BAYESIAN CLASSIFICATION AND SURVIVAL ANALYSIS WITH CURVE PREDICTORS

A Dissertation

by

XIAOHUI WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

December 2006

Major Subject: Statistics

BAYESIAN CLASSIFICATION AND SURVIVAL ANALYSIS WITH CURVE PREDICTORS

A Dissertation

by

XIAOHUI WANG

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Bani K. Mallick
Committee Members,	Michael Longnecker
	Suojin Wang
	Goong Chen
Head of Department,	Simon J. Sheather

December 2006

Major Subject: Statistics

ABSTRACT

Bayesian Classification and Survival Analysis
with Curve Predictors.
(December 2006)
Xiaohui Wang, B.E., University of Science and Technology at Beijing;
M.S., Beijing Jiaotong University;
M.S., Texas A&M University
Chair of Advisory Committee: Dr. Bani K. Mallick

We propose classification models for binary and multicategory data where the predictor is a random function. The functional predictor could be irregularly and sparsely sampled or characterized by high dimension and sharp localized changes. In the former case, we employ Bayesian modeling utilizing flexible spline basis which is widely used for functional regression. In the latter case, we use Bayesian modeling with wavelet basis functions which have nice approximation properties over a large class of functional spaces and can accommodate varieties of functional forms observed in real life applications. We develop an unified hierarchical model which accommodates both the adaptive spline or wavelet based function estimation model as well as the logistic classification model. These two models are coupled together to borrow strengths from each other in this unified hierarchical framework. The use of Gibbs sampling with conjugate priors for posterior inference makes the method computationally feasible. We compare the performance of the proposed models with the naive models as well as existing alternatives by analyzing simulated as well as real data. We also propose a Bayesian unified hierarchical model based on a proportional hazards

model and generalized linear model for survival analysis with irregular longitudinal covariates. This relatively simple joint model has two advantages. One is that using spline basis simplifies the parameterizations while a flexible non-linear pattern of the function is captured. The other is that joint modeling framework allows sharing of the information between the regression of functional predictors and proportional hazards modeling of survival data to improve the efficiency of estimation. The novel method can be used not only for one functional predictor case, but also for multiple functional predictors case. Our methods are applied to analyze real data sets and compared with a parameterized regression method.

To my lovely family

ACKNOWLEDGEMENTS

At the moment of finishing this dissertation, I am having a lot of warm feelings in my mind. First and foremost, I would like to thank my advisor, Dr. Bani K. Mallick, for his kind support, precious and valuable advice and patient guidance during my graduate study and in the preparation of this dissertation. I am very thankful that he led me to an interesting research field-Bayesian modeling, chose a challenging topic for me, and always shared his enthusiasm and knowledge in such a generous way.

I also want to thank my advisory committee members, Dr. Michael Longnecker, Dr. Suojin Wang and Dr. Goong Chen for their service and constructive comments on this dissertation. From each of their classes or seminars, I have learned a lot of things that will continue to influence my career.

I am very grateful to our former and current department heads, Dr. James Calvin and Dr. Simon Sheather, to Director of Graduate Studies, Dr. P. Fred Dahm, and to Lead Office Associate, Ms. Marilyn Randall for their always stand-by administrative assistance. A thank-you goes to Ms. Julie Hagen Carroll for her help with my teaching at A&M. Many of my classmates deserve special thanks for their help and friendship.

At the end, I want to thank my dear husband, Dr. Zhaosheng Feng, for his love, companionship and steady support over the years. I am indebted to two persons. My mother gave me unselfish and unconditional love, support and encouragement through out my life. Without her love, I doubt I would have accomplished my work successfully. My lovely son, Sebert Xi Feng, has brightened my whole life ever since he was born in the third year of my graduate study.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

FIGURE Page

FIGURE Page

CHAPTER I

INTRODUCTION

Functional data analysis has emerged as a new area of statistical research with wide range of applications. Functional measurements are ordered measurements on a regular grid, usually displayed using curves. A lot of data collected about cancer, growth, weather, goods fall into this category. Although the data is recorded on discrete points for each individual, the basic unit of information is the entire observed function rather than a string of numbers. The popular problems of interest for instance are smoothing, regression, curve classification and discrimination, and conditional functional quantiles (Ramsey and Silverman, 1997, 2002, Dimatteo *et al*, 2001, Kass *et al*, 2003). There are real challenging problems, both from methodological and applied points of view, in developing functional adaptation of usual techniques to these new kinds of problems.

Hierarchical modeling is a generalization of regression methods, in which regression coefficients themselves are also given a higher level model, whose parameters are also estimated from data. It can be used for a variety of purposes, including prediction, data reduction, and causal inference from experiments and observational studies (Kreft and De Leeuw, 1998, Snijders and Bosker, 1999, Raudenbush and Bryk, 2002, and Hox, 2002). Bayesian approach may have advantages to hierarchical modeling. In Bayesian paradigm, model assessment is more straightforward, computational implementation is typically much easier, and historical data can be easily incorporated

---

The format and style follow that of *Journal of the American Statistical Association*.

into the inference procedure. Because the Bayesian approach can capture all relevant sources of uncertainty, it has been developed to fit data much more realistically using hierarchical models with large number of parameters to model heterogeneity, interactions and nonlinearity (Gelman *et al* 2003, Gelman, 2004, Carlin and Louis, 1996, and Denison *et al*, 2002).

In this dissertation, our attention first is focused on classification of functional curves. Due to different types of functional curves, the challenges of classification come from different aspects. One type is irregularly and sparsely sampled curves so that only a fragment of each curve has been observed. This places popular analysis procedures such as linear discriminant analysis and support vector machine at inapplicable category so that the classification task is difficult. The other type is curves characterized by high dimension and many sharp local changes. The regression of the spiky curves requires careful investigation. We study both dichotomous and multicategory cases. Generally, in case of that the underlying function is smooth, spline-based method is a plausible choice and some summarization refer to Ruppert *et al* (2003). On the other hand, in functional context splines lack the ability to fit sharp localized changes in curves and there exists better alternatives such as wavelets. We propose a Bayesian hierarchical modeling method, which combines information from the curves predictors as well as from the associated categorical variables for classification by unifying functional regression and logistic classification models.

Except curve classification, the next topic of interest is time-to-event data analysis. There were some studies contributing to model time-to-event data with time-dependent covariates. However, it seems that not enough studies have focused on the case that covariates are functional curves measured on different time points. In this dissertation, we propose an efficient joint model using spline basis, in which the usage of the splines simplifies the parameterizations and the joint modeling framework

allows that regression model of functional curves and proportional hazards model of survival data exchange information with each other.

1.1 Irregular Curve Classification Problem

Classification using functional data is a relatively new concept. Recently curve classification has been studied in several scientific fields with significant applications like longevity status classification of medflies based on initial egg-laying curves (Muller and Stadtmuller, 2004), dynamic classification of genes for DNA microarray with repeated measurements (Alter *et al*, 2000), mutation detection (Pfeiffer *et al*, 2002) and serum proteomic pattern diagnostics for early detection of cardiotoxicity (Petricoin *et al*, 2004).

The case that only a fragment of each curve has been observed makes the classification even more difficult. In this situation the two common approaches to discriminant analysis, regularization and filtering methods, can break down (James and Hastie, 2001). James and Hastie (2001) proposed a functional linear discriminant analysis (FLDA) method to overcome the above difficulties. The procedure uses a spline curve plus random error to model observations from each individual. The spline is modeled using a basis function multiplied by a q-dimensional coefficient vector, which is modeled using a Gaussian distribution with common covariance matrix for all classes. In the literature, it seems that some kind of Bayesian methods for irregular curve classification have not been presented before.

A key component of splines, knot selection, requires sophisticated algorithms that can be computationally extensive. For example, Friedman's multivariate adaptive regression splines (MARS) algorithm (Friedman 1991 and Friedman 1993), Denison *et al*'s Bayesian MARS algorithm (Denison *et al*, 1998), and Smith and Kohn's Bayesian knot selector based on Gibbs sampling (Smith and Kohn, 1996) are dedicated to this

problem. However, when data is sparse and the span of curves is not too long, one can use a very fine latter as knots locations. Successful spline applications for various purposes without deeply involving knot selection are found in James (2002) and James and Hastie (2001). Spline-based method requires the choice of smoothing parameters. A standard approach for smoothing parameter estimation, generalized cross-validation (GCV), occasionally leads to instability of function estimation because it does no smoothing sometime (Carroll *et al*, 1999, Berry *et al*, 2002). There exists an alternative smoothing parameter selector that is to place prior distribution on smoothing parameter. Berry *et al* found that it is an automatic way of avoiding the possibility of gross undersmoothing. We also adopt this Bayesian smoothing parameter selector.

In most of the existing models, a naive approach is used, where the estimates from the regression model are simply plugged into the classification model. Thus the regression model is unaware of additional information in the categorical outcomes and completely overlooks the classification problem. The novelty of the proposed Bayesian model lies in its jointly modeling concept to draw information from the curves as well as from the associated categorical responses for classification by unifying spline-based functional regression and logistic classification models.

1.2 Spiky Curve Classification Problem

Classification of functional curves, especially spiky curves, is a relatively new challenging task. There seems no present work especially contributing to spiky curve classification although the precise classification for this type of curves is in demand. For example, proteomic methods simultaneously detect the expression of hundreds or thousands of different proteins in biological samples, and are gaining increased attention in biomedical research. In surface enhanced or matrix assisted laser desorption

and ionization technologies, usually an array surface is first created from the proteins of interest and then a mass spectrum is constructed using mass spectroscopy instrument. This mass spectrometry functional data has already shown promise in the identification of biomarker patterns for cancer diagnosis and classification (Conrads *et al* 2003, Hingorani *et al* 2003, Petricoin *et al* 2004). The spectrum functions are irregular, high dimensional and characterized by local jumps so wavelets are suitable basis functions to represent these curves with occasional singularities.

In a functional context, wavelets is better alternatives than splines in case of fitting sharp localized changes in curves. The Bayesian wavelet modeling used in this dissertation manages to take advantage of this fact as wavelets have nice approximation properties over a large class of functional spaces (Daubechies, 1992) that can accommodate almost all the functional forms observed in real life applications. Indeed, this richness of the wavelet representation provides the backbone for the popular frequentist wavelet shrinkage estimators of Donoho and Johnstone (1994,1995), which are the precursors of the more recent Bayesian wavelet estimation models (Abramovich *et al* 1998, Clyde *et al* 1998, Clyde and George 2000, Vidakovic 1998).

The novelty of our proposed Bayesian model is that it draws information from the functional data as well as from the associated categorical variables for classification by unifying wavelet-based functional regression and logistic classification models. In this process, it enjoys the advantages of Bayesian modeling in wavelet domain as well as the information from the classification indicator variables. On the other hand, a naive approach is used in most of the existing models, where the estimates from the regression model are simply plugged into the classification model. The disadvantage of the naive method is that the regression process completely overlooks the classification problem because it is unaware of additional information in the categorical variables.

1.3 Time to Event Data Analysis

The jointly hierarchical modeling idea can be extended to time-to-event data analysis with time-dependent covariates. Both parametric and semiparametric models are available to model survival data. Commonly used parametric models include the exponential and Weibull models, which are attractive in their simplicity and the easy interpretability of their components. In practice, however, semiparametric proportional hazards models are widely used, since they impose no particular shape on the survival curves. Especially in case of jointly modeling longitudinal and survival data, proportional hazards model is usually employed. For example, a general approach in Wulfsohn and Tsiatis (1997) combines a proportional hazards model for survival and a random effects model for regression. There are also existing Bayesian methods that use the same approach to construct the model as Wulfsohn and Tsiatis (1997). For example, Faucett and Thomas (1996) considered same random effects and proportional hazards model with noninformative priors on all parameters, while Ibrahim, Chen and Sinha (2004) modeled bivariate longitudinal and survival data by assuming both of two covariates measure a true unobservable univariate measure. There are various studies extended this type of work using either some kind of stochastic process (Wang and Taylor, 2001, Brown and Ibrahim, 2003) or standard computer packages (Guo and Carlin, 2003).

An apparent advantage of the joint modeling approach is that it can give efficient estimation by making a direct link between the survival and longitudinal covariate. However, the parametric form of the functional covariate may be inappropriate in some settings. Also, the assumption of independence over longitudinal measurements of same individual is a very strong assumption, which could be violated in most cases. These aforementioned Bayesian or non-Bayesian methods could not consider

these problems thoroughly.

We propose a relatively simple semi-parametric joint model using spline basis, in which the usage of the splines simplifies the parameterizations and the joint modeling framework allows the regression of functional predictors and proportional hazards modeling of survival data benefit from each other. The novel method can be used not only for one functional predictor case, but also for multiple functional predictors case. We consider survival data analysis in both situations. Another advantage of the proposed method is that regression coefficients are interpretable based on converting by spline basis.

1.4 Outline

The rest of this dissertation contains four main components, each of which is discussed in different chapters. The four components (parts) are:

1. Irregular curve classification using splines
2. High dimensional spiky curve classification using wavelets
3. Bayesian survival analysis using proportional hazards model and generalized linear regression
4. Conclusions

In the Chapter II, we first illustrate the motivating example, pediatric research about bone mineral acquisition, which lead us to irregular curve classification problem. Next we build the unified Bayesian spline-based classification model to solve the special type of classification problem. The model then is easily extended to multicategory classification case. Model choices and prediction procedure are also discussed. To illustrate the capability of our method, it is applied onto a simulated data set and

a real world data sets, and compared with several other methods. Chapter III further develops to address high-dimensional spiky curve classification problem. Because wavelets can fit spiky curve better than splines, we construct Bayesian wavelet-based classification model in a unified framework. Different model choices and extension to multicategory case are included. Applications and comparisons of several methods are conducted on a simulated data set and several real world data sets. In Chapter IV, we turn our eyes to survival analysis with irregular curve covariates, such as analysis of primary biliary cirrhosis (PBC) patients data. We develop Bayesian unified hierarchical model based on proportional hazard model and generalized linear model, which can be conveniently extended to multiple curve covariates. Bayes factor calculation is derived to select different models. We apply the proposed model on PBC patients data to study treatment effect and the relationship between survival status and two functional predictors, bilirubin and albumin levels. Finally, we give conclusions on presented work and discuss possible extensions in Chapter V.

CHAPTER II

IRREGULAR CURVE CLASSIFICATION USING SPLINES

2.1 Motivation Example

Despite the proliferation of pediatric research in the past two decades, there remain some controversies about bone mineral acquisition (Bachrach *et al*, 1999). For example, ethnic difference in bone mass have been observed in some (Gilsanz *et al*, 1998, Wang *et al*, 1997, Nelson *et al*, 1997) but not all (McCormick *et al*, 1991, Patel *et al*, 1992) studies. Similarly, there are discrepancies concerning the magnitude of gender differences in bone mass. The problem can be boiled down to determine how good the separation between ethnics or genders could be according to longitudinal measurements, such as bone mineral density curve.

Classifying highly correlated high-dimensional curves is a challenging topic because of the difficulty to estimate within-class covariance matrix. As pointed out in James and Hastie (2001), two common solutions exist to this problem. The first is regularization method, which uses some form of regularization, such as adding a diagonal matrix to the covariance matrix (Friedman, 1989, Hastie *et al*, 1995). The second is filtering method, which chooses a finite-dimensional basis and find the best projection of each curve onto this basis. The resulting basis coefficients can then be used as a finite dimensional representation. Then it is possible to use classification procedure such as linear discriminant analysis on the basis coefficients.

However, the case that only a fragment of each curve has been observed makes the classification even more difficult. The data illustrated in Figure 1 is such an example. These data is a subset of the data presented in Bachrach *et al* (1999) and was analyzed for classification purpose in James and Hastie (2001). The data consist

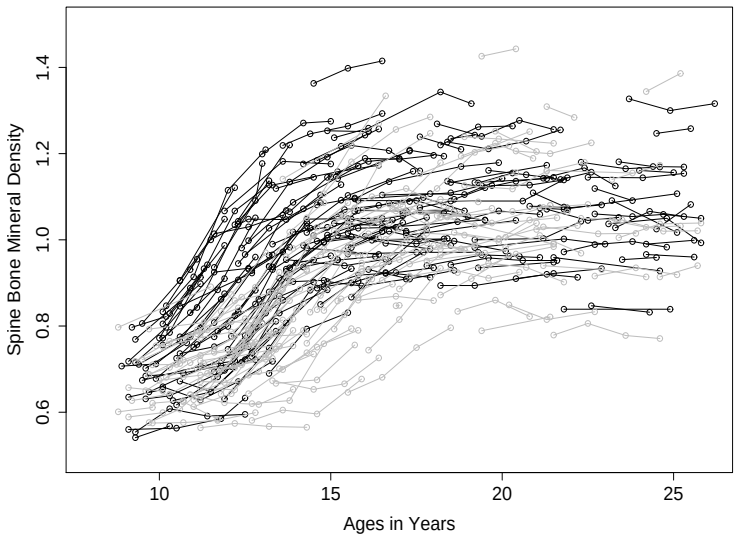

Figure 1: The spinal bone mineral density data. Black lines are 153 females and grey lines are 127 males.

of measurements of spinal bone mineral density for 280 people taken at various ages. For each person we only have two to four measurements, typically measured over no more than a couple of years. In this situation both of the common approaches to discriminant analysis can break down (James and Hastie, 2001). The regularization method fails because of the sparse characteristic of the data. The filtering method also gives its way to other methods due to several potential problems. Because the curves are measured at different time points so that the assumption of a common covariance matrix for each curves basis coefficients is not feasible. Another problem is that with extremely sparse data sets some of the basis coefficients may have infinite variance, making it impossible to estimate the entire curve.

James and Hastie (2001) proposed a functional linear discriminant analysis method (FLDA) to overcome the above difficulties. The FLDA method combines a regression fitting procedure and a linear discriminant analysis (LDA) using Bayes classifier. The regression procedure uses a spline curve $\mathbf{g}_{ij}$ plus random error to model observations

from each individual. The spline curve is further modeled using a basis function multiplied by a q-dimensional coefficient vector, $\boldsymbol{\eta}_{ij}$ so that the longitudinal measures for jth individual in ith classes, $\mathbf{Y}_{ij}$, can be expressed as

$$\mathbf{Y}_{ij} = \mathbf{S}_{ij}\boldsymbol{\eta}_{ij} + \boldsymbol{\varepsilon}_{ij}, i = 1, \ldots, K, \;\; j = 1, \ldots, m_i, \tag{2.1}$$

where $\mathbf{S}_{ij} = (\mathbf{s}(t_{ji1}), \ldots, \mathbf{s}(t_{ijn_{ij}}))^T$, and $\boldsymbol{\varepsilon}_{ij} \sim N(0, \sigma^2\mathbf{I})$. The spline coefficient vector is hierarchically parameterized by a Gaussian distribution with different mean vector $\boldsymbol{\mu}_i$ and common covariance matrix $\boldsymbol{\Gamma}$ for samples from all classes. Then, the rank constraints as in reduced-rank version of LDA (Anderson, 1951, Hastie and Tibshirani, 1996) are applied on those means. This gives the final form of the FLDA model

$$\mathbf{Y}_{ij} = \mathbf{S}_{ij}(\lambda_0 + \boldsymbol{\Lambda}\alpha_i + \boldsymbol{\gamma}_{ij}) + \boldsymbol{\varepsilon}_{ij} \tag{2.2}$$

where $\boldsymbol{\gamma}_{ij} \sim N(0, \boldsymbol{\Gamma})$. Finally, the classification using reduced-rank LDA is performed based on linear discriminant $\hat{\alpha}_Y$ and $\hat{\alpha}_i$, estimated from regression procedure.

In the literature, it seems that some kind of Bayesian methods for irregular curve classification have not been presented before. The novelty of the proposed Bayesian model lies in its ability to draw information from the curves as well as from the associated categorical responses for classification by unifying spine-based functional regression and logistic classification models. In this process, it enjoys the advantages of Bayesian modeling of functions with flexible spline basis as well as the simplicity of logistic classification models. In most of the existing models, a naive approach is used, where the estimates from the regression model are simply plugged into the classification model. Thus the regression model is unaware of additional information in the categorical outcomes and completely overlooks the classification problem.

2.2 Unified Bayesian Spline-based Classification Model

2.2.1 Regression Model for the Predictor

The data we observe for the ith subject or experimental unit are $\{\mathbf{Y}(\mathbf{t}_i), z_i\}$ where $\mathbf{Y}(\mathbf{t}_i)$ is the the predictor observed at time points $\mathbf{t}_i = (t_{i1}, ..., t_{im_i})$ as $\mathbf{Y}_i = \mathbf{Y}(\mathbf{t}_i) = (y_{t_{i1}}, \cdots, y_{t_{im_i}})$ and z_i is the binary response (class indicator) for $i = 1, \cdots, n$. For different subjects, the locations and number of time points are different. Although we only observe values at finite number of time points, the underlying unknown predictor curves, $f_1 \cdots, f_n$, are of interest. Assume they have been observed with white Gaussian noise as

$$\mathbf{Y}_i = f_i(\mathbf{t}_i) + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim MN(0, \boldsymbol{\Sigma}), i = 1, \ldots, n. \tag{2.3}$$

The measurement errors $\boldsymbol{\varepsilon}_i$ are assumed to be independent of the unknown predictor curves. The covariance structure for measurement errors $\boldsymbol{\varepsilon}_i$ is a key component to the estimation. A lot of effort has been made to better estimate covariance matrices (Daniels and Kass, 1999, 2001). Structured covariance matrix is attractive because of simplicity, but it may be inappropriate when the observations across each curve are from same individual and correlated. We adopt unstructured covariance matrix in this model. We also assume that the time points without observation are missing at random. Using a flexible basis to represent the functions is a common approach for modeling functional data (Ramsay and Silverman 1997). Natural cubic spline functions is employed in this paper because of their desirable mathematical properties and easy implementation (de Boor, 1978, Green and Silverman, 1994). Using a finite spline basis to represent the functions f_i, in a linear model notation we write

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \quad \boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{i}) \tag{2.4}$$

where $\mathbf{X}_i = \mathbf{X}(\mathbf{t}_i)$ is a spline basis of dimension q for ith individual, and $\boldsymbol{\beta}_i$ is the q-dimensional spline coefficients for function f_i after the transformation. In practice, the natural cubic spline basis can be generated based on B-spline basis matrix with certain degrees of freedom on a sequence of knots that should include at least all time points in the data set. Singular value decomposition is then applied to construct the orthogonal basis matrix. It is worth of pointing out that although the full matrix $\mathbf{X}$ is orthogonally formed, the basis matrix $\mathbf{X}_i$ for ith subject is not orthogonal.

The situation we are facing is that the basis set is not fixed across regressions. This type of regression is called "seemingly unrelated regressions (SUR)" (Zellner, 1962). Because the basis set $\mathbf{X}_j \neq \mathbf{X}_k$ for $j \neq k$, the regression are seemingly unrelated though they are actually related through the noise process $\boldsymbol{\epsilon}_i$. In conventional Bayesian linear regression models conjugate priors are usually adopted for the parameters because conjugacy aids the computational aspects of the modeling. In the Bayesian SUR model with different basis sets for each regression, there is no natural conjugate prior for $\boldsymbol{\beta}_i$ and $\boldsymbol{\Sigma}_i$. Hence, we adopt independent priors of the form $p(\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i) = p(\boldsymbol{\beta}_i)p(\boldsymbol{\Sigma}_i)$ and assign a higher level prior as,

$$\begin{aligned} \boldsymbol{\Sigma}_i &\sim IW(\mathbf{A}_i, b), \\ \boldsymbol{\beta}_i &\sim MN(0, \boldsymbol{\Omega}), \\ \boldsymbol{\Omega} &\sim IW(\mathbf{B}, d) \end{aligned} \tag{2.5}$$

where hyperparameters pairs $(\mathbf{A}_i, b)$ and $(\mathbf{B}, d)$ are scale matrices and degrees of freedom of inverse Wishart distribution. Here the covariance matrix $\boldsymbol{\Omega}$ serves as smoothing parameter that controls smoothness through the roughness penalty in the penalized sum of squares criterion, $\sum_{j=1}^{m_i}(Y_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_i)^2 + \boldsymbol{\beta}_i^t\boldsymbol{\Omega}^{-1}\boldsymbol{\beta}_i$ (see Berry *et al*, 2002). More precisely, the scale of diagonal elements of $\boldsymbol{\Omega}$ affect the smoothness with larger values resulting in smoother curves. There are at least two possible

methods for choosing the smoothing parameter for a smoothing spline. We assign prior distribution on $\mathbf{\Omega}$, Berry *et al* (2002) used a similar procedure. By placing a continuous density probability prior on $\mathbf{\Omega}$, we have automatically given zero prior probability to the possibility of doing no smoothing at all. Meanwhile, the way of adopting common covariance matrix $\mathbf{\Omega}$ for spline coefficients corresponding to each curve enables pooling of the information from each curve to achieve smoothness of the estimation. Therefore, it is possible to estimate the whole curve for each subject although only a fragment of the curve is observed.

2.2.2 Classification Model

Associated with each functional predictor $\mathbf{Y}_i$, there is a binary classification variable $z_i \in \{0, 1\}$ that takes unit value with unknown probability p_i. We have used the spline coefficients from equation (2.4) as classifiers. We develop a logistic classification model based on these coefficients $\boldsymbol{\beta}$ through a latent variable $T_i = logit(p_i)$ as

$$T_i = \boldsymbol{\beta}_i^t \boldsymbol{\theta} + \delta_i, \delta_i \sim N(0, \tau^2) \tag{2.6}$$

where $\boldsymbol{\theta}$ is $q \times 1$ vector of regression coefficients comprising a linear relation between the classification variables and the spline coefficients and δ_i is a random residual component. The use of a residual component is consistent with the belief that there may be unexplained sources of variation in the data perhaps due to nonlinear behavior of the classifiers.

Let $\mathbf{V} = diag(\mathbf{h})$, where $\mathbf{h}$ comprise the corresponding scaling parameters given by $h_j \sim IG(c_j, d_j)$, $j = 1, \ldots, q$, and (c_j, d_j) are hyperparameters. The effective joint prior for the coefficients and the model variance is

$$\boldsymbol{\theta}, \tau^2 | \mathbf{V} \sim NIG(0, \mathbf{V}, a_\tau, b_\tau), \tag{2.7}$$

where NIG denotes the normal-inverse gamma prior – the product of the conditionals

$\boldsymbol{\theta}|\tau^2, \mathbf{V} \sim MN(0, \tau^2\mathbf{V})$ and $\tau^2 \sim IG(a_\tau, b_\tau)$ with a_τ, b_τ as the usual hyperparameters for the inverse gamma (IG) prior.

To summarize the unified hierarchical Bayesian model, we have

$$\begin{aligned}
\textit{Random function } \mathbf{Y}_i &\sim MN(\mathbf{X}_i\boldsymbol{\beta}_i, \boldsymbol{\Sigma}_i) \\
\boldsymbol{\Sigma}_i &\sim IW(\mathbf{A}_i, b) \\
\boldsymbol{\beta}_i &\sim MN(0, \boldsymbol{\Omega}) \\
\boldsymbol{\Omega} &\sim IW(\mathbf{B}, d)
\end{aligned} \tag{2.8}$$

$$\begin{aligned}
\textit{Binary outcome } z_i &\sim Bernoulli(p_i) \\
T_i &\sim N(\boldsymbol{\beta}_i^t\boldsymbol{\theta}, \tau^2), \text{where } T_i = logit(p_i) \\
\boldsymbol{\theta}, \tau^2 &\sim NIG(0, \mathbf{V}, a_\tau, b_\tau), \text{where } \mathbf{V} = diag(\mathbf{h}) \\
h_j &\sim IG(c_j, d_j)
\end{aligned}$$

for $i = 1, ..., n$ and $j = 1, \ldots, q$.

2.3 Posterior Inference

As the joint posterior distribution of the parameters is not of explicit form, we have to depend on MCMC methods to simulate the parameters from this posterior distribution. In a Gibbs sampling framework (Gelfand and Smith, 1990), we need to derive the full conditional distributions. These conditional distributions are given below separately for the regression and the classification model. Because MCMC really is a standard tool in the literature, we leave the detail derivations out in this dissertation to avoid redundancy. Later chapters also exclude detail derivations for MCMC. For notation convenience, we let $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^n$, $\mathbf{T} = \{T_i\}_{i=1}^n$, $\mathbf{z} = \{z_i\}_{i=1}^n$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_i\}_{i=1}^n$.

2.3.1 Regression Model

The model variance matrix $\boldsymbol{\Sigma}_i$ is updated only using the regression likelihood as

$$\boldsymbol{\Sigma}_i|\boldsymbol{\beta}_i, \mathbf{Y}_i, \mathbf{X}_i \sim IW(\mathbf{A}_i^*, b^*), \tag{2.9}$$

where $\mathbf{A}_i^* = \mathbf{A}_i + (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^t$ and $b^* = b + 1$. The conditional distribution for the coefficients $\boldsymbol{\beta}_i$ follows from the model specifications and combination of information from both the regression and the classification segments,

$$\boldsymbol{\beta}_i \mid \mathbf{Y}_i, \boldsymbol{\Sigma}_i, \mathbf{X}_i, \boldsymbol{\Omega}, T_i, \boldsymbol{\theta}, \tau^2 \sim MN(\boldsymbol{\beta}_i^*, \tau^2\boldsymbol{\Omega}^*), \tag{2.10}$$

where $\boldsymbol{\Omega}^* = (\tau^2(\boldsymbol{\Omega}^{-1} + \mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{X}_i) + \boldsymbol{\theta}\boldsymbol{\theta}^t)^{-1}$ and $\boldsymbol{\beta}_i^* = \boldsymbol{\Omega}^*(\tau^2\mathbf{X}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{Y}_i + T_i\boldsymbol{\theta})$. It is worth of noting that the penalized least squares estimator, minimizing the penalized sum of square, is the mean of the posterior distribution of $\boldsymbol{\beta}_i$ when information from classification segment is excluded. In the next level, the covariance matrix $\boldsymbol{\Omega}$ is updated as

$$\boldsymbol{\Omega}|\boldsymbol{\beta} \sim IW(\mathbf{B}^*, d^*), \tag{2.11}$$

where $\mathbf{B}^* = \mathbf{B} + \sum_{i=1}^n \boldsymbol{\beta}_i\boldsymbol{\beta}_i^t$ and $d^* = d + n$.

2.3.2 Logistic Classification Model

The conditional distributions for the logistic classification model follow in a similar way, except now the detail coefficients $\boldsymbol{\beta}_i$ serve as the predictors of the latent variables T_i. The corresponding coefficients $\boldsymbol{\theta}$ are updated as

$$\boldsymbol{\theta}|\boldsymbol{\beta}, \tau^2, \mathbf{V}, \mathbf{T} \sim MN(\boldsymbol{\theta}^*, \tau^2\mathbf{V}^*) \tag{2.12}$$

where $\mathbf{V}^* = (\boldsymbol{\beta}\boldsymbol{\beta}^t + \mathbf{V}^{-1})^{-1}$ and $\boldsymbol{\theta}^* = \mathbf{V}^*\boldsymbol{\beta}\mathbf{T}$. The conjugate IG prior for τ^2 leads to its marginal conditional distribution as

$$\tau^2|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}, \mathbf{V} \sim IG(a_\tau^*, b_\tau^*) \tag{2.13}$$

where $a_\tau^* = a_\tau + n/2$ and $b_\tau^* = b_\tau + \left[\mathbf{T}^t\mathbf{T} - \mathbf{T}^t\boldsymbol{\beta}(\mathbf{V}^{-1} + \boldsymbol{\beta}^t\boldsymbol{\beta})^{-1}\boldsymbol{\beta}^t\mathbf{T}\right]/2$.

The scale parameters in this model h_j are again updated by

$$h_j|\boldsymbol{\theta}, \tau^2, \boldsymbol{\gamma} \sim IG(c_j^*, d_j^*) \tag{2.14}$$

where $c_j^* = c_j + 1/2$ and $d_j^* = d_j + \theta_{jk}^2/2\tau^2$. Finally, the latent variable vector $\mathbf{T}$ is updated from a non-standard posterior distribution by a Metropolis step,

$$f(\mathbf{T}|\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \mathbf{z}) \propto \exp\left\{-\frac{1}{2\tau^2}||\mathbf{T} - \boldsymbol{\beta}\boldsymbol{\theta}||^2\right\} \times \prod_{i=1}^{n} \frac{e^{T_i z_i}}{1 + e^{T_i}}. \tag{2.15}$$

2.4 Model Choice and Prediction

It might be the simplest way to classify using linear discriminant analysis based on the projection onto an adjusted spline basis, assuming independent identical distributed noises for each curve. For a comparative study, we also apply naive spline based classification model, which is the naive version of Bayesian spline-based method (BNSCC). Unlike the unified model, it separates the regression and classification models. It uses the regression model only to obtain the estimate of the spline coefficients and thereafter plug them in the classification model treating them as a set of classifiers. We also explore another naive method, which using the regression model as in naive Bayesian spline-based method to estimate the q-dimensional coefficients, and plugging the coefficients into support vector machine for classification.

Other model choices can be based on the investigation of different spline bases, which might involve aspects such as knot selection and determining the dimension of the spline basis. Except splines, other possible basis function can also be considered. These areas remain as ongoing research. A flexible natural cubic spline functions, evaluated at a fine lattice of points, could be a good choice because of their desirable mathematical properties and easy implementation (de Boor, 1978, Green and Silverman, 1994). We use natural cubic spline basis here.

To select from the different models, we will generally use classification results. For a new sample with predictor values $\mathbf{Y}_{new}$, the posterior predictive probability that its group type, denoted by z_{new} given the old data D is

$$p(z_{new}|\mathbf{Y}_{new}, D) = \int p(z_{new} = 1|\mathbf{Y}_{new}, \boldsymbol{\beta}_{new}, \boldsymbol{\phi})p(\boldsymbol{\beta}_{new}|\mathbf{Y}_{new}, \boldsymbol{\phi})p(\boldsymbol{\phi}|D)d\boldsymbol{\phi}, \quad (2.16)$$

where $\boldsymbol{\phi}$ is the vector of all the model parameters. Assuming conditional independence of the responses the integral can be approximated by the Monte Carlo estimate

$$\sum_{j=1}^{M} p(z_{new} = 1|\mathbf{Y}_{new}, \phi^{(j)})/M, \quad (2.17)$$

where $\phi^{(j)}$ $(j = 1 \ldots, M)$ are the MCMC posterior samples of the parameter $\boldsymbol{\phi}$.

When a test set is provided, we first obtain the posterior distributions of the parameters (training the model) based on the training data and use them to classify the test samples. For a new observation from the test set, say $\mathbf{z}_{i,tst}$, we will obtain the probability $p(\mathbf{z}_{i,tst} = 1|\mathbf{z}_{trn}, \mathbf{Y}_{trn}, \mathbf{Y}_{tst})$ by using an equation similar to (2.16), and approximate it by its Monte Carlo estimate as in equation (2.17). When this estimated probability exceeds .5, the new observation is classified as 1, otherwise, it is classified as 0.

For comparison purpose, we report the training error rate as classification result, as in James and Hastie (2001). The training error rate is given by using the data without classification as testing data after training the model with the data with classification.

2.5 Extension to Multicategory Classification

The Bayesian method can be easily extend to classification problems where the response is a categorical variable with more than two categories. Assume that response vector $\mathbf{z} = (z_1, ..., z_J)$, indicates the observed response data, with z_i taking one of j

possible categories, and let $p_{ij} = P(z_i = j)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, J$, be the probability that the ith observation falls into the jth category. These probabilities are related to the predictor curve f_i through a link function. Similar to the previous section, we span the function f with spline basis functions and use the regressed coefficients $\boldsymbol{\beta}$ as the classifiers or covariates in the link model. In the multinomial logit link function (McFadden, 1973) model we again introduce a latent variable T_{ij} and model the probabilities as

$$p_{i1} = \frac{1}{1 + \sum_{s=2}^{J} exp(T_{is})} \text{ and } p_{ij} = \frac{exp(T_{ij})}{1 + \sum_{s=2}^{J} exp(T_{is})}. \tag{2.18}$$

The generalized linear model based on spline curves can be expressed as

$$T_{ij} = \boldsymbol{\beta}_i^t \boldsymbol{\theta}_j + \delta_{ij}, \quad \delta_{ij} \sim N(0, \tau^2) \tag{2.19}$$

where T_{ij} is the latent variable corresponding to ith sample and jth category, $\boldsymbol{\beta}_i$ is the i'th wavelet coefficients curve and of size m by 1 and $\boldsymbol{\theta}_j$ is m by 1 regression coefficients vector. The MCMC training scheme is similar to the binary case and so are conditional distributions for posterior inference. We adopt the usual classification rule for multinomial logit model, which is to assign the new curve to group j if the estimated $T_j^* = argmax(\mathbf{T}^*)$.

2.6 Examples of Application

In this section, we apply the novel Bayesian spline-based classification method denoted by BSCC to analyze a simulation data and a real world data, spinal bone mineral density data. Except naive Bayesian methods results, the classification results by James and Hastie (2001) using functional linear discriminant analysis are included for comparison purpose.

All along in the Bayesian models, we wish to put proper but weak prior information, in the sense of bringing a lot of information to the problem. For inverse-Gamma

prior, we use the shape hyperparameter to be larger than 1, allowing the expectation of the IG distribution exists. For inverse-Wishart prior, we choose to use the degrees of freedom to be the smallest integer such that the expectation of the distribution exists. The scale matrix is specified as identity matrix. With the small degrees of freedom, the scale matrix is unlikely to be critical. Therefore, hyperparameters (a_τ, b_τ) are specified as (2,2), (c_j, d_j) are specified as (2,2), both $(\mathbf{A}_i, b)$ and $(\mathbf{B}, d)$ are specified as identity matrix and $1+rows$, where $rows$ is the number of rows of the corresponding scale matrix. We found the results insensitive to moderate modifications of these priors. Also we run the MCMC chain for 80,000 iterations and have thrown out first 20,000 burn in iterations. The results reported are average of 20 repeats. The different dimensions of the spline basis have very little effect on classification in our study. So we choose to use q equal to 6. Through out the analysis of the bone mineral density data, we use knots starting from smallest age of 8.8 (in years) of the involved subset and ending at largest age 26.2 (in years) with increments of 0.1.

2.6.1 Simulation Study

To illustrate effectiveness of proposed methods, first we apply them on a simulated data set. Similar to the motivating example, we generate 40 fragmental curves from each of two classes with different mean functions and evenly split them to form training and testing sets. The mean functions are $sin(1.8\pi x + 6.0) + cos(1.8\pi x)$ and $sin(2\pi x) + cos(2\pi x)$, as shown in Figure 2 (thick lines). The curves are corrupted by independent white Gaussian noise (signal to noise ratio is 4). As shown in Figure 3, the dimension of the curves are between 3 to 10, and those points on each curve are randomly selected with equal probability. The knots are equally spaced and divide the interval $[0, 1]$ into 100 pieces so that all predictor values in the data set are covered. We performed 20 repetitions of simulation and report the average CCR over these

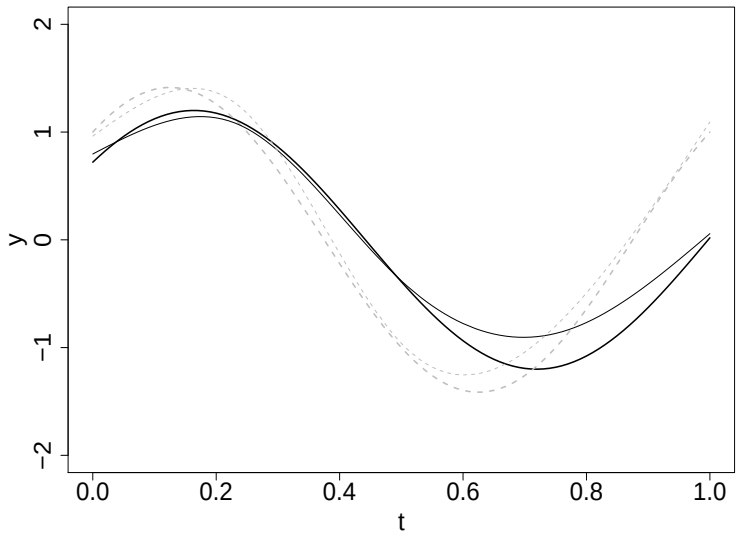

Figure 2: The mean functions (thick lines) for two classes in simulated data set, overlapped by estimated functions (thin lines) by unified Bayesian method. Color black and grey are used to represent two classes respectively.

20 replications. The highest correct classification rate is 92%, yielded by the unified Bayesian spline-based classification method. The CCRs are 84%, 82% and 85% for functional linear discriminant analysis method, the naive method with support vector machine and the naive Bayesian spline-based method. Classification results show that the unified model benefits from the combination of spline-based functional regression and logistics classification models. In the unified Bayesian method, the regression estimate of the functions in each class is overlapped in Figure 2.

2.6.2 Binary Classification Based on Gender or Ethnicity

The data is a subset of the data presented in Bachrach *et al* (1999) and was analyzed for classification purpose in James and Hastie (2001). These data consist of measurements of spinal bone mineral density for 280 people taken at various ages. For each person we only have two to four measurements, typically measured over no more than a couple of years. Although classification is not the primary goal of the spinal bone

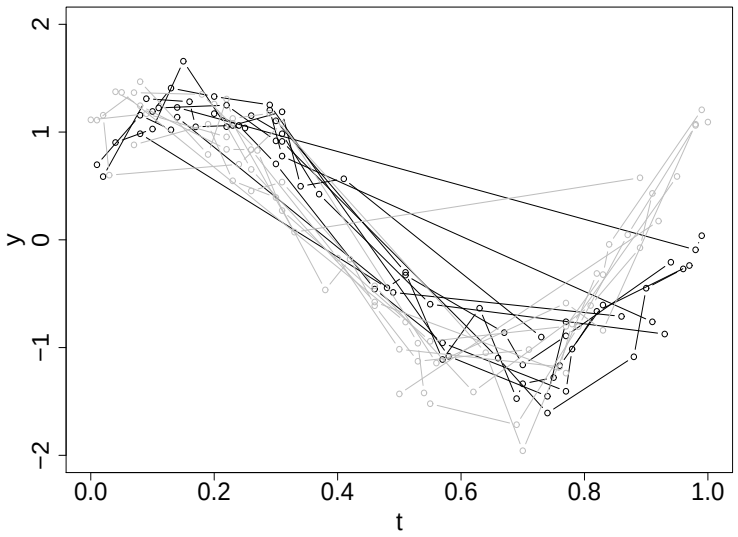

Figure 3: Examples of ten curves from each class are overlapped.

density data, we apply our methods to illustrate the irregular curve classification procedures. At the mean time, it might provide some kind of guidelines to address those controversy about bone mineral acquisition mentioned in Section 2.1. We first use gender as the categorical outcome variable. Out of those 280 people, 156 of them are female and 124 are male. From the data shown in Figure 1, we see that there is a weak overall separation of gender groups. It seems that female tends to have higher spinal bone mineral density than male when age is under 18 years (Figure 4). This pattern is not supported by densities measured after 18 years (Figure 5). Therefore, we consider to do the classification for three cases: overall ages, ages under 18 years and over 18 years. The following table give the results of different classification methods.

Estimations of spinal bone mineral densities for female, male and both groups, by the unified Bayesian method, are plotted in Figure 6. There is gender difference in spine bone mineral density when age is about below 18. During periods of ages

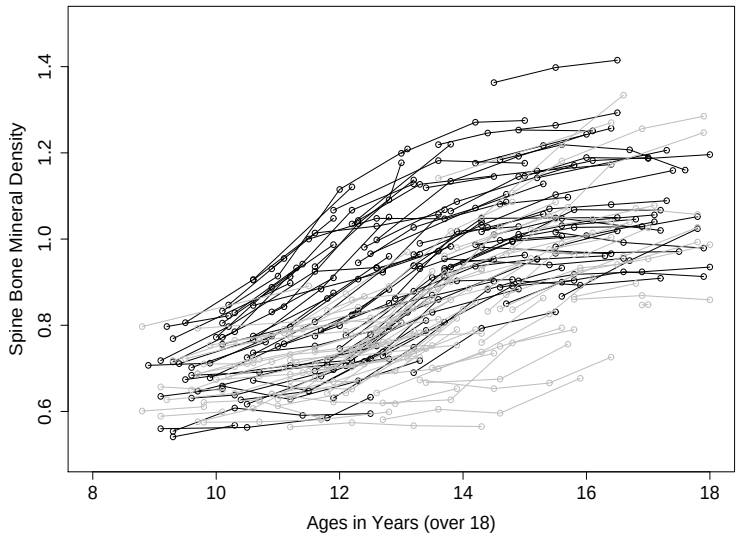

Figure 4: Younger (below 18) age group of spinal bone mineral density data. Black and grey lines represent females and males respectively.

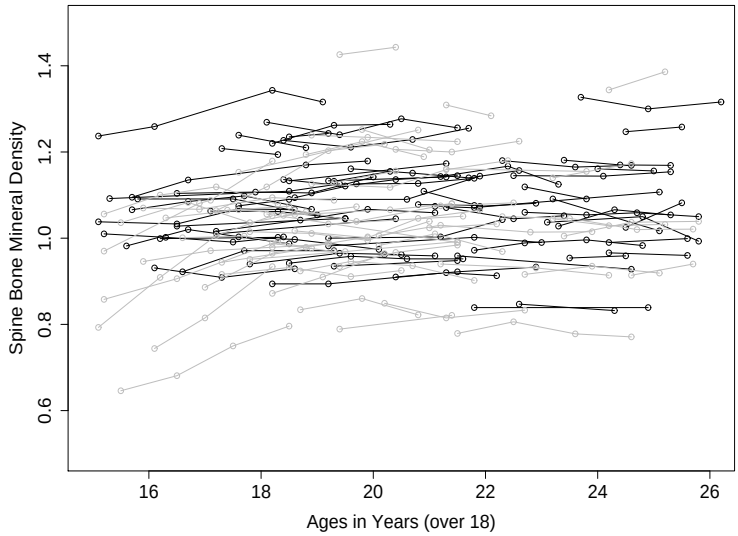

Figure 5: Elder (over 18) age group of spinal bone mineral density data. Black and grey lines represent females and males.

Table 1: The CCRs comparison of our method and other methods for analyzing spinal bone mineral data using gender as categorical response variable.

Methods	BSCC	BNSCC	BSRSVM	FLDA
CCR% for overall ages	75	65	64	71
CCR% for ages under 18	83	79	80	78
CCR% for ages above 18	57	54	55	56

Note: BNSCC and BSCC are the naive version and unified version of Bayesian spline-based classification methods. BSRSVM is the naive method simply stacking the Bayesian spline-based regression model and support vector machine. FLDA stands for the functional linear discriminant analysis in James and Hastie(2001).

younger than about 18, female has higher densities and reaches the peak density earlier than male. Once the spinal bone mineral density achieve at the peak level, it maintains at that level until to late twenties. There is no difference between two gender groups in the spinal bone mineral density after about age 18. For 281 individual curves, the maximum mean square errors are 0.006, 0.003 and 0.004 for three methods: BSCC, BNSCC and BSR. The errors of regression are small according to mean square error. Therefore the spline-based regression modeling is proper for these data. The classification results in Table 1 show that the unified version of Bayesian method yields best classification rate for all three cases. There is no obvious advantages among three other methods. The separation between two gender groups are more clear for ages below 18. For those above 18 years of age people, the correct classification rates are around fifty percentage, which agrees with the mix-up pattern of in Figure 5. Although the estimation of common covariance matrix, $\mathbf{\Omega}$, of the spline regression coefficients, $\boldsymbol{\beta}_i$'s, is not of direct interest, it does reflect the smoothness of the spline curve estimation. Also, the regression coefficient vector, $\boldsymbol{\theta}$, in logistic classification model indicates the effect of curve predictor on the categorical response. The posterior means and 90% credible intervals from both unified and naiver versions of Bayesian spline-based methods are given for parameters $\mathbf{\Omega}$ and $\boldsymbol{\theta}$ in Table 2. The

Table 2: The posterior means and 90% credible intervals for $\boldsymbol{\Omega}$ and $\boldsymbol{\theta}$.

Parameter	from BSCC			from BNSCC		
Ω_{11}	2.09	(1.57,	2.73)	2.08	(1.57,	2.76)
Ω_{21}	-0.40	(-0.77,	-0.06)	-0.40	(-0.79,	-0.06)
Ω_{22}	1.70	(1.32,	2.17)	1.68	(1.30,	2.15)
Ω_{31}	-0.31	(-0.61,	-0.02)	-0.31	(-0.63,	-0.01)
Ω_{32}	0.16	(-0.13,	0.46)	0.16	(-0.11,	0.47)
Ω_{33}	1.69	(1.33,	2.13)	1.70	(1.32,	2.18)
Ω_{41}	-0.09	(-0.35,	0.19)	-0.09	(-0.37,	0.17)
Ω_{42}	0.22	(-0.03,	0.49)	0.21	(-0.04,	0.48)
Ω_{43}	0.23	(-0.01,	0.48)	0.24	(-0.02,	0.51)
Ω_{44}	1.50	(1.19,	1.87)	1.50	(1.20,	1.86)
Ω_{51}	0.10	(-0.16,	0.34)	0.08	(-0.19,	0.35)
Ω_{52}	0.27	(0.03,	0.53)	0.26	(0.00,	0.54)
Ω_{53}	-0.11	(-0.37,	0.13)	-0.12	(-0.35,	0.11)
Ω_{54}	0.14	(-0.09,	0.38)	0.14	(-0.09,	0.37)
Ω_{55}	1.50	(1.21,	1.83)	1.49	(1.19,	1.86)
θ_1	0.21	(0.12,	0.28)	0.30	(0.15,	0.44)
θ_2	-0.02	(-0.11,	0.1)	-0.06	(-0.12,	0.13)
θ_3	-0.11	(-0.18,	-0.03)	-0.00	(-0.12,	0.1)
θ_4	0.03	(-0.06,	0.13)	0.00	(-0.09,	0.11)
θ_5	0.26	(0.16,	0.38)	0.12	(0.03,	0.22)

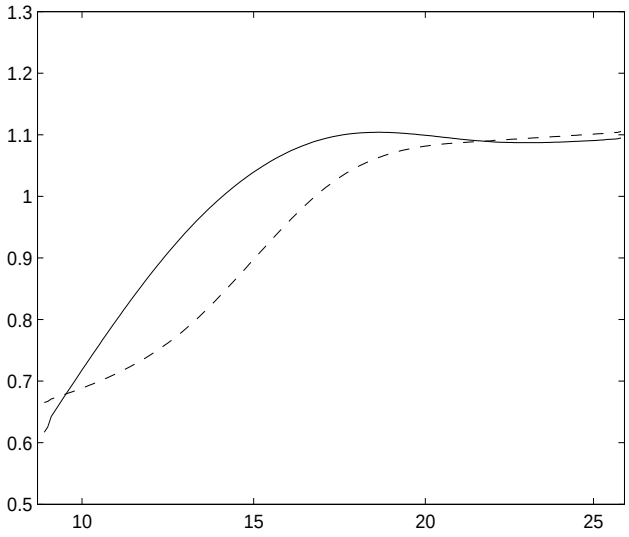

Figure 6: Estimations of spinal bone mineral densities for female (solid) and male (dash) groups.

relative large diagonal elements of $\mathbf{\Omega}$ indicates smoother estimated curves for both unified and naiver Bayesian methods. Although the estimated common covariance matrix $\mathbf{\Omega}$ from both unified and naive Bayesian methods are very similar to each other, the naive method produces a little wider 90% credible intervals that suggest more uncertainty of the spline coefficients. This implies that unified method does provide more precise estimation based on linkage between curve predictor and categorical response. According to the regression coefficient vector, the unified method, yielding three significant coefficients, incorporates more information from curve predictor for classification. On the other hand, the naive method only yields two significant coefficients.

To study ethnic effect on bone mineral densities, we let ethnic to be categorical response variable. For binary classification, we only consider two ethnics, Black and Asian. A subset of those data, all 78 female Asian and Blacks, are included in Figure 7. Blacks tend to have higher spinal bone mineral densities than Asians. The

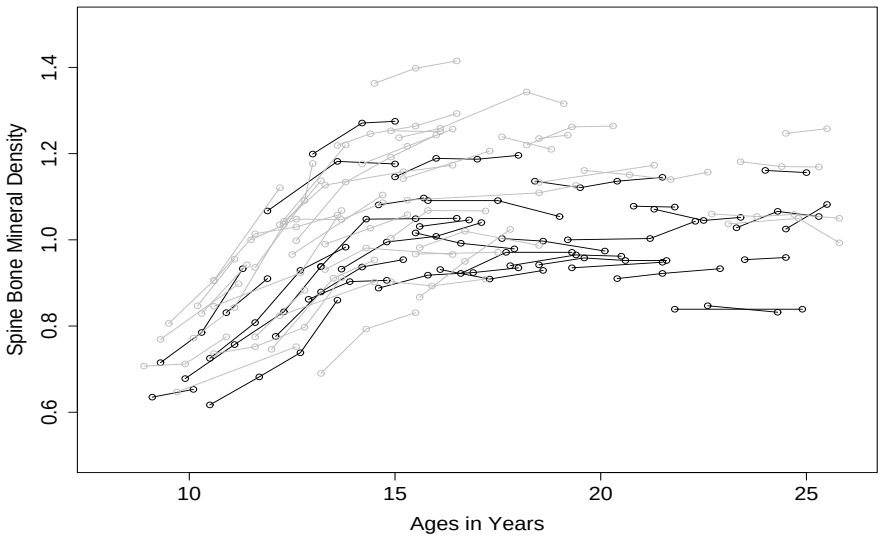

Figure 7: The spinal bone mineral density data grouped by ethnics. Thin black lines are 35 female Asians and thin grey lines are 43 female Blacks. Thick grey and black lines represent estimated spinal bone mineral densities for female Blacks and Asians.

Table 3: The CCRs comparison of our method and other methods for classifying female Blacks and Asians spinal bone mineral data.

Methods	BSCC	BNSCC	BSRSVM	FLDA
CCR%	82	78	78	75

Note: BSCC, BNSCC, BSRSVM and FLDA are same as in Table 1.

classification results are reported in Table 3. Our unified Bayesian classification method leads in the correct classification rates about 4 percentage more than the functional linear discriminant analysis by James and Hastie (2001), which in the second place. The two naive methods have tied results at about 78% CCR. Although all methods use natural cubic spline to smooth the curves and all regression errors are small, the model set-ups engaging differently with the spline basis make classification differ. The estimated spinal bone mineral densities of female Asians and Blacks, by the unified Bayesian method, are overlapped in Figure 7. There is very clear trend that Blacks have higher spine bone mineral density than Asians, no matter of age

period.

2.6.3 Multicategory Classification Based on Ethnicities

In this section, we illustrate that our method can be easily extend to multicategory classification case by applying the classification methods to classify four ethnics groups based on spinal bone mineral density curves. Out of 153 female individuals, there are 35 Asians, 43 Blacks, 27 Hispanics and 48 Whites. Table 4 gives classification results comparisons. Because the spinal bone mineral densities of four ethnicities are really mixed together, all correct classification rates are around forty to fifty percent. The unified Bayesian spline-based classification method performs best with 55% CCR. On the second place is the naive method using support vector machine, which has slightly higher overall CCR than FLDA method by James and Hastie (2001). Among four ethnicities, Hispanics and Whites are associated with low correct classification rates, while Asians and Blacks are relatively well classified by all methods. Estimations of spinal bone mineral densities for these four ethnicities, by the unified Bayesian method, are plotted in Figure 8. The estimated spinal bone mineral densities for Blacks are higher than other three ethnicities over the age span from nine to twenty five. Asians has lower spinal bone mineral densities than other three ethnicities after sixteen years old. The estimated spinal bone mineral densities for Asians and Blacks are separated from the rest of the female group. Therefore they are expected to be classified relatively better than other three ethnicities. The results in Table 4 support this finding.

Table 4: The CCRs comparison for multicategory classification: separating Asian, Black, Hispanic and White female individuals based on their spinal bone mineral density data.

Methods	BSCC	BNSCC	BSRSVM	FLDA
overall CCR%	55	40	45	43
CCR% for Asians	74	63	69	63
CCR% for Blacks	81	67	70	70
CCR% for Hispanics	33	11	22	19
CCR% for Whites	21	15	19	19

Note: BSCC, BNSCC, BSRSVM and FLDA are same as in Table 1.

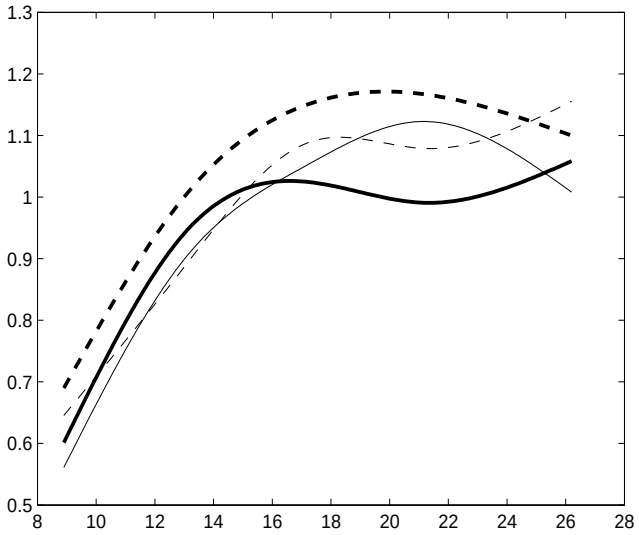

Figure 8: Estimated spinal bone mineral densities for female group: Asians(thick solid), Blacks(thick dash), Hispanics(thin dash) and Whites(thin solid).

CHAPTER III

HIGH DIMENSION SPIKY CURVE CLASSIFICATION USING WAVELETS

3.1 Motivation

In the previous chapter, we studied classifying irregular curves when the sample size is relatively larger than the dimension the curves. However, some curve classification problems involve high dimensional spiky curves, which pose a lot of difficulty on the task.

There are several existing approaches to curve classification, including the straightforward method of using summary quantiles, such as the mode, to perform classification (Pfeiffer *et al*, 2002). Parker (2002) performed classification by combining several simple algorithms such as moments, projections, convexity, slope histogram and angle-distance signature. Müller and Stadtmüller (2004) proposed a generalized functional linear regression model by approximating the predictor processes with a truncated Karhunen-Loéve expansion. James and Hastie (2001) developed a functional linear discriminant analysis method using splines to model the irregular curve functions. This spline model was later extended by James (2002) to predict survival status in the primary biliary cirrhosis data set by employing a functional logistic regression method. The performance of a spline model has to heavily rely on proper knot selection. Although there are approaches for adaptive knot selection, their use in an already involved model can be computationally infeasible.

In a functional context, splines lack the ability to fit sharp localized changes in curves and there exists better alternatives such as wavelets. The Bayesian wavelet modeling used in this paper manages to overcome these limitations as wavelets have

nice approximation properties over a large class of functional spaces (Daubechies, 1992) that can accommodate almost all the functional forms observed in real life applications. Indeed, this richness of the wavelet representation provides the backbone for the popular frequentist wavelet shrinkage estimators of Donoho and Johnstone (1994, 1995), which are the precursors of the more recent Bayesian wavelet estimation models (Abramovich *et al* 1998, Clyde *et al* 1998, Clyde and George 2000, Vidakovic 1998). Wavelets representations are also sparse and can be helpful in limiting the number of regressors.

The novelty of the proposed Bayesian model lies in its ability to draw information from the functional data as well as from the associated categorical outcome for classification by unifying wavelet-based functional regression and logistic classification models. In this process, it enjoys the advantages of Bayesian modeling of functions in wavelet domain as well as the simplicity of logistic classification models. In most of the existing models, a naive approach is used, where the estimates from the regression model are simply plugged into the classification model. Thus the regression model is unaware of additional information in the categorical outcomes and completely overlooks the classification problem. A simple example of a naive model would consist of a wavelet-based selection model - empirical Bayes thresholding method stacked over a classification scheme based on support vector machine or logistic regression.

3.2 Unified Bayesian Wavelet-based Classification Model

3.2.1 Regression Model for the Predictor Curves

Let the observation for the ith subject or experimental unit be $\{\mathbf{Y}_i, z_i\}$, where $\mathbf{Y}_i = (y_{i1}, \cdots, y_{im})$ is a vector of m sequential measurements and z_i is the corresponding binary classification variable. We write the observational equation with the

underlying function f_i as,

$$y_{i,k} = f_i(k/m) + \varepsilon_{i,k},\ \varepsilon_{i,k} \sim N(0, \sigma^2),\ k = 1, \ldots, m,\ i = 1, \ldots, n. \tag{3.1}$$

In nonparametric estimation, the functions are analyzed in the sequence space of coefficients in an orthonormal wavelet basis for $L_2([0,1])$. Wavelet representations are sparse for a wide variety of function spaces and their multi-resolution nature allow us to combine results from different resolutions and make conclusions for the estimation problem. In particular, the sparseness implies that when the wavelet basis is orthogonal and compactly supported (Daubechies, 1992), the i.i.d. normal noise affects all the wavelet coefficients equally, while the signal information remains isolated in a few coefficients. In shrinkage estimation, these small coefficients which are mostly noise are discarded to retrieve an effective reconstruction of the function. In terms of scaling and wavelet functions (φ, ψ), a wavelet expansion for f_i has the dyadic form

$$f_i(t) \approx \beta_{i00}\varphi_{00}(t) + \sum\nolimits_{j=1}^{J} \sum\nolimits_{k=0}^{2^{j-1}} \beta_{ijk}\psi_{jk}(t) \tag{3.2}$$

with β_{i00} as the scaling coefficient and the detail coefficients are β_{ijk}.

Using a finite orthonormal basis to represent the functions f_i, in a linear model notation we write

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i,\ \boldsymbol{\varepsilon}_i \sim MN(0, \sigma^2\mathbf{I}) \tag{3.3}$$

where $\mathbf{Y}_i = (y_{i,1}, \ldots, y_{i,m})$ is the vector of m observations from the i^{th} unit and $\boldsymbol{\beta}_i = (\beta_{i,1}, \ldots, \beta_{i,m})$ are the wavelet coefficients for f_i after the discrete wavelet transformation $\mathbf{X}$. For notational convenience, we let $\mathbf{Y} = \{\mathbf{Y}_i\}_{i=1}^n$ and $\boldsymbol{\beta} = \{\boldsymbol{\beta}_i\}_{i=1}^n$.

For this regression model, we assume variable selection priors for the wavelet coefficients that are used for nonlinear Bayesian wavelet modeling (DeCanditis and Vidakovic 2004, Vidakovic 1998). These priors are readily incorporated as a scale-mixture with latent indicator variables η_{jk} that equal 1 with probability π_j (Clyde

et al 1998, Clyde and George 2000, George and McCullogh 1993, DeCanditis and Vidakovic 2004) and comprise an effective strategy to adapt to the sparseness of the wavelet representation. Denote the diagonal matrix by $diag(\boldsymbol{\eta}_i)diag(\mathbf{g})$, where $\boldsymbol{\eta}_i = (\eta_{i00}, \eta_{i10}, \eta_{i20}, \eta_{i21}, \ldots)$ is a vector of latent indicator variables for selection of each coefficient and $\mathbf{g} = (g_0, g_1, g_2, g_2, \ldots)$ comprise the corresponding scaling parameters. Then the joint prior for the coefficients and the model variance is

$$\boldsymbol{\beta}_i, \sigma^2 | \boldsymbol{\eta}_i, \mathbf{g} \sim NIG(0, diag(\boldsymbol{\eta}_i)diag(\mathbf{g}), a_\sigma, b_\sigma) \tag{3.4}$$

where NIG denotes the normal-inverse gamma prior – the product of the conditionals $\boldsymbol{\beta}_i | \sigma^2, \boldsymbol{\eta}_i, \mathbf{g} \sim MN(0, \sigma^2 diag(\boldsymbol{\eta}_i)diag(\mathbf{g}))$ and $\sigma^2 \sim IG(a_\sigma, b_\sigma)$ with a_σ, b_σ as the usual hyperparameters for the inverse gamma (IG) prior. In the next layer, the prior distributions for each η_{ijk} and g_j are given by

$$\eta_{ijk} \sim Bernoulli(\rho_j) \text{ and } g_j \sim IG(u_j, v_j), \tag{3.5}$$

where ρ_j and (u_j, v_j) are hyperparameters specified levelwise, and $j \in \{0, 1, \ldots, \log_2 m\}$, $k \in \{0, \ldots, 2^j - 1\}$.

Setting the latent variables ρ_j to 1 leads to simple normal priors resulting in a pointwise Bayesian shrinkage of the wavelet coefficients. Alternatively, the wavelet coefficients can be specified by Laplace priors (Vidakovic, 1998). This is equivalent to use a penalized regression with a L_1 penalty term such as LASSO. It is convenient to express the Laplace prior as a scale mixture of normal where the scaling parameter is mixed by a exponential distribution as follows

$$\boldsymbol{\beta}_i | \sigma^2, \boldsymbol{\eta}_i, \mathbf{g} \sim MN(0, \sigma^2 diag(\boldsymbol{\eta}_i)diag(\mathbf{g})), \text{where} \tag{3.6}$$

$$\sigma^2 \sim IG(a_\sigma, b_\sigma) \text{ and } \; g_j \sim exp(\lambda_j/2). \tag{3.7}$$

Marginalizing the latent scale parameters g_j from the model lead to $\beta_{ijk} \sim Laplace(0, \sigma^2/\sqrt{\lambda_j})$, where λ_j is hyperparameters specified levelwise and $j \in \{0, 1, \ldots, \log_2 m\}$.

3.2.2 Classification Model

Associated with each functional predictor $\mathbf{Y}_i$, there is a binary classification variable $z_i \in \{0, 1\}$ that takes unit value with unknown probability p_i. The wavelet coefficients from equation (3.3) are used for classification. We develop a logistic classification model based on these wavelet coefficients $\boldsymbol{\beta}$ through a latent variable $T_i = logit(p_i)$ as

$$T_i = \boldsymbol{\beta}_i^t \boldsymbol{\theta} + \delta_i, \delta_i \sim N(0, \tau^2) \tag{3.8}$$

where $\boldsymbol{\theta}$ is $m \times 1$ vector of regression coefficients and δ_i is a random residual component. This produces a linear relationship between the classification variables and the regressed wavelet coefficients.

We assume a variable selection prior distribution for $\boldsymbol{\theta}$ similar to the priors (3.4) used in the regression model. This is a simple and effective way to reduce the dimensionality of the problem. We again write down the prior covariance matrix as $\mathbf{V} = diag(\boldsymbol{\gamma})diag(\mathbf{h})$, where $\boldsymbol{\gamma} = (\gamma_{00}, \gamma_{10}, \gamma_{20}, \gamma_{21}, \ldots)$ and $\mathbf{h} = (h_0, h_1, h_2, h_2, \ldots)$.

To summarize the unified hierarchical Bayesian model, we have

$$
\begin{aligned}
Random\ \ function\ \ \mathbf{Y}_i &\sim MN(\mathbf{X}\boldsymbol{\beta}_i, \sigma^2\mathbf{I}) \\
\boldsymbol{\beta}_i, \sigma^2 \mid \boldsymbol{\eta}_i, \mathbf{g} &\sim NIG(0, diag(\boldsymbol{\eta_i})diag(\mathbf{g}), a_\sigma, b_\sigma) \\
g_j &\sim IG(u_j, v_j) \\
\eta_{ijk} &\sim Bernoulli(\rho_j)
\end{aligned}
\tag{3.9}
$$

$$
\begin{aligned}
Binary\ \ outcome\ \ z_i &\sim Bernoulli(p_i) \\
T_i &\sim N(\boldsymbol{\beta}_i^t\boldsymbol{\theta}, \tau^2), \text{where } T_i = logit(p_i) \\
\boldsymbol{\theta}, \tau^2 \mid \boldsymbol{\gamma}, \mathbf{h} &\sim NIG(0, diag(\boldsymbol{\gamma})diag(\mathbf{h}), a_\tau, b_\tau) \\
h_j &\sim IG(c_j, d_j) \\
\gamma_{jk} &\sim Bernoulli(\pi_j)
\end{aligned}
$$

for $i = 1, ..., n$, $j = 0, \ldots, \log_2 m$ and $k = 0, ..., 2^j - 1$.

3.3 Posterior Inference

Again, we derive the full conditional distributions and depend on MCMC methods to simulate the parameters from this posterior distribution. The conditional distributions under mixture priors are given separately for the regression and the classification models.

3.3.1 Regression Model

The conditional distribution for the wavelet coefficients $\boldsymbol{\beta}_i$ follows from the conjugate model specifications and combination of information from both the regression and the classification segments. Let $\mathbf{U}_i = diag(\boldsymbol{\eta}_i)diag(\mathbf{g})$, then

$$
\boldsymbol{\beta}_i|\mathbf{Y}_i, T_i, \boldsymbol{\theta}, \sigma^2, \tau^2, \mathbf{U}_i \sim MN(\boldsymbol{\beta}_i^*, \sigma^2\tau^2\mathbf{U}_i^*) \tag{3.10}
$$

where $\mathbf{U}_i^* = (\tau^2(\mathbf{U}_i^{-1} + \mathbf{X}^t\mathbf{X}) + \sigma^2\boldsymbol{\theta}\boldsymbol{\theta}^t)^{-1}$ and $\boldsymbol{\beta}_i^* = \mathbf{U}_i^*(\tau^2\mathbf{X}^t\mathbf{Y}_i + \sigma^2 T_i\boldsymbol{\theta})$. However, the model variance σ^2 is updated only using the regression likelihood as

$$\sigma^2|\boldsymbol{\beta}, \mathbf{Y}, \mathbf{U} \sim IG(a_\sigma^*, d_\sigma^*) \tag{3.11}$$

where $a_\sigma^* = a_\sigma + mn/2$ and $b_\sigma^* = b_\sigma + \sum_{i=1}^n \left[\mathbf{Y}_i^t\mathbf{Y}_i - \mathbf{Y}_i^t\mathbf{X}(\mathbf{U}_i^{-1} + \mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}_i\right]/2$.

The next layer, consists of the scale parameters g_ℓ which are updated by

$$g_j|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma} \sim IG(u_j^*, v_j^*) \tag{3.12}$$

where $u_j^* = u_j + n\Big[\sum_{i,k}\eta_{ijk}\Big]/2$ and $v_j^* = v_j + \Big[\sum_{i,k}\eta_{ijk}\beta_{ijk}^2\Big]/2\sigma^2$. The indicator variables η_{ijk} are simply updated as

$$f(\eta_{ijk}|\eta_{-ijk}, \mathbf{U}_i, \boldsymbol{\theta}, \mathbf{Y}) \propto \frac{|\mathbf{A}_i^*|^{1/2}}{|\mathbf{U}_i|^{1/2}}(b_s^*)^{-a_s^*}\rho_j. \tag{3.13}$$

where $\mathbf{A}_i^* = (\mathbf{U}_i^{-1} + \mathbf{X}^t\mathbf{X})^{-1}$, $a_s^* = a_\sigma + m/2$ and $b_s^* = b_\sigma + \left[\mathbf{Y}_i^t\mathbf{Y}_i - \mathbf{Y}_i^t\mathbf{X}\mathbf{A}^*\mathbf{X}^t\mathbf{Y}_i\right]/2$.

3.3.2 Logistic Classification Model

The conditional distributions for the logistic classification model follow in a similar way, except now the detail coefficients $\boldsymbol{\beta}_i$ serve as the predictors of the latent variables T_i. The corresponding coefficients $\boldsymbol{\theta}$ are updated as

$$\boldsymbol{\theta}|\boldsymbol{\beta}, \tau^2, \mathbf{V}, \mathbf{T} \sim MN(\boldsymbol{\theta}^*, \tau^2\mathbf{V}^*) \tag{3.14}$$

where $\mathbf{V}^* = (\boldsymbol{\beta}\boldsymbol{\beta}^t + \mathbf{V}^{-1})^{-1}$, $\boldsymbol{\theta}^* = \mathbf{V}^*\boldsymbol{\beta}\mathbf{T}$ and $\mathbf{T} = (T_1, \ldots, T_n)^t$. The conjugate IG prior for τ^2 leads to its marginal conditional distribution as

$$\tau^2|\boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{T}, \mathbf{V} \sim IG(a_\tau^*, b_\tau^*) \tag{3.15}$$

where $a_\tau^* = a_\tau + n/2$ and $b_\tau^* = b_\tau + [\mathbf{T}^t\mathbf{T} - \boldsymbol{\theta}^{*t}(\mathbf{V}^*)^{-1}\boldsymbol{\theta}^*]/2$.

The scale parameters in this model h_j are again updated by

$$h_j|\boldsymbol{\theta}, \tau^2, \boldsymbol{\gamma} \sim IG(c_j^*, d_j^*) \tag{3.16}$$

where $c_j^* = c_j + \left[\sum_k \gamma_{jk}\right]/2$ and $d_j^* = d_j + \left[\sum_k \gamma_{jk}\theta_{jk}^2\right]/2\tau^2$. The indicator variables γ_{jk} are simply updated as

$$f(\gamma_{jk}|\boldsymbol{\gamma}_{-jk}, \mathbf{T}, \boldsymbol{\theta}, \mathbf{V}) \propto \frac{|\mathbf{V}^*|^{1/2}}{|\mathbf{V}|^{1/2}}(b_\tau^*)^{-a_\tau^*}\pi_j. \tag{3.17}$$

Finally, the latent variable vector $\mathbf{T}$ is updated from a non-standard posterior distribution by a Metropolis step,

$$f(\mathbf{T}|\boldsymbol{\beta}, \boldsymbol{\theta}, \tau^2, \mathbf{z}) \propto \exp\left\{-\frac{1}{2\tau^2}||\mathbf{T} - \boldsymbol{\beta}^t\boldsymbol{\theta}||^2\right\} \times \prod_{i=1}^{n} \frac{e^{T_i z_i}}{1 + e^{T_i}}. \tag{3.18}$$

3.3.3 Posterior Inference with Laplace Priors

Most of the posterior distributions follow from above two sections, except now we do not have the latent indicators ($\eta_{jk} = 1$) and the posterior distribution of the scaling parameters is given by

$$f(g_j|\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{g_j^{n2^{j-1}/2}} \exp\left[-\frac{1}{2}\left(\frac{\sum_{i=1,k=0}^{N,2^{j-1}} \beta_{ijk}^2}{g_j\sigma^2} + \lambda_j g_j\right)\right], \tag{3.19}$$

which is an inverse Gaussian distribution, $g_j \sim \text{InvGauss}(-\frac{n2^{j-1}}{2}+1, \sum_{i=1,k=0}^{N,2^{j-1}} \frac{\beta_{ijk}^2}{\sigma^2}, \lambda_j)$.

3.4 Extension to Multicategory Classification

Here we are interested in classification problems where the response is a categorical variable with more than two categories. Assume that response vector $\mathbf{z} = (z_1, ..., z_J)$, indicates the observed response data, with z_i taking one of j possible categories, and let $p_{ij} = P(z_i = j)$ for $i = 1, \cdots, n$ and $j = 1, \cdots, J$, be the probability that the ith observation falls into the jth category. These probabilities are related to the predictor curve f_i through a link function. Once again, we span the function f with wavelet basis functions and use the wavelet coefficients $\boldsymbol{\beta}$ as the classifiers or covariates in the link model. We introduce a latent variable T_{ij} in the multinomial logit link

function (McFadden, 1973) and model the probabilities as described in Section 2.4. The MCMC training scheme is similar to the binary case and so are conditional distributions for posterior inference.

3.5 Prediction and Model Choice

For a new sample with predictor values $\mathbf{Y}_{new}$, the posterior predictive probability that its group type, denoted by z_{new} given the old data D is

$$p(z_{new}|\mathbf{Y}_{new}, D) = \int p(z_{new} = 1|\mathbf{Y}_{new}, \boldsymbol{\beta}_{new}, \boldsymbol{\theta}, \tau^2, \mathbf{V}) p(\boldsymbol{\beta}_{new}|\mathbf{Y}_{new}, \sigma^2, \mathbf{U}) p(\boldsymbol{\phi}|D) d\boldsymbol{\phi}, \tag{3.20}$$

where $\boldsymbol{\phi}$ is the vector of all the model parameters. Assuming conditional independence of the responses the integral can be approximated by the Monte Carlo estimate

$$\sum_{j=1}^{M} p(z_{new} = 1|\mathbf{Y}_{new}, \phi^{(j)})/M, \tag{3.21}$$

where $\phi^{(j)}$ $(j = 1 \ldots, M)$ are the MCMC posterior samples of the parameter $\boldsymbol{\phi}$.

We use correct classification rates to compare performance of different classification methods in Section **??**. When a test set is provided, we first obtain the posterior distributions of the parameters (training the model) based on the training data and use them to classify the test samples. For a new observation from the test set, say $\mathbf{z}_{i,tst}$, we will obtain the probability $p(\mathbf{z}_{i,tst} = 1|\mathbf{z}_{trn}, \mathbf{Y}_{trn}, \mathbf{Y}_{tst})$ by using an equation similar to (3.20), and approximate it by its Monte Carlo estimate as in equation (3.21). When this estimated probability exceeds .5, the new observation is classified as 1, otherwise, it is classified as 0.

If there is no test set available, we use a hold-one-out cross-validation approach. We will exploit the technique described in Gelfand (1996) to simplify our computation. For the cross-validation predictive density, in general, writing $\mathbf{z}_{-i}$ as the vector of z_j's

minus z_i,

$$p(z_i|\mathbf{z}_{-i}) = \frac{p(\mathbf{z})}{p(\mathbf{z}_{-i})} = \left[\int \{p(z_i|\mathbf{z}_{-i},\phi)\}^{-1} p(\phi|\mathbf{z}) d\phi\right]^{-1}. \tag{3.22}$$

Monte-Carlo integration yields

$$\hat{p}(z_i|\mathbf{z}_{-i}) = M/\sum_{j=1}^{M} \left[p(z_i|\mathbf{z}_{-i},\phi^{(j)})\right]^{-1}, \tag{3.23}$$

where $\phi^{(j)}$, $j = 1,\ldots,M$ are the MCMC posterior samples of the parameter vector ϕ. This simple expression is due to the fact that z_i's are conditionally independent given ϕ_i's. If we wish to make draws from $p(z_i|\mathbf{z}_{-i,trn})$, then we need to use importance sampling (Gelfand, 1996).

3.6 Examples of Application

In this section, we apply the Bayesian wavelet-based classification method denoted by BWCC to analyze simulated data and several real data sets, including both binary and multicategory response cases. We analyze Medfly data containing smooth curves, leaf data with mild sharp curves and proteomics mass spectrometry data possessing many sharp curves. To put in weak but proper prior information for inverse-Gamma prior, we use the shape hyperparameter to be larger than 1, allowing the expectation of the IG distribution exists. So the hyperparameters (a, b) are specified as (2,2); and both (u, v) and (c, d) are specified as (2,2). In all the simulations, we run the MCMC for 80,000 iterations with a burn-in of the first 20,000 iterations.

For a comparative study, we also apply naive wavelet-based classification model (BNWCC) as well as two other naive plug-in methods to these data sets. Unlike the unified model, the naive version of Bayesian wavelet-based method (BNWCC) separates the regression and classification models. It uses the regression model to obtain the estimate of the wavelet coefficients that are later plugged in the classification

model treating them as a set of classifiers. Furthermore, we use wavelet-based empirical Bayes thresholding methods in the regression step following Silverman and Johnstone (2005) and the selected wavelet coefficients have been employed to two different classification methods, support vector machine and classical logistic regression. These two methods will be denoted as EBTSVM and EBTLOG respectively. For comparison purpose, we also apply a unified spline-based Bayesian method (SBCC), which comprises logistic regression and simply uses BIC to determine the number of evenly distributed knots. The following measures are used for comparison of the different methods. The correct classification rate (CCR), where

$$CCR = \frac{number\ \ of\ \ correctly\ \ classified\ \ samples}{total\ \ number\ \ of\ \ samples} \times 100\%,$$

is reported as the result of classification for all data sets. When data set includes disease group(s) versus control group, we also report the false discovery rate (FDR), where

$$FDR = \frac{number\ \ of\ \ samples\ \ falsely\ \ classified\ \ into\ \ disease\ \ group}{total\ \ number\ \ of\ \ samples\ \ classified\ \ into\ \ disease\ \ group} \times 100\%.$$

3.6.1 Application on Simulated Data

We conduct two simulation studies to illustrate the capabilities of our method. We want to simulate curves with very sharp peaks and have used the Bump and Heavisine functions (Donoho and Johnstone, 1994, 1995). The Bump functions corresponding to two classes are very similar except at two locations and separating them can be a difficult classification problem. Similarly for Heavisine functions, the first class contains a smooth function and the second class has single spike added to the smooth function. We have plotted the overlapping functions (for the two classes) in Figure 9. The curves are corrupted by additive Gaussian white noise $N(0, \sigma^2)$ with signal-to-noise ratio equal to 5. We generate 24 curves from two classes with different location

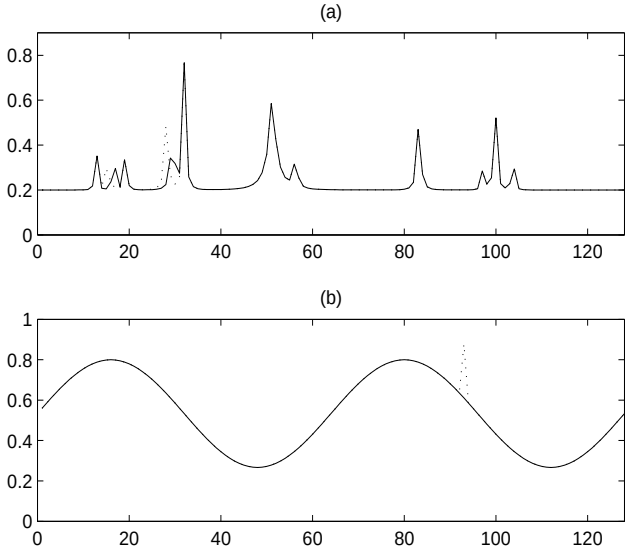

Figure 9: Above: Simulated Bump curves for the two classes. Below: Simulated Heavisine curves for the two classes. Solid line corresponding to the first class and dotted line corresponding to the second class.

parameters and evenly split them into training and testing sets. We have used three prior distributions Gaussian (G), Laplace (L), Mixture (M) for the wavelet coefficients ($\boldsymbol{\beta}$) as well as for the classification parameters ($\boldsymbol{\theta}$) in our unified (BWCC) and naive (BNWCC) wavelet based models. We have compared these models with the spline based Bayesian method (SBCC), and the naive method utilizing empirical Bayes wavelet thresholding with SVM classifier (EBTSVM). We performed 50 repetitions of simulation for both Bump curves and Heavisine curves and report the average CCR over these 50 replications.

Our results in Table 5 show that for both Bump and Heavisine curves, the naive and unified version of our wavelet-based methods yield best CCR with scale-mixture prior. Hence, we will focus on using scale-mixture prior for further applications in later sections. Classification results also show that the unified model benefits from the combination of wavelet functional regression and logistics classification models as

Table 5: The CCRs comparison of our methods and other methods for analyzing simulated Bumps and Heavisine curve data.

Methods	BWCC	BNWCC	SBCC	EBTSVM
Bump CCR%	82(G), 85(L), 86(M)	73(G), 77(L), 78(M)	66	75
Heavisine CCR%	88(G), 92(L), 92(M)	77(G), 83(L), 85(M)	73	79

Note: BNWCC and BWCC are the naive version and unified version of Bayesian wavelet-based classification methods. Three different priors are explored with this data set. SBCC is spline-based Bayesian curve classification method. EBTSVM is the naive method simply stacking empirical Bayes wavelet thresholding in Silverman and Johnstone (2005) and support vector machine.

it performs uniformly better than all the naive plug-in methods. Meanwhile, all the wavelet based methods performed better than the Bayesian spline-based method.

3.6.2 Application on Medfly Data

Even though our method is particularly useful for classification of wiggly functions nonetheless it performs well to classify smooth functions. To demonstrate this we consider Medfly data (Müller and Stadtmüller, 2004) where the predictor curves are smooth functions. It has been a long-standing problem in evolution and ecology to analyze the relationship between longevity and reproduction. The precise nature of the "cost of reproduction" remains elusive. Medfly data consists one thousand Mediterranean fruit files or medflies for short, described in Carey *et al* (1998). A fly is classified as long-lived if its lifetime is longer than 44 days, otherwise it is classified as short-lived. In addition to recording each fly lifetime, simple counts of daily eggs laid by that fly were also observed. For prediction of longevity, we use the egg-laying trajectories from 1 to 32 days as the predictor curves. Flies included in our analysis are those flies that lived past 34 days and were not barren during their first 32 days. Of those 511 flies passed this screening step, 246 were short-lived and 255 were long-lived. Randomly selected halves of each class form the training set and

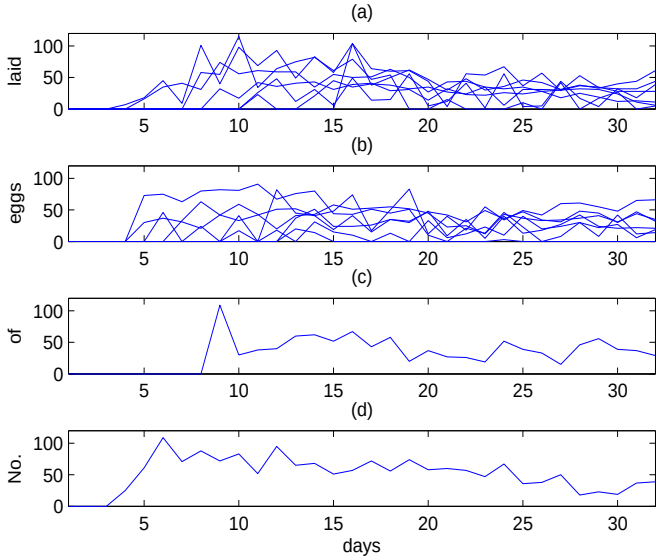

Figure 10: The egg-laying trajectories from 1 to 32 days for two classes in training data set, 123 of short-lived and 132 of long-lived, are shown in (a) and (b). Examples of single egg-laying trajectories, short- and long-lived, are in (c) and (d).

the other halves consist the testing set. Medfly data curves are shown in Figure 10. The top two plots shown in Figure 10 contains some randomly selected trajectories which reveal no clear distinction between the classes. Thus the classification task here is difficult. We repeat the splitting of training and testing sets 20 times and report the average CCR over the 20 repetitions in Table 6. The results indicate that even for a collection of smooth curves our unified wavelet based method performs marginally better than other methods. The linear classification boundary that enable validity of logistic regression is checked by residual plots as in Figure 11. There is no obvious non-linear trend in the residual plot so we claim that logistic regression model satisfactorily explains the relationship between the regressor, wavelet coefficients, and binary categorical outcomes.

We apply inverse wavelet transform onto the fitted regression coefficient $\boldsymbol{\theta}$ in wavelet domain to get the reconstructed regression coefficients for the original egg-

Table 6: The CCRs comparison of our method and other methods for testing Medfly data.

Methods	BWCC	BNWCC	SBCC	Müller and Stadtmüller		EBTSVM	EBTLOG
				logit	SPQR		
overall	62	57	58	58	59	58	57
short-lived	58	49	57	53	52	56	54
long-lived	68	65	61	63	65	62	62

Note: BNWCC, BWCC, SBCC and EBTSVM are same as in Table 1. SPQR stands for Müller and Stadmüller's semiparametric quasi-likelihood regression method. EBTLOG are the naive methods simply stacking empirical Bayes wavelet thresholding in Silverman and Johnstone (2005) and logistic regression.

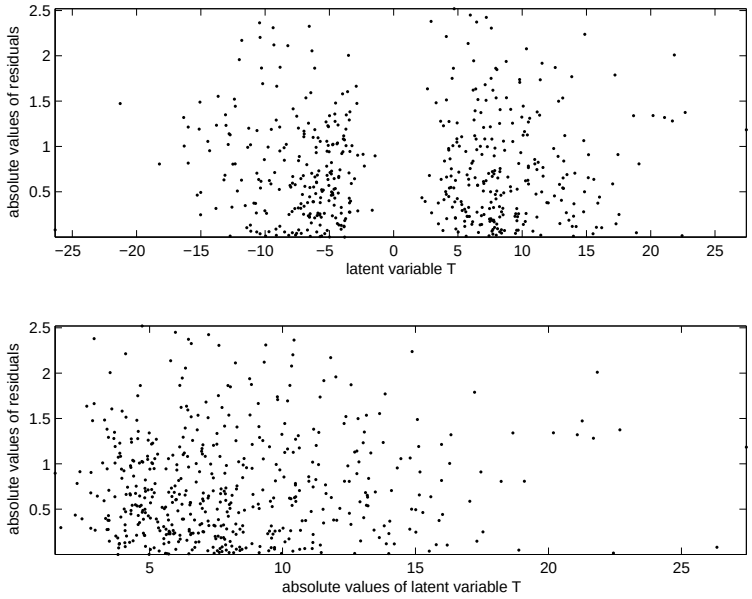

Figure 11: Residual plots of medfly data set. The top plot displays latent variable versus absolute value of residual, while the bottom one displays both latent variable and residual using absolute value scale.

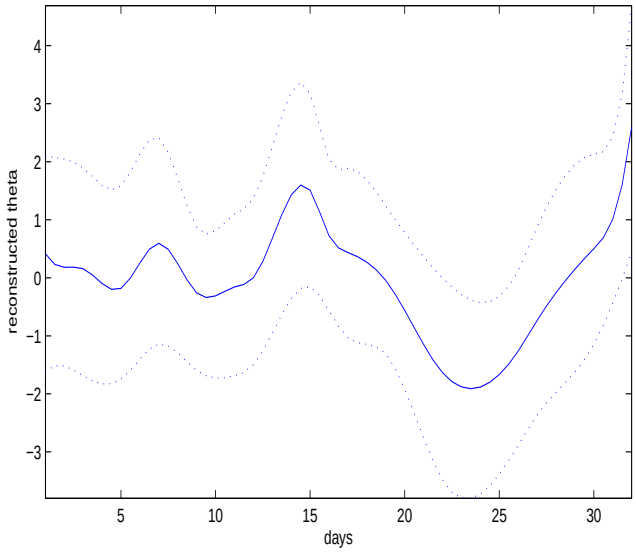

Figure 12: Reconstructed regression coefficient ($\boldsymbol{\theta}$) vs days. Dotted lines are 90% credible bands.

laying trajectory, Figure 12, which directly reflect the regression effect of reproduction on longevity. Therefore we can test the null hypothesis that reproduction has no linear regression effect on longevity. Since coefficients reach the highest end towards age 32 days, we reject the null hypothesis of no effect. Larger regression coefficients are associated with increased chance for longevity. More reproduction activity during about 13-18 days and past 28 days is associated with increased longevity. On the other hand, decreased reproduction between 9-11 days and 20-27 days results in decreased longevity. Late reproduction that may have a protective effect is most significantly associated with increased longevity in our analysis. Our conclusions also agree with those in Müller and Stadtmüller (2004).

3.6.3 Application on Leaf Data

We obtain a "pseudo time series" data set from Keogh and Folias (2002). The data set contains a collection of six different species of leaf images and was analyzed in

Table 7: The CCRs comparison of our methods and other methods for leaf data.

Methods	BWCC	BNWCC	SBCC	EBTSVM	EBTLOG
CCR%	94	74	70	75	73

Note: BWCC, BNWCC, SBCC, EBTSVM and EBTLOG are same as in Table 1 and 2.

Ratanamahatana and Keogh (2004a). The leaf image is converted into a "pseudo time series" by measuring the local angle of a trace of its perimeter. All time series are then interpolated into the same length, which is required to apply a type of distance measure for analysis utilized in Ratanamahatana and Keogh (2004a, 2004b) and Keogh *et al* (2004). After conversion and interpolation, each series was standardized to have mean zero and unit variance. The data set comprises four different species of maple and two species of oak, with 442 instances in total. For binary classification case, we only use a subset of leaf data set that comprises the two species (Circinatum (maple) and Garryana (oak)) with 150 instances. In our analysis, twenty two points are ignored from the end of each curve so that there are 128 points left for every curve. This last small part of the curve may carry similar information as the first small part because of approximate symmetry of the leaf image. Randomly selected 140 curves form the training set and the other 10 curves consist the testing set. We repeat the splitting of training and testing sets 20 times and report the average CCR over the 20 repetitions in Table 7. Our unified wavelet-based curve classification method outperforms all other method with the highest CCR, 94%. The naive plugging in method combining empirical Bayes thresholding and support vector machine is barely in the second place. Actually three naive wavelet-based methods yield very close classification results. For this leaf data set, all wavelet-based methods perform better than spline-based one. The possible reason could be that the separation between groups are emphasized when the wavelet-based methods achieve sparsity by

either mixture selection prior or thresholding. However, one may expect to improve the performance of spline-based method by some adaptive procedures such as knot tunning.

3.6.4 Wavelets for Unequispaced Design

For non-equispaced design, such as in the next two examples analyzing proteomics data sets, we use lifted wavelet transforms (Sweldens, 1997). These, unlike the traditional wavelet transforms, do not require regularly spaced samples. Traditional wavelet transforms (designed for equispaced samples) can be factored into a sequence of simpler transforms using the lifting scheme (Sweldens and Daubechies, 1996); and each lifting step is a refinement over the previous steps and represents an increase in the smoothness (or order) of the wavelet bases. These features can be extended to non-equispaced designs by allowing more flexible basis functions that are not simply translates or dilates of one fixed function and using the Lifting scheme to perform the construction in the time domain. The wavelets resulting from the lifting scheme still have all the powerful properties of traditional wavelets such as localization and good approximation. Despite these properties, the lifting scheme has been largely overlooked in recent literature and many authors have resorted to using interpolation for generating equispaced samples for their analysis.

The lifted construction used in the following examples involves two separate steps. The first step involves an *unbalanced* Haar transform, that is the usual Haar transform with adjustments for unequal distance between two successive observations. The coefficients from this transform are used as the input for a second lifting step that is an unbalanced version of a biorthogonal Spline wavelet. Thus the degree of the spline functions determines the smoothness of the overall basis. More details about such constructions can be found in Deloulle (2002). The wavelet transforms built in

this manner are not orthogonal as in the previous examples. This does not overly affect the posterior inference or the performance of our model as we ensure near orthogonality of transform within the lifting scheme. The posterior distributions as calculated in the appendix can be easily extended to the case where X is not orthogonal.

3.6.5 Application on Toxicoproteomics Data

Our next real data example is to classify toxicoproteomics data from surface enhanced laser desorption and ionization technology which was first analyzed in Petricoin *et al* (2004). This is an experiment on detection of doxorubincin induced cardiotoxicity and samples are from Spontaneously Hypertensive Rats with acute doxorubicin cardiotoxicity, subacute and saline alone controls. A mass spectrum is a curve where the x-axis is mass to charge (m/z) value, the ratio of the weight of a specific molecular to its electrical charge, and the y-axis is the relative signal intensity for the molecule, a measure of the abundance of the molecule in the sample. Apart from the serum spectra functional curve (see Figure 13), it has categorical response variables from cardiotoxicity and control groups. Pre-processing done by Petricoin *et al* (2004) is called binning process. The high-resolution spectra is binned using a function of 400 parts per million, e.g., the m/z bin sizes linearly increase from 0.28 at m/z 700 to 4.75 at m/z 12000. The m/z values in the spectra are not the actual m/z values from raw mass spectra but generated based on binned data by the high-resolution instrument. The binning process makes all samples have identical mass/charge (m/z) values and the number of data points condensed from 350000 to 7105 per sample. Binning can introduce coarseness and thus subtle trends or findings can then be masked. Therefore exploring of various binning techniques remains under investigation (Johann *et al*, 2004).

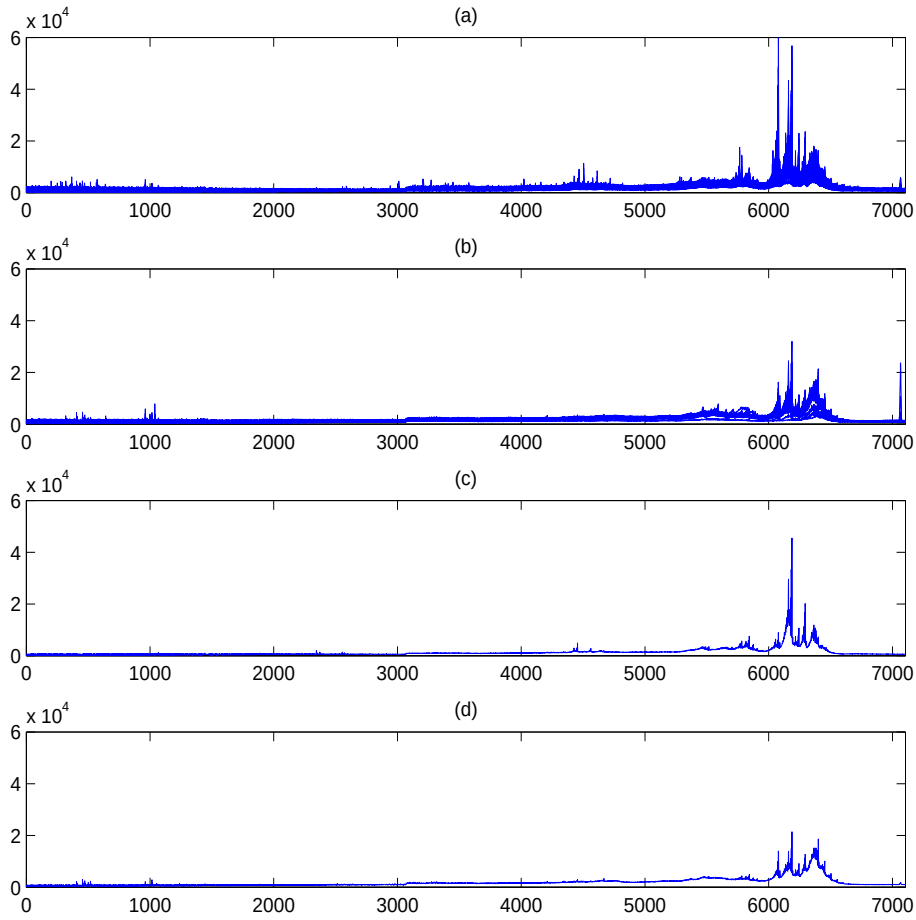

Figure 13: The spectra curves from two classes in training data set, 22 of cardiotoxicity and 14 of control, are shown in (a) and (b). Example of single original curves, cardiotoxicity and control, are in (c) and (d). Curves are shown based on spectra after binning process so there are total 7105 points in each curve.

Table 8: The CCRs and FDRs comparison of our method and other methods for analyzing toxicoproteomics data. BNWCC, BWCC, SBCC, and EBTSVM are same as in Table 1.

Methods	BWCC	BNWCC	SBCC	ProteomeQuest	EBTSVM
Test Set 1's CCR%	100	89	81	100	86
Test Set 1's FDR%	0	14	19	0	16
Test Set 2's CCR%	90	73	66	81	72

Note: BWCC, BNWCC, SBCC, and EBTSVM are same as in Table 1 and 2.

We have used this pre-processed data for classification purpose. There are 36 training samples and two sets of test samples. First set of test samples contains 36 observations which are very similar to the training data so easy to classify. The second test set contains 43 observations and is very different from the training data so is harder to classify than the previous one. Starting from those binned spectra, which are accessible to public, we further select 512 points through out the whole 7105 points by keeping every 12 other points so that each curve is well represented by reasonable number of points. Then the data has been transformed to the log-scale and further standardized to have mean zero and variance one. Testing data are treated exactly as the training data. Because of the non-equispaced m/z values for each spectrum, we apply the lifting scheme as stated at the beginning of this section when realizing our wavelet-based methods.

The results are presented in Table 8. We compare the classification results of our methods with those of three other methods: spline-based method, ProteomeQuest (Petricoin *et al*, 2004) and the naive plugging-in method using empirical Bayes thresholding and support vector machine. For test set 1, both unified wavelet-based method and ProteomeQuest correctly classify all samples for cardiotoxicity and control groups. Our naive wavelet-based method performs better than the plugging-in method and spline-based method in terms of both correct classification rate and false

discovery rate. All these methods also maintain the similar ranks according to false discovery rates. For test set 2, our unified wavelet-based method is the winner with 90% correct classification rate. Both naive wavelet-based methods generate similar results but still beat the spline-based method yielding unsatisfactory result. Rats in set 2 are older ones under long-term saline alone or dexrazoxane treatments. The difference in ages and treatments between rats in test set 2 and those in training or test set 1 makes the classification more difficult. Thus CCRs for test set 2 are lower than those of test set 1. All methods have very high false discovery rate for test set 2 as the samples are only from the control group.

To verify the normality assumption validity for proteomics data, we conduct a Bayesian analysis (Chaloner and Brant, 1988) of the residuals. The residuals in the regression model (3.3) are sampled from their posterior distributions, which are normal with mean $\mathbf{Y}_i - \mathbf{X}\boldsymbol{\beta}_i$ and covariance $\sigma^2\mathbf{I}$. A multivariate chi-square test is then performed to check the normality of sampled residuals for each curve. The p-values from all one hundred fifteen curves are provided in Figure 14. According to significance level of 0.05, we see that most of the curves satisfy the normality assumptions.

The assumption of independence across curve i could be not valid in some real applications like classification using proteomics mass spectra problems. We can induce correlations in the model by using correlated error structure (Johnstone and Silverman, 1997) or by exploiting a suitable random effect term within the wavelet model. This is a complex problem as the exact correlation structure is unknown in most of these proteomics studies and some type of simpler assumptions are needed about this structure.

3.6.6 Application on Multicategory Prostate Cancer Data

Classification of samples from multiple disease or cancer groups based on proteomics mass spectral curves is a challenging problem. We consider a multicategory prostate cancer mass spectral data which was previously analyzed by Adam *et al* (2002) and Wagner *et al* (2004). This data set, obtained at the Easter Virginia Medical School using SELDI-TOF mass spectrometry, consists of four categorical labels: unaffected healthy men, benign prostatic hyperplasia, organ-confined prostate cancer and non-organ-confined prostate cancer. There are 326 samples in the data set and 82 of them are unaffected healthy men, 77 men with benign prostatic hyperplasia, 84 patients have been diagnosed with organ-confined prostate cancer and 83 with non-organ-confined prostate cancer. Details of the pre-processing steps include peak detection and alignment (Adam *et al*, 2002). Pre-processing begins from selecting mass range between 2000 to 40000 Dalton because this range contained the majority of the resolved protein/peptides. Peak detection involves baseline subtraction, mass accuracy calibration and automatic peak detection, which are done by a software program through calculating noise, peak area and filter. Peaks are first sorted by mass and a mass error score, the measurement of mass difference between peak X and peak $X + 1$, is calculated for each peak. If the mass error score is small, peak X and peak $X + 1$ will be align into one peak, otherwise, they are considered distinct peaks. Finally, 779 peaks had been selected as input for analysis. We further select 512 peaks out of them by ignoring those having at most two samples with non-zero values, and standardize each spectrum to have mean zero and variance one. As in Wagner *et al* (2004), training and testing sets are formed by randomized 90/10 splits of each of the four classes. We repeat the splitting of training and testing sets 20 times and report the average CCR and FDR over the 20 repetitions. Due to the non-equispaced m/z

values for each spectrum, we apply the lifting scheme when realizing our wavelet-based methods. When applying the naive plugging-in method using support vector machine, we adopt the popular one-vs-all scheme for multicategory classifier.

Recently machine learning based methods have been used for cancer classification of binary and multi-class data (Ghosh *et al*, 2004; Chakraborty *et al*, 2005). Several flexible machine learning based classification methods like support vector machine, k-nearest neighbor, kernel method, quadratic discriminant rule were employed by Wagner *et al* (2004). Results of our wavelet-based methods together with those from Wagner *et al* (2004) are presented in Table 9. Our BWCC is clearly the winner with 92% CCR. Their support vector machine method is in the second place with 86% CCR. Spline-based Bayesian method fails to capture the spiky curves only yields 63% CCR. Our naive wavelet based method yields correct classification rate three percent more than simple plugging-in method with support vector machine. Table 9 also reports the overall false discovery rate that is calculated using control and benign prostatic hyperplasia groups as non-cancer group. Results according to false discovery rate suggest the naive plugging-in method is better than our naive wavelet-based method. In fact, these two naive methods are similar in two manners: they are wavelet-based and utilizing Bayesian modeling to select variables. The difference is that our method adopts mixture prior to achieve sparsity while empirical Bayes thresholding method uses different kind of threshold criterion. The main drawback of the machine learning methods is that they fail to recognize the functional nature of the data underlying the spectra curves.

The normality assumption for these prostate cancer data is verified by conducting a Bayesian analysis of the residuals, as aforementioned in Section 3.6.5. The p-values from all three hundred twenty six curves are provided in Figure 14. Most of them satisfy the normality assumptions, according to significance level of 0.05.

Table 9: The multicategory classification results comparison of our method and other methods for 4-category prostate cancer data.

Methods	overall CCR%	overall FDR%
BWCC	92	10
BNWCC	81	22
SBCC	63	30
EBTSVM	78	20
FCDA	84	–
SVM	86	–
Kernel	80	–
QDA	79	–
kNN	77	–

Note: BNWCC, BWCC, SBCC and EBTSVM are same as in Table 1. FCDA is Fisher's canonical (linear) discriminant analysis. SVM is linear support vector machine. Kernel is non-parametric discrimination method. QDA is quadratic discriminant rule. kNN is k-nearest neighbor method.

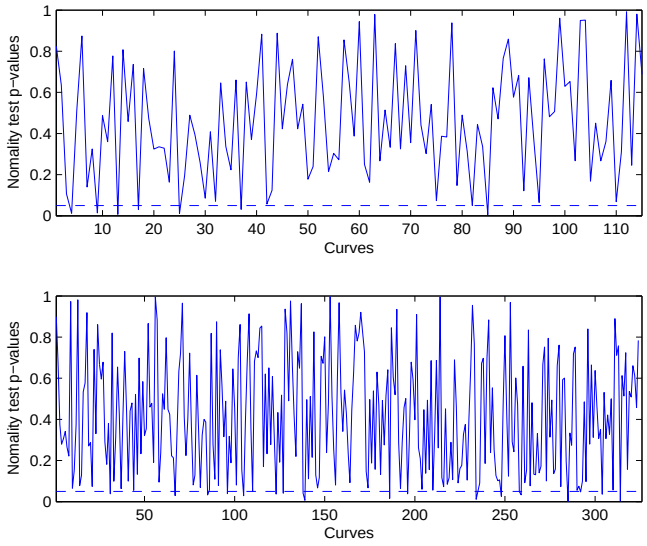

Figure 14: P-values for normality checking. The top plot is for 115 curves in toxicoproteomics data and the bottom one for 326 curves in prostate cancer data.

Methods and results in this chapter have been included in a submitted paper, Wang *et al* (2006).

CHAPTER IV

BAYESIAN SURVIVAL ANALYSIS USING PROPORTIONAL HAZARDS MODEL AND GENERALIZED LINEAR REGRESSION

4.1 Motivation

In previous two chapters, we were focus on classification task based on functional curves that are complex in sense of smooth or spiky, regularly or irregularly sampled. There are other issues of interest such as survival status that have relationships with time-dependent curve covariates. Usually, the curve covariates represent the process of disease marker. For example, for data collected by the Mayo Clinic between 1974 and 1984 (Fleming and Harrrington, 1991) on patients with primary biliary cirrhosis (PBC) of the liver, one wants to know how the potential predictors, serum bilirubin and albumin, are related to life expectancy and whether there is treatment effect of drug D-penicillamine. Notice that there are different numbers of measurements for each patient and they are taken at different times so it is not possible to use a standard multiple regression model. This characteristic pose most of existing methods in implausible situation. Both parametric and semiparametric models are available to model survival data. Commonly used parametric models include the exponential and Weibull models, which are attractive in their simplicity and the easy interpretability of their components. In practice, however, semiparametric proportional hazards models are widely used, since they impose no particular shape on the survival curves.

Among those frequentist methods (such as DeGruttola and Tu, 1994 and Hogan and Laird, 1997) for joint modeling of longitudinal and survival data, Wulfsohn and Tsiatis (1997) proposed a general approach that combines a proportional hazards

model for survival and a random effects model for regression. The random effects that consist of a linear functions of time in the form of $\theta_{0i}+\theta_{1i}t_{ij}+e_{ij}$ were assumed to have a bivariate normal distribution with a non-zero mean vector and covariance matrix. The hazards function is expressed by a baseline hazard term and an exponential function of the product of the linear function and regression coefficients. A Bayesian method was explored by Faucett and Thomas (1996), in which the same joint model was used and noninformative priors are assigned for all parameters. An apparent advantage of this approach is that it can give efficient estimation by making a direct link between the survival and longitudinal covariate. However, the linear parametric form of the functional covariate may be inappropriate in case of that the rate of change varies over the entire length of disease process. Also, the assumption of independence over longitudinal measurements of same individual is a very strong assumption, which could be violated in some settings.

There are also other existing Bayesian methods for this jointly type of modeling. They mainly differ in the ways of modeling longitudinal covariate. For example, Ibrahim, Chen and Sinha (2004) modeled bivariate longitudinal and survival data by assuming both of two covariates measure a true unobservable common measure that is modeled by an arbitrary function indexed by paramter vector. The relationships are illustrated as following

$$\begin{aligned}
\mathbf{Y}_{i1}(t) &= \mathbf{X}_i^*(t) + \boldsymbol{\varepsilon}_{i1}(t) \\
\mathbf{Y}_{i2}(t) &= \alpha_0 + \alpha_1 \mathbf{X}_i^*(t) + \boldsymbol{\varepsilon}_{i2}(t) \\
\mathbf{X}_i^*(t) &= g_{\gamma_i}(t)
\end{aligned}$$

The trajectory function was determined by exploratory analysis to take form of $\gamma_{1i} + \gamma_{2i}t + \gamma_{3i}t^2$. This approach partially overcomes the inappropriate parametric function format by leaving the trajectory function open to general structure. However, the

way to use underlying common measure for two covariates needs sound biological considerations for specific problems. Also, although the covariance between the two variables are modeled by a 2-by-2 matrix, the correlation across same individual is not considered. In another word, they couldn't avoid the independence assumption as in Wulfsohn and Tsiatis (1997) and Faucett and Thomas (1996). Motivated by the fact that the slope of CD4 for an individual can vary over time, Wang and Taylor (2000) introduced an integrated Ornstein-Uhlenbeck process into the longitudinal modeling, which is written as

$$\begin{aligned} Y_i(t_{ij}) &= Z_i(t_{ij}) + e_i(t_{ij}), \\ Z_i(t) &= a_i + bt + \beta X_i(t) + W_i(t). \end{aligned}$$

The term $W_i(t)$ is an IOU process with covariance function between values at times s and t given by

$$\frac{\sigma^2}{2\alpha^3}[2\alpha min(s,t) + exp(-\alpha t) + exp(-\alpha s) - 1 - exp(-\alpha \mid t - s \mid)].$$

This method also assumed independence across each longitudinal covariate. It is known that the IOU process greatly increases both the number of parameters and the computational complexity (Ibrahim, Chen and Sinha, 2001). Brown and Ibrahim (2003) started from similar model as in Wulfsohn and Tsiatis (1997) and Faucett and Thomas (1996) for their own Bayesian semiparametric joint modeling, but they used a quadratic form for longitudinal part and introduced nonparametric specification of the distribution of the random effects, β_i's, in longitudinal model. A Dirichlet process prior is used for those random effects to overcome concerns such as the distribution of β_i may vary over time or behave non-normally. However, the problems faced by Wulfsohn and Tsiatis (1997) and Faucett and Thomas (1996) were left unsolved. Guo and Carlin (2004) compared separate and joint modeling of longitudinal and

event time data and concluded that the joint Bayesian approach appears to offer significantly improved estimation and more efficient computation.

In the field of functional regression, basis function approach with splines is widely used for curve fitting. We propose a relatively simple joint model using spline basis, in which the usage of the splines simplifies the parameterizations and allows flexible non-linear pattern of the marker/predictor process. Joint model is more appropriate than separated model based on the fact that, generally, the longitudinal variable is correlated with survival response. Meanwhile, because information are shared between the regression and proportional hazards models, the joint modeling framework can improve the efficiency of estimation in both parts of the model. Additionally, we set up the model without the assumption of independence over longitudinal measurements of same individual, which fits better to real world problem settings. Our model can be easily expanded to include multiple functional covariates. We use Bayes factor to compare models with different covariates.

4.2 The Bayesian Unified Hierarchical Model

4.2.1 Regression Model for the Functional Covariates

In some survival analysis scenario, we observe time-dependent $\mathbf{Y}(\mathbf{t}_i)$ covariates curve and the pair (Z_i, δ_i) as response for each individual. Each individual has a lifetime T_i and an censor time C_i and they are related to response pair by the following way

$$Z_i = min(T_i, C_i) \quad \text{and} \quad \delta_i = \begin{cases} 1 & \text{if} \quad T_i \leq C_i \\ 0 & \text{if} \quad T_i > C_i \end{cases}$$

Assume that censoring is independent of all other survival and covariate information. For the covariates curve $\mathbf{Y}(\mathbf{t}_i)$, we have

$$\mathbf{Y}(\mathbf{t}_i) = \mathbf{f}(\mathbf{t}_i) + \boldsymbol{\epsilon}_i \quad \boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{i}) \tag{4.1}$$

where $\mathbf{t}_i = (t_1, ..., t_{p_i})$ is the time points when measurements were recorded for ith individual, $\mathbf{f}(\mathbf{t}_i)$ is the true functional covariates curve, and $\boldsymbol{\epsilon}_i$ represents noises. The covariates curves can often be used to predict the survival time or hazard function. However, the original covariates curves are usually not in proper condition to be employed in the prediction procedure. Hence, a generalized linear regression step for the covariates curves is necessary. We define $\mathbf{f}(\mathbf{t}_i) = \mathbf{X}(\mathbf{t}_i)\boldsymbol{\beta}_i$ so that the regression model becomes

$$\mathbf{Y}(\mathbf{t}_i) = \mathbf{X}(\mathbf{t}_i)\boldsymbol{\beta}_i + \boldsymbol{\epsilon}_i \quad \boldsymbol{\epsilon}_i \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_\mathbf{i}) \tag{4.2}$$

where $\mathbf{X}(\mathbf{t}_i)$ is a transform basis for ith individual, and $\boldsymbol{\beta}_i$ is the regressed covariates vector. If we adopt the spline basis as transform basis matrix, then the vector $\boldsymbol{\beta}_i$ is smoothed covariates coefficients. Natural cubic spline functions is employed here because of their desirable mathematical properties and easy implementation (de Boor, 1978, Green and Silverman, 1994). For notation convenience, we drop the time points part so that we use $\mathbf{Y}_i$ and $\mathbf{X}_i$ from now on. The natural cubic spline basis matrix $\mathbf{X}$ is evaluated on a fine lattice of points, then $\mathbf{X}_i$ is the basis matrix corresponding to time points included by ith individual. In practice, the natural cubic spline basis can be generated based on B-spline basis matrix with certain degrees of freedom on a sequence of knots that should include at least all time points in the data set. Singular value decomposition is then applied to construct the orthogonal basis matrix. It is worth of pointing out that although the full matrix $\mathbf{X}$ is orthogonally formed, the basis matrix $\mathbf{X}_i$ for ith subject is not orthogonal.

We can further concentrate the information from the covariates curve into one scalar variable w_i through a linear model

$$w_i = \boldsymbol{\beta}_i^t\boldsymbol{\theta} + e_i \quad e_i \sim N(0, \tau^2) \tag{4.3}$$

where $\boldsymbol{\theta}$ is the regression coefficient vector and e_i is error term. The benefit from

this linear model is that we overcome the computation difficulties by including the random error term.

For the regression model (2.4) with unstructured covariance Σ_i, we put prior distributions

$$\begin{aligned} \boldsymbol{\Sigma}_i &\sim IW(\mathbf{A}_i, b), && (4.4) \\ \boldsymbol{\beta}_i &\sim MN(0, \boldsymbol{\Omega}), && (4.5) \\ \boldsymbol{\Omega} &\sim IW(\mathbf{B}, d), && (4.6) \end{aligned}$$

where hyperparameters pairs $(\mathbf{A}_i, b)$ and $(\mathbf{B}, d)$ are scale matrices and degrees of freedom of inverse Wishart distribution. Again, as in Chapter II, the covariance matrix $\boldsymbol{\Omega}$ serves as smoothing parameter. Automatically coming from the Bayesian framework, the smoothing parameter selector is to place a continuous density probability prior on $\boldsymbol{\Omega}$ that allow us automatically put zero prior probability on the possibility of doing no smoothing at all. For the linear concentrating model (4.3), we use

$$\boldsymbol{\theta}, \tau^2 | \mathbf{V} \sim NIG(0, \mathbf{V}, a_\tau, b_\tau). \quad (4.7)$$

as prior distributions. Note that $V = diag(h_k)$ and $h_k \sim IG(c_k, d_k)$.

4.2.2 Cox Proportional Hazards (PH) Model

Cox PH model is often employed to study time-dependent covariates effects on survival responses. In those existing models (Ibrahim *et al*, 2004, Wang and Taylor, 2001), time dependent covariates and other baseline covariates such as gender and age are considered in the proportional hazards model. In this dissertation, we simply include the effects from time dependent covariates because other covariates may have some effects but they are not of main interest here. We plug the informative scalar w_i, which sometimes is called linear predictor and contains summarization of covariates

effects, into the PH model, so we have

$$h(t \mid \mathbf{Y}_i) = h_0(t)exp(w_i) \tag{4.8}$$

where $\mathbf{Y}_i$ is the ith individual covariates vector and $h_0(t)$ is the baseline hazard function free of the covariates. The baseline function can be approximated by a piece-wisely defined function

$$h_0(t) = \lambda_j \quad (s_{j-1} \leq t < s_j), \quad j = 1, \ldots, J \tag{4.9}$$

When the total number of intervals, J, is large, the step function approximates a smooth function. The value of J typically would be 10 or less. The prior distribution for parameters $\boldsymbol{\lambda} = \{\lambda_j\}$ in PH model is

$$\lambda_j \sim G(a_j, b_j) \tag{4.10}$$

where a_j and b_j are specified for each interval.

To summarize the hierarchical model set-up, we have

$$
\begin{aligned}
\textit{Random function } \mathbf{Y}_i &\sim MN(\mathbf{X}_i\boldsymbol{\beta}_i, \Sigma_i) \\
\boldsymbol{\Sigma}_i &\sim IW(\mathbf{A}_i, b) \\
\boldsymbol{\beta}_i &\sim N(0, \boldsymbol{\Omega}) \\
\boldsymbol{\Omega} &\sim IW(\mathbf{B}, d)
\end{aligned} \tag{4.11}
$$

$$
\begin{aligned}
\textit{Linear predictor } w_i &\sim MN(\boldsymbol{\beta}_i^t\boldsymbol{\theta}, \tau^2) \\
\boldsymbol{\theta}, \tau^2|\mathbf{V} &\sim NIG(0, \mathbf{V}, a_\tau, b_\tau), \ \ \textit{where } \mathbf{V} = diag(\mathbf{h}) \\
h_k &\sim IG(c_k, d_k)
\end{aligned}
$$

$$
\begin{aligned}
\textit{Hazard function } h(t \mid \mathbf{Y}_i) &= h_0(t)exp(w_i), \\
h_0(t) &= \lambda_j \ \ (s_{j-1} \leq t < s_j) \\
\lambda_j &\sim G(a_j, b_j)
\end{aligned}
$$

for $i = 1, ..., n$, $j = 1, \ldots, J$, and $k = 1, \ldots, q$.

4.3 Posterior Inference

MCMC methods are employed to simulate the parameters from joint posterior distribution which is not of explicit form. The full conditional distributions are given separately for the regression and PH models below.

4.3.1 Regression Model for the Functional Covariates

The conditional distribution for the ith regressed covariates vector $\boldsymbol{\beta}_i$ is updated using regression likelihood

$$
\boldsymbol{\beta}_i \mid \mathbf{X}_i, \mathbf{Y}_i, \Sigma_i, \boldsymbol{\Omega}, w_i, \tau^2, \boldsymbol{\theta} \sim MN(\boldsymbol{\beta}_i^*, \tau^2\boldsymbol{\Omega}^*) \tag{4.12}
$$

where $\boldsymbol{\Omega}^* = (\tau^2(\boldsymbol{\Omega}^{-1} + \mathbf{X}_i^t\Sigma_i^{-1}\mathbf{X}_i) + \boldsymbol{\theta}\boldsymbol{\theta}^t)^{-1}$ and $\boldsymbol{\beta}_i^* = \boldsymbol{\Omega}^*(\tau^2\mathbf{X}_i^t\Sigma_i^{-1}\mathbf{Y}_i + w_i\boldsymbol{\theta})$. The model covariance Σ_i is updated by

$$\boldsymbol{\Sigma}_i|\boldsymbol{\beta}_i, \mathbf{Y}_i, \mathbf{X}_i \sim IW(\mathbf{A}_i^*, b^*), \tag{4.13}$$

where $\mathbf{A}_i^* = \mathbf{A}_i + (\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\beta}_i)^t$ and $b^* = b + 1$. The common coefficient vector $\boldsymbol{\theta}$ is updated as

$$\boldsymbol{\theta} \mid \mathbf{w}, \boldsymbol{\beta}, \tau^2, \mathbf{V} \sim MN(\boldsymbol{\theta}^*, \tau^2\mathbf{V}^*) \tag{4.14}$$

where $\mathbf{V}^* = (V^{-1} + \sum_{i=1}^n \boldsymbol{\beta}_i\boldsymbol{\beta}_i^t)^{-1}$ and $\boldsymbol{\theta}^* = \mathbf{V}^*(\sum_{i=1}^n w_i\boldsymbol{\beta}_i)$. The conjugate IG prior for variance τ^2 leads to its conditional distribution as

$$\tau^2 \mid \boldsymbol{\theta}, \mathbf{V}, \mathbf{w}, \boldsymbol{\beta} \sim IG(a_\tau^*, b_\tau^*) \tag{4.15}$$

where $a_\tau^* = a_\tau + (q+n)/2$ and $b_\tau^* = b_\tau + \left[\boldsymbol{\theta}^t\mathbf{V}^{-1}\boldsymbol{\theta} + \sum_{i=1}^n (w_i - \boldsymbol{\beta}_i^t\boldsymbol{\theta})^2\right]/2$. The conditional distribution for the informative scalar w_i follows combination of information from both the regression and PH models. The likelihood of PH model lead to its non-standard form,

$$\begin{aligned} w_i \mid z_i, \delta_i, h_0(t), \boldsymbol{\beta}_i, \boldsymbol{\theta}, \tau^2 \quad \propto \quad & \Big[h_0(z_i)exp(w_i)\Big]^{\delta_i} \times \\ & exp\left\{-exp(w_i)\int_0^{z_i} h_0(u)du\right\} exp\left\{-\frac{(w_i^2 - 2w_i\boldsymbol{\beta}_i^t\boldsymbol{\theta})}{2\tau^2}\right\} \end{aligned} \tag{4.16}$$

which can be updated by a Metropolis step.

The next layer includes scale parameters h_k which is updated by

$$h_k \mid \boldsymbol{\theta}, \tau^2, \mathbf{V} \sim IG(c_k^*, d_k^*) \tag{4.17}$$

where $c_k^* = c_k + 1/2$ and $d_k^* = d_k + \theta_k^2/2\tau^2$, and the covariance matrix $\boldsymbol{\Omega}$ for spline coefficients which is updated as

$$\boldsymbol{\Omega}|\boldsymbol{\beta} \sim IW(\mathbf{B}^*, d^*), \tag{4.18}$$

where $\mathbf{B}^* = \mathbf{B} + \sum_{i=1}^n \boldsymbol{\beta}_i\boldsymbol{\beta}_i^t$ and $d^* = d + n$.

4.3.2 Cox Proportional Hazards (PH) Model

The parameters of baseline hazard step function $h_0(t)$, λ_j's, can be updated using PH model,

$$\lambda_j \mid \mathbf{Y}, \mathbf{Z}, \mathbf{w} \sim G(a_j^*, b_j^*) \tag{4.19}$$

where $a_j^* = a_j + \sum_{i=1}^{n} \delta_i I(s_{j-1} \leq z_i < s_j)$ and $b_j^* = b_j + \sum_{i=1}^{n} \Big[I(z_i > s_{j-1}) \times \int_{s_{j-1}}^{min(z_i, s_j)} exp(w_i) du \Big]$.

4.4 Bayesian Joint Model with Parametric Functional Regression

For comparison purpose, we also apply parametric regression model in the Bayesian joint modeling framework. We adopt the quadratic function format for the curve covariate and assign prior distributions, similarly in Ibrajim *et al* (2004). The model setup is summarized as below

$$\begin{aligned}
Random\ \ function\ \ \mathbf{Y}_i &\sim MN(\mathbf{T}_i \boldsymbol{\gamma}_i, \Sigma_i) \\
\boldsymbol{\Sigma}_i &\sim IW(\mathbf{C}_i, a) \\
\boldsymbol{\gamma}_i &\sim MN(\boldsymbol{\gamma}_0, \boldsymbol{\Phi}) \\
\boldsymbol{\gamma}_0 &\sim MN(0, \mathbf{V}) \\
\boldsymbol{\Phi} &\sim IW(\mathbf{D}, c)
\end{aligned} \tag{4.20}$$

$$\begin{aligned}
Hazard\ \ function\ \ h(t \mid \mathbf{Y}_i) &= h_0(t) exp\{\eta(\gamma_{1i} + \gamma_{2i} t + \gamma_{3i} t^2)\}, \\
h_0(t) &= \lambda_j \ \ (s_{j-1} \leq t < s_j) \\
\lambda_j &\sim G(a_j, b_j) \\
\eta &\sim N(0, \tau^2)
\end{aligned}$$

where $i = 1, ..., n,\ j = 1, \ldots, J,\ k = 1, \cdots, m_i,\ \boldsymbol{\gamma}_i = (\gamma_{1i}, \gamma_{2i}, \gamma_{3i})$ and the kth row of matrix $\mathbf{T}_i$ is $(1, t_{ik}, t_{ik}^2)$. Posterior distributions are given for regression model and

proportional hazards model separately. For the regression segment, we have

$$\begin{aligned}
\boldsymbol{\Sigma}_i \mid \cdot \sim IW(\mathbf{C}_i^*, a+1) \quad & where \quad \mathbf{C}_i^* = \mathbf{C}_i + (\mathbf{Y}_i - \mathbf{T}_i\boldsymbol{\gamma}_i)(\mathbf{Y}_i - \mathbf{T}_i\boldsymbol{\gamma}_i)^t, \\
\boldsymbol{\gamma}_i \mid \cdot \sim MN(\boldsymbol{\gamma}_0^*, \boldsymbol{\Phi}^*) \quad & where \quad \boldsymbol{\Phi}^* = (\mathbf{T}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{T}_i + \boldsymbol{\Phi}^{-1})^{-1} \\
& \qquad\qquad \boldsymbol{\gamma}_0^* = \boldsymbol{\Phi}^*(\mathbf{T}_i^t\boldsymbol{\Sigma}_i^{-1}\mathbf{Y}_i + \boldsymbol{\Phi}^{-1}\boldsymbol{\gamma}_0), \\
\boldsymbol{\gamma}_0 \mid \cdot \sim MN(\boldsymbol{\mu}^*, \mathbf{V}^*) \quad & where \quad \mathbf{V}^* = (n\boldsymbol{\Phi}^{-1} + \mathbf{V}^{-1})^{-1}, \boldsymbol{\mu}^* = \mathbf{V}^*\boldsymbol{\Phi}^{-1}\sum_{i=1}^{n}\boldsymbol{\gamma}_i, \\
\boldsymbol{\Phi} \mid \cdot \sim IW(\mathbf{D}^*, c+n) \quad & where \quad \mathbf{D}^* = \mathbf{D} + \sum_{i=1}^{n}(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_0)(\boldsymbol{\gamma}_i - \boldsymbol{\gamma}_0)^t.
\end{aligned}$$

For the PH model segment, the baseline hazard λ_j has gamma conditional posterior distribution, $G(a_j^*, b_j^*)$, where $a_j^* = a_j + \sum_{i=1}^{n}\delta_i I(s_{j-1} \leq z_i < s_j)$ and $b_j^* = b_j + \sum_{i=1}^{n}\left[I(z_i > s_{j-1}) \times \int_{s_{j-1}}^{min(z_i, s_j)} exp\{\eta(\gamma_{1i} + \gamma_{2i}u + \gamma_{3i}u^2)\}du\right]$. The conditional posterior of regression coefficient η does not have close form and is proportional to

$$exp\{-\frac{1}{2\tau^2}\eta^2\} \times \prod_{i=1}^{n}\left[\left(h_0(z_i)exp\{\eta(\gamma_{1i} + \gamma_{2i}z_i + \gamma_{3i}z_i^2)\}\right)^{\delta_i} \times exp\left\{-\int_0^{z_i} h_0(u)exp\left\{\eta(\gamma_{1i} + \gamma_{2i}u + \gamma_{3i}u^2)\right\}du\right\}\right] \tag{4.21}$$

The likelihood contribution of ith subject to posterior distribution of η is given by

$$\left(h_0(z_i)exp\{\eta(\gamma_{1i} + \gamma_{2i}z_i + \gamma_{3i}z_i^2)\}\right)^{\delta_i} exp\left\{-\int_0^{z_i} h_0(u)exp\left\{\eta(\gamma_{1i} + \gamma_{2i}u + \gamma_{3i}u^2)\right\}du\right\}.$$

Ibrahim *et al* (2004) used an approximation for the integral calculation. For computational convenience, we simply use approximation based on classical trapezoidal rule.

4.5 Extension to Multiple Covariates and Bayes Factor Calculation

In real world one may want to perform regression with multiple functional predictors. For example, measurements on both bilitubin and albumin levels in PBC data can be considered as functional predictors. Our proposed Bayesian unified hierarchical

model can be easily extended to multiple covariates case. For the ith individual, we observe ℓth functional covariate $\mathbf{Y}_{i\ell}$, the corresponding spline basis matrix is then $\mathbf{X}_{i\ell}$, and $\boldsymbol{\beta}_{i\ell}$ is regressed coefficients. So, instead of regression model (4.2), we have

$$\mathbf{Y}_{i\ell} = \mathbf{X}_{i\ell}\boldsymbol{\beta}_{i\ell} + \boldsymbol{\epsilon}_{i\ell} \quad \boldsymbol{\epsilon}_{i\ell} \sim MN(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{i}\ell}). \tag{4.22}$$

And the concentration linear model 4.3 becomes

$$w_i = \sum_{\ell=1}^{L} \boldsymbol{\beta}_{i\ell}^t \boldsymbol{\theta}_\ell + e_i \quad e_i \sim N(0, \tau^2). \tag{4.23}$$

Therefore, the prior distributions and posterior distributions are as same in Section 4.2 and 4.3. The MCMC scheme is similar to the single functional covariates case and so are conditional distributions for posterior inference.

To select from models with different functional covariates, we use Bayes factor that is the coherent way of comparing models in a Bayesian framework (Kass and Raftery, 1995). Let model M_1 includes only one functional covariates and model M_2 includes two. The Bayes factor is calculated using the ratio of posterior to prior odds, $B = \frac{\pi(\mathbf{D}|M_2)}{\pi(\mathbf{D}|M_1)}$ that is a measure of preference for a model M_2 against another model M_1 given data $\mathbf{D}$. If $2logB$ lies between the range 5 to 10, there is strong evidence in favor of model M_2. If it is larger than 10, there is very strong evidence for model M_2. The marginal density is an input to the computation of Bayes factor. When the marginal likelihood can not be obtained by $\pi(\mathbf{D} \mid M_i) = \int f(\mathbf{D} \mid \boldsymbol{\Theta})\pi(\boldsymbol{\Theta})d\boldsymbol{\Theta}$, one can compute the marginal density $m(\mathbf{D})$ (equivalent to marginal likelihood $\pi(\mathbf{D} \mid M_i)$ under model M_i) as

$$m(\mathbf{D}) = \frac{f(\mathbf{D} \mid \cdot)\pi(\cdot)}{\pi(\cdot \mid \mathbf{D})}.$$

The calculation of marginal likelihood has been proved extremely challenging and analytic evaluation of it is almost never possible (Chib, 1995, Chib and Jeliazkov, 2001). Due to the complexity of the likelihood in our proportional hazards model for

the survival part, it is impossible to derive the marginal likelihood in explicit form. We basically follow the technique for general case from Chib (1995) to calculate the marginal likelihood. The details of derivation are in the Appendix.

4.6 Applications to PBC Data

These data were obtained from StatLib. It is a follow-up to the original primary biliary cirrhosis (PBC) data set that were from the Mayo Clinic trial in PBC of the liver conducted between 1974 and 1984 (Fleming and Harrrington, 1991). The 312 patients participated in the randomized placebo controlled trial of the drug D-penicillamine have multiple laboratory results, which forms the first 312 cases in the original PBC file. Some baseline data values in this file differ from the original PBC file. At the time this data set was assembled, there was significantly more follow-up for many of the patients so that the time scope extended up to about fourteen years. For each patient we have a record of the time, in days, between the earlier of death or end of study ("End"), alive or dead ("Outcome"), whether they received the drug ("Drug"), day of each patient visit measured from registration ("Day"), serum bilirubin in mg/dl ("Bili") and albumin in mg/dl ("Alb"). Several other potential predictors were measured but for illustrative purposes we will restrict to these variables. Survival time, a right censored variable, is of interest. Note that each patient has multiple measurements of both bilirubin and albumin but only one time independent response. Furthermore, there are different numbers of measurements for each patient and they are taken at different times so it is not possible to use a standard multiple regression model. Figure 15 provides a typical example of a functional data set with unequally spaced observations.

The number of observations for each individual varied from one to sixteen. We use the data set after the following screening procedures: removing those patients who

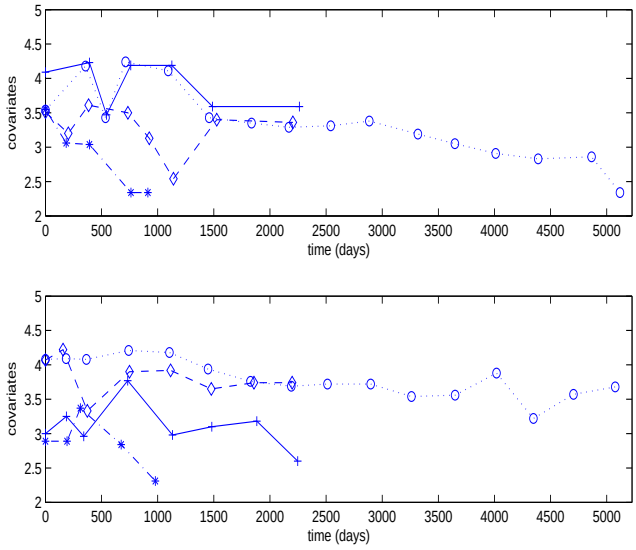

Figure 15: Some examples of the functional covariates (serum bilirubin in mg/dl) curves over time in PBC data. The above plot contains curves from control group and the bottom one from drug group.

had liver transplantation, removing those patients with fewer than four observations. Finally, of those 169 remaining patients, 65 died prior to the end of the study and 89 were from the drug group. To form the fine time lattice for natural cubic spline basis matrix in regression model, we use knots that equally divide the time interval (0, 14.115) into 5152 pieces so that each increment is actually one day in the unit of year. For the baseline hazard step function in PH model, we include 10 step intervals starting from day 0 to the last day. As discussed in Section 2.6, we use the following hyperparameters: (a_τ, b_τ) are specified as (2,2), (c_j, d_j) are specified as (2,2), both $(\mathbf{A}_i, b)$ and $(\mathbf{B}, d)$ are specified as identity matrix and $1 + rows$, where $rows$ is the number of rows of the corresponding scale matrix. Also we run the MCMC chain for 60,000 iterations and have thrown out first 20,000 burn in iterations. The results reported are average of 40 repeats.

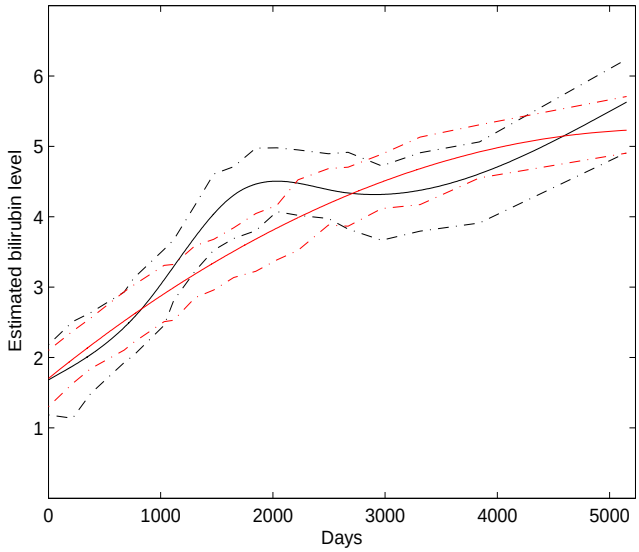

Figure 16: The estimated trajectory of bilirubin levels and the 90% credible bands.

4.6.1 Bilirubin Effect

To study effect of bilirubin levels on survival function, firstly we want to estimate the true trajectory of its process over time. Figure 16 gives the estimated average bilirubin level curve over the whole period, which shows a increasing pattern. Because increasing bilirubin level usually indicates liver failure, we see the population become sicker. However, the interpretation of the rate of increasing need careful consideration, especially toward the end of the period. Some patients having extreme high bilirubin levels might influence the estimation a lot, especially at the time period where only few patients were observed.

Although the hazard modeling utilized in above section shifts focus from survival times and survival time distribution to the hazard of failure, one can easily give the estimated survival function based on estimated hazards function. We superimpose the posterior estimates of survival curves with 5th and 95th credible intervals on the Kaplan-Meier estimates of survival functions in Figure 17. The fitness of our model

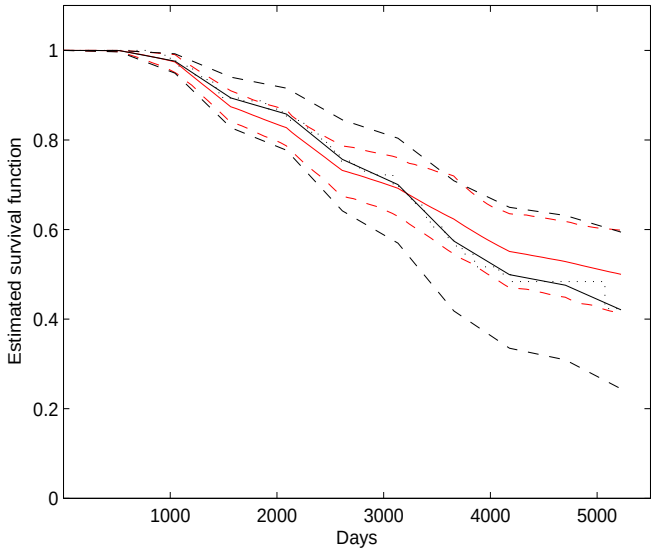

Figure 17: Survival curves: Kaplan-Meier (dotted line), our estimated survival curve based on bilirubin level (black solid line) and its 5th and 95th credible interval (black dash lines), and those estimations by Bayesian parametric model (red lines).

is satisfactory as we can see from comparing of the survival curves.

To test the null hypothesis that the level of bilirubin has no effect on survival time, we need to transform the coefficient vector $\boldsymbol{\theta}$ back to original time scale. Due to the orthogonality of the spline basis matrix, the linear model (4.23) can be written as $w_i = \boldsymbol{\beta}_i^t \mathbf{X}^t \mathbf{X}\boldsymbol{\theta} + e_i$. $\mathbf{X}\boldsymbol{\theta}$ is the converted coefficients and plotted in Figure 18, together with its 90% credible interval. We conclude that the level of bilirubin has effect on survival time based on that those credible intervals shift drastically away from zero. Liver failure is generally associated with high level of bilirubin. However, according to the converted coefficients, the time periods have slightly negative coefficients indicating lower hazards for high bilirubin levels between days 0 to 238. Similar to concerns in James (2002), this result needs to be interpreted carefully because patients with high levels in this time period will likely have high bilirubin levels at the early and late time periods also.

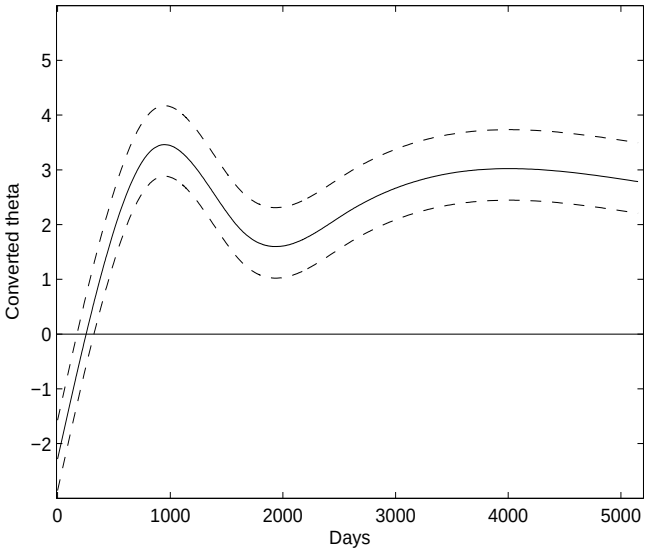

Figure 18: Converted coefficients for bilirubin levels over days. The dotted lines are 90% credible intervals.

The PBC data has two groups according to whether the patient receive the drug D-penicillamine. The drug effectiveness on survival is of interest. We apply our model on both control and drug groups respectively and compare the estimated survival curves. Figure 19 shows two superimposed survival curves based on our model and Kaplan-Meier method with 5th and 95th credible intervals for those two groups. There was no apparent improvement for those on the drug. In fact there was some evidence that the drug group may be performing worse than the control group because the estimated survival curve for drug group is a little lower than the one for control group. On the other hand, the estimated life expectancy is 4135 days for drug group and 4395 days for control group. The 90% credible bands are (3841, 4429) and (4114, 4676) for drug and control group. Overlapping of the two credible bands means that there is no significant difference of life expectancy for drug and control groups.

There are different ways for the regression of the functional predictors. For in-

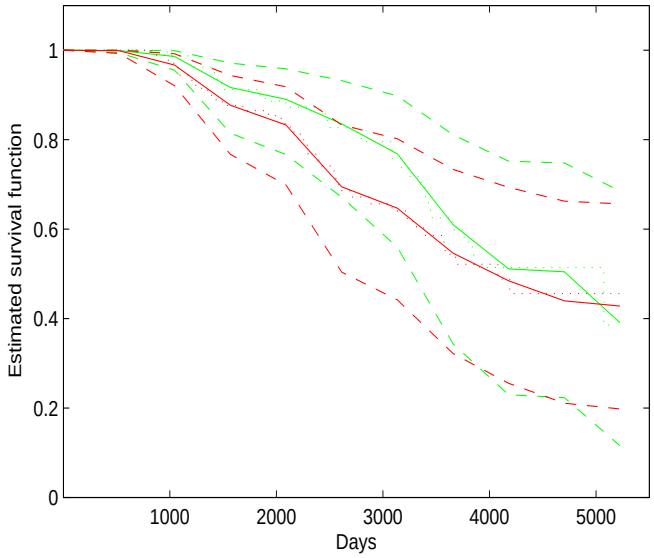

Figure 19: Survival curves for control and drug groups (green and red lines): Kaplan-Meier curve (dotted line), our estimated survival curve (solid line) and its 5th and 95th credible interval (dash lines).

stance, Section 4.4 gives an alternative one. The Bayesian parametric model has been applied to analyze bilirubin effect on PBC data. The estimated survival curves and trajectory of bilirubin levels were overlapped in Figure 17 and Figure 16 using red lines. The survival curve estimate is not as good as those from Bayesian unified hierarchical model although the 90% credible band is narrower. The estimated bilirubin trajectory can not reflect the bend shape as well as that from Bayesian unified hierarchical model. When drug effect is of interest, same conclusion as that of unified model can be reached based on the estimated survival curves for both control and drug groups (see Figure 20). However, comparing to the K-M estimates, survival curves estimates is not satisfiable, especially for the drug group. The estimated life expectancies are 4079 days for drug group and 4186 days for control group, with 90% credible intervals (3929, 4229) and (4056, 4340). For the Bayesian parametric model, the regression coefficient η can be used to test the no-effect null hypothesis for

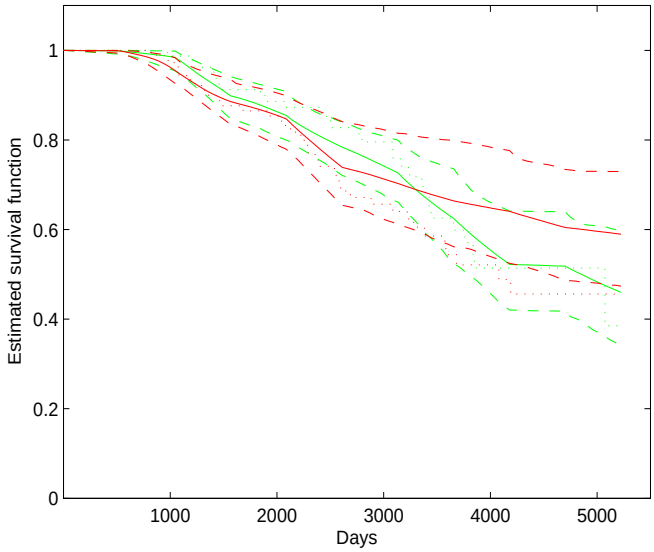

Figure 20: Estimated survival curves by Baysian parametric model for control (green lines) and drug (red lines) groups : Kaplan-Meier (dotted line), the estimated survival curve (solid line) and its 5th and 95th credible interval (dash lines).

bilirubin level. The estimated *eta* is 0.0135 and its 90% credible interval is (0.0083, 0.0191), which indicates significant bilirubin effect. Therefore, we conclude that the quadratic parametric function is not proper enough for high quality reference.

4.6.2 Bilirubin and Albumin Effects

To illustrate the extension capability of our method, we add another functional covariates, albumin, to the generalized linear model. The estimated average bilirubin and albumin curves over the whole period are plotted in Figure 21 and Figure 22 respectively. We see that the estimation for bilirubin levels are extremely similar to those in Section 4.6.1.

The average curve for albumin shows slow decreasing pattern. A healthy liver secretes albumin so the decreasing pattern indicates again that the population become sicker. Figure 23 shows the estimated survival curves using the two covariates model,

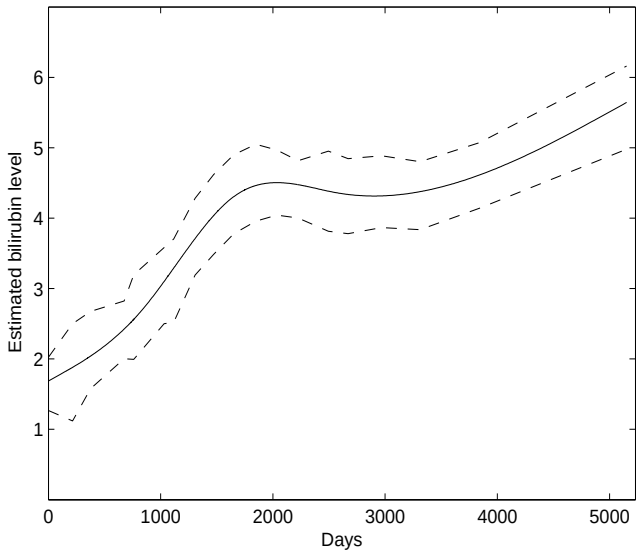

Figure 21: The estimated trajectory for bilirubin level from the model including two covariates and its 90% credible band.

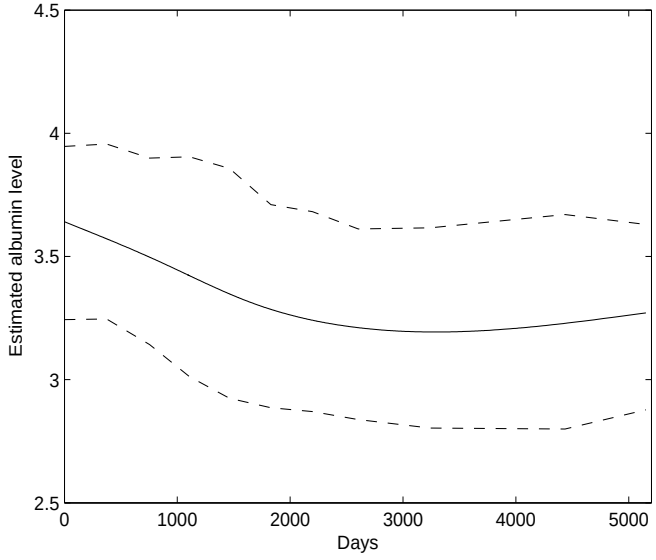

Figure 22: The estimated trajectory for albumin level from the model including two covariates and its 90% credible band.

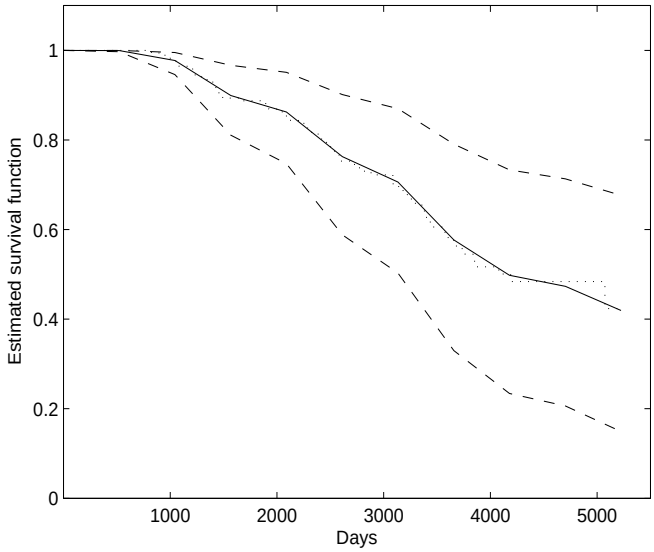

Figure 23: Estimated survival curves (solid lines) using both bilirubin and albumin as covariates. Dotted lines for Kaplan-Meier, and dash lines for 5th and 95th credible interval.

and the right plot shows the estimated survival curves separately for drug and control groups. We conclude that estimations are also satisfactory. Using the two covariates model, the estimated life expectancy is 4123 days for drug group and 4438 days for control group. The 90% credible bands are (3776, 4502) and (4087, 4798) for drug and control group. Comparisons based on survival curves (Figure 24) and life expectancies of drug and control groups reveal again that the two groups have no significantly different survival functions. These results are extremely similar to those of Section 4.5.1.

Next we compare the model with bilirubin and albumin as covariates and the one with bilirubin as covariates and study the effects of covariates on the survival time. $2log$(Bayes Factor) turns out to be 14.28, which shows very strong support to the model containing bilirubin and albumin as covariates. The converted coefficients for both covariates, bilirubin and albumin, are overlapped in Figure 25. The 90%

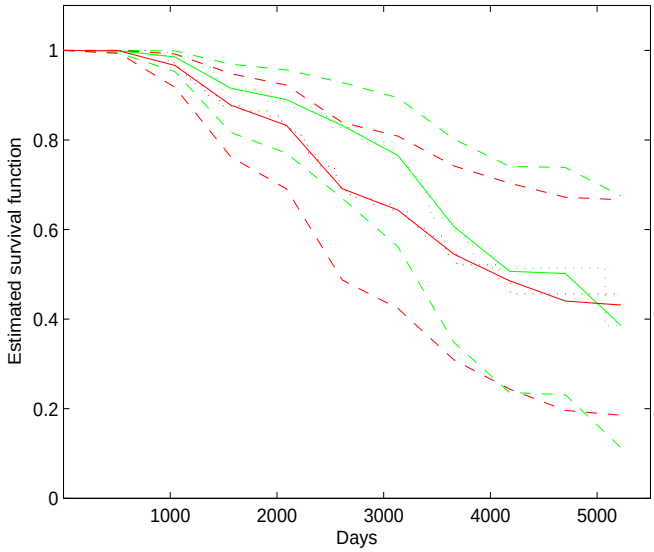

Figure 24: Control (green lines) and drug (red lines) groups estimated survival curves (solid lines) using both bilirubin and albumin as covariates. Dotted lines for Kaplan-Meier, and dash lines for 5th and 95th credible interval.

credible intervals strongly suggest that both bilirubin and albumin have effect on survival time. However, the bilirubin and albumin levels have inverted effects on survival time. High level of bilirubin generally implies liver failure, while high level of albumin indicates healthy liver. The converted coefficients for bilirubin are very similar to those in Figure 18. According to the converted coefficients, the negative coefficients means lower hazards and longer survival time for high albumin levels.

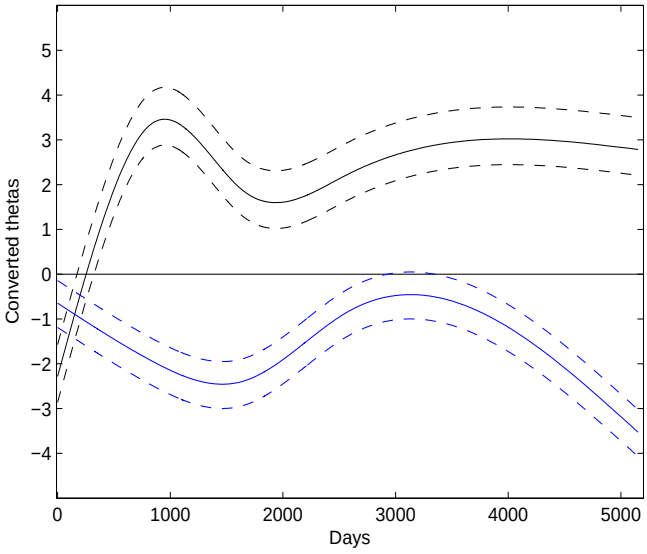

Figure 25: Converted coefficients vs days. The dotted lines are 90% credible intervals. The concave up curve is for bilirubin and the concave down one for albumin.

CHAPTER V

CONCLUSIONS

In Chapter II and IV, curve classification parts, we make comparisons in several aspects. For all comparisons, we use classification rates as criterion to judge the performance.

In irregular curve classification scenario, although all methods use natural cubic spline to smooth the curves predictors and all regression errors are small, the model set-ups engaging differently with the spline basis make classification differ. Compared with the naive version of Bayesian method, the unified model is always the winner because the regression procedure is aimed at classification by simultaneously drawing information from categorical response. The two naive methods combing either logistic regression or support vector machine have tied results most of the time. Another comparison is between existing frequentist hierarchical model, FLDA (James and Hastie, 2001) and our Bayesian spline-based methods. Our unified Bayesian method performs better than FLDA. Thus we conclude that the unified Bayesian spline-based method is appeared to be suitable to classify irregular sparse curves. The classification results shown in Chapter II have supported strongly to this point.

In spiky curve classification scenario, firstly, we compare different ways to do sparse regression in the wavelet domain. The Laplace prior (putting L^1 constraints) has been used as an alternative way, in addition to scale-mixture prior, to sparse regression in the wavelet domain. Secondly, we compare the wavelet and spline basis. The Bayesian spline based method is applied to all application examples. However, here we did not go into too many details for spline based method. The is possibility that spline based method can be drastically improved. Thirdly, we compare

difference classification technologies as well as the linear classification model. We report results for empirical Bayes methods, support vector machine, and simple logistic regression. For the data set examples, we also include both smooth curves and spiky curves. The classification results suggest that our wavelet based methods show more power when classifying spiky curves. Therefore we conclude that the unified Bayesian wavelet-based method is appeared to be suitable to classify sharp-peak curves. The classification results shown in Chapter III have supported strongly to this point.

When we apply joint hierarchical modeling for survival analysis with time-dependent covariates, the sparse characteristic of curve predictors leads us to employ splines for curve regression again. The results in Chapter IV shows that combing proportional hazard model and the generalized linear regression model provide a feasible and relatively simple way to study effects of both functional predictors and treatments on survival status.

We have witnessed that in the functional classification or survival analysis area, splines perform well for sparse smooth curves while wavelets suit high-dimensional spiky curves. Though spline-based method does not perform satisfactorily in the examples of spiky curves classification but proper tuning of the knot points or the selection of the smoothing parameters in an adaptive way may drastically improve the results. The spline-based methods for both irregular curve classification and survival analysis parts did not contain knot points and smoothing parameters selection either. These will be our future research topics. Some of simpler assumptions such as the assumption of independence across curve i could be not valid in some real applications but they are needed for the presented models. In the future study, we plan to consider to remove this strong assumption though it might be a complex problem as the exact correlation structure is unknown.

REFERENCES

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

Adam, B., Qu, Y., Davis, J.W., Ward, M.D., Clements, M.A., Cazares, L.H., Semmes, O.J., Schellhammer, P.F., Yasui, Y., Feng, Z. and Wright, G.L.Jr. (2002), "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men", *Cancer Research*, 62, 3609–3614.

Alter, O., Brown, P.O. and Boststein, D. (2000), "Singular value decomposition for genome-wide expression data processing and modeling", in *Proceedings of the National Academy of Sciences*, U.S.A. 97, 10101–10106.

Anderson, T.W. (1951), "Estimating linear restrictions on regression coefficients for multivariate normal distributions", *The Annals of Mathematical Statistics*, 22, 327–351.

Bachrach, L.K., Hastie, T.J., Wang, M.-C., Narasimhan, B. and Marcus, R. (1999),"Bone mineral acquisition in healthy Asian, Hispanic, Black, and Caucasian youth: a longitudinal study", *Journal of Clinical Endocrinology and Metabolism*, 84(12), 4702–4712.

Berry, S.M., Carroll, R.J. and Ruppert, D. (2002), "Bayesian smoothing and regression splines for measurement error problems", *Journal of the American Statistical Association*, 97, 160–169.

Brown, E.R., and Ibrahim, J.G. (2003), "A Bayesian semiparametric joint hierarchical model for longitudinal and survival data", *Biometrics*, 59, 221–228.

Carey, J., Liedo, P., M̀uller, H.G., Wang, J.L. and Chiou, J.M.(1998), "Relationship of age patterns of fecundity to mortality, longevity and lifetime reproduction in a large cohort of Mediterranean", *Journal of Gerontology*, 53A, 245–251.

Carlin, B.P. and Louis T.A., (1996), *Bayes and Empirical Bayes Methods for Data Analysis*, London: Chapman and Hall.

Carroll, R.J., Maca, J.D., and Ruppert, D. (1999), "Nonparametric regression with errors in covariates", *Biometrika*, 86, 541–554.

Chakraborty, S., Ghosh, M., Maiti, T. (2005), "Hierarchical Bayesian neural netwroks fo Bivariate binary data: an application to prostate cancer study", To appear, *Statistics in Medicine*.

Chaloner, K. and Brant, R. (1988), "A Bayesian approach to outlier detection and residual analysis", *Biometrika*, 75, 651–660.

Chib, S. (1995),"Mariginal likelihood from the Gibbs output", *Journal of the American Statistical Association*, 90, 1313–1321.

Chib, S. and Jeliazkov, I. (2001), "Mariginal likelihood from the Metropolis-Hastings output", *Journal of the American Statistical Association*, 96, 270–281.

Clyde, M., Parmigiani, G. and Vidakovic, B. (1998), "Multiple shrinkage and subset selection in wavelets", *Biometrika*, 85, 391–401.

Clyde, M. and George, E.I. (2000), "Flexible empirical Bayes estimation for wavelets",*Journal of the Royal Statistical Society, Series B*, 62(4), 681–698.

Conrads, T.P., Zhou, M., Petricoin, E.F. III, Liotta, L. and Veenstra, T. D. (2003), "Cancer diagnosis using proteomics patterns", *Expert Review of Molecular Diagnostics*, 3(4), 411–420.

Daniels, M.J. and Kass, R.E. (1999) "Nonconjugate Bayesian estimation of covariance matrices and its use in hierchical models", *Journal of the American Statistical Association*, 94, 1254–1263.

Daniels, M.J. and Kass, R.E. (2001) "Shrinkage estimators for covariance matrices", *Biometrics*, 57, 1173–1184.

Daubechies, I. (1992), *Ten lectures on wavelets*, Philadelphia, PA: Society for Industrial and Applied Mathematics.

de Boor, C. (1978), *A practical guide to splines*, New York: Springer-Verlag.

DeCanditis, D. and Vidakovic, B. (2004), "Wavelet Bayesian block shrinkage via mixture of normal-inverse gamma priors", *Journal of Computational and Graphical Statistics* 13, 383–398.

DeGruttola, V. and Tu, X.M. (1994), "Modeling progression of CD4-lymphocyte count and its relationship to survival time", *Biometrics*, 50, 1003–1014.

Delouille, V. (2002), *Nonparametric stochastic regression using design-adapted wavelets*, PhD dissertation, Institute of Statistics, Catholic University of Louvain, Belgium, 2002.

Denison, D., Holmes, C., Mallick, B. and Smith, A. (2002), *Bayesian methods for nonlinear classification and regression*, New York: Wiley.

Denison, D., Mallick, B. and Smith, A. (1998), "Bayesian MARS", *Statistics and Computing*, 8, 337–346.

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

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

Donoho, D.L. and Johnstone, I.M. (1995), "Adapting to unknown smoothness via wavelet shrinkage", *Journal of the American Statistical Association*, 90, 1200–1224.

Faucett, C.J. and Thomas, D.C. (1996), "Simultaneously modeling censored survival data and repeatedly measured covariates: a Gibbs sampling approach", *Statistics in Medicine*, 15, 1663–1685.

Fleming, T.R. and Harrington, D.P. (1991), "Counting Processes and Survival Analysis", New York: Wiley.

Friedman, J.H. (1989), "Regularized discriminant analysis", *Journal of the American Statistical Association*, 84, 165–175.

Friedman, J.H. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *Annals of Statistics*, 19, 1–141.

Friedman, J.H. (1993), "Fast MARS", Technical Report, Department of Statistics, Stanford University, Stanford, CA.

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

Gelfand, A. E. (1996), "Model determination using sampling-based methods", in *Markov Chain Monte Carlo in Practice*, eds. Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., London: Chapman & Hall, pp. 145–162.

Gelman, A., Carlin, J. B.,Stern, H. S., and Rubin, D. B. (2003), *Bayesian Data Analysis*, London: CRC Press.

Gelman, A., (2004), "Parameterization and Bayesian modeling", *Journal of the American Statistical Association*, 99, 537–545.

George, E. and McCulloch, R. (1993), "Variable selection via Gibbs sampling", *Journal of the American Statistical Association*, 88, 881–889.

Ghosh, M., Maiti, T., Kim, D., Chakraborty, S. and Tewari, A. (2004), "Bayesian neural network modeling in prostate cancer detection", *Journal of the American Statistical Association*, 99, 601–608.

Gilsanz V., Skaggs D.L., Kovanlikaya A., Sayre J., Loro M.L., Kaufman F., Korenman S.G. (1998), "Differential effect of race on the axial and appendicular skeletons of children", *Journal of Clinical Endocrinology and Metabolism*, 83, 1420–1427.

Green, P.J. and Silverman, B.W. (1994), *Nonparametric regression and generalized linear models: a roughness penelty approach*, London: Chapman and Hall.

Guo, X. and Carlin, B.P. (2004), "Separate and joint modeling of longitudinal and event time data using standatd computer packages", *The American Statistician*, 58(1), 16–24.

Hastie, T.J., Buja, A. and Tibshirani, R.J. (1995), "Penalized discriminant analysis", *Annals of Statistics*, 23, 73–102.

Hastie, T.J. and Tibshirani, R.J. (1996) "Discriminat analysis by Gaussian mixtures", *Journal of the Royal Statistical Society, Series B*, 58, 155–176.

Hingorani, S.R., Emanuel, F., Petricoin, E.F. III, Maitra, A., Rajapakse, V., King, C., Jacobetz, M.A., Ross, S., Conrads, T.P., Veenstra, T.D., Hitt, B.A., Kawaguchi,

Y., Zhou, Y., Johann, D., Liotta, L.A., Crawford, H.C., Putt, M.E., Jacks, T., Konieczny, S.F., Wright, C.E., Hruban, R.E., Lowry, A.M. and Tuveson D.A. (2003), "Preinvasive and invasive ductal pancreatic cancer and its early detection in the mouse", *Cancer Cell*, 10, 6–21.

Hogan, J.W. and Laird, N.M. (1997), "Mixture models for the joint distribution or repearted measures and event times", *Statistics in Medicine*, 16, 239–257.

Hox, J. (2002), *Multilevel Analysis: Techniques and Applications*, Mahwah, NJ: Lawrence Erlbaum Associates.

Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2004), "Bayesian methods for joint modeling of longitudinal and survival data with applications to cancer vaccine trials", *Statistica Sinica*, 14, 863–883.

Ibrahim, J.G., Chen, M.-H. and Sinha, D. (2001), *Bayesian survival analysis*, New York: Springer-Verlag.

James, G.M. and Hastie, T.J. (2001), "Functional linear discriminant analysis for irregularly sampled curves", *Journal of the Royal Statistical Society, Series B*, 63, 533–550.

James, G.M. (2002), "Generalized linear models with functional predictors",*Journal of the Royal Statistical Society, Series B*, 64, 411–432.

Johnstone, I. and Silverman, B. (1997), "Wavelet threshhold estimators for data with correlated noise", *Journal of the Royal Statistical Society, Series B*, 59, 319–351.

Johann, D.J., Jr., McGuiganb, M.D., Tomovb, S., Fusaroa, V.A., Rossa, S., Conradsc, T.P., Veenstrac, T.D., Fishmand, D.A., Whiteleye, G.R., Petricoinf E.F. and Liottaa, L.A. (2003), "Novel approaches to visualization and data mining reveals

diagnostic information in the low amplitude region of serum mass spectra from ovarian cancer patients", *Disease Markers*, 19, 197 207.

Kass,R.E. and Raftery,A.E. (1995), "Bayes factors", *Journal of the American Statistical Association*, 90, 773–795.

Kass, R.E., Ventura, V. and Cai, C. (2003), "Statistical smoothing of neuronal data", *Network: Computation in Neural Systems*, 14, 5–15.

Keogh, E. and Folias, T. (2002), "The UCR time series data mining archive", *http://www.cs.ucr.edu/ eamonn/TSDMA/index.html*, Riverside, CA, University of California, Computer Science and Engineering Department.

Keogh, E. and Lonardi, S. and Ratanamahatana, C. (2004), "Towards parameter-free data mining", in *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 206–215.

Kreft, I., and De Leeuw, J. (1998), *Introducing Multilevel Modeling*, London: Sage.

McCormick D.P., Ponder S.W., Fawcett H.D., and Palmer J.L. (1991), "Spinal bone mineral density in 335 normal and obese children and adolescents: evidence for ethnic and sex differences",*Journal of Bone Mineral Research*, 5, 507–513.

McFadden, D.L. (1973), "Conditional logit analysis of qualitative choice behavior", in *Frontiers in Economics*, eds. Zarembka P., New York: Academic Press, pp. 669–679.

Müller, H.G. and Stadtmüller, U. (2005), "Generalized functional linear models", *Annals of Statistics*, 33(2), 774–805.

Nelson D.A., Simpson P.M., Johnson C.C., Barondess D.A., and Kleerekoper M. (1997), "The accumulation of whole body skeletal mass in third- and fourth-grade

children: effects of age, gender, ethnicity, and body composition", *Bone*, 20, 73–78.

Parker, J.R. (2002), "Scientific curve classification by combining simple algorithms", in *1st IEEE International Conference on Cognitive Informatics*, pp. 222–228.

Patel D.N., Pettifor J.M., Becker P.J., Grieve C. and Leschner K. (1992), "The effect of ethnic group on appendicular bone mass in children", *Journal Bone Mineral Research*, 7, 263–272.

Petricoin,E.F., Rajapaske, V., Herman, E.H., Arekani, A.M., Ross, S., Johann, D., Knapton, A., Zhang, J., Hitt, B.A., Conrads, T.P., Veenstra, T.D., Liotta, L.A. and Sistare, F.D. (2004), "Toxicoproteomics: serum proteomics pattern diagnostics for early detection of drug induced cardiac toxicities and cardioprotection", *Toxicologic Pathology*, 32(Sippl. 1), 1–9.

Pfeiffer, R., Bura, E., Smith, A. and Rutter, J.L. (2002), "Two approaches to mutation detection based on functional data", *Statistics in Medicine*, 21(22), 3447–3464.

Ramsay, J.O. and Silverman, B.W. (1997), *Functional data analysis*, New York: Springer-Verlag.

Ramsay, J.O. and Silverman, B.W. (2002), *Applied Functional data analysis: Methods and Case Studies*, New York: Springer-Verlag.

Ratanamahatana, C. A. and Keogh, E. (2004a), "Everything you know about dynamic time warping is wrong", *Third Workshop on Mining Temporal and Sequential Data, in conjunction with the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2004)*, Seattle, WA.

Ratanamahatana, C. A. and Keogh, E. (2004b), "Making time-series classification more accurate using learned constraints", in *Proceedings of SDM Internatoinal Conference*, pp. 11–22.

Raudenbush, S. W., and Bryk, A. S. (2002), *Hierarchical Linear Models*, Thousand Oaks, CA: Sage.

Ruppert, D., Wand, M.P. and Carroll, R.J. (2003), *Semiparametric regression*, Cambridge: Cambridge University Press.

Silverman, B.W. and Johnstone, I. (2005), "Empirical Bayes selectionn of wavelet thresholds", *Annals of Statistics*, 33(4), 1700–1752.

Smith, M. and Kohn, R. (1996), "Nonparametric regression using Bayesian variable selection", *Journal of Econometrics*, 75, 317–344.

Snijders, T. A. B., and Bosker, R. J. (1999), *Multilevel Analysis*, London: Sage.

Sweldens, W. (1997), "The lifting scheme: a construction of second generation wavelets", *SIAM Journal on Mathematical Analysis*, 29(2), 511–546.

Vidakovic, B. (1998), "Non-linear wavelet shrinkage with Bayes rules and Bayes factors", *Journal of the American Statistical Association*, 93, 173–179.

Wagner, M., Naik, D.N., Pothen, A., Kasukurti, S., Devineni, R.R., Adam, B., Semmes, O.J., and Wright, G.L. Jr (2004), "Computational protein biomarker prediction: a case study for prostate cancer", *BMC Bioinformatics*, 5:26.

Wang M.-C., Aguirre M., Bhudhikanok G.S., Kendall C.G., Kirsch S., Marcus R. and Bachrach L.K. (1997), "Bone mass and hip axis length in healthy Asian, black, Hispanic, and white American youths", *Journal of Bone Mineral Research*, 12, 1922–1935.

Wang, X., Ray, S. and Mallick, B.K. (2006), "Bayesian curve classification using wavelets", submitted.

Wang, Y. and Taylor, J.M.G. (2000), "Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome", *Journal of the American Statistical Association*, 96, 895–905.

Wulfsohn, M.S. and Tsiatis, A.A. (1997), "A joint model for survival and longitudinal data measured with error", *Biometrics*, 53, 330–339.

Zellner, A. (1962), "An efficient method of estimating seemingly unrelated regressions and tests of aggregation bias", *Journal of the American Statistical Association*, 57, 348–368.

APPENDIX

BAYES FACTOR CALCULATION

Let $\boldsymbol{\phi}$ denote the parameters in the model $(\boldsymbol{\Omega}, \boldsymbol{\Sigma}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}, \tau^2, \mathbf{V}, \boldsymbol{\lambda})$, the logarithm marginal density is

$$log\hat{m}(\mathbf{D}) = logf(\mathbf{D} \mid \boldsymbol{\phi}^*) + log\pi(\boldsymbol{\phi}^*) - log\hat{\pi}(\boldsymbol{\phi}^* \mid \mathbf{D}).$$

Although $\boldsymbol{\phi}^*$ can be any $\boldsymbol{\phi}$ in its support $\boldsymbol{\Phi}$, the density is likely to be more accurately estimated at a high density point. We choose to use posterior mean provided that there is no concern that it is a low density point. Because there are several parameters with conjugate posterior distributions and one latent variable w_i in our model, we use Chib's general algorithm for arbitrary number of blocks. Rewrite the posterior density at the selected point as

$$\begin{aligned}
\pi(\boldsymbol{\phi}^* \mid \mathbf{D}) &= \pi(\boldsymbol{\Omega}^* \mid \mathbf{D}) \prod_{i=1}^{n} \left[\pi(\boldsymbol{\Sigma}_i^* \mid \mathbf{D}, \boldsymbol{\Omega}^*) \pi(\boldsymbol{\beta}_i^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}_i^*) \right] \\
&\quad \times \pi(\boldsymbol{\theta}^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*) \pi(\tau^{2*} \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*) \\
&\quad \times \pi(\mathbf{V}^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \tau^{2*}) \pi(\boldsymbol{\lambda}^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \tau^{2*}, \mathbf{V}^*).
\end{aligned}$$

It should be clear that the normalizing constants of all densities must be included in the integration for the above decomposition to be valid. The first term is the marginal ordinate that can be estimated from the full Gibbs run, by taking the average of the full conditional density with the posterior draws of $\boldsymbol{\beta}$, as

$$\hat{\pi}(\boldsymbol{\Omega}^* \mid \mathbf{D}) = G^{-1} \sum_{g=1}^{G} \pi(\boldsymbol{\Omega}^*, \mid \boldsymbol{\beta}^{(g)}).$$

To estimate the rest of those terms in $\pi(\boldsymbol{\phi}^* \mid \mathbf{D})$, we conduct several reduced complete conditional Gibbs runs. Respectively, we illustrate the estimation as

$$\hat{\pi}(\boldsymbol{\Sigma}_i^* \mid \mathbf{D}, \boldsymbol{\Omega}^*) = J^{-1} \sum_{j=1}^{J} \pi(\boldsymbol{\Sigma}_i^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\beta}^{(j)}, \boldsymbol{\theta}^{(j)}, \tau^{2(j)}, \mathbf{V}^{(j)}, \mathbf{w}^{(j)}, \boldsymbol{\lambda}^{(j)}),$$

$$\hat{\pi}(\boldsymbol{\beta}_i^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}_i^*) = J^{-1} \sum_{j=1}^{J} \pi(\boldsymbol{\beta}_i^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}_i^*, \boldsymbol{\theta}^{(j)}, \tau^{2(j)}, \mathbf{V}^{(j)}, w_i^{(j)}, \boldsymbol{\lambda}^{(j)}),$$

$$\hat{\pi}(\boldsymbol{\theta}^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*) = J^{-1} \sum_{j=1}^{J} \pi(\boldsymbol{\theta}^* \mid \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}_i^*, \boldsymbol{\beta}^*, \tau^{2(j)}, \mathbf{V}^{(j)}, \mathbf{w}^{(j)}, \boldsymbol{\lambda}^{(j)}),$$

$$\hat{\pi}(\tau^{2*} \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*) = J^{-1} \sum_{j=1}^{J} \pi(\tau^{2*} \mid \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}_i^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \mathbf{V}^{(j)}, \mathbf{w}^{(j)}, \boldsymbol{\lambda}^{(j)}),$$

$$\hat{\pi}(\mathbf{V}^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \tau^{2*}) = \pi(\mathbf{V}^* \mid \boldsymbol{\theta}^*, \tau^{2*}),$$

$$\hat{\pi}(\boldsymbol{\lambda}^* \mid \mathbf{D}, \boldsymbol{\Omega}^*, \boldsymbol{\Sigma}^*, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \tau^{2*}, \mathbf{V}^*) = J^{-1} \sum_{j=1}^{J} \pi(\boldsymbol{\lambda} \mid \mathbf{D}, \mathbf{w}^{(j)}).$$

The draws $\{\boldsymbol{\beta}_i^{(j)}, \boldsymbol{\theta}^{(j)}, \tau^{2(j)}, \mathbf{V}^{(j)}, \mathbf{w}^{(j)}, \boldsymbol{\lambda}^{(j)}\}$ is from the reduced complete conditional Gibbs runs, which is same as the full complete conditional Gibbs run except that it should exclude draws for $\boldsymbol{\Omega}$ and use $\boldsymbol{\Omega}^*$ everywhere. Similarly, the draws $\{\boldsymbol{\theta}^{(j)}, \tau^{2(j)}, \mathbf{V}^{(j)}, w_i^{(j)}, \boldsymbol{\lambda}^{(j)}\}$ is from the reduced complete conditional Gibbs run is same as the full run except that it exclude draws from $\boldsymbol{\Omega}, \boldsymbol{\Sigma}_i$ and use $\boldsymbol{\Omega}^*, \boldsymbol{\Sigma}_i^*$ everywhere. Other additional draws are collected similarly. Finally, the additional J iterations with densities $\pi(\boldsymbol{\lambda} \mid \mathbf{D}, \mathbf{w})$ and $\pi(\mathbf{w} \mid \mathbf{D}, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \tau^{2*})$ produce draws $\{\mathbf{w}^{(j)}\}$ from $\pi(\mathbf{w} \mid \mathbf{D}, \boldsymbol{\beta}^*, \boldsymbol{\theta}^*, \tau^{2*})$. Although this procedure leads to an increase in the number of iterations, it is worth of pointing out that it does not require new programming. Note that there is no need of the reduced conditional run when the complete conditional density is solely related to parameters in previous run, such as $\pi(\mathbf{V}^* \mid \boldsymbol{\theta}^*, \tau^{2*})$.

The marginal density for multiple functional covariates model can be computed in a very similar way. We can use these marginal density calculations to obtain the Bayes factor for model comparison.

VITA

Xiaohui Wang was born in Chifeng, China. She received a Bachelor of Engineering degree in mechanical engineering from University of Science and Technology at Beijing in 1994. She received a Master of Science degree in material management from Beijing Jiaotong University (formerly as Northern Jiaotong University) in 1998 and Master of Science in statistics from Texas A&M University in College Station, Texas, under the direction of Dr. Naisyin Wang in 2002. She continued her studies under the direction of Dr. Bani K. Mallick, and received a Doctor of Philosophy degree from Texas A&M University in August 2006. Her permanent address is Tianyi High School New Jiashu Building, Unit 3, Floor 3, Ningchen, Neimeng, 024200, China.