion while simultaneously reducing the number of principal component curves associated with the random effects. This was done by restricting \( t_k = t_{k'} \) and then sampling from \( p(k, t_k, r|y) \), where \( r \) is the number of principal components retained. This method of knot location and principal component reduction was proven effective when it was applied to a data set and studied through simulations.

5.2 Open issues

As with any research project, this work has raised just as many (if not more) questions as it has answered. This section addresses open issues and areas of further research.
As was mentioned in the conclusion of Chapter 2, it would be interesting to see how the shrinkage estimator behaves for other functions that can be specified parametrically. For the spectral densities of spatial processes, for example, methods similar to those in Chapter 2 could be developed. More innovative techniques may be required, however, since spatial processes operate in two dimensions (as opposed to one dimension for time). Methods might also have to be developed to accommodate irregularly spaced data.

Several issues in Chapter 3 remain unresolved. The first regards the penalty associated with the Laplace approximation, and how the magnitude of the smaller eigenvalues in $\Sigma_\gamma$, relative to $\sigma^2$, affect it. Simulation results presented in Chapter 3 suggest that the Laplace approximation is more accurate (correctly penalizes parameter rich models) as the magnitude of these eigenvalues increase relative to $\sigma^2$. A second issue to consider are alternative ways to sample from the posterior. One option would be to sample from the posterior $p(k, t_k, k', t_{k'}, \Sigma_\gamma, \sigma^2 | y)$, and from this sample, examine the marginal posterior distributions of $k$, $t_k$, $k'$, and $t_{k'}$. This technique, however, would require specifying a distribution from which candidate values of $\Sigma_\gamma$ and $\sigma^2$ would be generated. Specifying such a distribution is not a trivial task; the acceptance rate of the Markov chain can vary greatly with different selections. This technique was not used in this dissertation since we were unable to find a distribution that gave a reasonable acceptance rate.

Two issues in the fourth chapter which deserve further attention are: (1) the penalty associated with integrating out the eigenvalues, and (2) the estimate of $\Sigma_\gamma$. The penalty of the Laplace approximation is captured in the term $|S|^{-\frac{1}{2}}$, given in (4.7). This term depends on how the eigenvalues compare to $\sigma^2$, and how $m$ (the number of observations per subject) compares with $n$ (the number of subjects). These issues need further exploration. One avenue to explore may be the poor quality of the estimated eigenvalue. Perhaps they impair the behavior of the penalty. Improved estimates of the eigenvalues may be obtained by placing shrinkage priors on the covariance matrices $\Sigma_\gamma$ and $\Sigma_\gamma'$ (Yang and Berger, 1994; Daniels and Kass, 2001). The estimate of $\Sigma_\gamma$ is another concern in Chapter 4. This estimate is conditioned on $\alpha = \hat{\alpha}_{ml}$, where $\hat{\alpha}_{ml}$ maximizes the full-rank model. Ideally, of course, the estimate of $\alpha$ should maximize the reduced rank model. Regardless of which estimator is selected, asymptotically no difference should exist between the two since both are consistent for the truth.

5.3 References

6 ACKNOWLEDGEMENTS

This dissertation would not have been possible without the aid, support, and tremendous patience of my advisor, Mike Daniels. More than half of this degree belongs to him.

I would like to thank Mr. Goldman, Mrs. Hobbs, Mr. Nichols, Dr. Crist, Dr. Bachman, Dr. Vogt, Dr. Oweiss, Dr. Kulawiec, and Dr. Stern. I really really enjoyed going to all of your classes. You are wonderful teachers. (Mrs. Funk, I would have put you on this list, but I am still angry about that D.)

I would also like to thank Mr. Evans, Mrs. Weiser, Dr. Shulman, and Father Scotniki. The fact that I made it this far has a lot to do with the confidence you four gave me.

I also need to thank my friends that saw me through the good and bad times of the past six years: Brooke, Rhonda, Brown, Calhoon, Mike Case, Mike Eraas, Jeff, David, Jamie, and Winn.

Ultimately, I would like to thank my family (Torben, Mor, Far, and Marge). I dedicate this dissertation to you four.