BAYESIAN CURVE FITTING WITH ROUGHNESS PENALTY PRIOR DISTRIBUTIONS

by

MD MUSHFIQUR RAHMAN
M.Sc., University of Dhaka, 1996

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

July 2005

Abstract

In statistical research with populations having a multilevel structure, hierarchical models can play significant roles. The use of the Bayesian approach to hierarchical models has numerous advantages over the classical approach. For example, a spline with roughness penalties can easily be expressed as a hierarchical model and the model parameters can be estimated by the Bayesian techniques. Splines are sometimes useful to express the rapid fluctuating relationship between response and the covariate. In smoothing spline problems, usually one smoothing parameter (variance component in Bayesian context) is considered for the whole data set. But to deal with rapidly fluctuating or wiggly data sets, it is more logical to consider different smoothing parameters at different knot points in order to find more efficient estimates of the the regression functions under consideration. In this study, we have proposed the roughness penalty prior distribution considering local variance components at different knot points and call it Prior 2. Prior 2 is compared with Prior 1, where a single global variance component is considered for the whole data set, and with Prior 3, where no roughness penalty terms are considered (i.e., the parameters at different knot points are assumed independent). Performance of the proposed prior distributions are checked for three different data sets of different curvature. Similar performance of Prior 1 and Prior 2 is observed for all three data sets under the assumption of piecewise linear spline. The application has been extended to the case of natural cubic spline, where the modification of Prior 1 and Prior 3 are straightforward. However, for Prior 2, the modification becomes very tedious. We have proposed an approximate roughness penalty matrix for Prior 2.

Parameters corresponding to the smoothing splines are estimated using MCMC techniques. We carefully compare the inferential procedures in simulation studies and illustrate them for two data sets. Similarity among the curves produced by Prior 1 and Prior 2 are observed, and they are much smoother than the curve estimated by Prior 3 for both piecewise linear and natural cubic splines. Therefore, in the context of Bayesian curve fitting, both local and global roughness penalty priors produce equally smooth curves in dealing with wiggly data.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to warmly acknowledge the guidance and support of my supervisor, Dr. Paul Gustafson throughout my studies. I would also like to give many thanks to Professor Nancy E. Heckman, second reader of this thesis, for her valuable suggestions.

Special thanks to those faculty members who were amicable with me in their courses and my TA work. I appreciate the co-operation of Christine Graham as a Graduate Co-ordinator of this department.

Lastly, I am grateful to all the graduate students in this department especially to Jafar, Mikhail, Shahadut, Lawrence and Yiping for their warm friendship and constant encouragement throughout my studies.

MD. MUSHFIQUR RAHMAN

The University of British Columbia
July 2005

To my Parents.

Chapter 1

Bayesian Hierarchical Models

1.1 Introduction

Hierarchical models are at the heart of research in many real life situations where the population has hierarchical structure. The basic idea of hierarchical model (also known as variance component model, multilevel model, random effects model, or growth curve model) is to organize the lowest-level units into a hierarchy of successive higher-level units. For example, patients are in hospitals can be considered as level 1, hospitals are in cities can be considered as level 2, cities are in states can be considered as level 3 etc. We can then describe outcomes for an individual patient as a pooled effect for the individual patient, for her/his hospital, for the city and for the state. Each of these effects can often be regarded as one of an exchangeable collection of effects (e.g., all hospital-level effects) drawn from a distribution described by a variance component. There may also be regression coefficients at some or all of the levels.

Once a hierarchical model is specified, inferences can be drawn for the population means at any level (hospital, city, etc.) from available data in both frequentist and Bayesian approaches. In the Bayesian perspective the parameter estimates are referred to as the posterior means and variances, and in the frequentist perspective the parameter estimates are referred to as the *Best Linear Unbiased Predictors (BLUPs)* (Robinson, 1991). Both

approaches often have better properties than simple sample-based estimators using data only from the unit in question.

For multistage data structure hierarchical models are a good choice for the following reasons (Hossain, 2003):

(a) Hierarchical models permit the direct framing of the theories about the effect of structural change at each of the different levels of the hierarchy.

(b) They provide accurate adjustments to the uncertainty assessment based on simple random sampling when the data are gathered in a hierarchical fashion in the presence of strong intra-cluster correlations.

(c) Use of non-hierarchical models is inappropriate for hierarchical data because with a few parameters they usually can not fit the data accurately.

This chapter is organized as follows. In section 1.2 we give some real life examples where hierarchical models are often used. Section 1.3 is for the Bayesian approach to data analysis. There we discuss the Bayes' rules and the estimation procedures in Bayesian techniques. The Bayesian treatment for the hierarchical model is presented in section 1.4 where we discuss the computation techniques of posterior distributions. The procedures of drawing samples from the posterior distributions through Markov Chain Monte Carlo (MCMC) techniques are described in section 1.5. We end with a brief discussion in section 1.6.

1.2 Hierarchical Models in Real Life

Many authors worked on the data with hierarchical fashion and they explained the concept of hierarchical modelling with many nice examples. A few of them are as follows:

(i) Smoothing spline

In many practical situations, the relationship between a response and a covariate shows a rapid fluctuating curve. Examples include growth curves of organisms, learning

curves, and stock market trends. Then it is better to express the curve with the help of a piece-wise polynomial of certain degree (called a spline) over each subinterval of the range of the predictor considered. Nonparametric regression using linear and cubic splines is an attractive, flexible and widely applicable approach to curve estimation. Besides many existing classical approaches of smoothing splines a lot of studies concerning Bayesian curve fitting have been conducted. A spline with roughness penalties can easily be expressed as a hierarchical model and the model parameters can be estimated by Bayesian techniques. Denison *et al.* (1998), Smith and Kohn (1997), Altman and Casella (1995) and Silverman (1985) are some nice pieces of work for the Bayesian approach to smoothing splines.

(ii) Small Area Estimation

Hierarchical models can be used in small area estimation. Gosh and Rao (1994) introduced several techniques including Empirical Bayes and Hierarchical Bayes procedures for small area estimation. The term *small area* is defined as a small geographical area, such as a county, a municipality of a census division, or a small subpopulation such as a specific age-sex-race group of people within a large geographical area. The usual direct survey estimators for the totals and means for large domains from a small area, based on data only from the sample units in the area, are likely to yield unacceptably large standard errors due to the unduly small size of the sample in the area (Gosh and Rao, 1994). Small area estimation can be done efficiently by assuming a hierarchical small area model and then applying the Bayesian approach to hierarchical models. Datta and Ghosh (1991) applied the hierarchical Bayes approach to estimation of small area means under general mixed linear models and also discussed the computational aspects.

(iii) Random effects meta-analysis

Meta-analysis is defined to be the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings. In recent

years, random effects meta-analysis is widely used by researchers in medical and social sciences for many of its advantages. Interested readers can read Whitehead (2002). Similar to the random effects meta-analysis, in the Bayesian approach, all the model parameters are treated as random variables and hence one can account for the uncertainty of all relevant sources of variability in the model. Let us consider a meta-analysis with the log odds ratios from several studies. A full Bayesian random effects model for meta-analysis includes a three level hierarchy and can be written as follows:

Level (i) Observed individual log odds ratio given the true log odds ratio and the variance component follows a population distribution (usually normal).

Level (ii) True log odds ratio given the overall population mean log odds ratio and the variance component follows a population distribution (usually normal).

Level (iii) The overall mean log odds ratio and the variance component parameters are treated as independent random variables with some population distributions.

(iv) Repeated measures data

In longitudinal studies, we have repeated observations on the same individual under study at different occasions; for example, the monthly observations of the blood glucose levels of diabetic patients. This type of data can be arranged in a hierarchical model with two levels, where the measurement occasions (considered as level 1 units) are nested within individuals (considered as level 2 units).

(v) Hierarchical models in Biostatistics

In our previous patient-hospital and Bayesian meta-analysis examples we explained how the hierarchical models can be applied in biostatistics. Besides these, in some studies of offspring of animals and human beings we observe the data having hierarchical or clustered structure, where offspring are grouped within families. Offspring from the same parents tend to be more alike in their physical and mental characteristics than individuals chosen at random from the population at large. Thus, offspring

may be the level 1 units in a 2-level structure where the level 2 units are the families. Details are found in Goldstain *et al.* (1994).

1.3 Bayesian Approach to Data Analysis

Besides the observed data, sometimes, it is likely that the researcher has some knowledge about the parameter vector $\boldsymbol{\theta}$. Bayesian approach incorporates this information to the analysis through a density $\pi(\boldsymbol{\theta})$ with the observed data in the estimation process. The process of Bayesian data analysis can be divided into the following three steps (Gelman *et al.*, 1995):

1. Setting up a *full probability model* - a joint probability distribution for all observable and unobservable quantities in a problem (i.e., the likelihood function times the prior distributions).
2. Conditioning on observed data, calculating and interpreting the appropriate *posterior distribution* - the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data.
3. Evaluating the fit of the model and the implications of the resulting posterior distribution. If necessary, one can alter or expand the model and repeat these three steps.

1.3.1 Bayes' Rule

Bayesian inference about a parameter vector $\boldsymbol{\theta}$ is made in terms of probability statement. This probability statement is conditional on the observed value y and is called the posterior probability of $\boldsymbol{\theta}$ given y. Notationally we can write $\pi(\boldsymbol{\theta}|y)$. To calculate the posterior density we need two ingredients: $\pi(\boldsymbol{\theta})$, the *prior distribution* of $\boldsymbol{\theta}$, and $\pi(y|\boldsymbol{\theta})$, the *sampling distribution* of the observed data y which is also called the likelihood function of $\boldsymbol{\theta}$. The joint probability distribution of $\boldsymbol{\theta}$ and y can then be defined as

$$\pi(\boldsymbol{\theta}, y) = \pi(\boldsymbol{\theta})\pi(y|\boldsymbol{\theta}). \tag{1.1}$$

Applying the Bayes' rule for conditional probability we get the *posterior* density of $\boldsymbol{theta}$ as follows:

$$\begin{aligned} \pi(\boldsymbol{\theta}|y) &= \frac{\pi(\boldsymbol{\theta}, y)}{\pi(y)} \\ &= \frac{\pi(\boldsymbol{\theta})\pi(y|\boldsymbol{\theta})}{\pi(y)}, \end{aligned} \tag{1.2}$$

where

$$\pi(y) = \begin{cases} \sum_{\theta} \pi(\boldsymbol{\theta})\pi(y|\boldsymbol{\theta}), & \text{for discrete } \boldsymbol{\theta} \text{ or} \\ \int_{\theta} \pi(\boldsymbol{\theta})\pi(y|\boldsymbol{\theta})d\boldsymbol{\theta}, & \text{for continuous } \boldsymbol{\theta}, \end{cases}$$

is called the marginal distribution of y. Since $\pi(y)$ does not depend on $\boldsymbol{\theta}$ and, with fixed y can thus be considered as constant (also called normalization constant), an equivalent expression of (1.2) is as follows:

$$\pi(\boldsymbol{\theta}|y) \propto \pi(\boldsymbol{\theta})\pi(y|\boldsymbol{\theta}). \tag{1.3}$$

A Bayes' estimator of $\boldsymbol{\theta}$ is the mean of the posterior distribution of $\boldsymbol{\theta}$, $E(\boldsymbol{\theta}|y) = \int_{\boldsymbol{\theta}} \boldsymbol{\theta}\pi(\boldsymbol{\theta}|y)d\boldsymbol{\theta}$.

1.4 The Bayesian treatment of the hierarchical model

Let us consider the following example of the hierarchical model: in a study of the effectiveness of chemotherapy, in cancer patients, let us have the observed data y_{ij} from the ith individual and jth hospital with survival probability θ_j. Hence, it might be reasonable to expect that estimates of the θ_j's, which represent a sample of hospitals, should be related to each other. We can incorporate this similarity in a natural way if we use a prior distribution in which the θ_j's are viewed as a sample from a common population distribution. In such case the observed data y_{ij} can be used to estimate aspects of the population distribution of the θ_j's even though the values of θ_j's are not observed. It is natural to model such a phenomenon hierarchically by assuming that each of the parameters, the θ_j's, form an independent sample from a population distribution governed by some unknown parameter τ^2. Thus we write,

$$\pi(\boldsymbol{\theta}|\tau^2) = \prod_{j=1}^{p} \pi(\theta_j|\tau^2). \tag{1.4}$$

The main hierarchical part of these models is that τ^2 is not known and thus has its own prior distribution, $\pi(\tau^2)$. The joint prior distribution can be written as

$$\pi(\tau^2, \boldsymbol{\theta}) = \pi(\tau^2)\pi(\boldsymbol{\theta}|\tau^2),$$

and the joint posterior distribution is

$$\begin{aligned} \pi(\tau^2, \boldsymbol{\theta}|y) &\propto \pi(\tau^2, \boldsymbol{\theta})\pi(y|\tau^2, \boldsymbol{\theta}) \\ &= \pi(\tau^2)\pi(\boldsymbol{\theta}|\tau^2)\pi(y|\theta), \end{aligned}$$

since $\pi(y|\tau^2, \boldsymbol{\theta})$ depends only on $\boldsymbol{\theta}$; the hyperparameter τ^2 affects y only through $\boldsymbol{\theta}$. Now the problem is how to choose the hyperprior distribution for τ^2. If little is known about τ^2, we can assign a noninformative prior distribution (we will discuss more in chapter 3), but we must be careful when using an improper prior density to check that the resulting posterior distribution is proper, and we should assess whether our conclusions are sensitive to this simplifying assumption. After writing the joint posterior density $\pi(\tau^2, \boldsymbol{\theta}|y)$ in terms of the likelihood function and the joint prior distribution, the computational strategy for the above hierarchical models needs the following two additional steps: (i) we should calculate the conditional posterior density for $\boldsymbol{\theta}$ given the hyperparameter τ^2 and the observed data y, i.e., calculate $\pi(\boldsymbol{\theta}|\tau^2, y)$ and (ii) we should obtain the marginal posterior distribution $\pi(\tau^2|y)$.

1.5 Markov Chain Monte Carlo (MCMC)

1.5.1 Metropolis-Hastings Algorithm

Markov Chain Monte Carlo (MCMC) methods attempt to simulate direct draws from some distributions of interest. In recent years MCMC techniques have been increasingly used by researchers to simulate complex, nonstandard multivariate distributions. In the MCMC technique, one uses the previous sample values to randomly generate the next sample value and generate a Markov chain (as the transition probabilities between sample values are only

a function of the most recent sample value). Constructing such a Markov chain is surprisingly easy. We describe the form due to Hastings (1970), which is a generalization of the method first proposed by Metropolis *et al.* (1953).

Suppose our goal is to draw samples from some distribution $\pi(\theta)$. For the Metropolis-Hastings algorithm, at each time t, the next state θ_{t+1} is chosen by first sampling a candidate point θ^* from a proposal (or candidate-generating or jumping) distribution $q(\theta^*|\theta_t)$. Note that the proposal distribution may depend on the current point θ_t. The candidate point θ^* is then accepted with probability $\alpha(\theta_t, \theta^*)$ where

$$\alpha(\theta_t, \theta^*) = \min\left(1, \ \frac{\pi(\theta^*)q(\theta_t|\theta^*)}{\pi(\theta_t)q(\theta^*|\theta_t)}\right). \tag{1.5}$$

If the candidate point is accepted, the next state becomes $\theta_{t+1} = \theta^*$. If the candidate is rejected, the chain does not move, i.e., $\theta_{t+1} = \theta_t$. In practice the Metropolis-Hastings algorithm generates a sequence of draws from the distribution $\pi(\theta)$ with the following steps:

1. Start with any initial value θ_0, set $t = 0$.
2. Using the current θ value, sample a candidate point θ^* from some proposal distribution $q(\theta^*|\theta_t)$.
3. Sample a uniform (0,1) random variable U.
4. If $U \leq \alpha(\theta_t, \theta^*)$ set $\theta_{t+1} = \theta^*$, otherwise set $\theta_{t+1} = \theta_t$.

Repeat steps (2) to (4) n times to obtain a sample of size n. Convergence and mixing of the chain should be considered in order to select an appropriate sample size n. If the proposal distribution is symmetric, i.e., $q(\theta^*|\theta_t) = q(\theta_t|\theta^*)$, the Metropolis-Hastings algorithm becomes the Metropolis algorithm. Hence, for the Metropolis algorithm we need to calculate $\alpha(\theta_t, \theta^*)$, where

$$\alpha(\theta_t, \theta^*) = \min\left(1, \ \frac{\pi(\theta^*)}{\pi(\theta_t)}\right). \tag{1.6}$$

1.5.2 Gibbs Sampling

The Gibbs sampler (Geman and Geman, 1984) is a special case of Metropolis-Hastings sampling where the random value is always accepted (i.e., $\alpha = 1$). The main feature of the Gibbs sampler is that one only considers univariate conditional distributions. Such conditional distributions are far easier to simulate than complex joint distributions and usually have simple forms (e.g., normal, inverse gamma, or other common prior distributions). A very nice application of the Gibbs sampler to generalized linear models with random effects is found in Zeger and Karim (1991). Some other applications are found in Gelfand and Smith (1990) and Gelfand *et al.* (1990). To explain the Gibbs sampler algorithm, let us consider that there are k parameters in the model, denoted by $\theta_1, \cdots, \theta_k$, and that the conditional distributions $p(\theta_i|\theta_{j\neq i}, y)$, $i = 1, \cdots, k$, are available for sampling. Then, given a set of starting values $(\theta_1^{(0)}, \cdots, \theta_k^{(0)})$, for the first iteration we sample

$$\begin{aligned}
\theta_1^{(1)}|y &\text{ from } p(\theta_1|\theta_2^{(0)}, \cdots, \theta_k^{(0)}, y), \\
\theta_2^{(1)}|y &\text{ from } p(\theta_2|\theta_1^{(1)}, \theta_3^{(0)}, \cdots, \theta_k^{(0)}, y), \\
&\vdots \\
\theta_k^{(1)}|y &\text{ from } p(\theta_k|\theta_1^{(1)}, \cdots, \theta_{k-1}^{(1)}, y).
\end{aligned}$$

Then using the first set of samples $(\theta_1^{(1)}, \cdots, \theta_k^{(1)})$ we draw a second set and so on. The process continues until after n iterations when a sample $(\theta_1^{(n)}, \cdots, \theta_k^{(n)})$ is obtained. Again to select the sample size n, one should consider the convergence and mixing of the chain. The iterative process follows a Markov chain, which converges to its stationary distribution, that being the joint posterior distribution of the k parameters.

1.6 Discussion

In this chapter, we have discussed the concept of hierarchical models with some examples. The Bayesian approach to deal with hierarchical models is presented here. We have also described the techniques of drawing samples from conditional posterior distributions. Our

main objective of this study is Bayesian curve fitting using both piece-wise linear and natural cubic splines. Therefore, in the next chapter, we will introduce some basic concepts of splines and their classical computational aspects.

Chapter 2

Nonparametric Regression

2.1 Introduction

Regression analysis is one of the most popular and useful tools in data analysis. The goals of regression analysis are to describe the dependence of a response variable on one or more covariates and to predict the response in future. Suppose we have a response variable Y, and a predictor variable X. The dependence of Y on X can be expressed as the following functional form

$$y = f(x) + \epsilon, \tag{2.1}$$

where ϵ is the random noise and f is an unknown regression function that we wish to estimate. The value $f(X)$ is the conditional expectation of Y given the value X, so it can be used to predict the future values of Y for different measured values of X. There are two approaches to find the regression function f: (a) parametric approach and (b) nonparametric approach. In the parametric approach we have rigid parametric assumptions about the dependence between the response and the predictors. Linear regression, logistic regression, Poisson regression are examples of the parametric regression approach. In general, in the parametric approach the response variables are assumed to have a distribution in the exponential family. Within the exponential family the dependence of response on the covariates can be summarized under the framework of generalized linear models (GLM)(McCullah and

Nelder, 1989, Nelder and Wedderburn, 1972). The simplest form of GLM is the linear regression model $y = \beta_0 + \beta_1 x + \epsilon$, where the responses are assumed to be normally distributed.

Researchers use polynomial regression in situations where they know that curvilinear effects are present in the true response function. Polynomial regression is also useful for approximating functions in unknown and possibly very complex nonlinear relationships. But they suffer from various drawbacks. For example, individual observations can exert an influence, in unexpected ways, on remote parts of the curve. Also, owing to the global nature of polynomial fitting, there are problems in estimating wiggly curves (Denison *et al.* 1998).

Every parametric regression analysis requires rigid assumptions on the form of f that may not be true, and hence the usefulness of a non-parametric approach arises. In the non-parametric approach we do not have any rigid assumptions about the dependence between the response and the predictors. By making the relatively weak assumption that whatever the true relationship might be, it is a smooth curve, it is possible to let the data tell the analyst what the pattern truly is. Smoothing methods provide a bridge between making no assumptions on formal structure (non-parametric approach) and making very strong assumptions (parametric approach) (Simonoff, 1996). The assumption of smoothness can be interpreted as meaning that the dependence of the mean of Y on X should not change much if X does not change much. We expect an estimate $\hat{f}$ of f which is not much wigglier than f. The non-parametric approach involves the choice of a smoothing parameter which controls the balance between goodness of fit and smoothness of the estimated regression function. The running-mean, running-line, smoothing spline, kernel, and regression spline smoothers are all linear smoothers (Hastie and Tibshirani, 1990). Some smoothers are nonlinear, that is, the fit $\hat{f}$ cannot be written as $\hat{f} = Sy$ for any smoother matrix S independent of y. All the smoothers mentioned above are nonlinear if the selection of the smoothing parameter is based on the data y.

Since we compare Bayesian curve fitting using different roughness penalty prior distributions in splines, we describe the spline-based roughness penalty approach and associated techniques in the next few sections. In section 2.2 we give the definition of splines. In section 2.3 we discuss how the design matrices can be computed in case of piecewise linear splines. We briefly discuss the natural cubic spline approach to curve fitting in section 2.4, where we describe how to plot a natural cubic spline and how to use the cross-validation method to choose the smoothing parameter in the classical curve fitting approach.

2.2 Splines

Univariate and polynomial splines are piecewise polynomials of some degree d. The breakpoints marking a transition from one polynomial to the next are referred to as *knots* (Hansen and Kooperberg, 2002). For example, a linear spline consists of several piecewise linear functions and a cubic spline consists of several piecewise polynomials of degree three. Typically, a spline will also satisfy smoothness constraints describing how the different pieces are to be joined. These restrictions are specified in terms of the number of continuous derivatives, m, exhibited at the joins of the piecewise polynomials. Hence, in case of a cubic spline, we make the constraint of continuous first and second derivatives. In this study we denote the vector of p knots as $t = (t_1, \cdots, t_p)$. Mathematically, a function f is called a spline of degree d if

1. the domain of f is an interval $[a, b]$ and
2. $f, f', f'', \cdots, f^{(d-1)}$ are all continuous functions on $[a, b]$. The knots $t_i,\ i = 1, \cdots, p$, satisfy $a = t_0 < t_1 < \cdots < t_p < t_{p+1} = b$, and f is a polynomial of degree at most d on each subinterval $[t_i,\ t_{i+1}],\ i = 0, \cdots, p$.

2.3 Piecewise linear spline

Higher-order polynomial interpolation is rarely used for practical purposes because of a polynomial wiggle, i.e., high swings of the interpolating polynomials between the data points.

Instead, most computational algorithms use the idea of splines, i.e., piecewise interpolation. In this research we consider experiments that generate scatter plot data $(y_i, x_i), i = 1, \cdots, n$, where the relationship can be linear in some places and curvilinear in other places. In other words, we consider scatter plot data where the regression function can easily be expressed by a piecewise linear spline. The primary objective of this thesis is to fit the scatter plot data through the Bayesian hierarchical model technique.

For the given data, we want to fit equation (2.1) where $f(x) = f(x; \boldsymbol{\theta})$ is modelled as a piecewise linear function, with knots $\boldsymbol{t} = (t_0, t_1, \cdots, t_p, t_{p+1})$ satisfying $a = t_0 < t_1 < \cdots < t_p < t_{p+1} = b$, and function values at the knots $\theta_0, \theta_1, \cdots, \theta_p, \theta_{p+1}$, such that the coordinate (t_i, θ_i) represents the location at which two line segments meet. To keep things simple, we assume that:

(i) The knots are equally spaced.

(ii) The value of θ_is are equal to zero in two exterior knots t_0 and t_{p+1}, i.e., $\theta_0 = 0$ and $\theta_{p+1} = 0$, so that $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_p)'$ is the vector of unknown parameters.

(iii) The linear trend is being considered in the model separately and the estimation will be done by the backfitting procedure.

In matrix notation we can express our model as

$$Y = D\boldsymbol{\delta} + A\boldsymbol{\theta} + \boldsymbol{\epsilon}, \tag{2.2}$$

where Y is an $n \times 1$ vector of observations, $\boldsymbol{\theta}$ is a $p \times 1$ vector of parameters, $\boldsymbol{\delta} = (\delta_0, \delta_1)'$ is the vector of linear regression parameters, $D = (\mathbf{1}, \boldsymbol{x})$ is an $n \times 2$ matrix, A is an $n \times p$ design matrix and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of random error terms. The basis matrix or the design matrix A for the piece-wise linear spline can be computed from the corresponding x values as follows:

$$A_{ij} = \begin{cases} \frac{x_i - t_{j-1}}{t_j - t_{j-1}}, & \text{if } t_{j-1} \leq x_i \leq t_j, \\ \frac{t_{j+1} - x_i}{t_{j+1} - t_j}, & \text{if } t_j \leq x_i \leq t_{j+1} \quad \text{and} \\ 0, & \text{otherwise} \end{cases}$$

for $i = 1, \cdots, n$ and $j = 1, \cdots, p$. It is evident from here that for $t_j \leq x_i \leq t_{j+1}$, $A_{i,j+1} = 1 - A_{ij}$, and that makes easy computation of the basis matrix. We assume that the random error $\boldsymbol{\epsilon}$ follows a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\sigma^2 I$. Hence, we can write $Y \sim N(D\boldsymbol{\delta} + A\boldsymbol{\theta}, \ \sigma^2 I)$, i.e., the observation Y follows a multivariate normal distribution with the mean vector $D\boldsymbol{\delta} + A\boldsymbol{\theta}$ and the covariance matrix $\sigma^2 I$. Here the quantity $D\boldsymbol{\delta}$ represents the linear part and the quantity $A\boldsymbol{\theta}$ represents the smooth part of the curve.

2.4 Natural Cubic Spline (NCS) with Roughness Penalty

A cubic spline, f, on an interval $[a, b]$ is said to be a natural cubic spline if its second and third derivatives are zero at the boundaries a and b. These conditions are known as natural boundary conditions. This implies that f is linear on the two extreme intervals $[a, t_1]$ and $[t_p, b]$ (Green and Silverman, 1994). There is much literature on natural cubic splines including the books of Green and Silverman (1994), Hastie and Tibshirani (1990) and Wahba (1990). The purpose of the nonparametric regression through natural cubic splines is to summarize the trend of a response measurement as a function of one or more predictors assuming a piecewise polynomial of degree three. In general it is always desirable to have an estimated curve (trend) which provides a good fit without being too wiggly. The roughness penalty approach makes a compromise between these two opposing factors.

Roughness or wigglyness of a twice-differentiable curve f defined on $[a, b]$ is measured by calculating its integrated squared second derivative $\int_a^b \{f''(x)\}^2 dx$. The roughness penalty

approach to curve estimation is easily done by defining the penalized sum of squares

$$S = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \tau^2 \int_a^b \{f''(x)\}^2 dx. \tag{2.3}$$

The penalized least square estimator $\hat{f}$ is obtained by minimizing S over all twice-differentiable functions f. An important point here is that the minimizer of S is a natural cubic spline with knots at the x_i's. The addition of the roughness penalty term $\tau^2 \int_a^b \{f''(x)\}^2 dx$ in equation (2.3) ensures that both the residual sum of squares $\sum_{i=1}^{n} \{y_i - f(x_i)\}^2$ and the roughness $\tau^2 \int_a^b \{f''(x)\}^2 dx$ are used to determine the cost S. In this context, the smoothing parameter τ^2 represents the rate of exchange between residual error and local variation. It also gives the amount in terms of sum of squared residual error that corresponds to one unit of integrated squared second derivative. For the given value of τ^2, minimizing S will give the best compromise between smoothness and goodness-of-fit. Choice of the optimum value of τ^2 is very important, and usually it is done by cross validation that we will discuss in later sections.

2.4.1 Interpolating and Plotting an NCS

Detailed calculations for interpolation and plotting an NCS are found in Green and Silverman (1994). For convenience, we use the same notation as Green and Silverman. Let f be a natural cubic spline with knots $t_1 < t_2 < \cdots < t_p$ and we define $f_i = f(t_i)$ and $\gamma_i = f''(t_i)$ for $i = 1, \cdots, p$. By the definition of NCS, $\gamma_1 = \gamma_p = 0$. Let $\boldsymbol{f} = (f_1, \cdots, f_p)^T$ and $\boldsymbol{\gamma} = (\gamma_2, \cdots, \gamma_{p-1})^T$. The vectors $\boldsymbol{f}$ and $\boldsymbol{\gamma}$ specify the curve f completely with the two band matrices Q and R such that

$$Q^T \boldsymbol{f} = R\boldsymbol{\gamma}. \tag{2.4}$$

Appendix C.1 contains the calculations for Q and R matrices. The roughness penalty term $\int_a^b \{f''(t)\}^2 dt$ can be expressed as,

$$\int_a^b \{f''(t)\}^2 dt = \boldsymbol{\gamma}^T R \boldsymbol{\gamma} = \boldsymbol{f}^T K \boldsymbol{f}, \tag{2.5}$$

where the matrix K is defined by

$$K = QR^{-1}Q^T. \tag{2.6}$$

A natural cubic spline can be plotted as $\hat{Y} = A\hat{\boldsymbol{\theta}}$, where A is an $n \times p$ basis matrix. The j^{th} row ($j = 1, \cdots, n$) of A, as shown in Appendix C.1, can be written as:

$$a_{j,i+1} = \frac{(t-t_i)}{h_i} - \frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i+1} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i+1}\right\},$$

$$a_{ji} = \frac{(t_{i+1}-t)}{h_i} - \frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i}\right\},$$

$$\text{and } a_{j,i'} = -\frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i'} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i'}\right\},$$

where $i' = 1, 2, \cdots, i-1, i+2, \cdots, p$, for any t between t_i and t_{i+1}, and v_{ij} is the $(i,j)^{th}$ element of $R^{-1}Q^T$.

2.4.2 Choosing the Smoothing Parameter for Spline Smoothing

The problem of choosing the smoothing parameter in curve estimation is very important because of the dependency of the estimated curve on the parameter. One can choose the parameter value subjectively. If τ^2 in equation (2.3) is very large then the main component in S will be the roughness penalty term and hence the minimizer $\hat{f}$ will display very little curvature. Therefore, as τ^2 tend to infinity, the term $\int_a^b \{f''(x)\}^2 dx$ will be forced to zero and the curve $\hat{f}$ will approach the linear regression fit. On the other hand, if τ^2 is relatively small, the main contribution to S will be the residual sum of squares, and the curve estimate $\hat{f}$ will not represent the curvature of the data well. There are many different data driven methods to select the value of the smoothing parameter τ^2, among them cross-validation (CV) and generalized cross-validation (GCV) are the widely used techniques.

The Cross-validation and the Generalized Cross-validation

Silverman (1985) and Green and Silverman (1994) discussed the 'leave-one-out' principle to choose the smoothing parameter and this is known as cross-validation. The basic principle of cross-validation is to leave the data points out one at a time and to choose the value of τ^2

under which the missing data points are best predicted by the remainder of the data. Let $\hat{f}^{-i}(x;\tau^2)$ be the smoothing spline calculated from all the data points except (x_i, y_i), using the value of τ^2 as the smoothing parameter. The cross-validation choice of τ^2 is then the value of τ^2 that minimizes the cross-validation score defined as

$$CV(\tau^2) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i - \hat{f}^{-i}(x_i;\tau^2)\right\}^2. \tag{2.7}$$

For computational ease the CV score defined above can be modified as follows:

$$CV(\tau^2) = \frac{1}{n}\sum_{i=1}^{n}\left\{\frac{y_i - \hat{f}(x_i;\tau^2)}{1 - A_{ii}(\tau^2)}\right\}^2, \tag{2.8}$$

where $\hat{f}$ is the spline smoother calculated from the full data set (t_i, y_i) with the smoothing parameter τ^2 and $A(\tau^2) = (I - \tau^2 Q R^{-1} Q^T)^{-1}$. The equation (2.8) shows that, provided that the diagonal entries $A_{ii}(\tau^2)$ are known, the cross-validation score can be calculated from the residuals $\{y_i - \hat{f}(t_i)\}$ about the spline smoother calculated from the full data set.

Generalized cross-validation (GCV) is a modified form of CV and is a popular method for choosing the smoothing parameter (Craven and Wahba, 1979). It is obtained by replacing the term $1 - A_{ii}(\tau^2)$ in the denominator of equation (2.8) by $1 - n^{-1}\mathrm{tr}A(\tau^2)$ as follows:

$$GCV(\tau^2) = \frac{1}{n}\frac{\sum_{i=1}^{n}\{y_i - \hat{f}(x_i;\tau^2)\}^2}{\{1 - n^{-1}\mathrm{tr}A(\tau^2)\}^2}. \tag{2.9}$$

As in ordinary CV, the GCV choice of smoothing parameter is then carried out by minimizing the function $GCV(\tau^2)$ over τ^2.

Chapter 3

Bayesian Curve Fitting

3.1 Introduction

The purpose of any curve fitting is to determine the true functional relationship between the response and the covariates. Before fitting any curve, we may have some clues about the relationship, or we may have some ideas about what we think our approximation should look like. Obviously, we should think of the following issues which may influence how we proceed in determining unknown relationships. First, we should think about the type of the approximating function. Commonly used approximating functions include linear models, generalized linear models, smoothing splines, neural networks, wavelets, decision trees and kernel smoothers. All of these provide explicit models for the relationship between the response and the predictors (Denison *et al.*, 2002). In this thesis we will consider the spline based relationships. The second issue is that we will need some criterion to find the best approximation of the truth. Since in the real data analysis situation, no single model can completely explain the true relationship between the response and the predictors, we may think of the best model among a set of alternatives as the one which most closely explains the true relationship for the particular purpose of interest. Thirdly, we should think of improving the quality of the models by incorporating available qualitative or quantitative knowledge *a priori*. This approach naturally leads us to Bayesian methods where we assign

prior distributions to the parameters in the model and then update the priors in the light of data. These yield a posterior distribution via Bayes' theorem:

$$\text{posterior} \propto \text{prior} \times \text{likelihood}.$$

In this chapter we mainly describe the prior and posterior distributions for curve fitting assuming the regression function is either piece-wise linear spline or a natural cubic spline. In section 3.2 we discuss simple roughness priors. There we discuss three different roughness penalty priors that we will use to compare the smoothness of the fitted curve. The hierarchical structures of our Bayesian data analysis are presented in section 3.3. We calculate the conditional posteriors for all the model parameters in section 3.4. Finally, we give a brief discussion in section 3.5.

3.2 Simple Roughness Priors

To perform Bayesian inference, the scientific question is how to specify a prior distribution for the parameter θ. Prior distributions can be either informative and non-informative. In the informative case the prior distribution represents a population of possible parameter values from which θ has been drawn. But in many practical situations, there is no perfectly relevant population of θ's from which the current θ has been drawn. Some authors suggest that the prior distribution should include all plausible values of θ, but the distribution need not be realistically concentrated around the true value, because often the information about θ contained in the data will far outweigh that of any reasonable prior probability specification. When prior distributions have no population basis, they can be difficult to construct, and there has long been a desire for prior distributions that can be guaranteed to play a minimal role in the posterior distribution. Such distributions are sometimes called 'reference prior distributions', and the prior density is described as vague, flat, diffuse or noninformative (Gelman *et al.* 1995). Jeffreys' invariance principle is sometimes used to define noninformative prior distributions.

The property that the posterior distribution follows the same parametric form as the prior distribution is called conjugacy. In this case we call the prior distribution a conjugate prior distribution. For instance, the beta distribution is a conjugate prior for the binomial likelihood function, and the normal distribution is a conjugate prior for the normal likelihood function. Conjugate prior distributions have some practical advantages and computational convenience. The probability distributions that belong to an exponential family have natural conjugate prior distributions.

As mentioned earlier, we are interested in a Gaussian response that can be expressed as $\boldsymbol{Y} \sim N(D\boldsymbol{\delta}+A\boldsymbol{\theta},\ \sigma^2 I)$, i.e., the response vector $\boldsymbol{Y}$ follows a multivariate normal distribution with mean $D\boldsymbol{\delta}+A\boldsymbol{\theta}$ and covariance matrix $\sigma^2 I$, where the quantity $D\boldsymbol{\delta}$ represents the linear part and the quantity $A\boldsymbol{\theta}$ represents the smooth part of the curve. Hence, the likelihood function can be written as:

$$\pi(\mathbf{Y}|\boldsymbol{\delta},\boldsymbol{\theta},\sigma^2,\mathbf{X}) \propto |\Omega|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})'\Omega^{-1}(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})\right\}, \tag{3.1}$$

where $\Omega = \sigma^2 I$. Since the main focus of this study is to estimate the parameters $\boldsymbol{\delta}$ and $\boldsymbol{\theta}$ with the Bayesian technique, we need to specify the prior distributions of the parameters. Let us assume that the relationship between the response and the covariate can be expressed by a piecewise linear spline and we are interested in comparing the following roughness penalty prior distributions of $\boldsymbol{\theta}$.

Prior 1. Let us be interested in estimating $\boldsymbol{\theta}$ using p equally spaced knot points and hence $\boldsymbol{\theta}$ contains p unknown parameters. By the conditional probability law the joint prior of $\boldsymbol{\theta}$ and τ^2 can be written as $\pi(\boldsymbol{\theta},\tau^2) = \pi(\boldsymbol{\theta}|\tau^2)\pi(\tau^2)$, and we write the roughness penalty prior for $\boldsymbol{\theta}$ given τ^2 as

$$\pi(\boldsymbol{\theta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{p/2} \exp\left\{-\frac{1}{2}\frac{\sum_{i=1}^{p}\left\{\theta_i - \frac{[\theta_{i-1}+\theta_{i+1}]}{2}\right\}^2}{\tau^2}\right\}. \tag{3.2}$$

Here the proportionality (or normalizing) constant is not shown because it does not depend

on τ^2. The term $\sum_{i=1}^{p}\{\theta_i - [\theta_{i-1}+\theta_{i+1}]/2\}^2$ in equation (3.2) can be expressed in matrix notation as $\boldsymbol{\theta}' M \boldsymbol{\theta}$, where

$$M = \begin{pmatrix} \frac{5}{4} & -1 & \frac{1}{4} & 0 & \cdots & 0 & 0 \\ -1 & \frac{3}{2} & -1 & \frac{1}{4} & \ldots & 0 & 0 \\ \frac{1}{4} & -1 & \frac{3}{2} & -1 & \ldots & 0 & 0 \\ 0 & \frac{1}{4} & -1 & \frac{3}{2} & \ldots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \ldots & \frac{3}{2} & -1 \\ 0 & 0 & 0 & 0 & \ldots & -1 & \frac{5}{4} \end{pmatrix}$$

is a $p \times p$ symmetric matrix. Therefore, expression (3.2) can be written as

$$\begin{aligned} \pi(\boldsymbol{\theta}|\tau^2) &\propto \left(\frac{1}{\tau^2}\right)^{p/2} |M^{-1}|^{-1/2} \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\theta}' M \boldsymbol{\theta}\right\} \\ &= |B|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}' B \boldsymbol{\theta}\right\}, \end{aligned} \tag{3.3}$$

where $B = (1/\tau^2)M$ and B^{-1} represents the covariance matrix of $\boldsymbol{\theta}$. The expression on the right hand side of (3.3) looks like the pdf of a multivariate normal distribution without the normalizing constant and hence we can say that $\boldsymbol{\theta}$ given τ^2 follows multivariate normal distribution with mean vector $\mathbf{0}$ and precision matrix B. Here B measures the overall roughness or wigglyness of the curve. For the variance component τ^2, the most commonly used prior distribution is the inverse Gamma distribution, and this is the conjugate prior distribution for the normal distribution. Hence, for τ^2, we assume an inverse gamma distribution with parameters α and β and therefore, the pdf of τ^2 can be written as

$$\pi(\tau^2) = \frac{\beta^\alpha}{\Gamma(\alpha)}\left(\frac{1}{\tau^2}\right)^{(\alpha+1)} \exp\left\{-\frac{\beta}{\tau^2}\right\}, \quad \text{for } \tau^2 > 0. \tag{3.4}$$

We assume that both α and β are known and hence we can write the above prior distribution as $\pi(\tau^2) \propto (1/\tau^2)^{(\alpha+1)} \exp\{-\beta/\tau^2\}$ where the proportionality constant involves only α and β, and does not have any effect in calculating the posterior distribution of τ^2.

Prior 2. Instead of considering overall roughness of the curve, if we consider the local roughness at each knot point to estimate the curve we may think of a vector of variance components say, $\boldsymbol{\tau}^2 = (\tau_1^2 \cdots \tau_p^2)'$. Therefore, the joint prior of $\boldsymbol{\theta}$ and $\boldsymbol{\tau}^2$ can be written as $\pi(\boldsymbol{\theta}, \boldsymbol{\tau}^2) = \pi(\boldsymbol{\theta}|\boldsymbol{\tau}^2)\pi(\tau_1^2)\pi(\tau_2^2)\cdots\pi(\tau_p^2)$, where

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}^2) \propto h\left(\frac{1}{\tau_1^2}, \cdots, \frac{1}{\tau_p^2}\right) \exp\left\{-\frac{1}{2}\sum_{i=1}^{p} \frac{\left\{\theta_i - \frac{[\theta_{i-1}+\theta_{i+1}]}{2}\right\}^2}{\tau_i^2}\right\}, \tag{3.5}$$

where $h(\cdot)$ is some function of $\left(\frac{1}{\tau_1^2}, \cdots, \frac{1}{\tau_p^2}\right)$ related to the determinant of the covariance matrix of $(\theta_1, \cdots, \theta_p)$. Similar to the case of Prior 1, we can express the term $\sum_{i=1}^{p}\left\{\frac{\theta_i - \frac{[\theta_{i-1}+\theta_{i+1}]}{2}^2}{\tau_i^2}\right\}$ in matrix notation as $\boldsymbol{\theta}' M_{\boldsymbol{\tau}} \boldsymbol{\theta}$, where $M_{\boldsymbol{\tau}}$ is a symmetric $p \times p$ matrix as follows

$$\begin{pmatrix}
\frac{1}{\tau_1^2}+\frac{1}{4\tau_2^2} & -\frac{1}{2}\left(\frac{1}{\tau_1^2}+\frac{1}{\tau_2^2}\right) & \frac{1}{4\tau_2^2} & \cdots & 0 & 0 \\
-\frac{1}{2}\left(\frac{1}{\tau_1^2}+\frac{1}{\tau_2^2}\right) & \frac{1}{4\tau_1^2}+\frac{1}{\tau_2^2}+\frac{1}{4\tau_3^2} & -\frac{1}{2}\left(\frac{1}{\tau_2^2}+\frac{1}{\tau_3^2}\right) & \cdots & 0 & 0 \\
\frac{1}{4\tau_2^2} & -\frac{1}{2}\left(\frac{1}{\tau_2^2}+\frac{1}{\tau_3^2}\right) & \frac{1}{4\tau_2^2}+\frac{1}{\tau_3^2}+\frac{1}{4\tau_4^2} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \frac{1}{4\tau_{p-2}^2}+\frac{1}{\tau_{p-1}^2}+\frac{1}{4\tau_p^2} & -\frac{1}{2}\left(\frac{1}{\tau_{p-1}^2}+\frac{1}{\tau_p^2}\right) \\
0 & 0 & 0 & \cdots & -\frac{1}{2}\left(\frac{1}{\tau_{p-1}^2}+\frac{1}{\tau_p^2}\right) & \frac{1}{4\tau_{p-1}^2}+\frac{1}{\tau_p^2}
\end{pmatrix}$$

and therefore $\pi(\boldsymbol{\theta}|\boldsymbol{\tau}^2)$ in (3.5) can be written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{\tau}^2) \propto |M_\tau|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}' M_\tau \boldsymbol{\theta}\right\}. \tag{3.6}$$

Therefore, we can say that $\boldsymbol{\theta}$ follows a multivariate normal distribution with mean vector $\mathbf{0}$ and precision matrix M_τ. Here M_τ^{-1} measures the local variation of the data at each knot point. As we are assuming different τ_i^2 at different knot points, each $\pi(\tau_i^2)$ might be assigned the same inverse gamma distribution with hyperparameters α and β. We are assuming this kind of prior rather than Prior 1 because, in some real life situations curves show different roughness in different places of the curve. More precisely, the data variances are equal but the smoothness of the curve differs for some of the knot points. Hence there is a chance of getting less efficient estimates if we consider the equal variance component for all the knot

points of the data. Instead, the inferences can be improved if we use a joint prior of $\boldsymbol{\theta}$ and τ^2 which can account for local variation of the data, and hence we choose Prior 2.

To estimate the parameters with the heteroscedastic situation we express τ_i^2 as

$$\tau_i^2 \;=\; w_i\tau^2 \;\text{ for } i = 1, \cdots, p.$$

This is in some sense similar to the weighted least squares (WLS) problem of linear regression. Here we consider both w_i and τ^2 to be random samples form inverse gamma distributions. One can think of the scalar τ^2 as governing 'overall' roughness, while $\boldsymbol{w} = (w_1, \cdots, w_p)$ governs local deviations from the overall roughness. By considering random w_i we always have the flexibility to let the data select the appropriate values for the local variation and hence produce better estimates. So now we have prior $\pi(\boldsymbol{\theta}|\boldsymbol{w}, \tau^2)$ instead of prior $\pi(\boldsymbol{\theta}|\boldsymbol{\tau}^2)$, which penalizes roughness of the curve. Thus the prior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{w}$ and τ^2 can be written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{w}, \tau^2) \propto \left(\frac{1}{\tau^2}\right)^{p/2} h\left(\frac{1}{w_1^2}, \cdots, \frac{1}{w_p^2}\right) \exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{p}\frac{\left\{\theta_i - \frac{[\theta_{i-1}+\theta_{i+1}]}{2}\right\}^2}{w_i}\right\}. \tag{3.7}$$

Similar to the earlier cases we can express the term $\sum_{i=1}^{p}\left\{\frac{\theta_i - \frac{[\theta_{i-1}+\theta_{i+1}]^2}{2}}{w_i}\right\}$ in matrix notation as $\boldsymbol{\theta}' M_{\boldsymbol{w}}\boldsymbol{\theta}$, where $M_{\boldsymbol{w}}$ is a $p \times p$ symmetric matrix as follows

$$\begin{pmatrix}
\frac{1}{w_1}+\frac{1}{4w_2} & -\frac{1}{2}\left(\frac{1}{w_1}+\frac{1}{w_2}\right) & \frac{1}{4w_2} & \cdots & 0 & 0 \\
-\frac{1}{2}\left(\frac{1}{w_1}+\frac{1}{w_2}\right) & \frac{1}{4w_1}+\frac{1}{w_2}+\frac{1}{4w_3} & -\frac{1}{2}\left(\frac{1}{w_2}+\frac{1}{w_3}\right) & \cdots & 0 & 0 \\
\frac{1}{4w_2} & -\frac{1}{2}\left(\frac{1}{w_2}+\frac{1}{w_3}\right) & \frac{1}{4w_2}+\frac{1}{w_3}+\frac{1}{4w_4} & \cdots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \cdots & \frac{1}{4w_{p-2}}+\frac{1}{w_{p-1}}+\frac{1}{4w_p} & -\frac{1}{2}\left(\frac{1}{w_{p-1}}+\frac{1}{w_p}\right) \\
0 & 0 & 0 & \cdots & -\frac{1}{2}\left(\frac{1}{w_{p-1}}+\frac{1}{w_p}\right) & \frac{1}{4w_{p-1}}+\frac{1}{w_p}
\end{pmatrix}$$

and hence $\pi(\boldsymbol{\theta}|\boldsymbol{w}, \tau^2)$ can be written as

$$\pi(\boldsymbol{\theta}|\boldsymbol{w}, \tau^2) \propto |B_{\boldsymbol{w},\tau}|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}' B_{\boldsymbol{w},\tau}\boldsymbol{\theta}\right\}, \tag{3.8}$$

where $B_{\boldsymbol{w},\tau} = (1/\tau^2)M_{\boldsymbol{w}}$. And therefore, we can say that $\boldsymbol{\theta}$ given $\boldsymbol{w}$ and τ^2 follows a multivariate normal distribution with mean vector $\mathbf{0}$ and precision matrix $B_{\boldsymbol{w},\tau}$. As a second-stage prior, let us assume that τ^2 has an inverse gamma distribution with parameters α and β to match with the first prior, and we assume that each of $w_1, \cdots, w_p$ follows an inverse gamma distribution with both scale and shape parameters equal to $1/\lambda$, i.e., $w_i \sim IG(1/\lambda, 1/\lambda)$, $i = 1, \cdots, p$. Here we see that the mean of w_i is $1/(1-\lambda)$ and the fact that as λ goes to 0, the mean of w_i goes to 1. Hence w_i's are centered at 1 for very small λ. Another interesting feature of this prior specification is that as λ decreases, the variance of w_i decreases, that is, as λ goes to 0, the second prior converges to the first prior.

Prior 3. To compare the above two priors, we consider a third prior for $\boldsymbol{\theta}$ which indicates neither local nor global specification for the roughness of the curve. We write,

$\pi(\boldsymbol{\theta}, \tau^2) = \pi(\boldsymbol{\theta}|\tau^2)\pi(\tau^2)$, where

$$\pi(\boldsymbol{\theta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{p/2} \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\theta}'\boldsymbol{\theta}\right\}. \tag{3.9}$$

That is, $\boldsymbol{\theta}$ is assumed to have a multivariate normal distribution with mean vector $\mathbf{0}$ and precision matrix $(1/\tau^2)I$ and as usual we assume that $\pi(\tau^2)$ is an inverse gamma density with hyperparameters α and β.

Many studies suggest that a good choice of prior distribution for the data variance σ^2 is inverse gamma and the uniform prior for the linear regression parameters works well. Hence, for all the three cases, we assume that the data variance σ^2 has an inverse gamma distribution with known parameters a and b, i.e., $\sigma^2 \sim IG(a, b)$. Although the linear regression parameters can take both positive and negative values, for simplicity, we assume that δ_0 and δ_1 are uniform over the interval (0,1).

3.3 Levels of Hierarchy

To verify the performance of the proposed prior distributions we need to conduct simulation studies for the three-level hierarchical models for each of the priors. For Prior 1 the hierarchical structure can be written as:

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\delta},\boldsymbol{\theta},\sigma^2 &\sim N(D\boldsymbol{\delta}+A\boldsymbol{\theta},\Omega), \quad \text{where} \quad \Omega=\sigma^2\mathbf{I}, \\ \boldsymbol{\theta}|\tau^2 &\sim N\left(\mathbf{0},\frac{1}{\tau^2}M^{-1}\right), \quad \sigma^2\sim IG(a,b) \quad \text{and} \quad \boldsymbol{\delta}\sim U(0,1), \\ \tau^2 &\sim IG(\alpha,\beta). \end{aligned}$$

Similarly for Prior 2 the 3-level hierarchical structure can be expressed as follows:

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\delta},\boldsymbol{\theta},\sigma^2 &\sim N(D\boldsymbol{\delta}+A\boldsymbol{\theta},\Omega); \quad \text{where} \quad \Omega=\sigma^2\mathbf{I}, \\ \boldsymbol{\theta}|\mathbf{w},\tau^2 &\sim N(\mathbf{0},B_{w,\tau}^{-1}), \quad \sigma^2\sim IG(a,b) \quad \text{and} \quad \boldsymbol{\delta}\sim U(0,1), \\ \tau^2 &\sim IG(\alpha,\beta) \quad \text{and} \quad w_i\sim IG\left(\frac{1}{\lambda},\frac{1}{\lambda}\right), \quad i=1,\cdots,p. \end{aligned}$$

And finally, the hierarchical structure for Prior 3 is as follows:

$$\begin{aligned} \mathbf{Y}|\boldsymbol{\delta},\boldsymbol{\theta},\sigma^2 &\sim N(D\boldsymbol{\delta}+A\boldsymbol{\theta},\Omega); \quad \text{where } \Omega=\sigma^2\mathbf{I}, \\ \boldsymbol{\theta}|\tau^2 &\sim N\left(\mathbf{0},\frac{1}{\tau^2}I\right), \quad \sigma^2\sim IG(a,b) \quad \text{and} \quad \boldsymbol{\delta}\sim U(0,1), \\ \tau^2 &\sim IG(\alpha,\beta). \end{aligned}$$

For the simplicity of computation, we now assume that the 4^{th} level hyperparameters α, β, a, b and λ are known constants.

3.4 Posterior Distributions and Simulation Study

Under the Bayesian approach, prior beliefs about parameters are combined with sample information to create updated, or posterior, beliefs about the parameters. In the following subsections we present the conditional posterior distributions for all the model parameters in the cases of our proposed prior distributions.

3.4.1 Conditional Posteriors for Prior 1

Combining the sampling distribution for the observable Y's and the prior distributions yields the joint posterior distribution of all the parameters and hyperparameters. For our Prior 1 the joint posterior distribution can be written as:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \sigma^2 | \mathbf{Y}, \mathbf{X}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{X})\pi(\boldsymbol{\theta}|\tau^2)\pi(\tau^2)\pi(\boldsymbol{\delta})\pi(\sigma^2). \tag{3.10}$$

For posterior sampling through MCMC, we need to find the conditional posterior distributions for each of the model parameters given all others, i.e., we have to calculate the marginal distributions, $\pi(\boldsymbol{\theta}|\boldsymbol{\delta}, \tau^2, \sigma^2, \mathbf{Y}, \mathbf{X})$, $\pi(\boldsymbol{\delta}|\boldsymbol{\theta}, \tau^2, \sigma^2, \mathbf{Y}, \mathbf{X})$, $\pi(\tau^2|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{Y}, \mathbf{X})$ and $\pi(\sigma^2|\boldsymbol{\delta}, \tau^2, \mathbf{Y}, \mathbf{X})$ from the joint posterior distribution (3.10). Details of the calculations are presented in Appendix A. From Appendix A.1, it can be written that the conditional posterior distribution of $\boldsymbol{\theta}$ given $\boldsymbol{\delta}$, τ^2, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, is multivariate normal with mean vector $((1/\tau^2)M + (A'A)^{-1}/\sigma^2)^{-1} A'\mathbf{Z}_1/\sigma^2$ and covariance matrix $((1/\tau^2 M + (A'A)^{-1}/\sigma^2)^{-1}$, where $\mathbf{Z}_1 = \mathbf{Y} - D\boldsymbol{\delta}$. By Appendix A.2, we find that the conditional posterior distribution for the variance component τ^2 given $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, is inverse gamma with parameters $\alpha + p/2$ and $(\boldsymbol{\theta}' M \boldsymbol{\theta})/2 + \beta$. The conditional posterior distribution of the linear regression parameter $\boldsymbol{\delta}$ given $\boldsymbol{\theta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, is multivariate normal with mean vector $(D'D)^{-1}D'\mathbf{Z}_2$ and the covariance matrix $(D'D)^{-1}\sigma^2$, where $\boldsymbol{Z}_2 = Y - A\boldsymbol{\theta}$, and is shown in Appendix A.3. Finally form Appendix A.4, we can say that the conditional posterior distribution of the data variance σ^2 given $\boldsymbol{\delta}$, $\boldsymbol{\theta}$, τ^2, $\mathbf{Y}$ and $\mathbf{X}$, is inverse gamma with parameters $a + n/2$ and $(\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta})'(\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta}) + b$. Since the closed form solutions are found for all the conditional posterior distributions, the Gibbs sampler algorithm will be useful to draw the posterior samples.

3.4.2 Conditional Posteriors for Prior 2

Similar to Prior 1, the joint posterior distribution for Prior 2 can be written as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \boldsymbol{w}, \sigma^2 | \mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{X})\pi(\boldsymbol{\theta}|\boldsymbol{w}, \tau^2)\pi(\tau^2)\pi(w_1)\cdots\pi(w_p)\pi(\boldsymbol{\delta})\pi(\sigma^2). \tag{3.11}$$

As shown in Appendix B.1, the conditional posterior $\pi(\boldsymbol{\theta}|\boldsymbol{\delta}, \tau^2, \boldsymbol{w}, \sigma^2, \mathbf{Y}, \mathbf{X})$ is a multivariate normal density with mean vector of $(B_{\boldsymbol{w},\tau} + (A'A)^{-1}/\sigma^2)^{-1} A'\mathbf{Z}_1/\sigma^2$ and covariance matrix of $(B_{\boldsymbol{w},\tau} + (A'A)^{-1}/\sigma^2)^{-1}$, and the conditional posterior distribution for τ^2 given $\boldsymbol{\theta}$, $\boldsymbol{w}$, $\boldsymbol{\delta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, is inverse gamma with the parameters $\alpha + p/2$ and $1/2(\boldsymbol{\theta}' B_{\boldsymbol{w},\tau}\boldsymbol{\theta}) + \beta)$. The joint conditional posterior distribution for the vector $\boldsymbol{w}$ given $\boldsymbol{\theta}$ and τ^2 (as in Appendix B.2) is

$$\pi(\boldsymbol{w}|\boldsymbol{\theta}, \tau^2) \;\propto\; |B_{\boldsymbol{w},\tau}|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}' B_{\boldsymbol{w},\tau}\boldsymbol{\theta}\right\} \times \prod_{i=1}^{p} \left[\frac{(\frac{1}{\lambda})^{\frac{1}{\lambda}}}{\Gamma(\frac{1}{\lambda})} \left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)} \exp\left\{-\frac{1}{\lambda w_i}\right\}\right].$$

Apparently, neither $\pi(\boldsymbol{w}|\boldsymbol{\theta}, \tau^2)$ can be written in a closed form density function, nor we can calculate the marginal conditional posterior densities $\pi(w_i|w_1, \cdots, w_{i-1}, w_{i+1}, \cdots, w_p, \boldsymbol{\theta}, \tau^2)$, for $i = 1, \cdots, p$, easily. This is because $\pi(\boldsymbol{w}|\boldsymbol{\theta}, \tau^2)$ involves the determinant term $|B_{\boldsymbol{w},\tau}|^{1/2}$. The conditional posterior distribution of $\boldsymbol{\delta}$ given $\boldsymbol{\theta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, and the conditional posterior distribution of σ^2 given $\boldsymbol{\delta}$, $\boldsymbol{\theta}$, τ^2, $\mathbf{Y}$ and $\mathbf{X}$, are the same as we have found in the case of Prior 1, that is, $\pi(\boldsymbol{\delta}|\boldsymbol{\theta}, \sigma^2, \mathbf{Y}, \mathbf{X}) \sim N(\{D'D\}^{-1}D'\mathbf{Z}_2,\ \{D'D\}^{-1}\sigma^2)$ and $\pi(\sigma^2|\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \mathbf{Y}, \mathbf{X}) \sim IG(a + n/2,\ \{\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta}\}'\{\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta}\} + b)$.

Ideally, one may think that the hyperparameter λ is fixed, so the posterior samples can be drawn by using the Gibbs sampler algorithm from the conditional posteriors that have closed forms, and using the Metropolis Hastings algorithm from the posterior density $\pi(\boldsymbol{w}|\boldsymbol{\theta}, \tau^2)$. However, in practice, we do not know which value of λ will give us better estimates of the parameters of interest and hence, we should adopt some trial and error simulations to find the optimum value for λ. On the other hand, we may consider some noninformative priors for λ and let the data decide the value of λ instead of using fixed values. We consider both fixed and random λ approaches for posterior sampling. Assuming that λ is uniform within $(0, \lambda_0)$ we calculate the conditional posterior distribution for λ given $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, $\boldsymbol{w}$, τ^2, $\mathbf{Y}$ and $\mathbf{X}$ in Appendix B.3. Daniels (1999) suggested an uniform shrinkage prior for the hyperparameter λ, and in our usual notation, the uniform shrinkage prior can be written as $\pi(\lambda) \sim \sigma^2/(\sigma^2 + \lambda)$, where σ^2 is the data variance. In the study we also consider the uniform shrinkage prior. The conditional posterior distribution for this case is calculated

in Appendix B.3. Since there are no closed form solutions found in both uniform and uniform shrinkage λ cases, we will apply random walk Metropolis-Hastings algorithm to draw posterior samples for the hyperparameter λ.

3.4.3 Conditional Posteriors for Prior 3

For Prior 3, the joint posterior distribution can be written as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \sigma^2|\mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{X})\pi(\boldsymbol{\theta}|\tau^2)\pi(\tau^2)\pi(\boldsymbol{\delta})\pi(\sigma^2). \tag{3.12}$$

As shown in Appendix B.4, the conditional posterior for $\boldsymbol{\theta}$ given $\boldsymbol{\delta}$, τ^2, σ^2, $\mathbf{Y}$ and $\mathbf{X}$ is a multivariate normal density with mean vector $((1/\tau^2)I + (A'A)^{-1}/\sigma^2)^{-1} A'\mathbf{Z}_1/\sigma^2$ and the covariance matrix $((1/\tau^2)I + (A'A)^{-1}/\sigma^2)^{-1}$. The conditional posterior distribution for τ^2 given $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$ is inverse gamma with parameters $\alpha + \frac{p}{2}$ and $\frac{1}{2}(\boldsymbol{\theta}'\boldsymbol{\theta}) + \beta$. The other two conditional distributions, $\pi(\boldsymbol{\delta}|\boldsymbol{\theta}, \sigma^2, \mathbf{Y}, \mathbf{X})$ and $\pi(\sigma^2|\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \mathbf{Y}, \mathbf{X})$, remain the same as before. That is, $\pi(\boldsymbol{\delta}|\boldsymbol{\theta}, \sigma^2, \mathbf{Y}, \mathbf{X})$, is a multivariate normal density with mean vector $(D'D)^{-1}D'\mathbf{Z}_2$, and covariance matrix $(D'D)^{-1}\sigma^2$, and $\pi(\sigma^2|\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \mathbf{Y}, \mathbf{X})$ is an inverse gamma density with parameters $a + \frac{n}{2}$ and $(\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta})'(\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta}) + b$. Hence for all the conditional posterior distributions we may apply the Gibbs Sampler algorithm in order to draw the posterior samples.

3.5 Discussion

In this chapter we have discussed our three choices of roughness penalty prior distributions. The posterior distribution for every prior distribution has been calculated. Closed form solutions are found for all the conditional posterior distributions in Prior 1 and Prior 3, and for all the parameters in Prior 2 except $\boldsymbol{w}$ and λ. Hence, in Prior 2 both Gibbs sampler and random walk Metropolis-Hastings algorithms are applied to generate the posterior samples. In practice, we do not know how rough (wiggle) the true curve is. If the curve is not very rough, Prior 1 can be used for Bayesian curve estimation. On the other hand, if the true curve

is rough in some places but linear in other places, our conjecture is that Prior 2 may produce better estimates of the parameters. In the next chapters, we will check the performance of our proposed priors as well as the validity of our assertion about the roughness penalty priors in Bayesian curve fitting through several examples.

Chapter 4

MCMC Simulation for Piecewise Linear Spline

4.1 Introduction

As our approaches to curve fitting are to allow the regression function f to be either *piecewise linear* or *cubic spline* functions, in this chapter, we provide a Bayesian version which models f by a piecewise straight line with a known number of equidistant knots. Some studies regarding Bayesian curve fitting has been done assuming an unknown number of knots at unknown locations. Denison, Mallick and Smith (1998) described a Bayesian methodology of curve fitting by a piecewise polynomial assuming a large collection of possible knots instead of choosing a single collection of knots. Their Bayesian knot selection procedure was done by selecting posterior samples for both the number of knots and their locations using the Markov chain Monte Carlo (MCMC) simulation technique of reversible jumps (Green, 1995).

In this chapter we mainly discuss the simulation studies to check the performance of our proposed priors in the case of piecewise linear splines. Since we consider both linear and smooth effects of the same covariate on the response we need to apply the Bayesian back-fitting algorithm to fit the curve. We discuss the Bayesian backfitting algorithm in Section

4.2. Section 4.3 and 4.4 present the simulated results. Finally, a brief discussion is given in Section 4.5.

4.2 Backfitting Algorithm

Hastie and Tibshirani (1990) discussed the backfitting procedure to fit the additive model $Y = \alpha + \sum_{i=1}^{p} f_i(X_i) + \epsilon$, where the errors ϵ are independent of the X_is, $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2$. The Bayesian version of the Backfitting algorithm is also proposed by Hastie and Tibshirani (2000) and they called it Bayesian backfitting procedure and they applied it for posterior sampling from additive and generalized additive models.

In this study we assume the additive model as

$$\boldsymbol{y} = \delta_0 + \delta_1 x + f(\boldsymbol{x} : \boldsymbol{\theta}) + \epsilon, \tag{4.1}$$

where ϵ is a random error term that follows normal distribution with mean 0 and constant variance σ^2. The fitted form of equation (4.1) can be written as $\hat{\boldsymbol{y}} = \hat{\delta}_0 + \hat{\delta}_1 \boldsymbol{x} + A\hat{\boldsymbol{\theta}}$. So in our backfitting procedure we first guess the initial values of δ_0 and δ_1 and then calculate $\boldsymbol{y} - \hat{\delta}_0 + \hat{\delta}_1 \boldsymbol{x}$ and draw posterior samples for $\boldsymbol{\theta}$. Then we update δ_0 and δ_1 by drawing posterior samples using $\hat{\boldsymbol{\theta}}$ in the equation $\boldsymbol{y} - A\hat{\boldsymbol{\theta}} = \hat{\delta}_0 + \hat{\delta}_1 \boldsymbol{x}$. We continue this process until a posterior sample of size n is obtained.

4.3 Simulated Results: Example 1

To illustrate how the priors work we apply them to several data sets. Our first data set, called Simulated Motorcycle Accident (Silverman, 1985) is taken from the S-plus data base, and is presented in Figure (4.1). The data frame consists of a series of $n = 133$ accelerometer readings taken through time in an experiment on the efficacy of crash helmets. In the data, the time points are not regularly spaced and there are multiple observations at some time

points, and all the observations are subject to random error. Considering time as a predictor variable, we fit the additive model (4.1) with all choices of prior distributions.

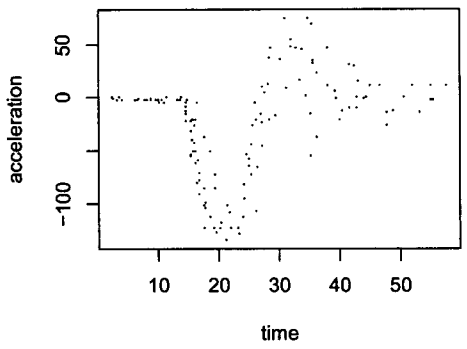

Figure 4.1: *The motor-cycle data.*

Let us fix the number of knots $p = 20$. It is clear from Figure (4.1) that the smoothness of the underlying curve is not constant. To fit this data we choose the Bayesian method that makes the use of our proposed prior distributions. The model parameters $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, τ^2, $\boldsymbol{w}$, σ^2 and λ have been estimated and throughout the study we use the statistical software 'R' for computational purposes.

In a Bayesian analysis a common question is what will be the value of the hyperparameters. As the posterior sampling depends on the values of the hyperparameters, one should be careful in choosing them. In this study, we have to set the values for the hyperparameters a and b (for σ^2), α and β (for τ^2), and λ (for Prior 2). For the data variance σ^2 we choose both the hyperparameters equal to 0.001, i.e., we consider $\sigma^2 \sim IG(0.001, 0.001)$. Such a prior distribution is referred to as non-informative. In a related context, Gustafson *et al.* (2003) suggested an inverse gamma prior for the variance component τ^2 with mode equal to $(0.01)^2$ and coefficient of variation equal to 1. This implies an inverse gamma distribution with a shape parameter of 3 and a scale parameter of $(0.02)^2$. In this study we choose the shape parameter $\alpha = 3$ and scale parameter $\beta = 0.01$. We can not make β less than 0.01 because of some computational difficulty. The smaller value of β forces τ^2

to take very small values and when we compute the covariance matrix for $\boldsymbol{\theta}$ by using the term $(B_{\boldsymbol{w},\tau} + (A'A)^{-1}/\sigma^2)^{-1} A'\mathbf{Z}_1/\sigma^2$, we get the matrix $(B_{\boldsymbol{w},\tau} + (A'A)^{-1}/\sigma^2)$ being close to non-invertible for Prior 2. Before considering random λ, we run MCMC to simulate posterior samples for some fixed values of λ. Among the estimates, it is observed that $\lambda = 0.33$ (i.e., when the coefficient of variation of $w_i = 1$) gives better estimates of the parameters, and hence, we consider Prior 2 with $\lambda = 0.33$ for comparing estimates obtained using other priors. A noninformative prior for the parameter λ is assumed uniform within the interval $(0, \lambda_0)$. Here we need not assume any value for the hyperparameter λ_0 because it cancels out both in the numerator and in the denominator when we calculate the acceptance ratio α in Metropolis algorithm (1.6). The uniform shrinkage prior for the hyperparameter λ is also assumed. Both the priors for λ are proper and hence lead to proper posterior distributions. Hence, depending on the different choices of λ (fixed, uniform and uniform shrinkage), we have 3 different cases for Prior 2, and with Prior 1 and Prior 3 the total choices of prior distributions has increased to five.

4.3.1 Parameter Estimates

For the parameter estimates we have run MCMC simulation techniques. The Gibbs sampler technique has been adopted for posterior simulation of $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, τ^2 and σ^2, and random walk Metropolis-Hastings algorithm for λ and $\boldsymbol{w}$. We present the estimates of the parameters τ^2, $\boldsymbol{\delta}$ and σ^2 for all the five specifications of the prior distributions in five tables (Tables 4.1-4.5). For the parameter vector $\boldsymbol{\theta}$ we have 20 estimates and, therefore, instead of presenting the estimates in a tabular form we prefer to present them graphically. Figure (4.2) shows the 60 posterior realizations (grey curves) for the parameter vector $\boldsymbol{\theta}$ with the means (dark curves) for all the five choices of the prior distributions. The 60 posterior realizations are chosen systematically from the last 3000 samples i.e., take one after every 50 samples. Although no big differences among the estimates have been noticed, by having a close look we observe that the mean curves produced by Prior 1 and Prior 2 are better than the mean curve produced by Prior 3 with respect to smoothness.

Similarity among the estimates of the data variance σ^2 is observed in all the five choices of the prior distributions. Estimates of the overall variance component τ^2 differ greatly. For the case of Prior 3 we observe the biggest τ^2. The smallest τ^2 is observed for Prior 2 with fixed λ. For the Prior 2 with uniform λ the estimate of τ^2 is approximately equal to the estimate obtained in Prior 1 case. Although the estimates of the linear parameters δ_0 and δ_1 differ slightly in all the cases, they have been considered as statistically insignificant since the 90% credible intervals include zero.

Table 4.1: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 1.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	177.92	69.41	94.92	311.14
δ_0	0.69	21.73	-36.76	35.13
δ_1	0.25	0.64	-0.77	1.35
σ^2	509.82	67.11	410.25	627.95

Table 4.2: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with fixed lambda ($\lambda = 0.33$).

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	103.36	51.37	40.10	205.44
δ_0	-2.88	19.20	-33.93	29.24
δ_1	0.22	0.55	-0.66	1.10
σ^2	512.71	67.35	412.00	630.89

4.3.2 Monitoring the Convergence of MCMC Simulation

Before making any valid inferences about the estimates we need to check both the mixing and the convergence of MCMC run for them. We present the MCMC run for the parameters

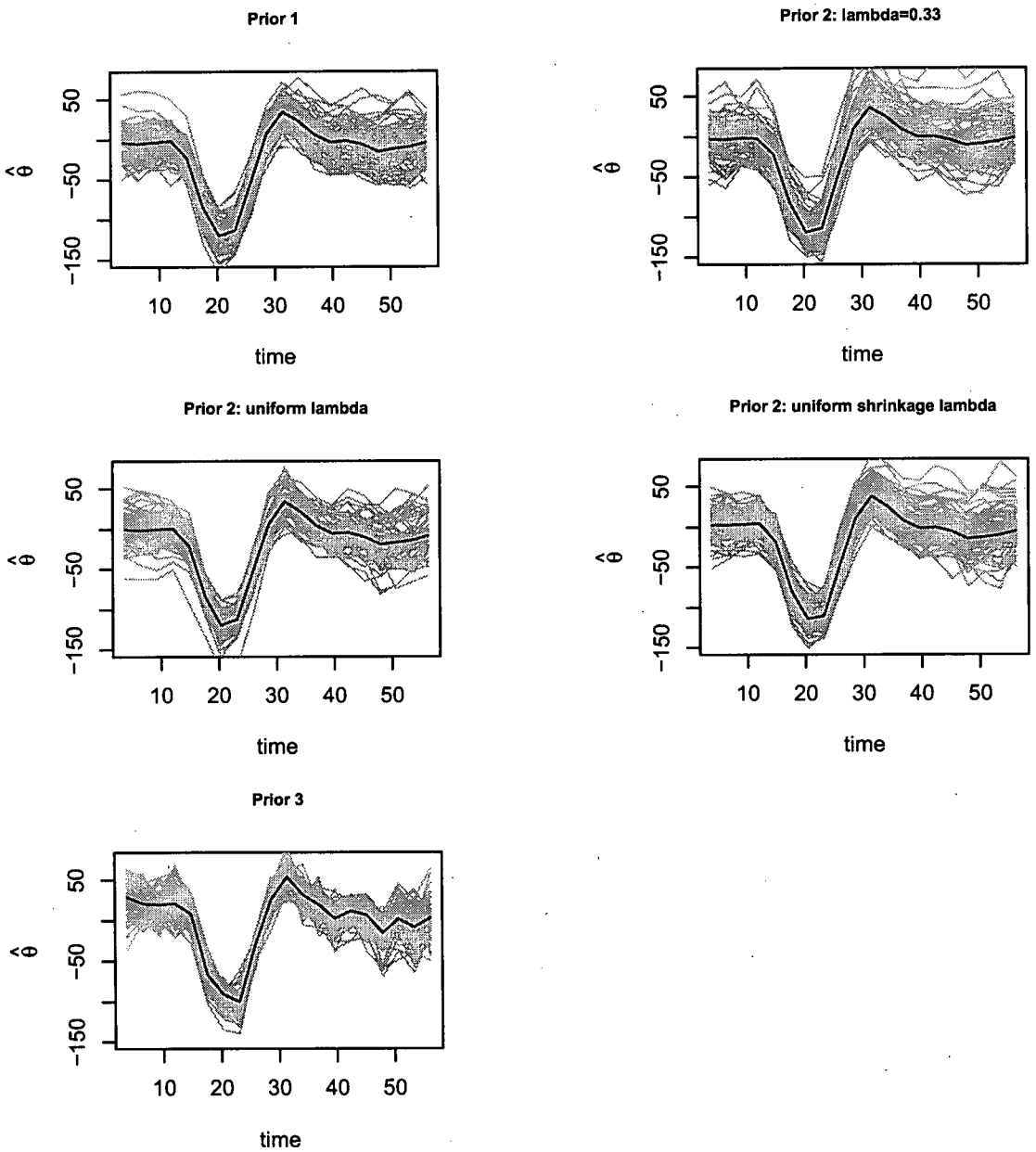

Figure 4.2: *Sixty posterior realizations (grey curves) for the parameter vector* $\boldsymbol{\theta}$. *Dark curves show the posterior means.*

Table 4.3: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform lambda.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	174.05	73.46	87.11	310.32
δ_0	-12.91	19.63	-43.24	20.67
δ_1	0.20	0.62	-0.77	1.29
σ^2	511.82	67.32	412.92	629.77

Table 4.4: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform shrinkage lambda.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	166.52	68.31	82.79	293.66
δ_0	-2.87	21.05	-36.54	32.20
δ_1	0.10	0.55	-0.83	0.97
σ^2	508.79	68.32	405.95	633.42

Table 4.5: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 3.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	1533.47	516.51	890.90	2538.64
δ_0	-31.85	16.61	-57.15	-3.99
δ_1	0.60	0.47	-0.18	1.37
σ^2	522.54	69.94	420.45	648.39

τ^2 and σ^2 in Figures (4.3) and (4.5) respectively. The corresponding histograms based on the samples of iterations 2001-5000 are presented in Figure (4.4) and Figure (4.6). For each of the model parameters we have run MCMC simulations for 5000 iterations, among which the first 2000 iterations have been deleted as burn-in period of the chain and the remaining 3000 samples are used for inference purposes.

For updating λ and $\boldsymbol{w}$ we perform random walk Metropolis-Hastings algorithm in an exponential scale. In other words, let λ_0 be the initial guess for λ, then its candidate states are taken as

$$\lambda^* = \lambda_0 \times \exp(N(0, v)), \tag{4.2}$$

where v is the jump size for updating λ. We need to check for the convergence as well as mixing of the MCMC run for λ. In practice, it is always desirable to have a Markov chain which can mix and converge well at the same time to ensure the accurate inference about the target posterior distribution. To have such a chain we need to adjust the jump size through monitoring the output. There is no hard and fast rule to determine the appropriate jump size and usually a trial and error method is adopted. Many research findings, for example, Gilks *et al.* (1996), suggest that for better mixing and convergence a desirable acceptance rate is 50%. So, while implementing an MCMC simulation it is necessary to plot the run to monitor the mixing and the convergence of the chain, and calculate the acceptance rate. We plot the MCMC chains for λ and the corresponding histograms in Figure (4.7). In case of uniform λ the posterior mean is 0.02, standard deviation is 0.01, the fifth percentile is 0.006 and the 95^{th} percentile is 0.04, where the acceptance ratio for the MCMC run is 50 percent. We find almost similar estimates for λ when we consider the uniform shrinkage prior. In this case the posterior mean is 0.025, standard deviation is 0.02, the fifth and the 95^{th} percentiles are 0.006 and 0.07, respectively, with the acceptance rate of 52 percent.

At each knot point we have 5000 posterior samples for the w's (in other words, we have 5000 posterior samples for the vector $\boldsymbol{w}$) in case of Prior 2. We have performed the same random walk Metropolis-Hastings algorithm as shown in Equation (4.2) for the parameter vector $\boldsymbol{w}$. To check the convergence of the MCMC run, we take average $\bar{w} = \sum_{i=1}^{20} w_i$ and plot the 5000 posterior realizations of $\bar{w}$ in Figure (4.8). The histograms for the last 3000 samples are also plotted in the same graph. It is known that for small values of λ, the prior distribution of w_i's is centered at 1, so we expect small values for $\bar{w}$. From Figure (4.8), it is clear that for both uniform and uniform shrinkage λ priors, $\bar{w}$ converges to 1 after approximately 2000

iterations. The fitted curves with the data points are plotted in Figure (4.9). Since the true curve is unknown, we can only make the decision of better performance by visual inspection. We see that Prior 1 and Prior 2 with both uniform and uniform shrinkage λ give similar but better fits and produce more smooth curves. From the mathematical point of view, we know that as λ goes to 0, w_i goes to unity, and consequently, Prior 2 converges to Prior 1. Our simulation results support this argument empirically when we compare the five curves for all five different choices of prior distributions in Figure (4.10).

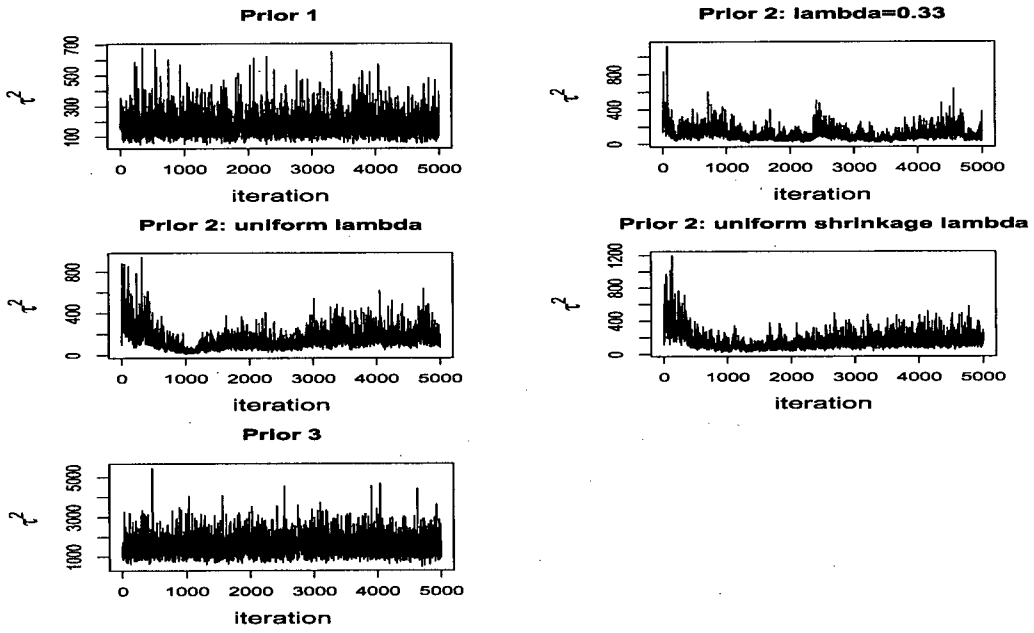

Figure 4.3: *Plots of 5000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

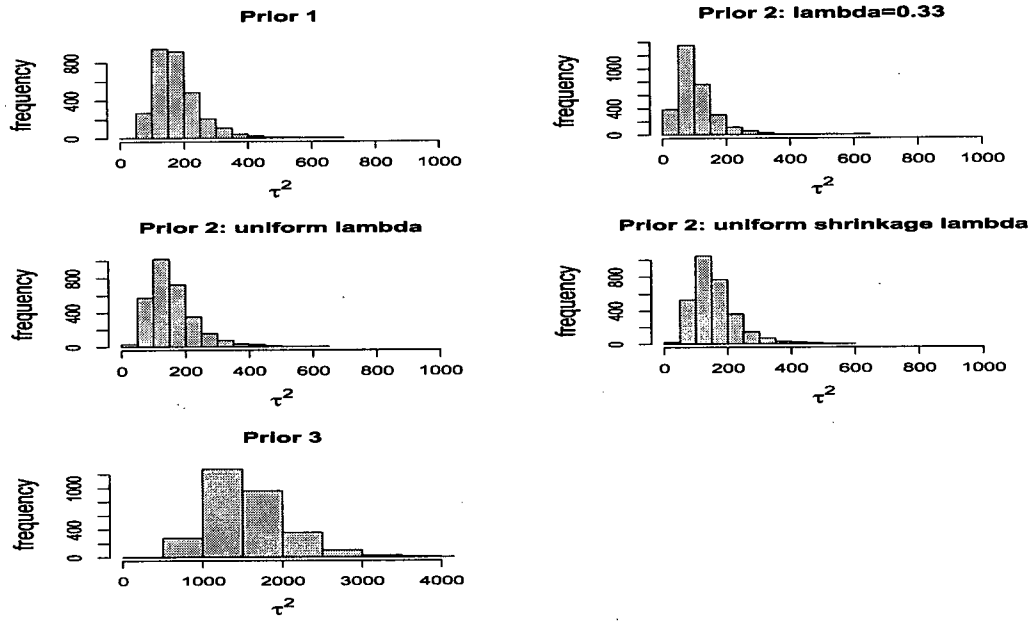

Figure 4.4: *Histograms of the last 3000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

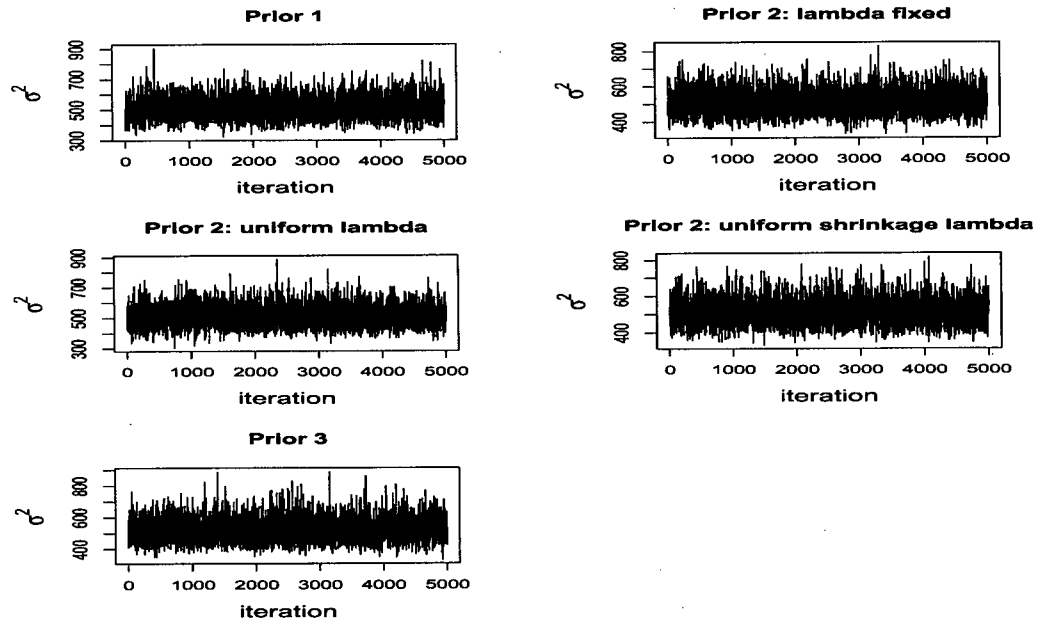

Figure 4.5: *Plots of 5000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

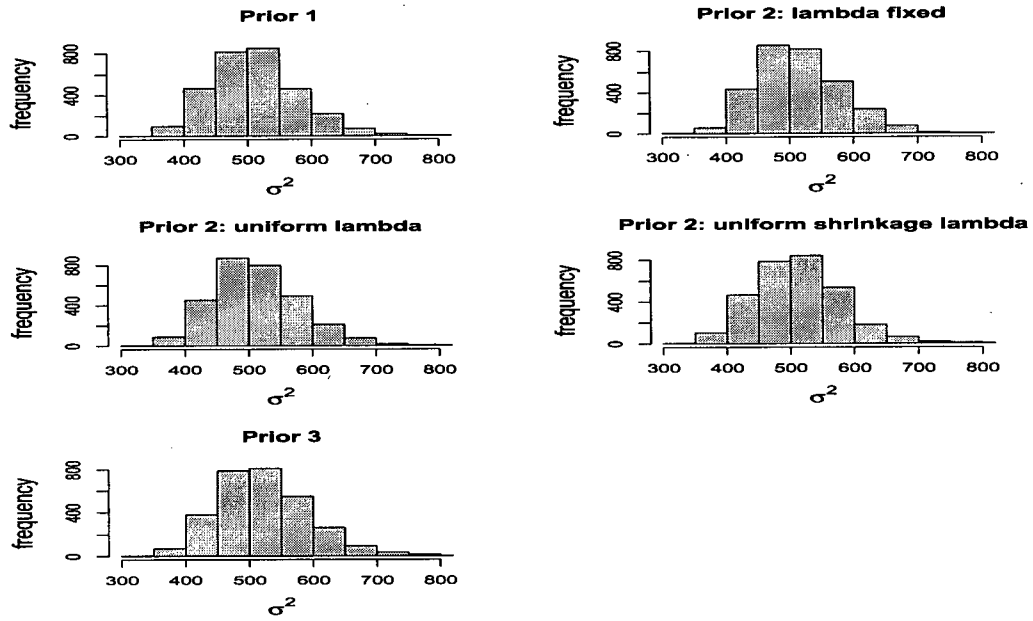

Figure 4.6: *Histograms of the last 3000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

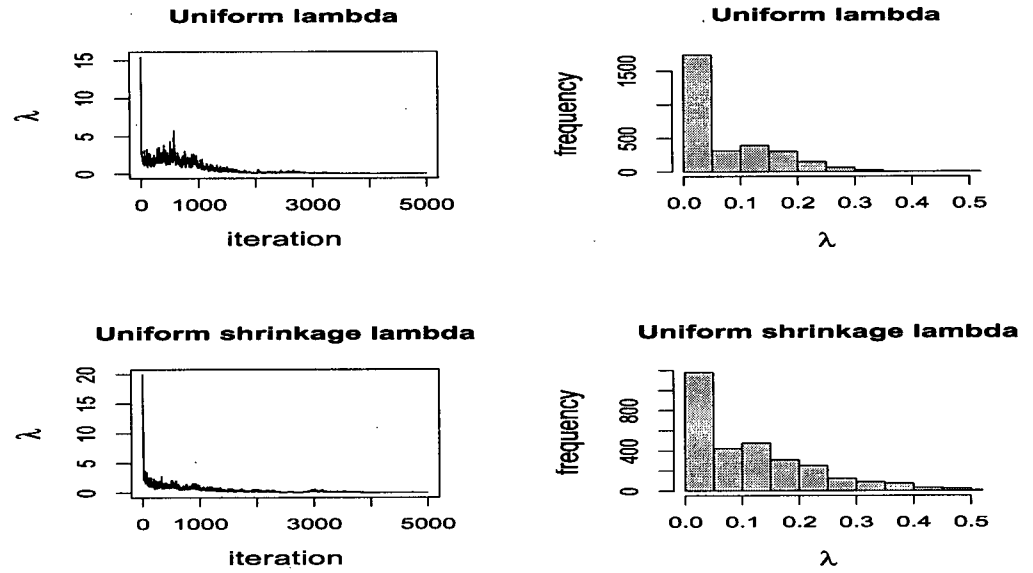

Figure 4.7: *Plots of 5000 posterior samples of the parameter λ and the histograms for the last 3000 samples.*

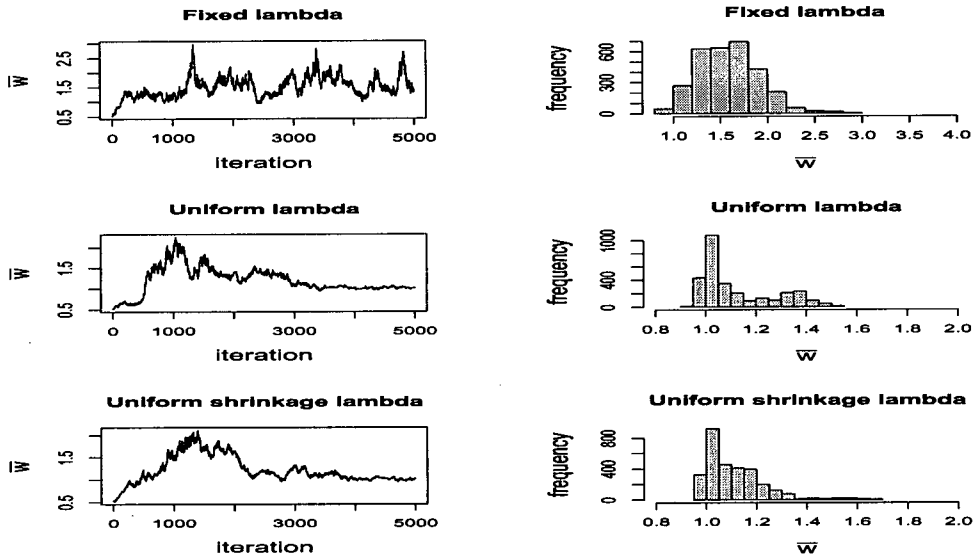

Figure 4.8: *Plots of 5000 posterior samples of mean w and the histograms for the last 3000 samples.*

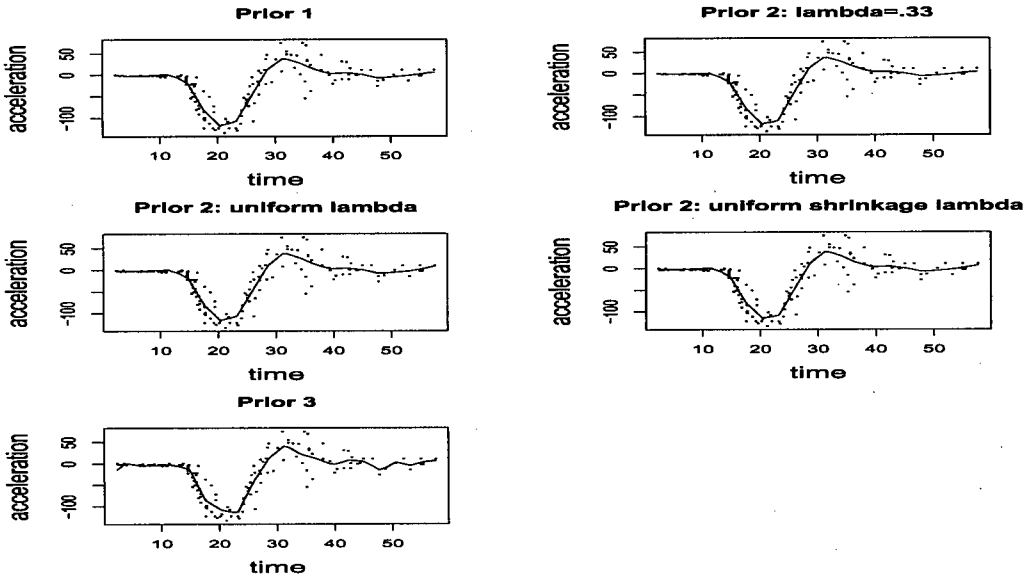

Figure 4.9: *Estimated curves for all the five choices of the prior distributions.*

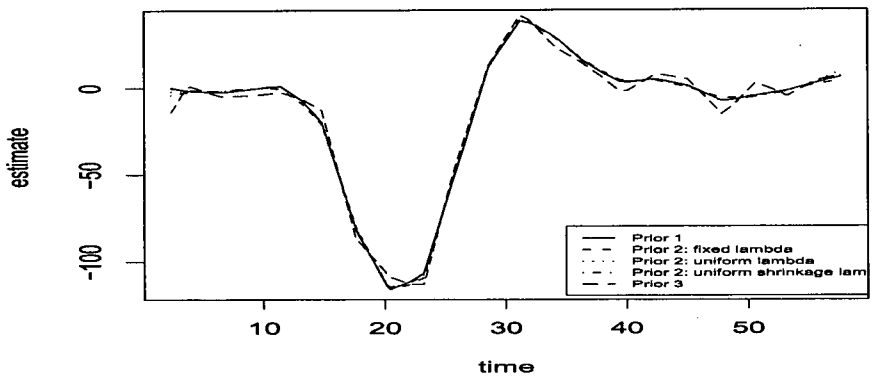

Figure 4.10: *All the five estimated curves from the five prior specifications.*

4.4 Simulated Results: Example 2

Our second data set is simulated in the following way:

$$y = \begin{cases} \sin\left(30 + \frac{x}{7.5}\right) \times 3x + \epsilon & \text{for } x = 1, 2, \cdots, 100, \\ \frac{x}{10} + \epsilon & \text{for } x = 101, 102, \cdots, 175 \text{ and} \\ 30 - 0.2 \times x + \epsilon & \text{for } x = 171, 172, \cdots, 250, \end{cases}$$

where ϵ is a random error term which follows a normal distribution. In particular, we choose $\epsilon \sim N(0, 100)$. Hence we have a data set of size $n = 250$ which is plotted in Figure (4.11). The response is generated in such a way that some parts are linear and some parts are curvilinear. From the generated data, the model parameters $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, $\boldsymbol{w}$, τ^2, σ^2 and λ are estimated by the Bayesian technique that makes the use of our proposed prior distributions for known and fixed number of knots $p = 20$. We use the same hyperparameters as in Example 1, i.e., we use $\tau^2 \sim IG(3,\ 0.01)$ and $\sigma^2 \sim IG(0.001,\ 0.001)$.

4.4.1 Parameter Estimates

As in the case of Example 1, the Gibbs sampler technique has been applied for posterior simulation of $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, τ^2 and σ^2, and a random walk Metropolis-Hastings algorithm has been

used for $\boldsymbol{w}$ and λ. We present the parameter estimates for all the five specifications of the prior distributions in five tables (Tables 4.6-4.10). Estimates of the parameter vector $\boldsymbol{\theta}$ are presented graphically. Figure (4.12) shows the 60 (random) posterior realizations (grey curves) for the parameter vector $\boldsymbol{\theta}$ and the means (dark curves) for all the five choices of the prior distributions. Apparently, from the plots we can say that the estimates from Prior 1 to Prior 2 do not differ. Although Prior 3 gives a less smooth mean curve, the standard errors of the estimates are lower (i.e., less variation among the 60 estimates) as compared to the other four cases. For this data set we also see that the estimates of the data variance σ^2 are approximately equal for all five choices of the priors. Estimates of τ^2 differ greatly: the non-roughness penalty prior (i.e., Prior 3) gives the biggest τ^2; on the other hand, Prior 2 with fixed λ gives the smallest τ^2. Since the 90% credible intervals contain zero, we can say that the estimates of the linear regression parameters δ_0 and δ_1 are statistically insignificant in all five cases.

4.4.2 Monitoring the Convergence of MCMC Simulation

In this example we also run MCMC simulations for 5000 iterations to get posterior samples for all the model parameters, from which the first 2000 iterations are discarded as burn-in period of the chain. Figure (4.13) and (4.15) present 5000 posterior realizations for the

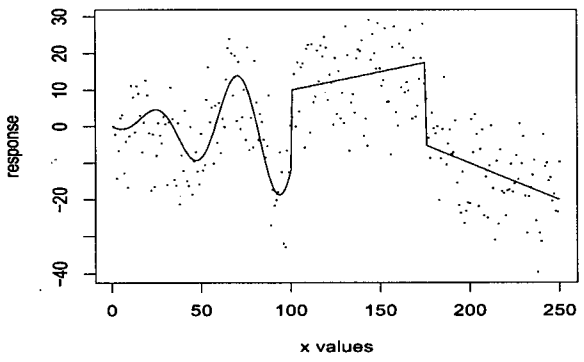

Figure 4.11: *The simulated data 1.*

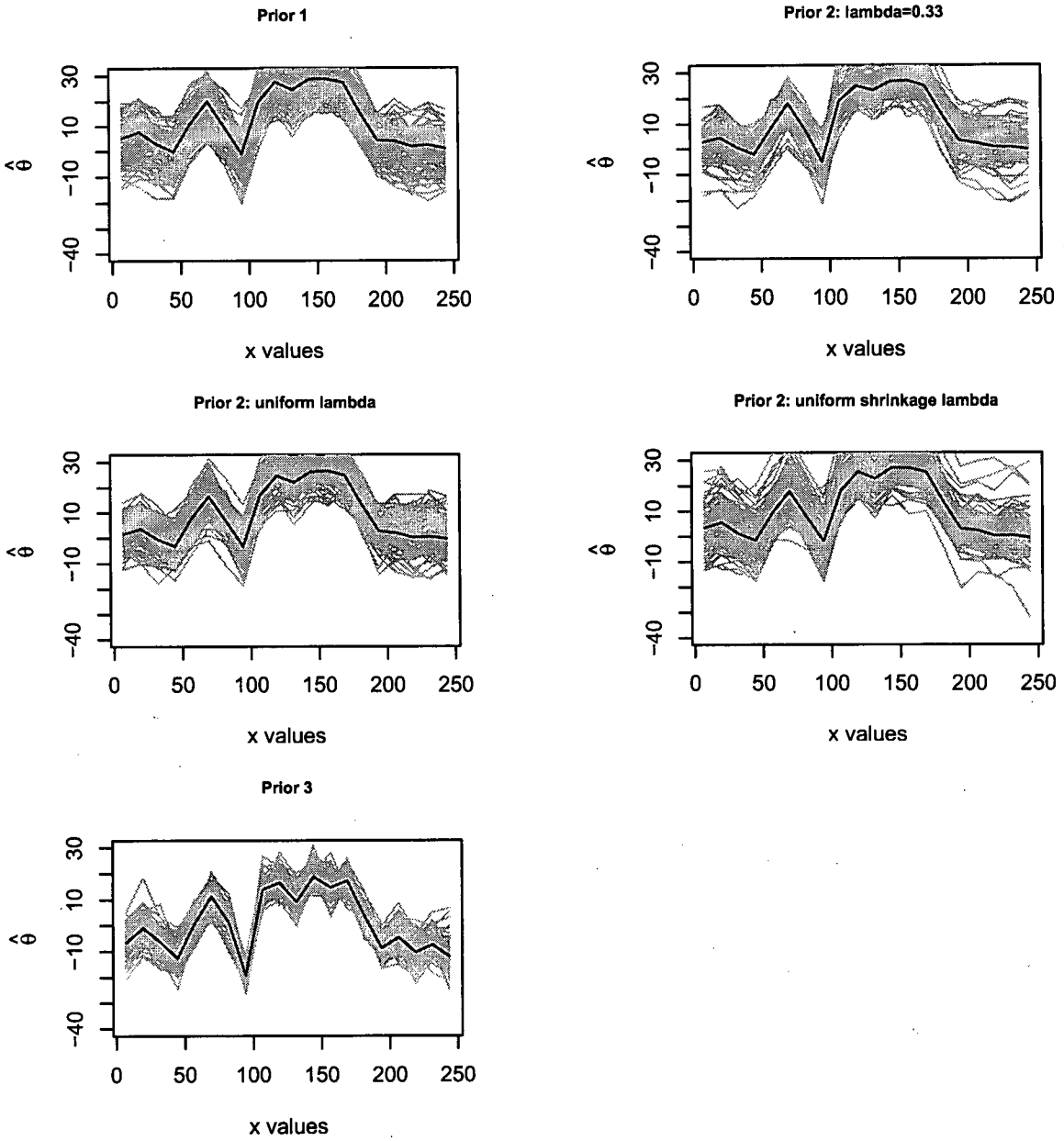

Figure 4.12: *Sixty posterior realizations (grey curves) for the parameter vector* $\boldsymbol{\theta}$. *Dark curves show the posterior means.*

Table 4.6: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 1.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	33.37	17.07	12.31	65.42
δ_0	-6.69	7.68	-18.48	6.85
δ_1	-0.05	0.04	-0.11	0.01
σ^2	93.49	9.45	78.81	110.05

Table 4.7: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with fixed λ ($\lambda = 0.33$).

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	18.75	10.07	6.41	37.43
δ_0	-3.93	6.65	-14.82	7.16
δ_1	-0.56	0.04	-0.12	0.02
σ^2	92.69	9.09	78.79	108.10

parameters τ^2 and σ^2, respectively. We also plot the corresponding histograms for the last 3000 posterior samples in Figure (4.14) and Figure (4.16). The MCMC chains for λ and the corresponding histograms are shown in Figure (4.17). In this figure, we observe that after approximately 1000 iterations, λ nicely converges to 0. In this example, we also get the similar estimates of λ for both the uniform and uniform shrinkage priors. In case of uniform prior, the posterior mean for the parameter λ is 0.03, standard deviation is 0.06, the fifth percentile is 0.006 and the 95^{th} percentile is 0.16. The acceptance rate for the MCMC run is 52 percent. We calculate $\bar{w}$ for all the three cases of Prior 2 and plot the posterior samples in Figure (4.18). Similar to the case of Example 1, it is also observed that $\bar{w}$ converges to unity after approximately 2000 iterations.

The true and fitted curves with the data points are plotted in Figure (4.19). Since the true curve is known, we can make the decision of better performance by comparing the estimated

Table 4.8: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform λ.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	29.48	16.33	9.18	59.97
δ_0	-2.90	6.27	-12.42	7.75
δ_1	-0.06	0.03	-0.11	0.0002
σ^2	94.50	9.91	79.76	111.82

Table 4.9: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform shrinkage λ.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	29.73	18.14	6.85	62.72
δ_0	-3.22	7.77	-15.91	9.24
δ_1	-0.06	0.04	-0.11	0.01
σ^2	95.23	10.19	80.30	113.07

curves with it. We have generated this kind of wiggly data set in order to have a clear snapshot of the behaviors of the prior distributions. Since our assertion is that Prior 2 will produce better curves for estimating wiggly functions, we will have a more smooth curve from it. Figure (4.20) shows that Prior 2 with both uniform and uniform shrinkage λ give better estimates of the true curve, although the differences of the estimates obtained from Prior 1 and Prior 2 are very small.

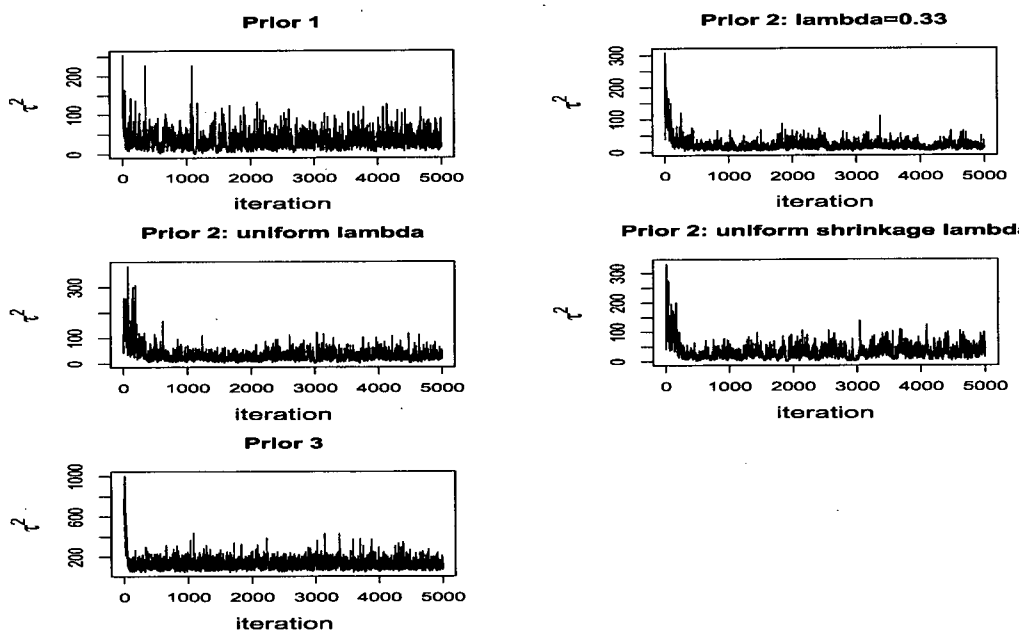

Figure 4.13: *Plots of 5000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

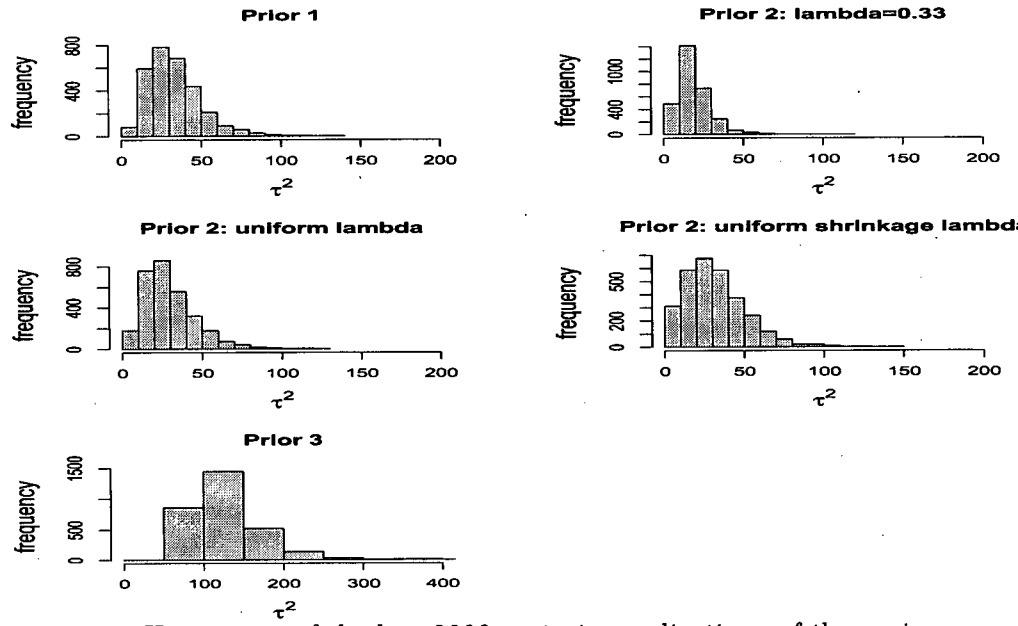

Figure 4.14: *Histograms of the last 3000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

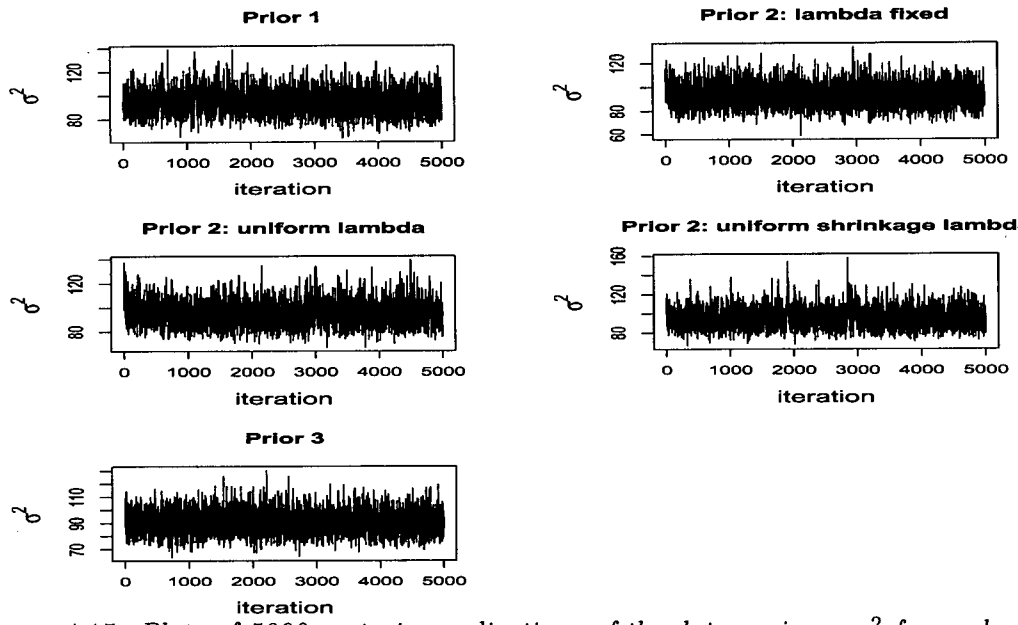

Figure 4.15: *Plots of 5000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

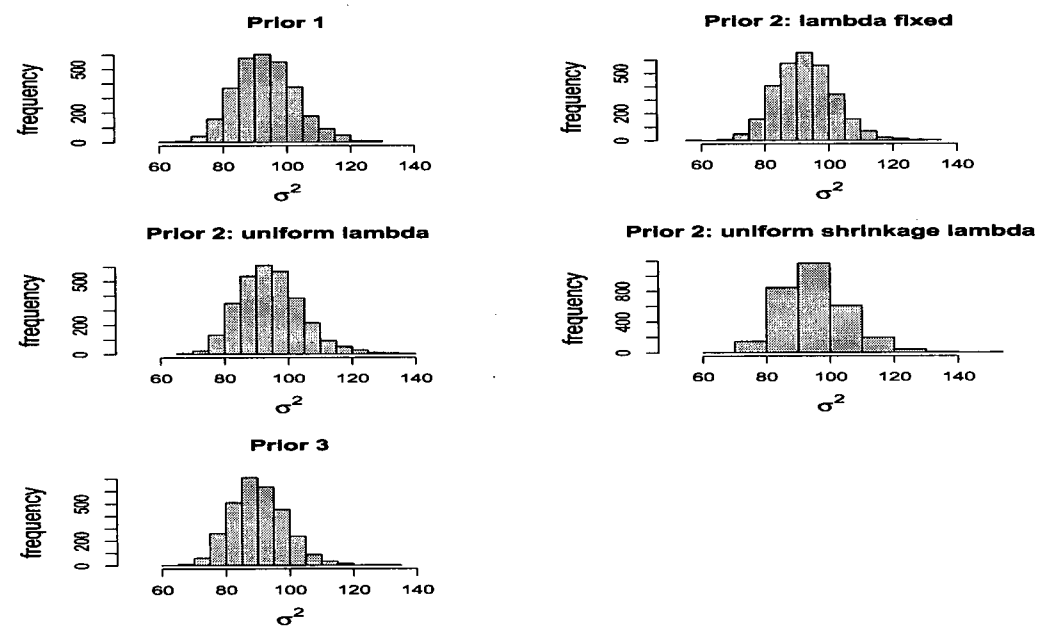

Figure 4.16: *Histograms of the last 3000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

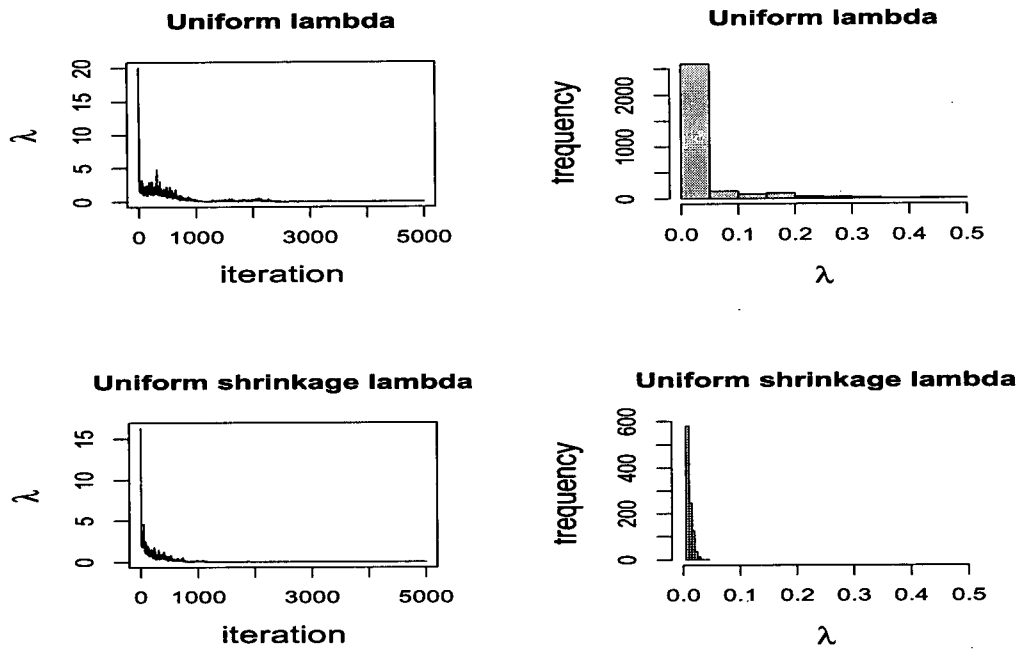

Figure 4.17: *Plots of 5000 posterior samples of the parameter λ and the histograms for the last 300 samples.*

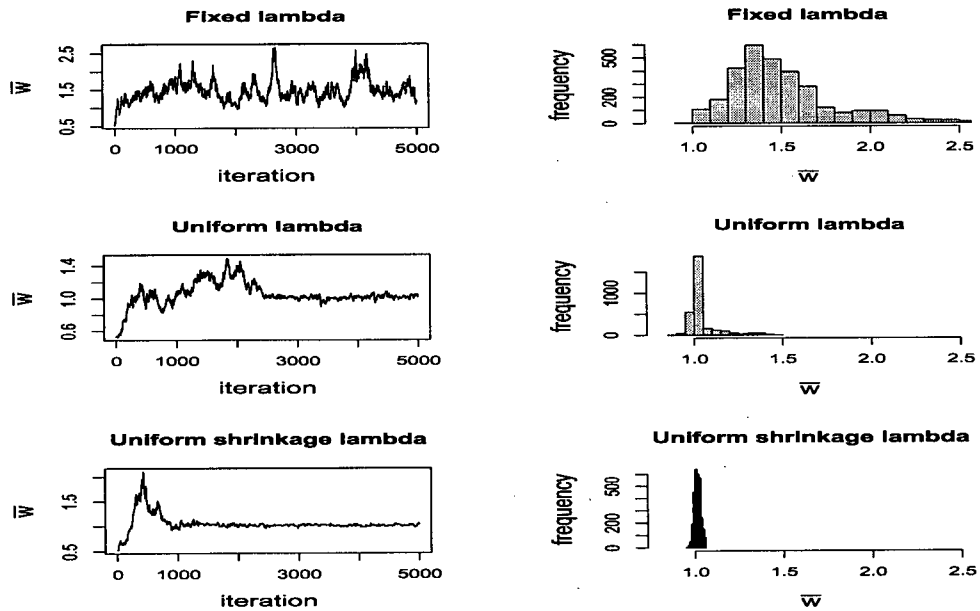

Figure 4.18: *Plots of 5000 posterior samples of mean w and the histograms of the last 3000 samples.*

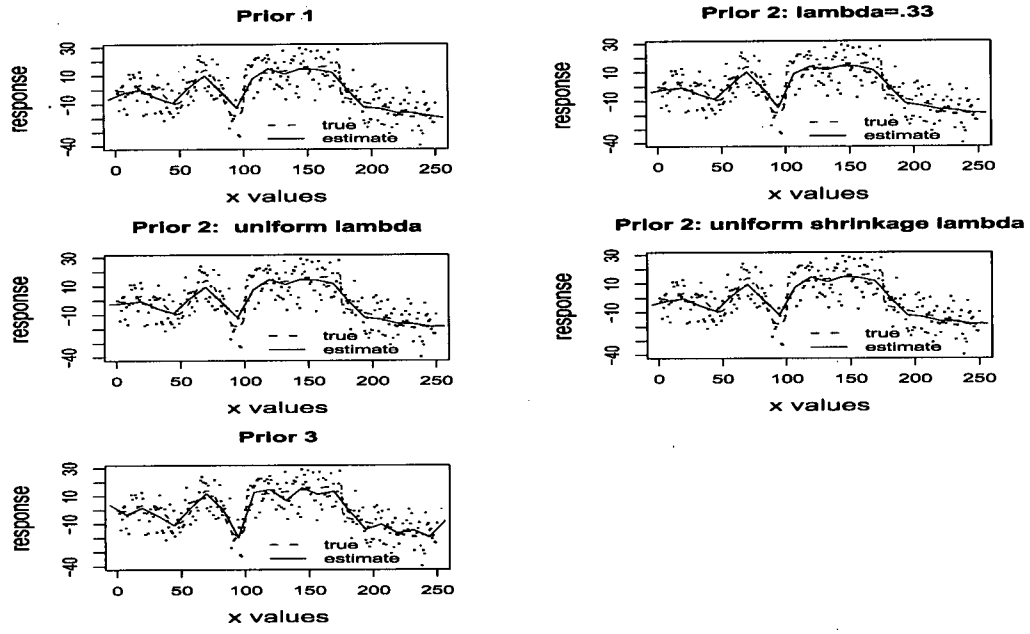

Figure 4.19: *True and the estimated curves of all five choices of the prior distributions.*

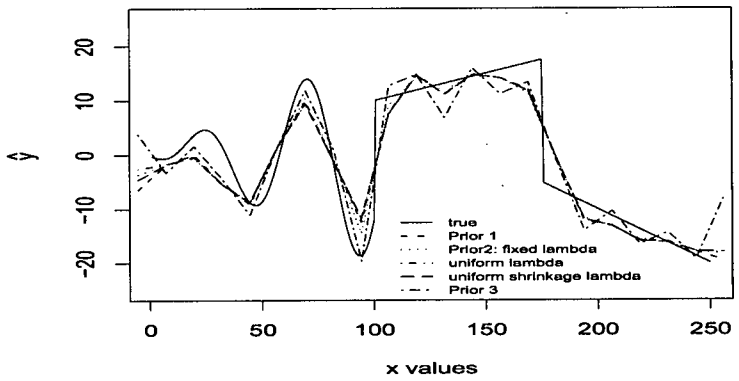

Figure 4.20: *All five estimated curves with the true curve.*

Table 4.10: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 3.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	126.29	43.87	71.61	207.16
δ_0	3.61	5.39	-6.29	11.91
δ_1	-0.04	0.03	-0.09	0.01
σ^2	90.20	8.42	77.19	104.51

4.5 Simulated Results: Example 3

In the previous two examples we have considered data sets with much wiggleness. That is, the underlying function fluctuates at different knot points. For comparison of the proposed priors we now consider a third data set with very small data variance and the underlying curve is not so rough. The data are generated in the following way:

$$y = (x - 3.5)^2 + \epsilon \text{ for } 0 \leq x \leq 8,$$

where ϵ is a random error term and follows standard normal distribution. In this case we have a random sample of size $n = 81$ and we plot the data in Figure (4.21). From the figure it is clear that the data variance is almost the same and the underlying curve is not wiggly. We set the number of knots $p = 10$ and estimate all the model parameters with the Bayesian technique using the proposed prior distributions. We use the same specifications for the hyperprior parameters as in the previous two examples.

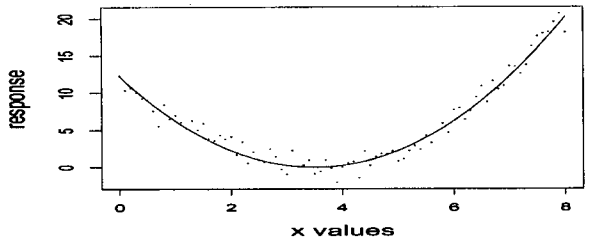

Figure 4.21: *The simulated data 2.*

4.5.1 MCMC Run and Parameter Estimates

We also draw 5000 posterior samples for all the model parameters by performing MCMC runs. In this example, we are not showing all the parameter estimates and the figures regarding the convergence and mixing of MCMC runs. We only present 60 posterior samples (systematically chosen from the last 3000 samples) of the parameter vector $\boldsymbol{\theta}$ with the mean curves in Figure (4.22), the true and the fitted curves with the data points in Figure (4.23), and all the estimated curves with the true curve in Figure (4.24). Although all the priors give similar estimates as the true curve, by a close look at the figures, we can say that both Prior 1 and Prior 2 work better than Prior 3. As we see in the previous chapters, Prior 2 needs much more computation than Prior 1 and hence, in case of simple curve estimation, we would rather use Prior 1 for Bayesian curve estimation technique.

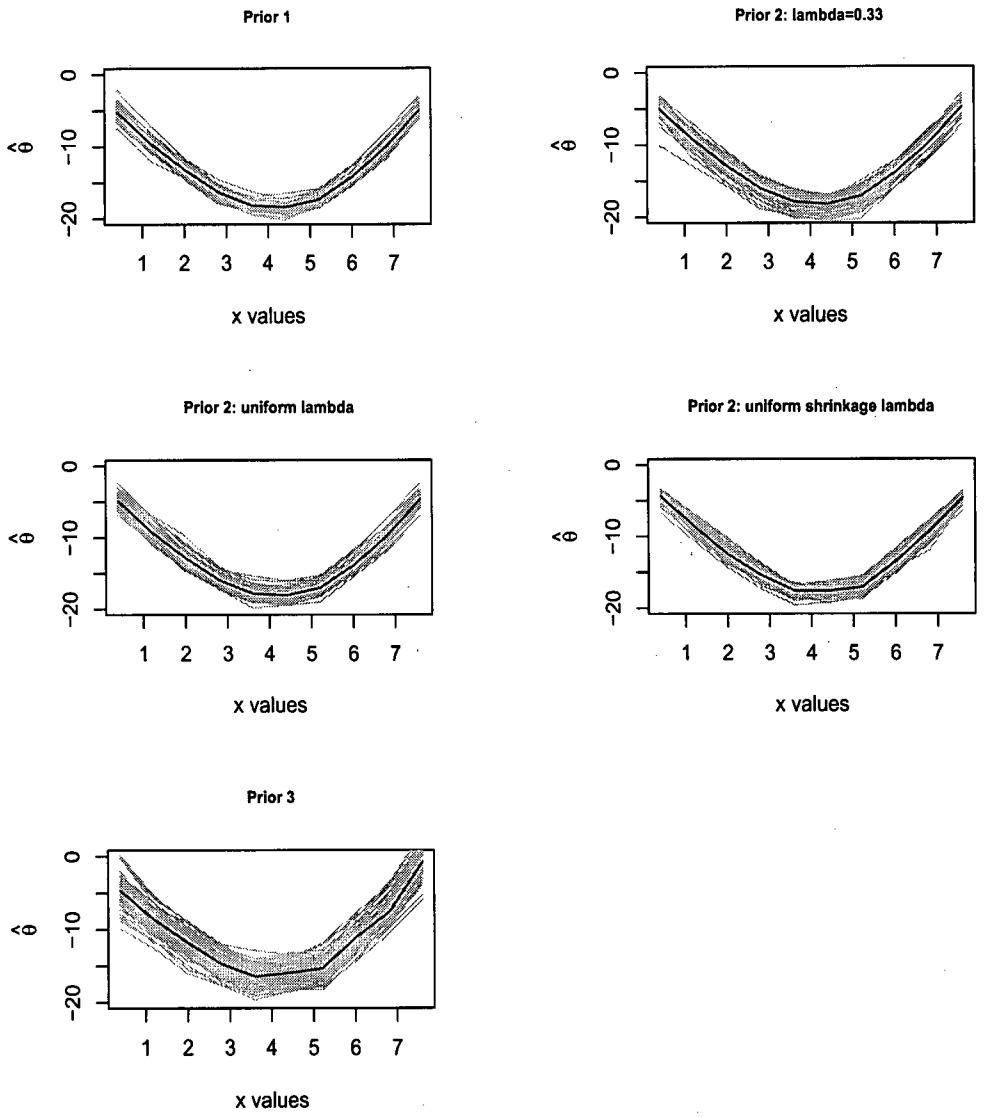

Figure 4.22: *Sixty posterior realizations (grey curves) for the parameter vector* $\boldsymbol{\theta}$. *Dark curves show the posterior means.*

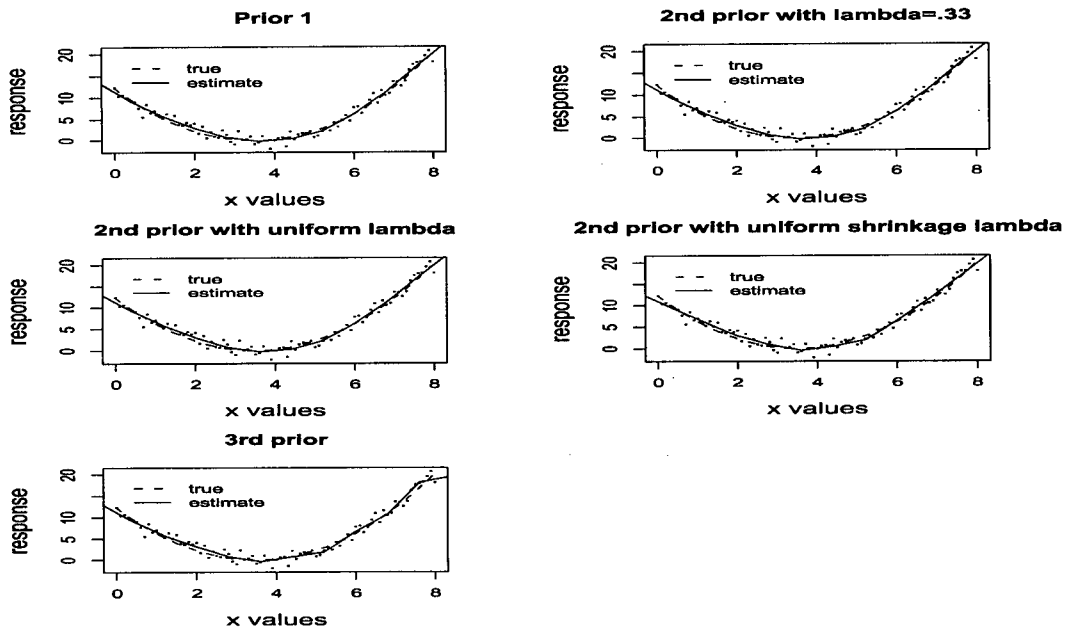

Figure 4.23: *True and estimated curves of five choices of the prior distributions.*

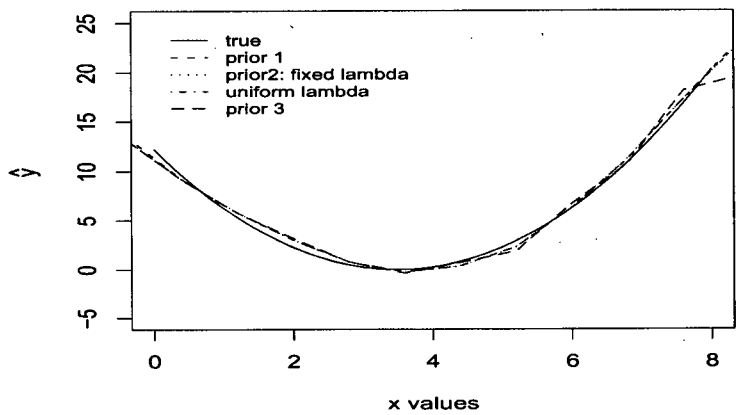

Figure 4.24: *All five estimated curves with the true curve.*

4.6 Discussion

In this chapter three simulation studies have been done for the Bayesian curve fitting using hierarchical models with our three roughness penalty prior distributions. We perform MCMC to simulate 5000 posterior samples for all the model parameters. The first 2000 iterations can be regarded as the burn-in period of the chain, and therefore, they have been discarded for the chain. The remaining 3000 samples are used for parameter estimation and curve fitting. From the plots of posterior realizations we observe good convergence and mixing of the chains. In all the three examples, we find similar performance of the proposed prior distributions, where Prior 1 and Prior 2 show similar but better performance than Prior 3. However, we have different expectations regarding the performance of the proposed prior distributions. This is because Prior 1 estimates the curve considering the overall roughness penalty of the whole data set, while Prior 2 estimates the curve considering the local roughness penalty at each knot point, but Prior 3 does not consider any roughness penalty to estimate the curve. Therefore, if the data is not too rough, we do not expect much variation among the performance of the three priors. But, if the data are very rough (wiggle), Prior 3 should produce the worst estimate among the three. Also, if the wiggleness of the data varies at different knot points, Prior 2 should produce better estimates than Prior 1. On the basis of the examples we reject our initial hypothesis that in the case of much wiggleness of the data, Prior 2 produces the best estimates among the three, but accept the fact that both local and global roughness penalty priors produce the same smooth fit of the curve in case of Bayesian curve fitting using hierarchical models. In this chapter, all the inferences are drawn under the piecewise linear spline assumption. In the next chapter, we will continue the application of Bayesian hierarchical models under the assumption of natural cubic spline with roughness penalty approach.

Chapter 5

MCMC Simulation for Natural Cubic Spline

5.1 Introduction

The Bayesian curve fitting with our proposed prior distributions under the assumption of piecewise linear splines has been discussed in Chapter 4. In this chapter, we also provide a Bayesian curve fitting which models the regression function f by a natural cubic spline (NCS) with a known number of equidistant knots. For our three different roughness penalty prior distributions of the parameter $\boldsymbol{\theta}$, discussed in Chapter 3, we have found three precision matrices, B, $B_{w,\tau}$ and I, in the case of piecewise linear splines. Modifications in the precision matrices B and $B_{w,\tau}$ are necessary in order to apply the prior distributions to the natural cubic spline. In section 5.2 we present the modified prior distributions and the corresponding posterior distributions. The simulation results of two examples are presented in section 5.3 and 5.4, and we give a brief discussion in section 5.5.

5.2 Modified Prior and Posterior Distributions

In the case of a natural cubic spline with roughness penalty, the penalized sum of squares can be written as

$$S = \sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \frac{1}{\tau^2} \int_a^b \{f''(x)\}^2 dx. \tag{5.1}$$

Here we use the smoothing parameter $1/\tau^2$ instead of τ^2 in equation (2.3) of chapter 2 for our convenience. In matrix notation equation (5.1) can be written as

$$\begin{aligned} S &= (\boldsymbol{Y} - \boldsymbol{f})'(\boldsymbol{Y} - \boldsymbol{f}) + \frac{1}{\tau^2} \boldsymbol{f}' K \boldsymbol{f} \\ &= (\boldsymbol{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta})'(\boldsymbol{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta}) + \frac{1}{\tau^2} \boldsymbol{f}' K \boldsymbol{f} \\ &\approx (\boldsymbol{Z} - A\boldsymbol{\theta})'(\boldsymbol{Z} - A\boldsymbol{\theta}) + \frac{1}{\tau^2} \boldsymbol{\theta}' K \boldsymbol{\theta}, \end{aligned}$$

where K is the $(p+2) \times (p+2)$ roughness penalty matrix (Green and Silverman, 1994) and $\boldsymbol{Z} = \boldsymbol{Y} - D\boldsymbol{\delta}$. We get the roughness penalty term $(1/\tau^2)\boldsymbol{\theta}' K \boldsymbol{\theta}$ from $(1/\tau^2)\boldsymbol{f}' K \boldsymbol{f}$ by separating the linear and smooth parts of f and constraining the smooth part to be zero at the exterior knots.

5.2.1 Prior 1

For NCS, Prior 1 can be written as $\pi(\boldsymbol{\theta}; \tau^2) = \pi(\boldsymbol{\theta}|\tau^2)\pi(\tau^2)$ where $\pi(\boldsymbol{\theta}|\tau^2) \sim N(0, V^{-1})$, $V = \frac{1}{\tau^2} K_1$, and K_1 is the $p \times p$ matrix constructed from the matrix K excluding the rows and colums corresponding to the two exterior knots (i.e., deleting the first row and the first column, and the last row and the last column). The prior distributions for the parameters τ^2, σ^2 and $\boldsymbol{\delta}$ remain the same as before. Hence, by replacing $(1/\tau^2)M$ by V, we calculate the conditional posterior distributions. We find that $\boldsymbol{\theta}$ given $\boldsymbol{\delta}$, τ^2, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, is multivariate normal with mean vector $(V + (A'A)^{-1}/\sigma^2)^{-1} A'\mathbf{Z}_1/\sigma^2$ and covariance matrix $(V + (A'A)^{-1}/\sigma^2)^{-1}$, where $\mathbf{Z}_1 = \mathbf{Y} - D\boldsymbol{\delta}$. The conditional posterior distribution for the variance component τ^2 given $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, is inverse gamma with parameters $\alpha + p/2$ and $(\boldsymbol{\theta}' V \boldsymbol{\theta})/2 + \beta$. The conditional posterior distribution of the linear regression parameter $\boldsymbol{\delta}$ given $\boldsymbol{\theta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, and the conditional posterior distribution of the data variance σ^2

given $\boldsymbol{\delta}$, $\boldsymbol{\theta}$, τ^2, $\mathbf{Y}$ and $\mathbf{X}$ remain same as in the case of piecewise linear splines. We get the closed form solutions for all the conditional posterior distributions. Hence we will apply the Gibbs Sampler algorithm to draw the posterior samples.

5.2.2 Prior 2

In the case of the second prior we have assumed different variance components at different knot points. In particular, we have variances $\tau_1^2, \tau_2^2, \ldots, \tau_p^2$ for the parameters $\theta_1, \theta_2, \ldots, \theta_p$ where τ_i^2 measures the roughness of the curve at knot t_i. Hence it is logical to modify the roughness penalty term as follows:

$$\sum_{i=0}^{p+1} \frac{1}{\tau_i^2} \int_{t_{i-1}}^{t_i} \{f''(x_i)\}^2 dx_i. \tag{5.2}$$

Unlike the case of piecewise linear splines, it is not easy to express (5.2) in matrix form. We are not going to do that tedious mathematical operation. However, in the light of our expression in the case of piecewise linear splines, we can approximate (5.2) as:

$$\sum_{i=0}^{p+1} \frac{1}{\tau_i^2} \int_{t_{i-1}}^{t_i} \{f''(x_i)\}^2 dx_i \approx \boldsymbol{f}' K_\tau \boldsymbol{f}, \tag{5.3}$$

where K_τ is a $(p+2) \times (p+2)$ matrix obtained by pre- and post- multiplications of the K matrix by a diagonal Δ_τ matrix as follows:

$$\begin{aligned} K_\tau &= \begin{bmatrix} \frac{1}{\tau_0} & 0 & \ldots & 0 \\ 0 & \frac{1}{\tau_1} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \frac{1}{\tau_{p+1}} \end{bmatrix} \begin{bmatrix} k_{00} & k_{01} & \ldots & k_{0,(p+1)} \\ k_{10} & k_{11} & \ldots & k_{1,(p+1)} \\ \vdots & \vdots & & \vdots \\ k_{(p+1),0} & k_{(p+1),1} & \ldots & k_{(p+1),(p+1)} \end{bmatrix} \begin{bmatrix} \frac{1}{\tau_0} & 0 & \ldots & 0 \\ 0 & \frac{1}{\tau_1} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & \frac{1}{\tau_{p+1}} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1}{\tau_0^2} k_{00} & \frac{1}{\tau_0 \tau_1} k_{01} & \cdots & \frac{1}{\tau_0 \tau_{(p+1)}} k_{0,(p+1)} \\ \frac{1}{\tau_1 \tau_0} k_{10} & \frac{1}{\tau_1^2} k_{11} & \cdots & \frac{1}{\tau_1 \tau_{(p+1)}} k_{1,(p+1)} \\ \vdots & \vdots & & \vdots \\ \frac{1}{\tau_{(p+1)} \tau_0} k_{(p+1),0} & \frac{1}{\tau_{(p+1)} \tau_1} k_{(p+1),1} & \cdots & \frac{1}{\tau_{(p+1)}^2} k_{(p+1),(p+1)} \end{bmatrix}. \end{aligned}$$

Since the variance components at the two exterior knots do not make any sense, we delete the first row and the first column and the last row and the last column of K_τ so that the roughness penalty matrix is

$$K_\tau = \begin{bmatrix} \frac{1}{\tau_1^2}k_{11} & \frac{1}{\tau_1\tau_2}k_{12} & \cdots & \frac{1}{\tau_1\tau_p}k_{1p} \\ \frac{1}{\tau_2\tau_1}k_{21} & \frac{1}{\tau_2^2}k_{22} & \cdots & \frac{1}{\tau_2\tau_p}k_{2p} \\ \vdots & \vdots & & \vdots \\ \frac{1}{\tau_p\tau_1}k_{p1} & \frac{1}{\tau_p\tau_2}k_{p2} & \cdots & \frac{1}{\tau_p^2}k_{pp} \end{bmatrix}.$$

Hence we assume that $\boldsymbol{\theta}$ given $\boldsymbol{\tau}^2$ has the multivariate normal distribution with mean $\mathbf{0}$ and precision matrix K_τ. The interesting feature of this parameterization is that if $\tau_1^2 = \tau_2^2 = \ldots = \tau_p^2 = \tau^2$, we get $K_\tau = (1/\tau^2)K_1$. That is if all the τ_i^2s are equal, Prior 2 is exactly equal to Prior 1. For our reparameterization $\tau_i^2 = w_i\tau^2$ we modify the precision matrix as $V_{\boldsymbol{w},\tau} = (1/\tau^2)K_{\boldsymbol{w}}$, where

$$K_{\boldsymbol{w}} = \begin{bmatrix} \frac{1}{w_1}k_{11} & \frac{1}{\sqrt{w_1w_2}}k_{12} & \cdots & \frac{1}{\sqrt{w_1w_p}}k_{1p} \\ \frac{1}{\sqrt{w_2w_1}}k_{21} & \frac{1}{w_2}k_{22} & \cdots & \frac{1}{\sqrt{w_2w_p}}k_{2p} \\ \vdots & \vdots & & \vdots \\ \frac{1}{\sqrt{w_pw_1}}k_{p1} & \frac{1}{\sqrt{w_pw_2}}k_{p2} & \cdots & \frac{1}{w_p}k_{pp} \end{bmatrix}.$$

Therefore, $\boldsymbol{\theta}$ given $\boldsymbol{w}$ and τ^2 follows a multivariate normal distribution with mean vector $\mathbf{0}$ and precision matrix $V_{\boldsymbol{w},\tau}$. For the prior distributions $\tau^2 \sim IG(\alpha,\beta)$ and $w_i \sim IG(1/\lambda, 1/\lambda)$ the conditional posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\delta},\tau^2,\boldsymbol{w},\sigma^2,\mathbf{Y},\mathbf{X})$ is a multivariate normal density with mean vector $(V_{\boldsymbol{w},\tau} + (A'A)^{-1}/\sigma^2)^{-1}A'\mathbf{Z}_1/\sigma^2$ and the covariance matrix $(V_{\boldsymbol{w},\tau} + (A'A)^{-1}/\sigma^2)^{-1}$, and the conditional posterior distribution for τ^2 given $\boldsymbol{\theta}$, $\boldsymbol{w}$, $\boldsymbol{\delta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, is inverse gamma with the parameters $\alpha + p/2$ and $(\boldsymbol{\theta}'V_{\boldsymbol{w},\tau}\boldsymbol{\theta})/2 + \beta)$. To calculate the conditional posterior distribution of $\boldsymbol{w}$ we write $\pi(\boldsymbol{w}|\boldsymbol{\theta},\tau^2) \propto \pi(\boldsymbol{\theta}|\tau^2,\boldsymbol{w})\pi(\boldsymbol{w})$. Since the w_i's are independent and identically distributed with inverse gamma, the joint posterior

distribution of $w_1, \cdots, w_p$, given $\boldsymbol{\theta}$ and τ^2, can be written as

$$\begin{aligned}\pi(w_1, \cdots, w_p|\boldsymbol{\theta}, \tau^2) &\propto \pi(\boldsymbol{\theta}|\tau^2, \boldsymbol{w})\pi(w_1) \cdots \pi(w_p) \\ &\propto |V_{\boldsymbol{w},\tau}|^{\frac{1}{2}} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}' V_{\boldsymbol{w},\tau}\boldsymbol{\theta}\right\} \\ &\times \prod_{i=1}^{p}\left[\frac{(\frac{1}{\lambda})^{\frac{1}{\lambda}}}{\Gamma(\frac{1}{\lambda})}\left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)} \exp\left\{-\frac{1}{\lambda w_i}\right\}\right]. \qquad (5.4)\end{aligned}$$

We have $V_{\boldsymbol{w},\tau} = (1/\tau^2)K_{\boldsymbol{w}}$, and the matrix $K_{\boldsymbol{w}}$ can be expressed as $\Delta_{\boldsymbol{w}} K_1 \Delta_{\boldsymbol{w}}$ with

$$\Delta_{\boldsymbol{w}} = \begin{bmatrix} \frac{1}{\sqrt{w_1}} & 0 & \dots & 0 \\ 0 & \frac{1}{\sqrt{w_2}} & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{w_p}} \end{bmatrix}.$$

Hence the determinant $|V_{\boldsymbol{w},\tau}|$ can be expressed as $\prod_{i=1}^{p}\left(\frac{1}{w_i}\right)|K_1|$. The quadratic form $\boldsymbol{\theta}' V_{\boldsymbol{w},\tau}\boldsymbol{\theta}$ is equivalent to $\mathrm{tr}(V_{\boldsymbol{w},\tau}\boldsymbol{\theta}\boldsymbol{\theta}')$ and after some mathematical manipulation it can be written as $(1/\tau^2)\mathrm{tr}(K_1(\Delta_w\boldsymbol{\theta})(\boldsymbol{\theta}'\Delta_w))$ which is equal to $\frac{1}{\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{k_{ij}\theta_i\theta_j}{\sqrt{w_i w_j}}$, where k_{ij} is element corresponding to the i^{th} row and the j^{th} column of matrix K_1. Therefore, the conditional posterior distribution in expression (5.4) can be written as

$$\begin{aligned}\pi(w_1, \cdots, w_p|\boldsymbol{\theta}, \tau^2) &\propto \left(\frac{1}{\tau^2}\right)^{p/2}\prod_{i=1}^{p}\left(\frac{1}{w_i}\right)^{1/2}|K_1|^{\frac{1}{2}}\exp\left\{-\frac{1}{2\tau^2}\sum_{i=1}^{p}\sum_{j=1}^{p}\frac{k_{ij}\theta_i\theta_j}{\sqrt{w_i w_j}}\right\} \\ &\times \prod_{i=1}^{p}\left[\frac{(\frac{1}{\lambda})^{\frac{1}{\lambda}}}{\Gamma(\frac{1}{\lambda})}\left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)}\exp\left\{-\frac{1}{\lambda w_i}\right\}\right].\end{aligned}$$

From the conditional joint distribution of $w_1, \cdots, w_p$, it is difficult to find the conditional posterior distribution of the i^{th} element w_i, given all other w_js ($j \neq i$), $\boldsymbol{\theta}$ and τ^2. Hence we decide to draw posterior samples for $\boldsymbol{w}$ from the conditional joint distribution using a random walk Metropolis Hastings Algorithm. All other conditional posteriors such as $\pi(\boldsymbol{\delta}|\boldsymbol{\theta}, \sigma^2, \mathbf{Y}, \mathbf{X})$, $\pi(\sigma^2|\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \mathbf{Y}, \mathbf{X})$ and the conditional posterior for λ remain the same as found in section 3.4.2. It is obvious that all the conditional posterior distributions for the case of Prior 3 remain the same as we find in section 3.4.3.

5.3 Example 1

To illustrate the methodologies that we have described above, our first simulation study has been done for the Simulated Motorcycle Accident (Silverman, 1985) data. The additive model (4.1) is fitted with all five choices of prior distributions assuming fixed number of knots $p = 20$, and the model parameters $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, τ^2, σ^2,$\boldsymbol{w}$ and λ have been estimated. For the same specifications of the hyperparameters as defined in the previous chapter, we perform Gibbs sampling and random walk Metropolis algorithms for 5000 iterations. More specifically, we have performed Gibbs sampling technique for posterior simulation of $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, τ^2 and σ^2, and random walk Metropolis algorithm for $\boldsymbol{w}$ and λ. The estimates of the parameters τ^2, $\boldsymbol{\delta}$ and σ^2 for all five specifications of the prior distributions are presented in five tables (Tables 5.1-5.5). Twenty estimates for the parameter vector $\boldsymbol{\theta}$ are shown graphically. Figure (5.1) shows the 60 posterior samples (grey curves) for the parameter vector $\boldsymbol{\theta}$ with the means (dark curves) for all five choices of the prior distributions. The differences among the estimated mean curves produced by Prior 1 and Prior 2 are very small. However, the curves are more smooth than that of Prior 3. It may also be mentioned that Prior 3 gives estimates with small standard errors. We observe that the estimates for all the parameters are similar for all three specifications of Prior 2. Estimates of the data variance σ^2 are above 500 with standard deviation more than 60 in all the cases. Estimates of τ^2 differ greatly: for the non-roughness penalty prior (i.e., Prior 3) case τ^2 is largest, on the other hand, the smallest τ^2 is observed for Prior 2. The estimates of the linear regression parameters δ_0 and δ_1 are statistically insignificant (i.e., 90% credible intervals include zero).

5.3.1 Monitoring the Convergence of MCMC Simulation

To check the mixing and convergence we present the MCMC run for the parameters τ^2 and σ^2 in Figures (5.2) and (5.4), respectively, and to have an idea about their distribution, the corresponding histograms are plotted in Figures (5.3) and (5.5). For each of the model parameters we have run MCMC simulations for 5000 iterations, from which the first 2000

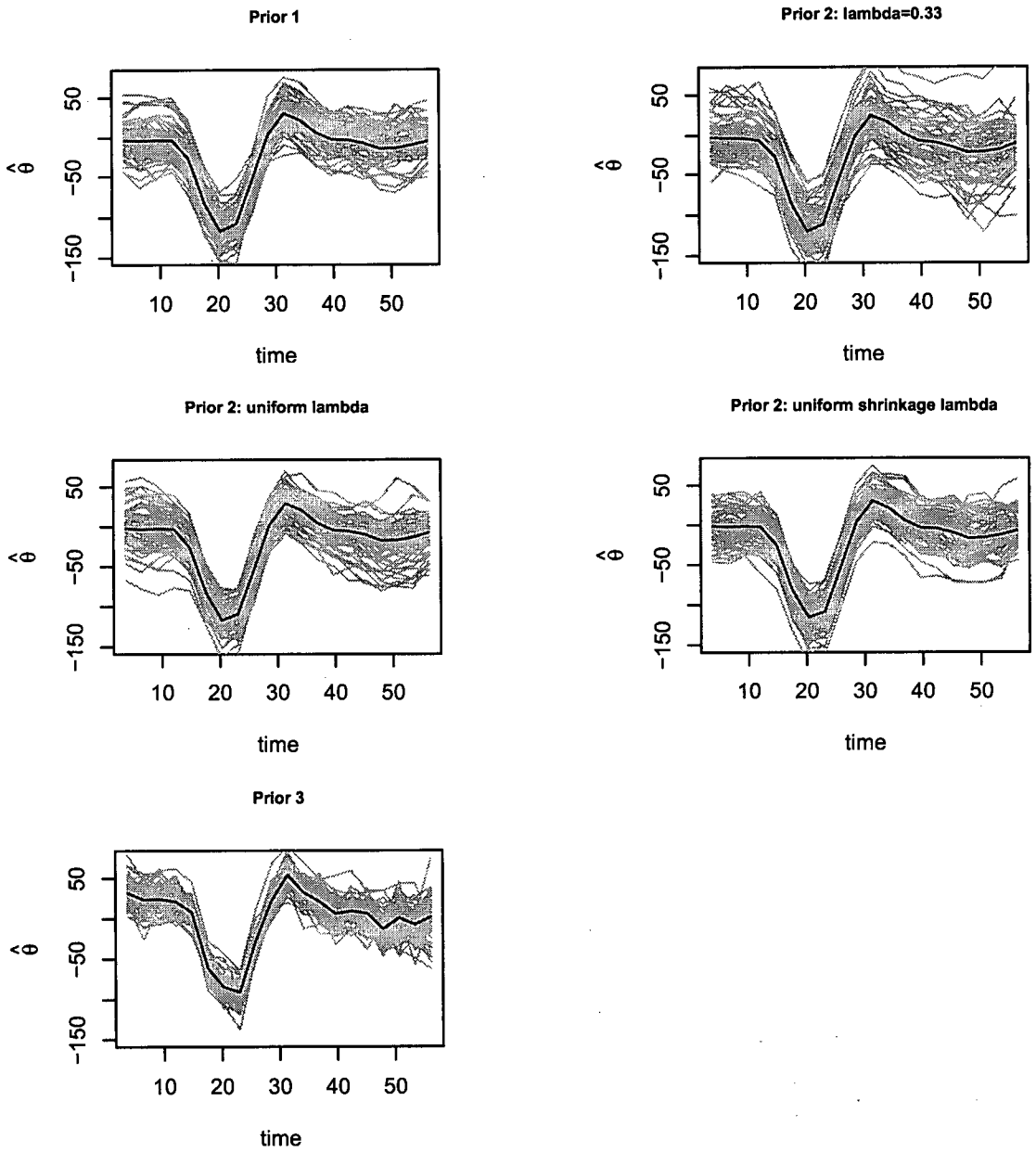

Figure 5.1: *Sixty posterior realizations (grey curves) for the parameter vector* $\boldsymbol{\theta}$. *Dark curves show the posterior means.*

Table 5.1: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 1.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	34.88	14.67	17.52	62.64
δ_0	-5.71	23.60	-45.88	31.32
δ_1	0.23	0.64	-0.87	1.26
σ^2	520.40	69.56	416.96	642.15

Table 5.2: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with fixed λ ($\lambda = 0.33$).

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	23.95	12.52	8.78	48.39
δ_0	15.10	25.98	-24.69	61.41
δ_1	-0.02	0.61	-1.07	0.92
σ^2	519.93	72.05	414.24	648.17

iterations have been deleted as burn-in period of the chain and the remaining 3000 samples are used to draw inferences. The MCMC chains for λ and the corresponding histograms are shown in Figure (5.6). In the case of uniform λ, the posterior mean is 0.02, standard deviation is 0.04, the 5^{th} percentile is 0.006 and the 95^{th} percentile is 0.1. The acceptance ratio for the MCMC run is 51 percent. We find similar estimates for λ when we consider the uniform shrinkage prior for it.

Five thousand posterior samples of $\bar{w}$ and the histograms of the last 3000 samples are plotted in Figure (5.7). The empirical results indicate that the distribution of the w_i's is centered at 1, as we see, $\bar{w}$ converges to 1 after approximately 2000 iterations for both uniform and uniform shrinkage priors of λ. On the other hand, for the fixed λ case, MCMC run does not show any convergence in 5000 iterations with an acceptance rate of 43 percent. The fitted curves with the data points are plotted in Figure (5.8). It is apparent that Prior 1 and

Table 5.3: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform λ.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	31.91	13.92	15.58	57.64
δ_0	-1.47	22.99	-41.56	32.68
δ_1	0.21	0.58	-0.73	1.13
σ^2	516.21	68.99	417.29	639.29

Table 5.4: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform shrinkage λ.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	32.98	14.95	15.47	60.89
δ_0	-1.30	20.81	-38.08	31.53
δ_1	0.18	0.60	-0.79	1.18
σ^2	516.12	67.47	414.16	631.02

Prior 2 produce exactly the same smooth curve, better than the curve produced by Prior 3, when we see all the estimates together in Figure (5.9). Therefore, we are commented on the similar behavior of Prior 1 and Prior 2 under the assumption of natural cubic spline regarding to estimation and smoothness.

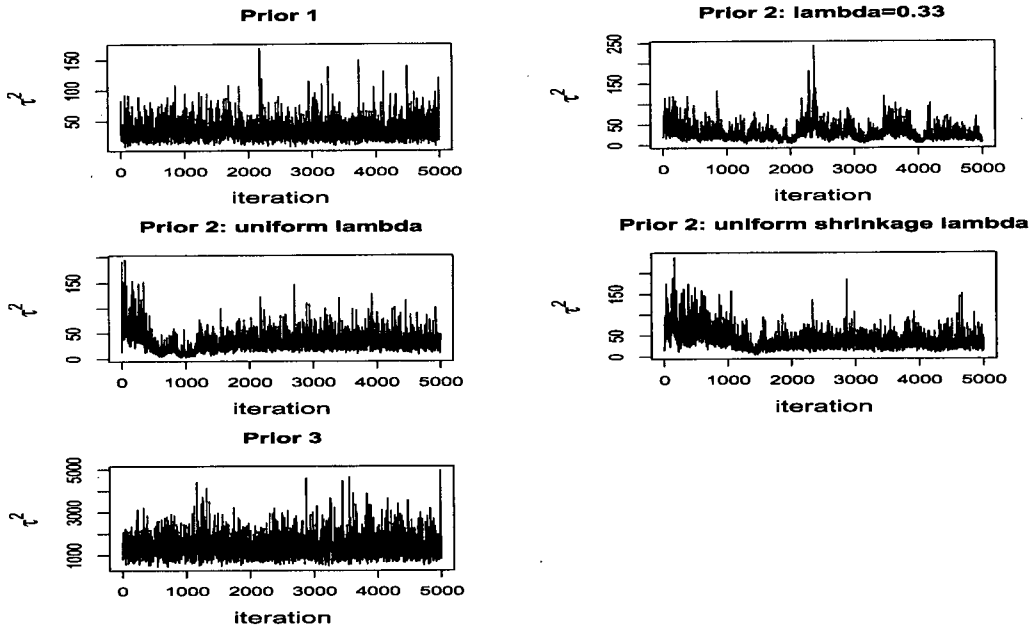

Figure 5.2: *Plots of 5000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

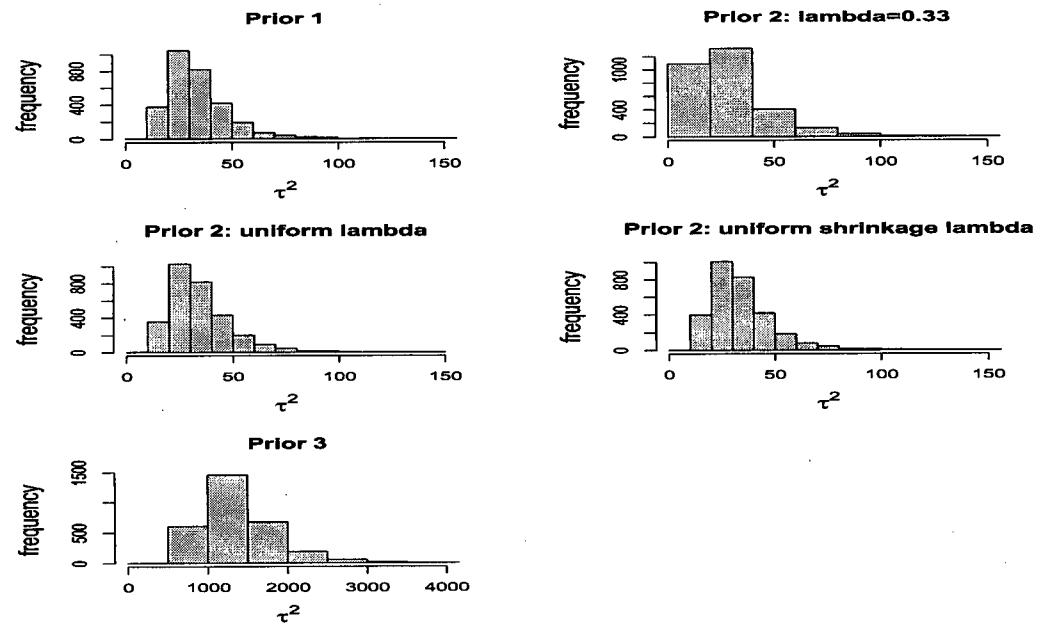

Figure 5.3: *Histograms of the last 3000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

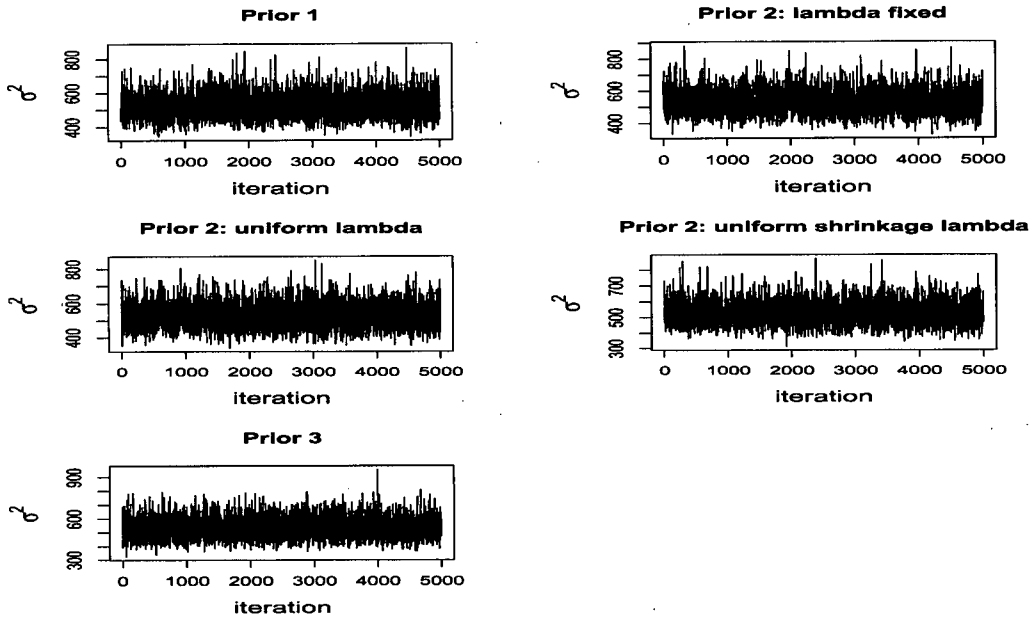

Figure 5.4: *Plots of 5000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

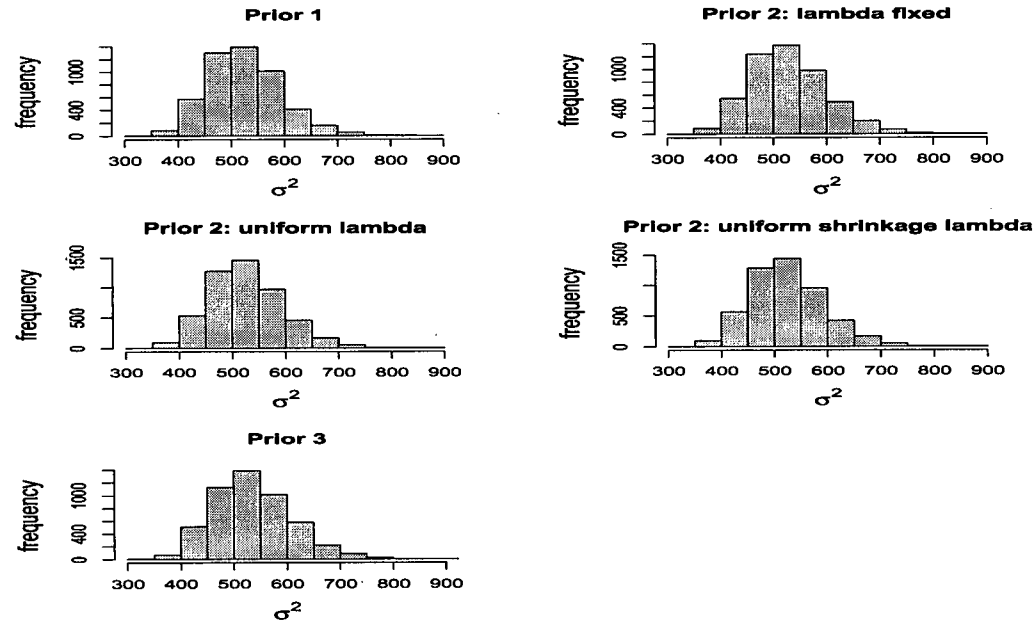

Figure 5.5: *Histograms of the last 3000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

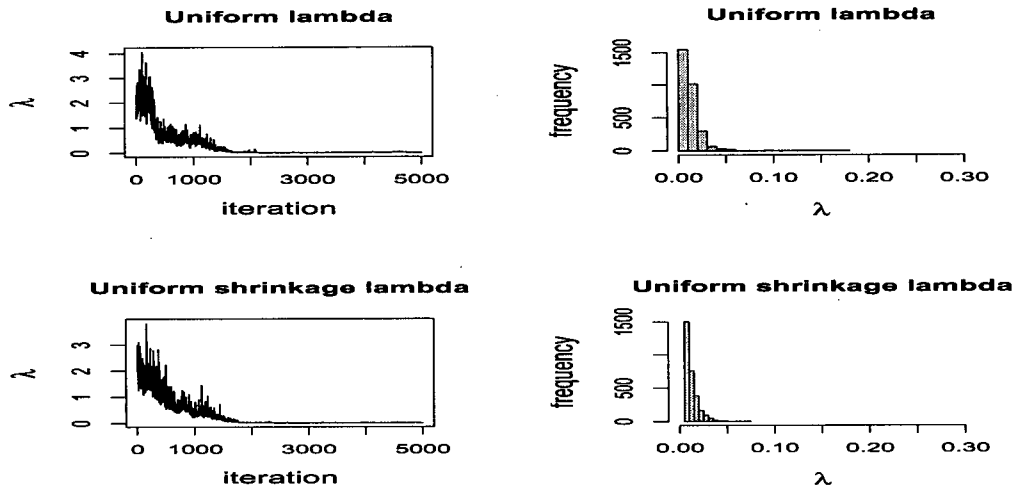

Figure 5.6: *Plots of 5000 posterior samples of the parameter λ and histograms for the last 3000 samples.*

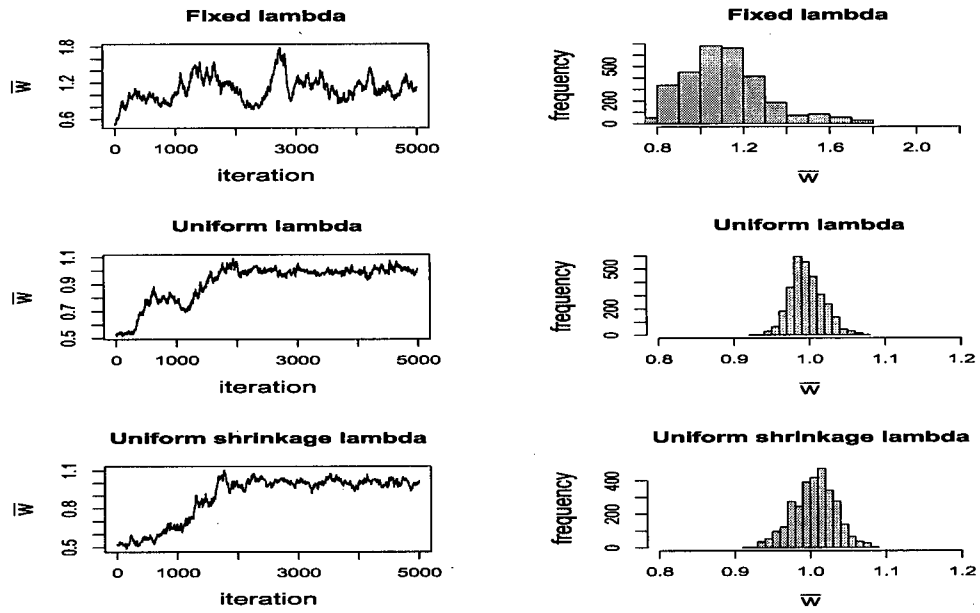

Figure 5.7: *Plots of 5000 posterior samples of mean w and histograms for the last 3000 samples.*

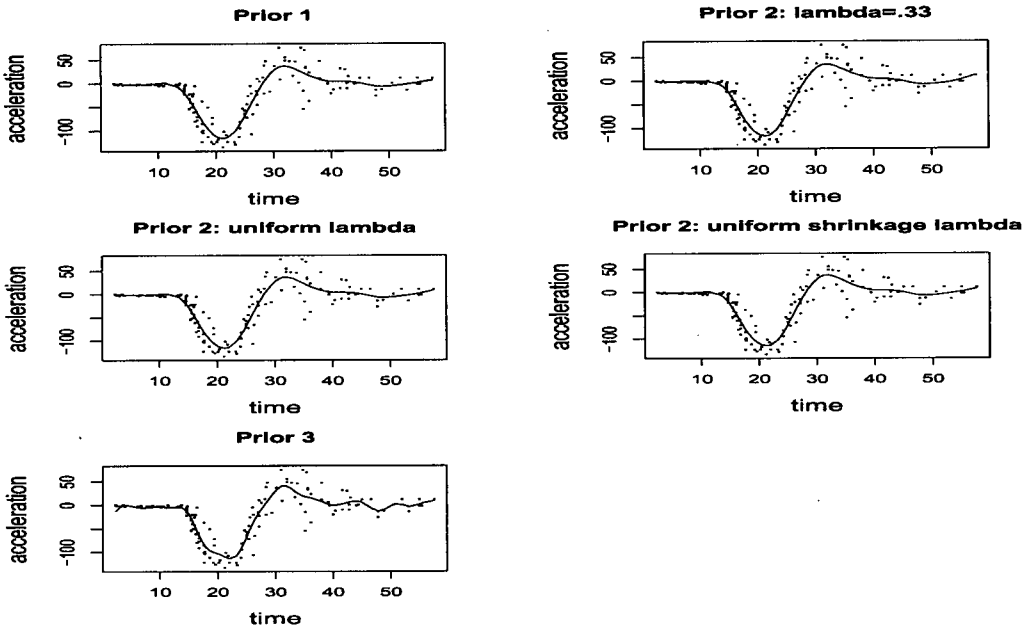

Figure 5.8: *Estimated curves for all five choices of the prior distributions.*

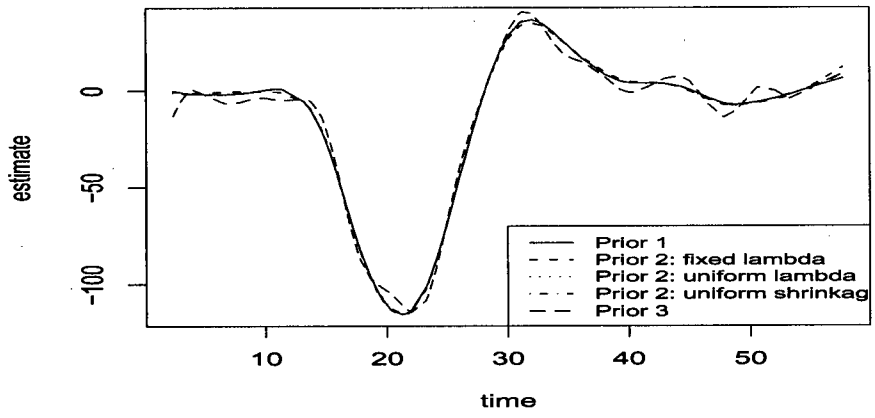

Figure 5.9: *All five estimated curves from the five prior specifications.*

Table 5.5: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 3.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	1366.78	442.48	802.83	2166.88
δ_0	-32.53	15.35	-56.11	-7.22
δ_1	0.64	0.48	-0.18	1.41
σ^2	533.12	71.91	427.13	662.09

5.4 Example 2

Here we fit the same data set of size $n = 250$ that we have used in Example 2 of Chapter 4 for the fixed number of knots $p = 20$ in order to have a second chance of comparing the proposed prior distributions. The same hyperparameters are being used, i.e., we use $\tau^2 \sim IG(3,\ 0.01)$ and $\sigma^2 \sim IG(0.001,\ 0.001)$. Similar to the case of Example 1, both the Gibbs sampler technique and a random walk Metropolis-Hastings algorithm are applied for posterior simulation of the parameters. Estimates of τ^2, $\boldsymbol{\delta}$ and σ^2 for all five specifications of the prior distributions are presented in five tables (Tables 5.6-5.10). Estimates of the parameter vector $\boldsymbol{\theta}$ are presented graphically. Figure (5.10) shows the 60 posterior realizations (grey curves) for the parameter vector $\boldsymbol{\theta}$ with the means (dark curves) for all five choices of the prior distributions. We observe the similar performance as we find in example 1. To be specific, both Prior 1 and Prior 2 produce approximately the same smooth mean curves. The estimates of the data variance σ^2 are similar in all the five choices of the priors. The variance component τ^2 is the only parameter whose estimates differ greatly for different choices of the prior distributions, and the largest τ^2 is found in the case of Prior 3. The regression parameters remain statistically insignificant in this example as well.

5.4.1 Monitoring the Convergence of MCMC Simulation

In our 5000 iterations of MCMC simulations, a clear picture of a burn-in period is found in the first 2000 iterations. Figures (5.11) and (5.13) present 5000 posterior realizations

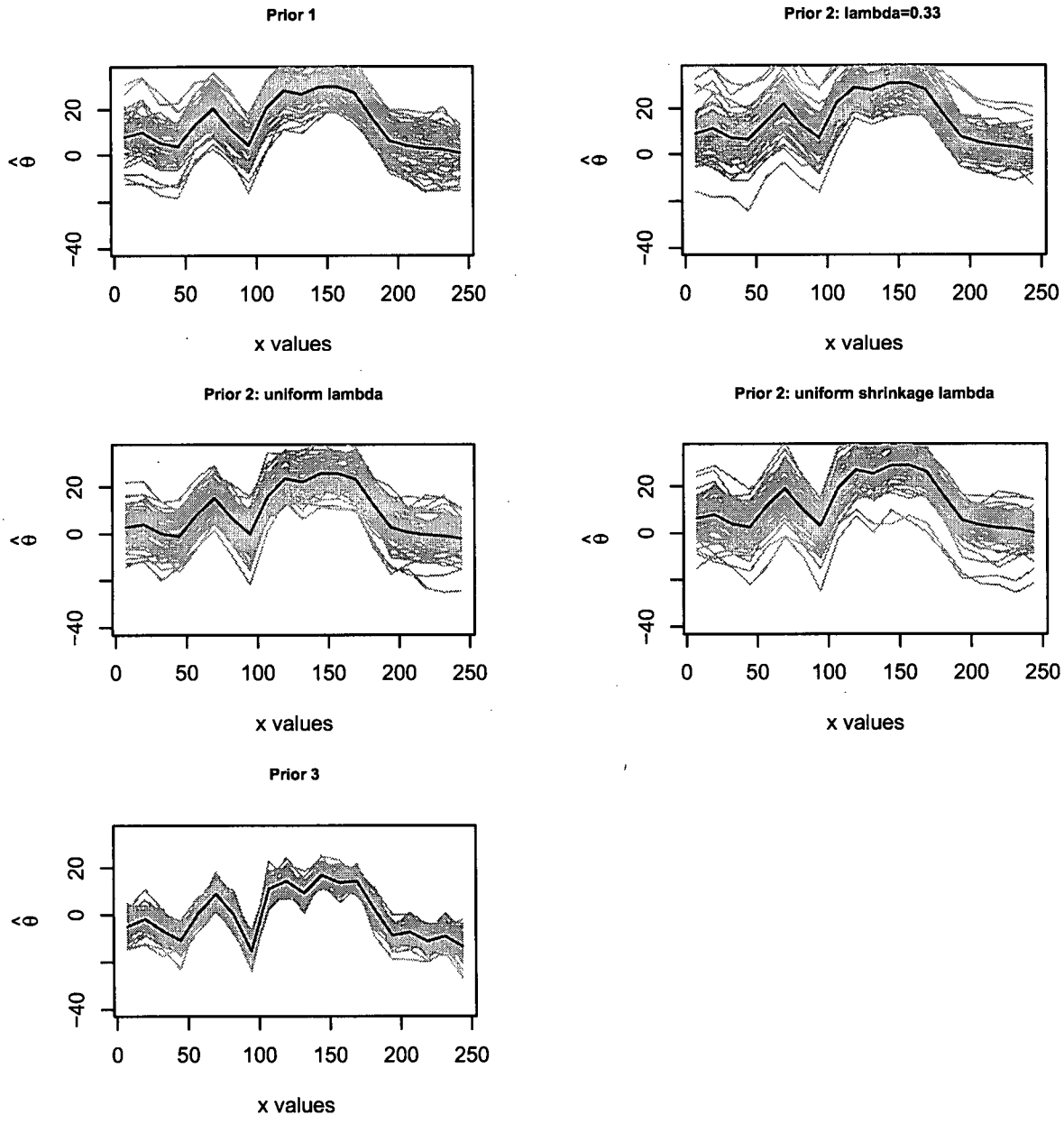

Figure 5.10: *Sixty posterior realizations (grey curves) for the parameter vector* $\boldsymbol{\theta}$. *Dark curves show the posterior means.*

Table 5.6: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 1.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	0.07	0.04	0.02	0.14
δ_0	-9.54	8.99	-24.36	6.79
δ_1	-0.04	0.05	-0.11	-0.05
σ^2	94.79	9.49	80.14	111.18

Table 5.7: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with fixed λ ($\lambda = 0.33$).

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	0.04	0.03	0.007	0.102
δ_0	-11.49	9.92	-30.03	3.60
δ_1	-0.03	0.04	-0.10	0.04
σ^2	96.87	9.86	81.67	114.10

Table 5.8: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform λ.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	0.06	0.04	0.02	0.14
δ_0	-4.07	8.05	-17.25	9.03
δ_1	-0.05	0.04	-0.12	0.03
σ^2	95.53	9.87	80.84	112.36

Table 5.9: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 2 with uniform shrinkage λ.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	0.06	0.04	0.02	0.13
δ_0	-8.00	8.79	-22.48	6.50
δ_1	-0.04	0.04	-0.11	0.03
σ^2	95.20	9.38	80.98	112.39

Table 5.10: Posterior distributions from Gibbs sampling based on iterations 2001-5000 for Prior 3.

Parameter	Mean	Standard deviation	5^{th} centile	95^{th} centile
τ^2	102.91	34.61	60.63	172.05
δ_0	3.25	4.32	-3.56	10.59
δ_1	-0.03	0.03	-0.08	0.01
σ^2	91.13	8.51	77.88	105.56

for the parameters τ^2 and σ^2, respectively. We also plot the corresponding histograms in Figures (5.12) and (5.14), respectively. The MCMC chains for λ and the corresponding histograms are shown in Figure (5.15). We observe a nice convergence and a good mixing of the MCMC run. In this example, similar estimates of λ for both the uniform and uniform shrinkage priors are found. For the uniform prior, the posterior mean for the parameter λ is 0.015, the standard deviation is 0.01, the 5^{th} percentile is 0.006 and the 95^{th} percentile is 0.033. The acceptance ratio for the MCMC run is 53 percent. Posterior samples of $\bar{w}$ for the Prior 2 cases are plotted in Figure (5.16). Similar to the case of Example 1, $\bar{w}$ converges to 1 after approximately 2000 iterations for both uniform and uniform shrinkage λ cases. True and the fitted curves with the data points are plotted in Figure (5.17). To compare the performance of the five priors together, all the five estimated curves with the true curve are plotted in Figure (5.18). Similar conclusion as in Example 1 is drawn, that is Prior 1 and Prior 2 perform equally better than Prior 3 in regard to smoothness.

5.5 Discussion

Two simulation studies have been done in this chapter for the Bayesian curve fitting using hierarchical models with our proposed prior distributions under the assumption of natural cubic spline. We perform MCMC to simulate 5000 posterior samples for all the model parameters. The plots of posterior realizations reflects well convergence as well as mixing of the chains. Although the concepts of three prior distributions are completely different, in

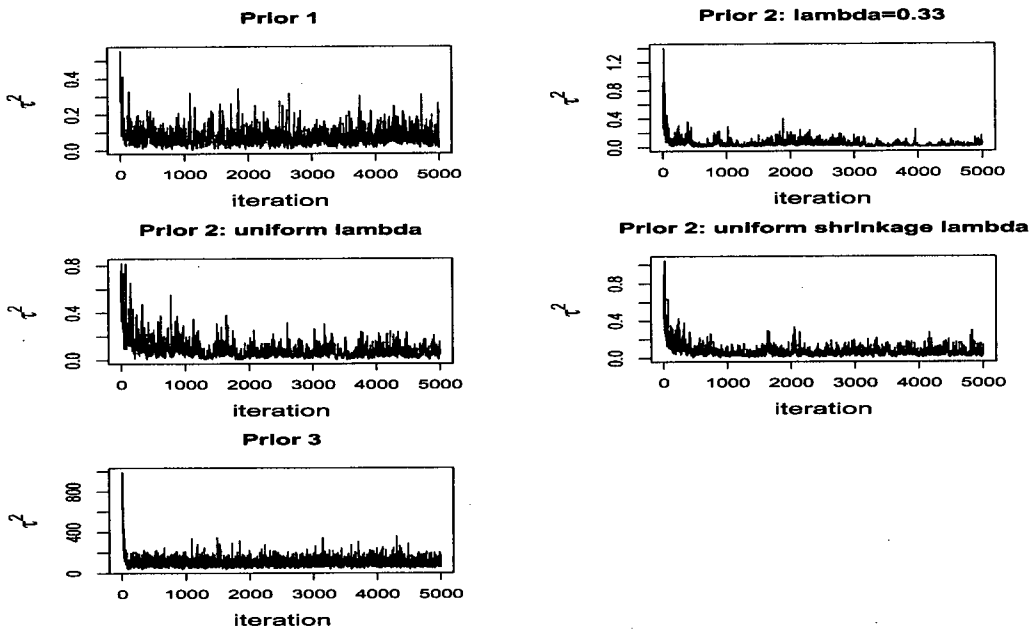

Figure 5.11: *Plots of 5000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

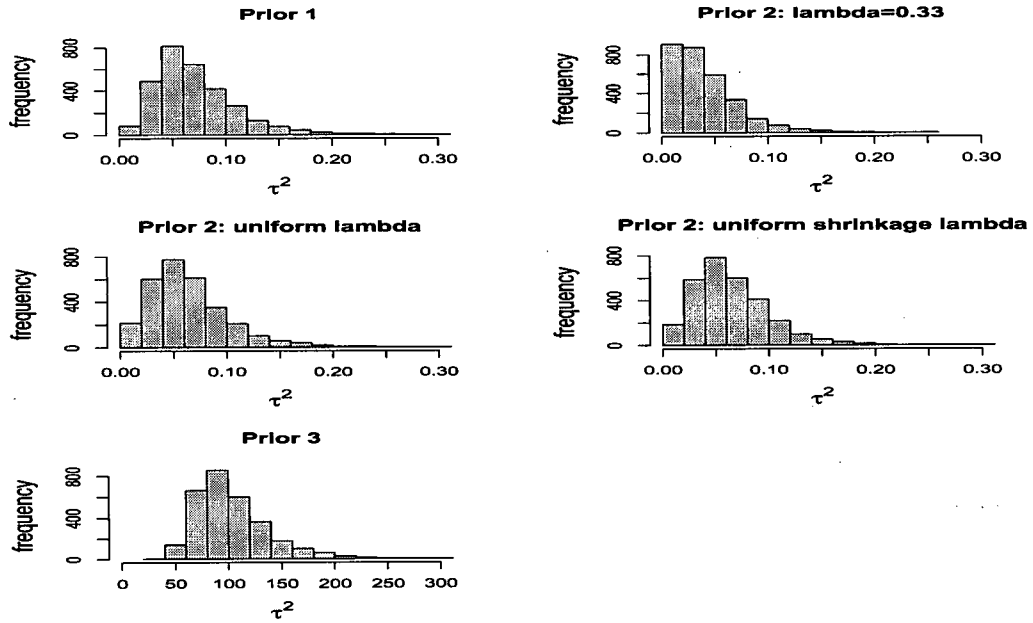

Figure 5.12: *Histograms of the last 3000 posterior realizations of the variance component τ^2 for each of the prior distributions.*

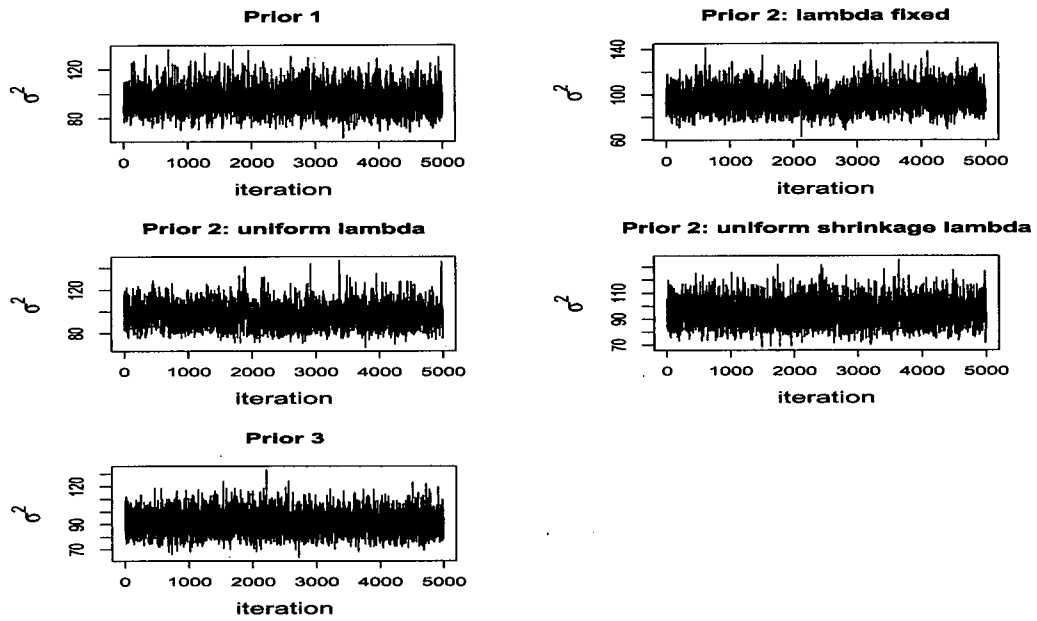

Figure 5.13: *Plots of 5000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

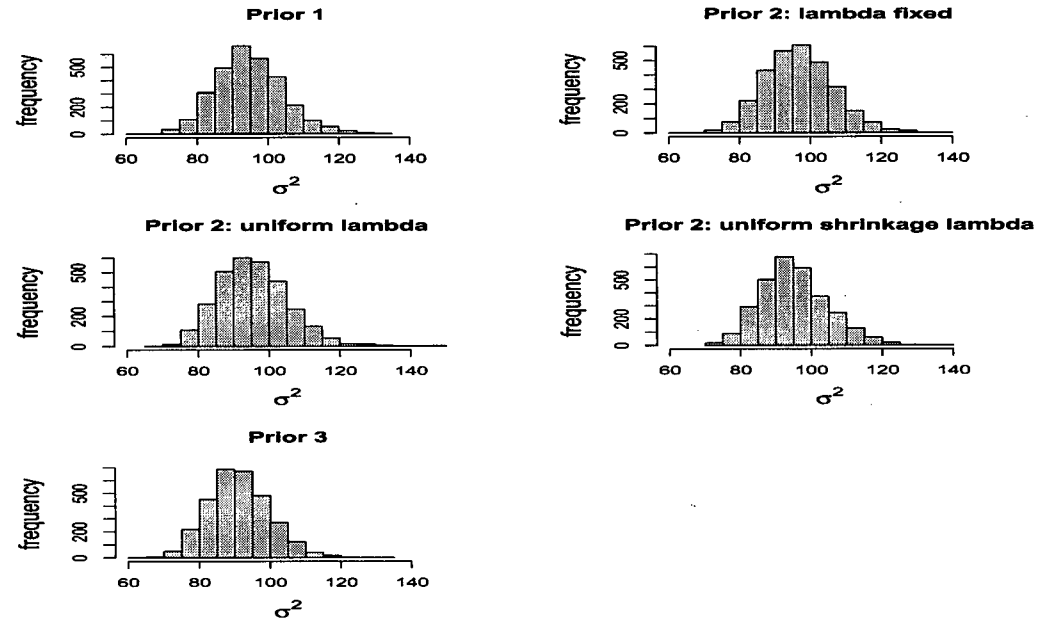

Figure 5.14: *Histograms of the last 3000 posterior realizations of the data variance σ^2 for each of the prior distributions.*

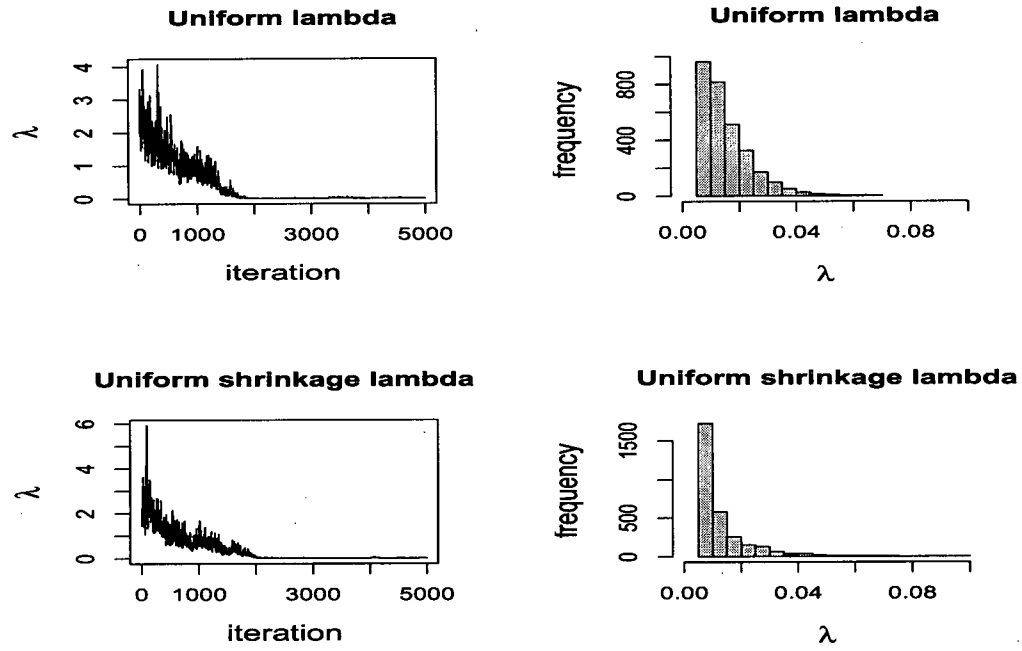

Figure 5.15: *Plots of 5000 posterior samples of the parameter λ and histograms for the last 3000 samples.*

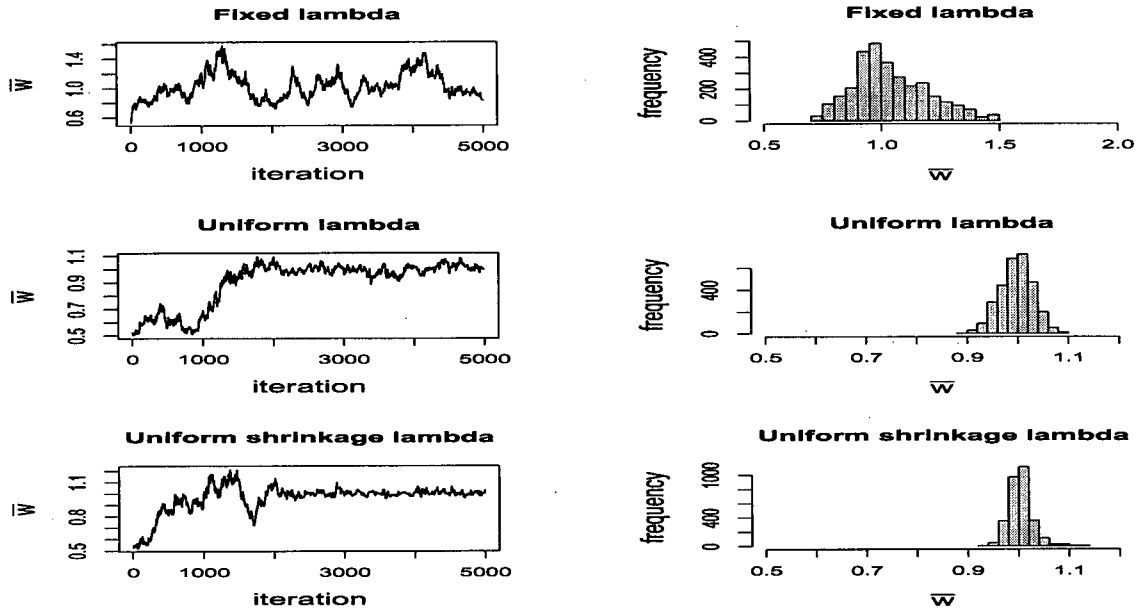

Figure 5.16: *Plots of 5000 posterior samples of $\bar{w}$ and histograms for the last 3000 samples.*

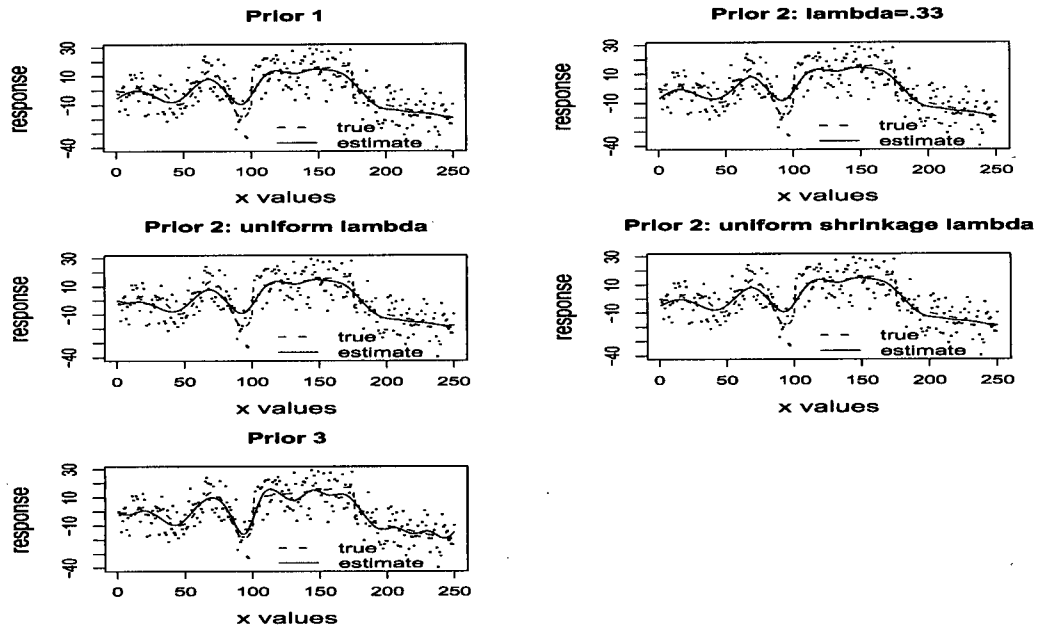

Figure 5.17: *True and the estimated curves of all the five choices of the prior distributions.*

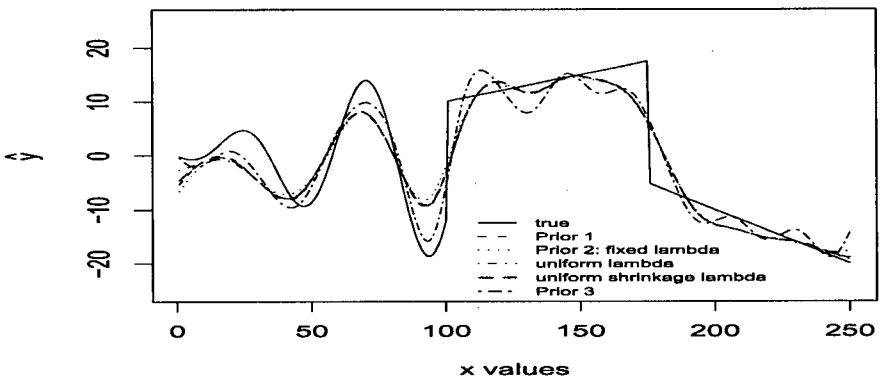

Figure 5.18: *All the five estimated curves with the true curve.*

Example 1, we do not observe any big difference among the estimates. To be specific, both Prior 1 and Prior 2 have produced exactly the same smooth curves which are slightly better than the curve produced by Prior 3. In Example 2, the upshot remain the same. Despite of the non-smoothing behavior of Prior 3, it gives estimates with less standard errors as we see in Figures (5.1) and (5.10). Therefore, we accept the fact that Prior 1 and Prior 2 perform in a similar way for estimating wiggly functions under the natural cubic spline assumption.

Chapter 6

Conclusion and Future Work

The main focus of this dissertation is to compare three roughness penalty prior distributions for the parameters of smoothing splines in Bayesian hierarchical models. These models are useful in many real life situations and some of them have been discussed in Chapter 1. The MCMC simulation technique is essential for the Bayesian curve fitting and we have briefly explained the Gibbs sampler and the Metropolis-Hastings algorithms in Chapter 1.

In order to estimate a curve in the context of a smoothing spline with roughness penalty approach, the usual prior distribution for the regression function considers only one variance component (or smoothing parameter) for the entire function and measures the global roughness of the curve. In this thesis we call it Prior 1. Our scientific interest is to explore more smooth curve from a wiggly data set, and hence we may rather think of different variance components for the parameters at different knot points. This kind of local variance component concept may produce more smooth curves in dealing with wiggly data. To apply the concept of measuring local roughness of the underlying curve, we propose a second prior and call it Prior 2. In connection with our prior search, to estimate a rapidly fluctuating function, we come up with the idea of comparing the above two priors with a non-roughness penalty prior distribution, in which all the parameters are independently distributed with a common global variance component, and we call it Prior 3. Our nonparametric approaches

to curve fitting include both piecewise linear and natural cubic splines. Hence for simulation we need to calculate the basis matrices for both piecewise linear and natural cubic spline cases. We have briefly discussed the procedures of calculating basis matrices in Chapter 2.

Under the assumption of a piecewise linear spline, we have formulated the hierarchical models for the Bayesian curve fitting with roughness penalty prior distributions, i.e., for Prior 1 and Prior 2, and also for the non-roughness penalty prior, Prior 3. We have calculated the posterior distributions for all the model parameters. In most of the cases, we have found closed form solutions for the posterior distributions and so we have applied Gibbs sampling technique. For the parameters $\boldsymbol{w}$ and λ we have not obtained any closed form solutions, and hence we have adopted random walk Metropolis Hastings algorithms in exponential scale for simulating these parameters.

In Chapter 4, the Bayesian backfitting algorithm has been applied to draw posterior samples. This is because our regression function consists of both linear and smooth parts. Three different data sets of different curvatures are chosen for comparing the performance of the proposed prior distributions. Necessary steps have been taken to make the MCMC run and mixing of the posterior samples good. We have also calculated the standard errors and the credible intervals for some of the model parameters. In Chapter 4, simulation has been done only for the piecewise linear spline assumption.

After observing the results from the piecewise linear spline, we feel encourage to compare the prior distributions in the case of natural cubic spline. The formulation of the Bayesian hierarchical model in the case of the global variance component prior distribution (Prior 1) is well established and found in many books and articles, on the other hand, the concept of local variance component prior distribution (Prior 2) is new and there is no such reference available. Since the calculation of the roughness penalty matrix under the assumption of NCS becomes very tedious, an approximate roughness penalty matrix has been proposed for

Prior 2 in order to calculate the posterior distributions. And no changes have been necessary in calculating the posterior distributions for Prior 3 in the case of NCS.

For Prior 2, we have three different sets of posterior distributions depending on the hyperparameter λ (fixed, uniform and uniform shrinkage). However, they did not play any significant role to make the estimated curve different. Although the estimates of $\bar{w}$ do not show any convergence for the fixed λ case, this does not show any effect on the estimated regression function. We carefully checked the estimation process and necessary steps are taken to make good mixing as well as convergence of the MCMC run. In Chapters 4 and 5 we have given brief discussions about our estimated curves. In recapitulating what we have found in our simulation studies we can say that both Prior 1 and Prior 2 perform in a similar way, better than Prior 3, in wiggly curves estimation. So, inclusion of the roughness penalty term is necessary for rapidly fluctuating curve estimation, and the local and global variance component concept does not make any difference in the estimation. Even though this argument is plausible, we still would like to have some more application to real life wiggle data. This may need some future research. Comparisons of the estimated curves have been made through visual examinations. Some deviance measures in the Bayesian context may be developed in order to have better comparisons. Although some of the classical techniques of fitting splines have been discussed in this thesis, we did not get any chance to compare them with our Bayesian approach. Throughout the study, the response has been assumed to be normal. In conclusion, we can say that future work can also be done on the application of these approaches to binary outcome data, count data, etc. i.e., to the exploration of the concept in the case of generalized additive models.

Appendix A

A.1 Conditional Posterior for θ in Prior 1

Let $\mathbf{Y}$ be the vector of responses and $\mathbf{X}$ be the vector of predictors in a regression analysis context such that $\boldsymbol{Y} \sim N(D\boldsymbol{\delta} + A\boldsymbol{\theta},\ \sigma^2 I)$, where $\boldsymbol{\theta}$, $\boldsymbol{\delta}$ and σ^2 are the parameters of interest, and A and D are the basis matrices calculated from the x values. Hence, the likelihood function can be written as

$$\pi(\mathbf{Y}|\boldsymbol{\delta}, \boldsymbol{\theta}, \sigma^2, \mathbf{X}) \propto |\Omega|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta})'\Omega^{-1}(\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta})\right\}, \tag{A.1}$$

where $\Omega = \sigma^2 I$. In a Bayesian perspective, we consider all the above parameters as random variables and have probability distributions (called prior distributions). We assume that the density functions of the above parameters have the following forms (ignoring the normalizing constants)

$$\begin{aligned} &\pi(\boldsymbol{\theta}|\tau^2) \propto \left(\tfrac{1}{\tau^2}\right)^{(p/2)} \exp\left\{-\tfrac{1}{2}\boldsymbol{\theta}' B \boldsymbol{\theta}\right\}, \\ &\pi(\sigma^2) \propto \left(\tfrac{1}{\sigma^2}\right)^{(a+1)} \exp\{-b/\sigma^2\} \ \text{ and} \\ &\pi(\boldsymbol{\delta}) \propto 1, \end{aligned} \tag{A.2}$$

where τ^2, a and b are called hyperparameters. Let us assume that a and b are known and τ^2 has an inverse gamma distribution with the density function

$$\pi(\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{(\alpha+1)} \exp\{-\beta/\tau^2\}, \tag{A.3}$$

where α and β are konwn constants. By applying Bayes' theorem we can derive the posterior distribution from the prior distribution and the likelihood function. The joint posterior

distribution of the parameters $\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2$ and σ^2 given $\mathbf{Y}$ and $\mathbf{X}$ can be written as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \sigma^2 | \mathbf{Y}, \mathbf{X}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{X}) \pi(\boldsymbol{\theta}|\tau^2) \pi(\tau^2) \pi(\boldsymbol{\delta}) \pi(\sigma^2). \tag{A.4}$$

With the likelihood function, prior densities and the joint posterior distribution defined above, the conditional posterior distribution for $\boldsymbol{\theta}$ given $\boldsymbol{\delta}$, τ^2, σ^2, $\mathbf{Y}$ and $\mathbf{X}$ can be calculated as

$$\begin{aligned}
\pi(\boldsymbol{\theta}|\boldsymbol{\delta}, \tau^2, \sigma^2, \mathbf{Y}, \mathbf{X}) &\propto \pi(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{X}) \pi(\boldsymbol{\theta}|\tau^2) \\
&\propto |\Omega|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta})' \Omega^{-1} (\mathbf{Y} - D\boldsymbol{\delta} - A\boldsymbol{\theta})\right\} \\
&\times \left(\tfrac{1}{\tau^2}\right)^{(p/2)} \exp\left\{-\tfrac{1}{2}\boldsymbol{\theta}' B \boldsymbol{\theta}\right\} \\
&= |\Omega|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{Z}_1 - A\boldsymbol{\theta})' \Omega^{-1} (\mathbf{Z}_1 - A\boldsymbol{\theta})\right\} \left(\tfrac{1}{\tau^2}\right)^{(p/2)} \exp\left\{-\tfrac{1}{2}\boldsymbol{\theta}' B \boldsymbol{\theta}\right\} \\
&\propto \exp\left\{-\tfrac{1}{2}(\boldsymbol{\theta} - (B + A'\Omega^{-1}A)^{-1} A'\Omega^{-1}\mathbf{Z}_1)' (B + A'\Omega^{-1}A)^{-1} \right. \\
&\left. \times (\boldsymbol{\theta} - (B + A'\Omega^{-1}A)^{-1} A'\Omega^{-1}\mathbf{Z}_1)\right\},
\end{aligned} \tag{A.5}$$

where $\mathbf{Z}_1 = \mathbf{Y} - D\boldsymbol{\delta}$. It follows that the density $\pi(\boldsymbol{\theta}|\boldsymbol{\delta}, \tau^2, \sigma^2, \mathbf{Y}, \mathbf{X})$ is multivariate normal with mean vector $\boldsymbol{\mu}_1$ and covariance matrix Ω_1 where

$$\begin{aligned}
\boldsymbol{\mu}_1 &= (B + A'\Omega^{-1}A)^{-1} A'\Omega^{-1}\mathbf{Z}_1 = \left(\frac{1}{\tau^2}M + \frac{(A'A)^{-1}}{\sigma^2}\right)^{-1} \frac{A'\mathbf{Z}_1}{\sigma^2} \quad \text{and} \\
\Omega_1 &= (B + A'\Omega^{-1}A)^{-1} = \left(\frac{1}{\tau^2}M + \frac{(A'A)^{-1}}{\sigma^2}\right)^{-1}.
\end{aligned}$$

A.2 Conditional Posterior Distribution for τ^2 in Prior 1

From the definitions of Appendix A.1, the conditional posterior distribution for τ^2 given $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$ is equivalent to the conditional posterior distribution of τ^2 given $\boldsymbol{\theta}$ since other densities do not contain τ^2. So

$$\begin{aligned}
\pi(\tau^2|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{Y}, \mathbf{X}) = \pi(\tau^2|\boldsymbol{\theta}) &\propto \pi(\tau^2)\pi(\boldsymbol{\theta}|\tau^2) \\
&\propto \left(\tfrac{1}{\tau^2}\right)^{(\alpha+1)} \exp\{-\beta/\tau^2\} \left(\tfrac{1}{\tau^2}\right)^{(p/2)} \exp\left\{-\tfrac{1}{2}\boldsymbol{\theta}' B \boldsymbol{\theta}\right\} \\
&\propto \left(\tfrac{1}{\tau^2}\right)^{(\alpha+\frac{p}{2}+1)} \exp\left\{-\tfrac{1}{2\tau^2}(\boldsymbol{\theta}' M \boldsymbol{\theta}) - \tfrac{\beta}{\tau^2}\right\}.
\end{aligned} \tag{A.6}$$

The right hand side of the above expression says that the density $\pi(\tau^2|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{Y}, \mathbf{X})$ is inverse gamma with parameters $\alpha + \frac{p}{2}$ and $\frac{1}{2}(\boldsymbol{\theta}' M \boldsymbol{\theta}) + \beta$.

A.3 Conditional Posterior Distribution of δ in Prior 1

From Appendix A.1, the conditional posterior distribution of $\boldsymbol{\delta}$ given $\boldsymbol{\theta}$, σ^2,$\mathbf{Y}$ and $\mathbf{X}$ is obtained as follows:

$$\begin{aligned}
\pi(\boldsymbol{\delta}|\boldsymbol{\theta},\sigma^2,\mathbf{Y},\mathbf{X}) &\propto \pi(\mathbf{Y}|\boldsymbol{\theta},\boldsymbol{\delta},\sigma^2,\mathbf{X})\pi(\boldsymbol{\delta}) \\
&\propto |\Omega|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})'\Omega^{-1}(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})\right\} \\
&\propto |\Omega|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{Z}_2-D\boldsymbol{\delta})'\Omega^{-1}(\mathbf{Z}_2-D\boldsymbol{\delta})\right\} \\
&\propto \exp\{-\tfrac{1}{2}(\boldsymbol{\delta}-(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\mathbf{Z}_2)'(D'\Omega^{-1}D)^{-1} \\
&\quad\times(\boldsymbol{\delta}-(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\mathbf{Z}_2)\},
\end{aligned} \tag{A.7}$$

where $\mathbf{Z}_2 = \mathbf{Y} - A\boldsymbol{\theta}$. Hence we can say that $\pi(\boldsymbol{\delta}|\boldsymbol{\theta},\sigma^2,\mathbf{Y},\mathbf{X})$ is a multivariate normal density with mean vector $(D'\Omega^{-1}D)^{-1}D'\Omega^{-1}\mathbf{Z}_2$ and covariance matrix $(D'\Omega^{-1}D)^{-1}$. Since $\Omega = \sigma^2 I$, we can write the mean vector as $(D'D)^{-1}D'\mathbf{Z}_2$ and the covariance matrix as $(D'D)^{-1}\sigma^2$.

A.4 Conditional Posterior Distribution of σ^2 in Prior 1

The conditional posterior distribution of σ^2 given $\boldsymbol{\delta}$, $\boldsymbol{\theta}$, τ^2 $\mathbf{Y}$ and $\mathbf{X}$ is obtained as per the definitions in Appendix A.1 as follows:

$$\begin{aligned}
\pi(\sigma^2|\boldsymbol{\theta},\boldsymbol{\delta},\tau^2,\mathbf{Y},\mathbf{X}) &\propto \pi(\mathbf{Y}|\boldsymbol{\theta},\boldsymbol{\delta},\sigma^2,\mathbf{X})\pi(\sigma^2) \\
&\propto \left(\tfrac{1}{\sigma^2}\right)^{n/2} \exp\left\{-\tfrac{1}{2\sigma^2}(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})'(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})\right\} \\
&\quad \left(\tfrac{1}{\sigma^2}\right)^{(a+1)} \exp\{-\tfrac{b}{\sigma^2}\} \\
&\propto \left(\tfrac{1}{\sigma^2}\right)^{a+\frac{n}{2}+1} \exp\left\{-\tfrac{1}{2\sigma^2}(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})'(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})+b\right\}.
\end{aligned} \tag{A.8}$$

This is an $IG(a+\frac{n}{2},\ (\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})'(\mathbf{Y}-D\boldsymbol{\delta}-A\boldsymbol{\theta})+b)$ density without the normalizing constants.

Appendix B

B.1 Conditional Posterior for θ and τ^2 in Prior 2

In addition to the parameters defined in Appendix A.1, let us have a vector $\boldsymbol{w} = (w_1, \cdots, w_p)'$ of p parameters and λ. By inclusion of the new parameters the density $\pi(\boldsymbol{\theta}|\tau^2)$ has been changed to $\pi(\boldsymbol{\theta}|\tau^2, \boldsymbol{w})$ with the functional form:

$$\pi(\boldsymbol{\theta}|\tau^2, \boldsymbol{w}) \propto |B_{\boldsymbol{w},\tau}|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}' B_{\boldsymbol{w},\tau}\boldsymbol{\theta}\right\},$$

where $B_{\boldsymbol{w},\tau} = \frac{1}{\tau^2} M_{\boldsymbol{w}}$. Here we assume that w_i's are independent and identically distributed random variables with density function

$$\pi(w_i) = \frac{(\frac{1}{\lambda})^{\frac{1}{\lambda}}}{\Gamma(\frac{1}{\lambda})} \left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)} \exp\left\{-\frac{1}{\lambda w_i}\right\}; \quad i = 1, \cdots, p. \tag{B.1}$$

And the prior distribution of λ is assumed as uniform over the interval $(0, \lambda_0)$. The joint probability distribution for $\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \boldsymbol{w}$ and σ^2 given $\mathbf{Y}$ and $\mathbf{X}$ can be written as

$$\pi(\boldsymbol{\theta}, \boldsymbol{\delta}, \tau^2, \boldsymbol{w}, \sigma^2|\mathbf{Y}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta}, \boldsymbol{\delta}, \sigma^2, \mathbf{X})\pi(\boldsymbol{\theta}|\boldsymbol{w}, \tau^2)\pi(\tau^2)\pi(w_1)\cdots\pi(w_p)\pi(\boldsymbol{\delta})\pi(\sigma^2). \tag{B.2}$$

As we have already shown the calculation for the conditional pdf $\pi(\boldsymbol{\theta}|\boldsymbol{\delta}, \tau^2, \sigma^2, \mathbf{Y}, \mathbf{X})$ in (A.5), in the similar manner, we can show that the condition posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{\delta}, \tau^2, \boldsymbol{w}, \sigma^2, \mathbf{Y}, \mathbf{X})$ is multivariate normal with mean vector $\boldsymbol{\mu}_2$ and the covariance matrix Ω_2 where

$$\begin{aligned}
\boldsymbol{\mu}_2 &= (B_{\boldsymbol{w},\tau} + A'\Omega^{-1}A)^{-1}A'\Omega^{-1}\mathbf{Z}_1 = \left(\frac{1}{\tau^2}M_{\boldsymbol{w}} + \frac{(A'A)^{-1}}{\sigma^2}\right)^{-1}\frac{A'\mathbf{Z}_1}{\sigma^2} \quad \text{and} \\
\Omega_2 &= (B_{\boldsymbol{w},\tau} + A'\Omega^{-1}A)^{-1} = \left(\frac{1}{\tau^2}M_{\boldsymbol{w}} + \frac{(A'A)^{-1}}{\sigma^2}\right)^{-1}.
\end{aligned}$$

Also in the similar way to (A.6), we obtain the conditional posterior distribution for τ^2 given $\boldsymbol{\theta}$, $\boldsymbol{w}$, $\boldsymbol{\delta}$, σ^2, $\mathbf{Y}$ and $\mathbf{X}$, and which is inverse gamma with the parameters $\alpha + \frac{p}{2}$ and $\frac{1}{2}(\boldsymbol{\theta}' B_{\boldsymbol{w},\tau} \boldsymbol{\theta}) + \beta)$.

B.2 Conditional posterior distribution of w in Prior 2

With the definitions in Appendix B.1, the Conditional posterior distribution of $\boldsymbol{w}$ given $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, τ^2, σ^2, $\mathbf{Y}$ and $\mathbf{X}$ is equivalent to the conditional posterior distribution of $\boldsymbol{w}$ given $\boldsymbol{\theta}$ and τ^2. Therefore, we can write $\pi(\boldsymbol{w}|\boldsymbol{\theta}, \tau^2) \propto \pi(\boldsymbol{\theta}|\tau^2, \boldsymbol{w})\pi(\boldsymbol{w})$. Since w_i's are independent and identically distributed with inverse gamma, the joint posterior distribution of $w_1, \cdots, w_p$ given $\boldsymbol{\theta}$ and τ^2 can be written as

$$\begin{aligned}
\pi(w_1, \cdots, w_p|\boldsymbol{\theta}, \tau^2) &\propto \pi(\boldsymbol{\theta}|\tau^2, \boldsymbol{w})\pi(w_1) \cdots \pi(w_p) \\
&\propto |B_{\boldsymbol{w},\tau}|^{1/2} \exp\left\{-\frac{1}{2}\boldsymbol{\theta}' B_{\boldsymbol{w},\tau}\boldsymbol{\theta}\right\} \times \prod_{i=1}^{p}\left[\frac{(\frac{1}{\lambda})^{\frac{1}{\lambda}}}{\Gamma(\frac{1}{\lambda})}\left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)} \exp\left\{-\frac{1}{\lambda w_i}\right\}\right].
\end{aligned}$$

B.3 Conditional Posterior Distribution for λ in Prior 2

Let us consider the prior distribution of λ is uniform over the interval $(0, \lambda_0)$, i.e., $\pi(\lambda) \propto \frac{1}{\lambda_0}$, then the conditional probability distribution for λ given $\boldsymbol{\theta}$, $\boldsymbol{\delta}$, $\boldsymbol{w}$, τ^2, $\mathbf{Y}$ and $\mathbf{X}$ as per the definitions in Appendix B.1 is as follows:

$$\begin{aligned}
\pi(\lambda|\boldsymbol{\theta}, \tau^2, \boldsymbol{w}, \sigma^2, \boldsymbol{\delta}, \mathbf{Y}, \mathbf{X}) &\propto \pi(\boldsymbol{w}|\lambda)\pi(\lambda) \\
&= \prod_{i=1}^{p}\left[\frac{(\frac{1}{\lambda})^{\frac{1}{\lambda}}}{\Gamma(\frac{1}{\lambda})}\left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)} \exp\left\{-\frac{1}{\lambda w_i}\right\}\right]\frac{1}{\lambda_0} \\
&= \frac{(\frac{1}{\lambda})^{\frac{p}{\lambda}}}{(\Gamma\frac{1}{\lambda})^p}\prod_{i=1}^{p}\left[\left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)}\right] \exp\left\{-\frac{1}{\lambda}\sum_{i=1}^{p}\frac{1}{w_i}\right\}\frac{1}{\lambda_0}.
\end{aligned}$$

Let us consider another probability density function for λ as, $\pi(\lambda) \sim \frac{\sigma^2}{\sigma^2+\lambda}$. For this case the conditional posterior is obtained as

$$\begin{aligned}
\pi(\lambda|\boldsymbol{\theta},\tau^2,\boldsymbol{w},\sigma^2;\boldsymbol{\delta},\mathbf{Y},\mathbf{X}) &\propto \pi(\boldsymbol{w}|\lambda)\pi(\lambda|\sigma^2)\pi(\sigma^2) \\
&= \prod_{i=1}^{p}\left[\frac{(\frac{1}{\lambda})^{\frac{1}{\lambda}}}{\Gamma(\frac{1}{\lambda})}\left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)}\exp\left\{-\frac{1}{\lambda w_i}\right\}\right]\frac{1}{\lambda_0} \\
&= \frac{(\frac{1}{\lambda})^{\frac{p}{\lambda}}}{(\Gamma\frac{1}{\lambda})^p}\prod_{i=1}^{p}\left[\left(\frac{1}{w_i}\right)^{(\frac{1}{\lambda}+1)}\right]\exp\left\{-\frac{1}{\lambda}\sum_{i=1}^{p}\frac{1}{w_i}\right\} \\
&\times \frac{\sigma^2}{(\sigma^2+\lambda)^2}\left(\frac{1}{\sigma^2}\right)^{a+1}\exp\left(-\frac{b}{\sigma^2}\right).
\end{aligned}$$

B.4 Conditional posteriors for $\boldsymbol{\theta}$ and τ in Prior 3

For Prior 3 the joint posterior distribution can be written as:

$$\pi(\boldsymbol{\theta},\boldsymbol{\delta},\tau^2,\sigma^2|\mathbf{Y},\mathbf{X}) \propto \pi(\mathbf{Y}|\boldsymbol{\theta},\boldsymbol{\delta},\sigma^2,\mathbf{X})\pi(\boldsymbol{\theta}|\tau^2)\pi(\tau^2)\pi(\boldsymbol{\delta})\pi(\sigma^2), \tag{B.3}$$

where the densities $\pi(\tau^2)$, $\pi(\sigma^2)$ and $\pi(\delta)$ are defined previously in Appendix A.1 except that $\pi(\boldsymbol{\theta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{(p/2)}\exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\theta}'\boldsymbol{\theta}\right\}$ and hence, similar calculations as in (A.5) show that $\pi(\boldsymbol{\theta},\boldsymbol{\delta},\tau^2,\sigma^2|\mathbf{Y},\mathbf{X})$ is a multivariate normal density with mean vector $\boldsymbol{\mu}_3$ and the covariance matrix Ω_3 where

$$\boldsymbol{\mu}_3 = \left(\frac{1}{\tau^2}I + \frac{(A'A)^{-1}}{\sigma^2}\right)^{-1}\frac{A'\mathbf{Z}_1}{\sigma^2} \text{ and } \Omega_3 = \left(\frac{1}{\tau^2}I + \frac{(A'A)^{-1}}{\sigma^2}\right)^{-1}.$$

Also similar to (A.6) we can show that $\pi(\tau^2|\boldsymbol{\theta},\boldsymbol{\delta},\sigma^2,\mathbf{Y},\mathbf{X})$ is inverse gamma with parameters $\alpha+\frac{p}{2}$ and $\frac{1}{2}(\boldsymbol{\theta}'\boldsymbol{\theta})+\beta$.

Appendix C

C.1 Interpolating and Plotting an NCS

Let f be a natural cubic spline with knots $t_1 < t_2 < ... < t_p$ and we define $f_i = f(t_i)$ and $\gamma_i = f''(t_i)$ for $i = 1, ..., p$. By the definition of NCS, $\gamma_1 = \gamma_p = 0$. Let $\boldsymbol{f} = (f_1, ..., f_p)^T$ and $\boldsymbol{\gamma} = (\gamma_2, ..., \gamma_{p-1})^T$. The vectors $\boldsymbol{f}$ and $\boldsymbol{\gamma}$ specify the curve f completely with the two band matrices Q and R such that

$$Q^T f = R\gamma. \tag{C.1}$$

The Q and R matrices can be calculated as follows: Let $h_i = t_{i+1} - t_i$ for $i = 1, \cdots, p-1$. The Q matrix is a $p \times (p-2)$ matrix with entries q_{ij}, $i = 1, \cdots, p$; $j = 2, \cdots, p-1$, defined as

$$q_{j-1,j} = h_{j-1}^{-1}, \quad q_{jj} = -h_{j-1}^{-1} - h_j^{-1}, \quad q_{j+1,j} = h_j^{-1},$$

for $j = 2, \cdots, p-1$ and $q_{ij} = 0$ for $|i-j| \geq 2$. The symmetric matrix R is $(p-2) \times (p-2)$ with elements r_{ij}, for $i, j = 2, \cdots, (p-1)$, defined as

$$\begin{aligned} r_{ii} &= \frac{1}{3}(h_{i-1} + h_i) \text{ for } i = 2, \cdots, p-1, \\ r_{i,i+1} &= r_{i+1,i} = \frac{1}{6}h_i \text{ for } i = 2, \cdots, p-2, \text{ and} \\ r_{ii} &= 0 \text{ for } |i-j| \geq 2. \end{aligned}$$

We define a matrix K by

$$K = QR^{-1}Q^T \tag{C.2}$$

with the property that

$$\int_a^b f''(t)^2 dt = \boldsymbol{\gamma}^T R \boldsymbol{\gamma} = \boldsymbol{f}^T K \boldsymbol{f}. \tag{C.3}$$

The value of the cubic spline at a knot point t can be calculated as:

$$f(t) = \frac{(t-t_i)f_{i+1} + (t_{i+1}-t)f_i}{h_i} - \frac{1}{6}(t-t_i)(t_{i+1}-t)\left\{\left(1+\frac{t-t_i}{h_i}\right)\gamma_{i+1} + \left(1+\frac{t_{i+1}-t}{h_i}\right)\gamma_i\right\} \tag{C.4}$$

for $h_i = t_{i+1} - t_i$; $t_i \leq t \leq t_{i+1}$; $i = 1, \cdots, p-1$. We know that $\boldsymbol{\gamma} = R^{-1}Q^T\boldsymbol{f}$ is a $(p-2)$ vector and if we write the ith element of $\boldsymbol{\gamma}$ as $v_{i1}f_1 + v_{i2}f_2 + \cdots + v_{ip}f_p$, we can write equation (C.4) as

$$\begin{aligned}
f(t) &= \frac{(t-t_i)f_{i+1}}{h_i} - \frac{(t-t_i)(t_{i+1}-t)f_{i+1}}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i+1} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i+1}\right\} \\
&+ \frac{(t_{i+1}-t)f_i}{h_i} - \frac{(t-t_i)(t_{i+1}-t)f_i}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i}\right\} \\
&- \frac{(t-t_i)(t_{i+1}-t)f_1}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,1} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,1}\right\} \\
&- \frac{(t-t_i)(t_{i+1}-t)f_2}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,2} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,2}\right\} \\
&\vdots \\
&- \frac{(t-t_i)(t_{i+1}-t)f_p}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,p} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,p}\right\}.
\end{aligned} \tag{C.5}$$

In matrix notation, equation (C.5) can be written as

$$(f(t_1), f(t_2), \cdots, f(t_p)) = A\boldsymbol{f} \tag{C.6}$$

where A is a matrix of order $n \times p$ whose jth row $(j = 1, \cdots, n)$ is obtained as

$$\begin{aligned}
a_{j,i+1} &= \frac{(t-t_i)}{h_i} - \frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i+1} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i+1}\right\}, \\
a_{ji} &= \frac{(t_{i+1}-t)}{h_i} - \frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i}\right\}, \\
\text{and } a_{j,i'} &= \frac{(t-t_i)(t_{i+1}-t)}{6}\left\{\left(1+\frac{t-t_i}{h_i}\right)v_{i+1,i'} + \left(1+\frac{t_{i+1}-t}{h_i}\right)v_{i,i'}\right\},
\end{aligned}$$

where $i' = 1, 2, \cdots, i-1, i+2, \cdots, p$, for any t between t_i and t_{i+1}, and v_{ij} is the (i,j)th element of $R^{-1}Q^T$ and p is the number of knot points for which f needs to be estimated.

Bibliography

[1] N. S. Altman and G. Casella. Nonparametric empirical Bayes growth curve analysis. *Journal of the American Statistical Association*, **90**:508–515, 1995.

[2] P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of Generalized Cross-Validation. *Numerische Mathematika*, **31**:377–403, 1979.

[3] G. S. Datta and M. Ghosh. Bayesian prediction in linear models: applications to small area estimation. *Ann. Statist.*, **19**:1748–1770, 1991.

[4] M. J. Deniels. A prior for the variance in hierarchical models. *The Canadian Journal of Statistics*, **27**:567–578, 1999.

[5] D. G. T. Denison, C. C. Holmes, B. K. Mallick, and A. F. M. Smith. *Bayesian Method for Nonlinear Classification and Regression.* John Wiley & Sons, Ltd., England, 2002.

[6] D. G. T. Denison, B. K. Mallick, and A. F. M. Smith. Automatic Bayesian curve fitting. *J. R. Statist. Soc. Ser.B*, **60**:333–350, 1998.

[7] A. E. Gelfand, S. E. Hills, A. Racine-Poon, and A. F. M. Smith. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**:972–985, 1990.

[8] A. E. Gelfand and A. F. M. Smith. Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**:398–409, 1990.

[9] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis.* Chapman and Hall, London, 1995.

[10] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**:721–741, 1984.

[11] M. Ghosh and J. N. K. Rao. Small area estimation: an appraisal. *Statistical Science*, **9**:55–76, 1994.

[12] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman and Hall/CRC, New York, 1996.

[13] H. Goldstein, M. J. R. Healy, and J. Rasbash. Multilevel time series models with applications to repeated measures data. *Statistics in Medicine*, **13**:1643–1655, 1994.

[14] P. J. Green. Reversible jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**:711–732, 1995.

[15] P. J. Green and B. W. Silverman. *Nonparametric Regression and Generalized Linear Models.* Chapman and Hall, London, 1994.

[16] P. Gustafson, D. Aeschliman, and A. R. Levy. A simple approach to fitting Bayesian Survival models. *Lifetime Data Analysis*, **9**:5–19, 2003.

[17] M. H. Hansen and C. Koopberg. Spline adaptation in extended linear models. *Statistical Science*, **17**:2–51, 2002.

[18] T. Hastie and R. Tibshirani. *Generalized Additive Models.* Chapman and Hall, London, 1990.

[19] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**:97–109, 1970.

[20] M. S. Hossain. A conservative prior for Bayesian hierarchical models in Biostatistics. Master's thesis, University of British Columbia, 2003.

[21] P. McCullagh and J. A. Nelder. *Generalized Linear Models (2nd ed.).* Chapman and Hall, London, 1989.

[22] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, and A. H. Teller. Equations of sate calculations by fast computing machines. *Journal of Chemical Physics*, **21**:1087–1091, 1953.

[23] J. A. Nelder and R. W. M. Wedderburn. Generalized Linear Models. *J. R. Statist. Soc., Ser. A*, **135**:370–384, 1972.

[24] G. K. Robinson. That BLUP is a good thing: the estimation of random effects. *Statistical Science*, **6**:15–51, 1991.

[25] B. W. Silverman. Some aspects of the spline smoothing approach to non-parametric curve fitting. *J. R. Statist. Soc. Ser. B*, **47**:1–52, 1985.

[26] J. S. Simonoff. *Smoothing Methods in Statistics.* Springer, New York, 1996.

[27] M. Smith and R. Kohn. A Bayesian approach to nonparametric bivariate regression. *Journal of the American Statistical Association*, **92**:1522–1535, 1997.

[28] G. Wahba. *Spline Models for Observational Data.* CBMS-NSF, Philadelphia, Pennsylvania, 1990.

[29] A. Whitehead. *Meta-Analysis of controlled clinical trials.* John Wiley and Sons, Ltd., England, 2002.

[30] S. L. Zeger and M. R. Karim. Generalized linear models with random effects; a gibbs sampling approach. *Journal of the American Statistical Association*, **86**: 79–86, 1991.