Bayesian Smoothing Spline Analysis of Variance Models

A Dissertation Presented to

the Faculty of the Graduate School

University of Missouri

In Partial Fulfillment Of the

Requirements for the Degree

Doctor of Philosophy

---

by

Chin-I Cheng

Dr. Paul Speckman
Dissertation Supervisor

August 2009

The undersigned, appointed by the Dean of the Graduate School, have examined the dissertation entitled.

Bayesian Smoothing Spline Analysis of Variance Models

Presented by Chin-I Cheng

A candidate for the degree of Doctor of Philosophy

And hereby certify that in their opinion it is worthy of acceptance.

Dr. Paul Speckman _____

Dr. Sounak Chakraborty _____

Dr. Athanasios Micheas _____

Dr. Shawn Xiaoguang Ni _____

Dr. Dongchu Sun _____

To my father and the memory of my mother.

Acknowledgements

I would like to express my sincere appreciation to my advisor, Dr. Paul Speckman, for his patience, encouragement and guidance throughout my graduate study, which has been a rewarding journey to me. This dissertation could not have been written without his facilitation.

I am also grateful to other members of my advisory committee: Dr. Sounak Chakraborty, Dr. Athanasios Micheas, Dr. Shawn Xiaoguang Ni and Dr. Dongchu Sun for their valuable inputs and supports.

I would also like to express gratitude to the Statistics Department faculty members, staffs and my graduate fellows for their support, assistance and friendships that had made my graduate study in our department a great experience.

At the last, I would like to thank my family members, my husband, my son, my parents and my siblings, for their love and patience.

Table of Contents

List of Tables

List of Figures

Bayesian Smoothing Spline Analysis of Variance Models

Chin-I Cheng

Dr. Paul Speckman

Dissertation Supervisor

Abstract

Based on the pioneering work by Wahba (1990) in smoothing splines for nonparametric regression, Gu (2002) decomposed the regression function based on a tensor sum decomposition of inner product spaces into orthogonal subspaces so the estimated functions from each subspaces can be viewed separately. This is based on an ANOVA type decomposition and is called the smoothing spline ANOVA (SSANOVA) model. Current research related to smoothing spline ANOVA focuses on the frequentist approach for statistical inference in estimation and prediction. In this dissertation, we apply a fully Bayesian approach in SSANOVA to extend statistical inference not only for estimation and prediction but to model testing and selection. The prior selected for the smoothing parameter in level effects is a variant of the Zellner-Siow prior. Two sets of priors, the Pareto and the scaled χ^2_1, are used for the smoothing parameters corresponding to smooth effects. We study this fully Bayesian SSANOVA model for Gaussian response variables and also extend it to generalized additive models with binary response variables. Bayesian

SSANOVA methods are illustrated by simulated examples and also by application to real datasets, potassium measurement on dogs and a Wisconsin epidemiological study of diabetic retinopathy. The flexibility of hypothesis testing provides a powerful tool in statistical inference when dealing with real datasets to come up with the most parsimonious models.

Chapter 1

Introduction

1.1 Introduction

Additive regression models with fixed effects and smooth effects are among the most popular models used in practice. Their usage in semiparametric regression (e.g. Ruppert et al., 2003) and in nonparametric regression (e.g. Stone, 1985; Hastie and Tibshirani, 1999) has broadened the applications of regression. For nonparametric regression in additive models, Buja et al. (1989) proposed backfitting to fit the model. For variable selection and estimation in nonparametric regression, there are several popular proposals such as CART (Breiman et al., 1984), TURBO (Friedman and Silverman, 1989), BRUTO (Hastie, 1989), MARS (Friedman, 1991), Luo and Wahba (1997), Wahba (1990) and Eubank (1988).

In the Bayesian approach, Eubank (1988) describes Bayesian polynomial regression and its close relationship to splines and time series analysis. Some of the proposed solutions are Smith and Kohn (1996), Kleinman and Ibrahim (1998), Speckman and Sun (2003), Smith et al. (1998) and others. Dey et al. (1998) provide another resource for Bayesian nonparametric and semipara-

metric regression methods. One of the directions for nonparametric regression is focused on full Bayesian analysis by simulation techniques such as Markov Chain Monte Carlo (MCMC) (Gelfand and Smith, 1990).

Based on the pioneering work by Wahba (1990) in smoothing splines for nonparametric regression, Gu (2002) decomposed the regression function based on a tensor sum decomposition of inner product spaces into orthogonal subspaces so the estimated function from each subspace can be viewed separately. This is based on an ANOVA-type decomposition and is called the smoothing spline ANOVA (SSANOVA) model. The decomposition of the estimated function into main effects and interaction effects provides not only flexility to the fitted model but also makes it possible to select a parsimonious model from a large class of semiparametric additive models.

Estimating unknown functions has attracted many researchers' attention. Luo (1998) applied backfitting in SSANOVA, and Karcher and Wang (2001) proposed a Markov Chain Monte Carlo (MCMC) method, which leads to asymptotically consistent estimates for the SSANOVA model. To popularize the application of SSANOVA, Wang (1997) has developed a user-friendly package as an R function called ASSIST. Several different methods have been proposed for variable selection and model building in SSANOVA models. In the Gaussian regression setting, Gu (1992) proposed using cosine diagnostics as model checking tools after model fitting. For regression in exponential families, Zhang et al. (2004) proposed likelihood basis pursuit. Zhang and Lin (2006) used the COSSO-type penalized likelihood method to develop a computational algorithm for variable selection in SSANOVA for exponential families.

Current research related to smoothing spline ANOVA focuses on the fre-

quentist approach. We apply a Bayesian approach in SSANOVA to extend the statistical inference to model selection. After providing suitable prior distributions to the parameters of the model, the posterior distribution provides sufficient information for statistical inference. Model selection can be done through Bayes factors.

In the next section, we start with an introduction of smoothing splines in Section 2.1. Calculating the smoothing spline requires some fundamental theory of reproducing kernel Hilbert spaces, which will be covered in Section 2.2. The reproducing kernel solution to smoothing splines and the Bayesian interpretation of that solution is in Section 2.4. Details for a Bayesian approach to SSANOVA will be in Section 2.7.

Chapter 2

Smoothing Spline Models

2.1 Smoothing Spline for One Variable

Consider a nonparametric regression model with one independent variable x_i and response variable y_i,

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $x_i \in [0, 1]$, the ε_i are independent $N(0, \delta_0)$ random variables and f is an unknown smooth function. The smoothing spline $\widehat{f}$ minimizes the penalized likelihood

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_0^1 (f^{(m)}(x))^2 dx, \tag{2.2}$$

where $f^{(m)} = d^m f / dx^m$, and λ is a smoothing parameter. The first term measures closeness to the data, while the second term penalizes curvature in the function. As $\lambda \to 0$, $\widehat{f}$ converges to an interpolating spline. As $\lambda \to \infty$, $\widehat{f}$ approaches the least squares linear regression curve of degree $m-1$. Those two extreme cases have the function varying from very rough to very smooth, and

the choice of $\lambda \in (0, \infty)$ indexes an interesting class of functions in between. The focus of our project is in the cubic smoothing spline, where $\widehat{f}$ minimizes the penalized likelihood,

$$\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int_0^1 (f''(x))^2 dx. \tag{2.3}$$

The unknown function f can be approximated by a linear combination of terms from an appropriate basis. For a regular function, the number of basis terms doesn't need to be large, while still ensuring the estimated function is nearly without bias (Smith and Kohn, 1996). Choices of bases include reproducing kernel bases (Wahba, 1990), cubic polynomial splines (Friedman and Silverman, 1989; Friedman, 1991; Smith and Kohn, 1996), linear natural splines (Wahba, 1990), mixed radial bases, and others.

The exact solution for equation (2.3) is constructed based on reproducing kernel Hilbert spaces (Wahba, 1990), discussed in the next section.

2.2 Reproducing Kernel Hilbert Space

A space $\mathcal{L}$ is called *linear* if $x, y \in \mathcal{L}$ implies that $\alpha x + \beta y \in \mathcal{L}$ for all $\alpha, \beta \in \mathbb{R}$. A bilinear form (x, y) in a linear space $\mathcal{L}$ satisfies $(\alpha x + \beta y, z) = \alpha(x, z) + \beta(y, z)$, and $(x, \alpha y + \beta z) = \alpha(x, y) + \beta(x, z)$ for all $x, y, z \in \mathcal{L}$ and all $\alpha, \beta \in \mathbb{R}$. A linear space is often equipped with an inner product, a positive definite bilinear form with notation $\langle \cdot, \cdot \rangle$. A Hilbert space $\mathcal{H}$ is a complete inner product linear space. Consider a Hilbert space $\mathcal{H}$ of functions on domain $\mathcal{X}$. If the evaluation function $[x]f = f(x)$ is continuous in $\mathcal{H}$ for all $x \in \mathcal{X}$, then $\mathcal{H}$ is called a reproducing kernel Hilbert space. By the Riesz

representation theorem, there exists $R_x \in \mathcal{H}$, the representer of the evaluation functional $[x](\cdot)$, such that $\langle R_x, f\rangle = f(x)$ for all $f \in \mathcal{H}$. The symmetric bivariate function $R(x,y) = R_x(y) = \langle R_x, R_y\rangle$ has the reproducing property $\langle R(x,\cdot), f(\cdot)\rangle = f(x)$ and is called the reproducing kernel of the space $\mathcal{H}$.

The reproducing kernel is unique when exists. It can be shown that for every non-negative definite function $R(x,y)$ on $\mathcal{X}$, there corresponds a unique reproducing kernel Hilbert space $\mathcal{H}$ that has $R(x,y)$ as its reproducing kernel. So one can construct a reproducing kernel Hilbert space by specifying a non-negative definite function as its reproducing kernel (Gu, 2002).

If the reproducing kernel R of a space $\mathcal{H}$ on domain $\mathcal{X}$ can be decomposed into $R = R_0 + R_1$, where R_0 and R_1 are both non-negative definite, $R_0(x,\cdot), R_1(x,\cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$, and $\langle R_0(x,\cdot), R_1(y,\cdot)\rangle$=0 for all $x, y \in \mathcal{X}$, then the spaces $\mathcal{H}_0$ and $\mathcal{H}_1$ corresponding respectively to R_0 and R_1 form a tensor sum decomposition of $\mathcal{H}$ (Gu, 2002) that is introduced in the next section. The following introduces examples of reproducing kernels for each of $\mathcal{X} = [0,1]$ and $\mathcal{X} = \{1,\ldots,K\}$.

Example 2.1

For $f \in C^{(m)}[0,1]$, the standard Taylor expansion with integral remainder gives

$$f(x) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!} f^{(m)}(0) + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!} f^{(m)}(u)\, du, \tag{2.4}$$

where $(\cdot)_+ = \max\{0,\cdot\}$. Define the inner product

$$\langle f, g\rangle = \sum_{\nu=0}^{m-1} f^{(\nu)}(0) g^{(\nu)}(0) + \int_0^1 f^{(m)} g^{(m)} dx. \tag{2.5}$$

The representor of evaluation $[x]f = f(x)$ is

$$\begin{aligned} R_x(y) &= R_0(x,y) + R_1(x,y) \\ &= \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!}\frac{y^\nu}{\nu!} + \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}\,du. \end{aligned}$$

Note that

$$R_0(x,y) = \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!}\frac{y^\nu}{\nu!}, \tag{2.6}$$

$$R_1(x,y) = \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}\,du. \tag{2.7}$$

Now take derivatives with respect to y for both $R_0(x,y)$ and $R_1(x,y)$, evaluated at $y = 0$ for $R_0(x,y)$, evaluated at y for $R_1(x,y)$.

$$\begin{aligned} \frac{\partial^{m-1}}{\partial y^{m-1}} R_0(x,y) \mid_{y=0} &= \frac{x^{(m-1)}}{(m-1)!} \\ \frac{\partial}{\partial y} R_1(x,y) &= \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-2}}{(m-2)!}\,du, \\ \frac{\partial^{m-1}}{\partial y^{m-1}} R_1(x,y) &= \int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}(y-u)_+^0\,du = \int_0^y \frac{(x-u)_+^{m-1}}{(m-1)!}\,du, \\ \frac{\partial^m}{\partial y^m} R_1(x,y) \mid_y &= \frac{(x-y)_+^{m-1}}{(m-1)!}. \end{aligned}$$

Then

$$\begin{aligned} R_0^{(\nu)}(x,0) &= \frac{x^{(\nu)}}{\nu!}, \quad \nu = 0,\ldots,m-1, \\ R_1^{(m)}(x,y) &= \frac{(x-y)_+^{m-1}}{(m-1)!}. \end{aligned}$$

Now set $g = R_x$ and plugging the g into equation (2.5), we obtain equation (2.4). Thus $\langle R_x, f\rangle = f(x)$, which proves that R_x is the reproducing kernel.

Moreover, the non-negative definite functions $R_0(x, y)$ and $R_1(x, y)$ are the reproducing kernels for spaces $\mathcal{H}_0$ and $\mathcal{H}_1$ respectively. The kernel R_0 corresponds to the space of polynomials $\mathcal{H}_0 = \{f \in \mathcal{H} : f^{(m)} = 0\}$ with inner product $\langle f, g\rangle_{\mathcal{H}_0} = \sum_{\nu=0}^{m-1} f^{(\nu)}(0)g^{(\nu)}(0)$, and R_1 corresponds to the orthogonal complement of $\mathcal{H}_0$,

$$\mathcal{H}_1 = \{f \in \mathcal{H} : f^{(\nu)}(0) = 0, \nu = 0, 1, \ldots, m-1, \int_0^1 (f^{(m)}(x))^2 dx < \infty\}$$

with inner product $\langle f, g\rangle_{\mathcal{H}_1} = \int_0^1 f^{(m)} g^{(m)} dx$.

Example 2.2

For a function on the discrete domain $\mathcal{X} = \{1, \ldots, K\}$, write $\mathcal{H} = \mathcal{R}^K$. For any $f \in \mathcal{H}$, let $f = (f(1), \ldots, f(K))'$. Set $e_x = (0, \ldots, 0, 1, 0, \ldots, 0,)'$, the xth unit vector. With an inner product

$$\langle f, g\rangle = f^{'} g,$$

we have

$$\langle f, e_x\rangle = f^{'} e_x = f(x).$$

Thus the representor of evaluation $[x]f = f(x)$ is e_x. Hence, the reproducing kernel is given by $R(x, y) = \langle e_x, e_y\rangle = I_{\{x=y\}}$. Consider a decomposition of the reproducing kernel, $R(x, y) = 1/K + (I_{\{x=y\}} - 1/K) = R_0(x, y) + R_1(x, y)$. Since $(11^{'}/K)(I - 11^{'}/K) = 0_{K\times K}$, the inner product $\langle R_0(x, \cdot), R_1(y, \cdot)\rangle = 0$, for all x, y. This decomposition defines a tensor sum decomposition of the space $\mathcal{R}^K = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0 = \{f : f(1) = \cdots = f(K)\}$ and $\mathcal{H}_1 = \{f :$

$\sum_{x=1}^{K} f(x) = 0\}$. The inner product for $\mathcal{H}_0$ is $\langle f, g\rangle_{\mathcal{H}_0} = f^{'}(11^{'}/K)g$ and for $\mathcal{H}_1$ is $\langle f, g\rangle_{\mathcal{H}_1} = f^{'}(I - 11^{'}/K)g$.

2.3 Tensor Sum Decomposition of Inner Product Spaces

The distance between a point $f \in \mathcal{H}$ and a closed linear subspace $\mathcal{G} \subset \mathcal{H}$ is defined by $D[f, \mathcal{G}] = \inf_{g\in\mathcal{G}} \|f - g\|$. Since $\mathcal{G}$ is closed, there exists an $f_{\mathcal{G}} \in \mathcal{G}$, called the projection of f in $\mathcal{G}$, such that $\|f - f_{\mathcal{G}}\| = D[f, \mathcal{G}]$. It can be shown that $(f - f_{\mathcal{G}}, g) = 0$ for all $g \in \mathcal{G}$. The linear subspace $\mathcal{G}^c = \{f : (f, g) = 0, \forall g \in \mathcal{G}\}$ is called the orthogonal complement of $\mathcal{G}$. It can be verified that $\|f - f_{\mathcal{G}} - f_{\mathcal{G}^c}\|^2 = 0$, where $f_{\mathcal{G}} \in \mathcal{G}$ and $f_{\mathcal{G}^c} \in \mathcal{G}^c$ are the projections of f in $\mathcal{G}$ and $\mathcal{G}^c$, respectively. Therefore, there exists a unique decomposition $f = f_{\mathcal{G}} + f_{\mathcal{G}^c}$ for every $f \in \mathcal{H}$. It is clear that $(\mathcal{G}^c)^c = \mathcal{G}$. The decomposition $f = f_{\mathcal{G}} + f_{\mathcal{G}^c}$ is called the tensor sum decomposition (Murray and Von Neumann, 1936) and is denoted by $\mathcal{H} = \mathcal{G} \oplus \mathcal{G}^c$, $\mathcal{G}^c = \mathcal{H} \ominus \mathcal{G}$, or $\mathcal{G} = \mathcal{H} \ominus \mathcal{G}^c$. Multiple term tensor sum decompositions can be derived similarly.

A non-negative definite bilinear form $J(f, g)$ in a linear space $\mathcal{H}$ defines a semi-inner product in $\mathcal{H}$ that induces a seminorm $J(f) = J(f, f)$. Unless $J(f, g)$ is positive definite, the null space $\mathcal{N}_J = \{f : J(f, f) = 0, f \in \mathcal{H}\}$ is a linear subspace of $\mathcal{H}$ containing elements other than 0. Now suppose there is another non-negative definite bilinear form $\widetilde{J}(f, g)$ in $\mathcal{H}$ satisfying the following conditions: (i) it is positive definite when restricted to $\mathcal{N}_J$, so $\widetilde{J}(f) = \widetilde{J}(f, f)$ defines a full norm in $\mathcal{N}_J$ and (ii) for every $f \in \mathcal{H}$, there exists a $g \in \mathcal{N}_J$ such that $\widetilde{J}(f - g) = 0$. With such an $\widetilde{J}(f, g)$, it can be verified that $J(f, g)$ is positive definite in the linear subspace $\mathcal{N}_{\widetilde{J}} = \{f : \widetilde{J}(f, f) = 0, f \in \mathcal{H}\}$ and that $(J + \widetilde{J})(f, g)$ is positive definite in $\mathcal{H}$. Hence, a semi-inner product

can be made a full inner product either via restriction to a subspace or via augmentation by an extra term, a new inner product on its null sapce. So if $\mathcal{H}$ is complete under the norm induced by $(J+\widetilde{J})(f,g)$, then $\mathcal{N}_J$ and $\mathcal{N}_{\widetilde{J}}$ form a tensor sum decomposition of $\mathcal{H}$.

Example 2.3 refer to Gu (2002)

All square integrable functions on $[0,1]$ form a Hilbert space

$$\mathcal{L}_2[0,1] = \{f : \int_0^1 f^2 dx < \infty\}$$

with inner product $\langle f,g\rangle = \int_0^1 fg\,dx$. The space

$$\mathcal{G} = \{f : f = gI_{[x\leq .5]}, g \in \mathcal{L}_2[0,1]\}$$

is a closed linear subspace with orthogonal complement

$$\mathcal{G}^c = \{f : f = gI_{[x\geq .5]}, g \in \mathcal{L}_2[0,1]\}.$$

The bilinear form $J(f,g) = \int_0^{0.5} fg\,dx$ defines a semi-inner product in $\mathcal{L}_2[0,1]$, with null space

$$\mathcal{N}_{\widetilde{J}} = \mathcal{G}^c = \{f : f = gI_{[x\geq .5]}, g \in \mathcal{L}_2[0,1]\}.$$

Define $\widetilde{J}(f,g) = C\int_{0.5}^1 fg\,dx$, with $C>0$ a constant; one has an inner product $\langle f,g\rangle = (J+\widetilde{J})(f,g) = \int_0^{0.5} fgdx + C\int_{0.5}^1 fgdx$ on $\mathcal{L}_2[0,1]$. On $\mathcal{G} = \mathcal{L}_2 \ominus \mathcal{N}_J$, $J(f,g)$ is a full inner product.

Example 2.4

For a function on the discrete domain $\mathcal{X} \in \{1, \ldots, K\}$, consider the inner product

$$\langle f, g \rangle = \sum_{x=1}^{K} f(x)g(x) = f^{'} g.$$

From Example 1.2, the space $\mathcal{G} = \mathcal{H}_0 = \{f : f(1) = \cdots = f(K)\}$ is a closed linear subspace with orthogonal complement $\mathcal{G}^c = \mathcal{H}_1 = \{f : \sum_{x=1}^{K} f(x) = 0\}$. The bilinear form $J(f, g) = \langle f, g \rangle_{\mathcal{H}_1} = f^{'}(\boldsymbol{I} - \mathbf{1}\mathbf{1}^{'}/K)g$ defines a semi-inner product with null space $\mathcal{N}_J = \mathcal{G} = \mathcal{H}_0 = \{f : f(1) = \cdots = f(K)\}$. Define $\widetilde{J}(f, g) = c\boldsymbol{f}^{'}(\mathbf{1}\mathbf{1}^{'}/K)\boldsymbol{g}$, with $c > 0$ a constant; one has an inner product,

$$\langle f, g \rangle = (J + \widetilde{J})(f, g) = \boldsymbol{f}^{'}(\boldsymbol{I} + \frac{c-1}{K}\mathbf{1}\mathbf{1}^{'})\boldsymbol{g},$$

which reduces to the Euclidean inner product when $c = 1$. On $\mathcal{G}^c = \mathcal{H}_1 = \{f : \sum_{x=1}^{K} f(x) = 0\}$, $J(f, g)$ is a full inner product.

The application of tensor sum decompositions in smoothing splineF ANOVA will be illustrated in later sections.

2.4 Reproducing Kernel Solution to Smoothing Spline

Suppose that

$$\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1, \tag{2.8}$$

where $\mathcal{H}_0$ is a finite dimensional space with basis functions $\phi_1(t), \ldots, \phi_M(t)$, and $\mathcal{H}_1$ is a reproducing kernel Hilbert space with reproducing kernel $R_1(s, t)$. To compute the minimizer $\widehat{f}$ for equation (2.2), we consider the following

penalized least squares equation,

$$\sum_{i=1}^{n}\{y_i - L_i f\}^2 + \lambda\|P_1 f\|^2, \tag{2.9}$$

where L_i is the evaluation functional at observed points: $L_i f = f(t_i) = \langle \xi_i, f\rangle$ with the representer ξ_i, P_1 is the orthogonal projection operator of f onto $\mathcal{H}_1$ in $\mathcal{H}$ and λ is the smoothing parameter. The representer $\xi_i(t) = L_i R_1(t,\cdot)$ is exactly the reproducing kernel $R_1(t,t_i)$. It can be shown (Wahba, 1990) that the minimizer of equation (2.9) lies in the span of the null space $\{1,\phi_1(t),\ldots,\phi_M(t)\}$ plus the evaluator functionals $\{\xi_1,\ldots,\xi_n\}$. Thus

$$\widehat{f}(t) = \sum_{i=0}^{M} d_i\phi_i(t) + \sum_{j=1}^{n} c_j\xi_j(t). \tag{2.10}$$

Define the matrix

$$\boldsymbol{\Sigma} = \{\langle \xi_i, \xi_j\rangle\}_{n\times n} = \{R_1(t_i,t_j)\},$$

and let $\boldsymbol{c} = (c_1,\ldots,c_n)'$. Note that

$$\begin{aligned}\|P_1\widehat{f}\|^2 &= \|P_1(\sum_{i=0}^{M} d_i\phi_i(t) + \sum_{j=1}^{n} c_j\xi_j(t))\|^2 = \|\sum_{j=1}^{n} c_j\xi_j(t)\|^2 \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} c_i c_j\langle\xi_i,\xi_j\rangle = \boldsymbol{c'\Sigma c},\end{aligned}$$

which is a seminorm induced by the semi-inner product in $\mathcal{H}$. If $\boldsymbol{d} = (d_1,\ldots,d_M)'$ and

$$\boldsymbol{T} = (\phi_i(t_j))_{n\times M}, \tag{2.11}$$

equation (2.9) can be written as

$$\|\boldsymbol{y} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{\Sigma}\boldsymbol{c}\| + \lambda \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}. \tag{2.12}$$

Then the estimated $\widehat{\boldsymbol{f}} = (\widehat{f}(t_1), \ldots, \widehat{f}(t_n))'$ has the following form shown by (Kimeldorf and Wahba, 1971):

$$\widehat{\boldsymbol{f}} = \boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c}. \tag{2.13}$$

Equation (2.13) is over parameterized. Following Wahba (1990), one solution is obtained by solving

$$(\boldsymbol{\Sigma} + \lambda \boldsymbol{I})\boldsymbol{c} + \boldsymbol{T}\boldsymbol{d} = \boldsymbol{y}, \tag{2.14}$$

$$\boldsymbol{T}'\boldsymbol{c} = \boldsymbol{0}. \tag{2.15}$$

Again following Wahba (1990), consider the spectral decomposition of $\boldsymbol{T}\boldsymbol{T}'$, i.e.,

$$\boldsymbol{T}\boldsymbol{T}' = \boldsymbol{F}\boldsymbol{\Lambda}\boldsymbol{F}', \tag{2.16}$$

where $\boldsymbol{F}$ is an orthogonal matrix of eigenvectors and $\boldsymbol{\Lambda}$ is a diagonal matrix with eigenvalues as the elements. We write $\boldsymbol{F}$ as

$$\boldsymbol{F} = (\boldsymbol{F}_1, \boldsymbol{F}_2), \tag{2.17}$$

where $\boldsymbol{F}_1$ is the $n \times M$ matrix of vectors spanning the column space of $\boldsymbol{T}$ and $\boldsymbol{F}_2$ has dimension $n \times (n - M)$. So $\boldsymbol{T} = \boldsymbol{F}_1\boldsymbol{R}$, where $\boldsymbol{R} = \boldsymbol{F}_1^{'}\boldsymbol{T}$ is an $M \times M$ nonsingular matrix. Because $\boldsymbol{F}$ is orthonormal, $\boldsymbol{F}_1^{'}\boldsymbol{F}_2 = \boldsymbol{0}$ and $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_{n\times n}$. Since $\boldsymbol{T}'\boldsymbol{c} = \boldsymbol{0}$,

$$\boldsymbol{c} = \boldsymbol{F}_2\boldsymbol{\gamma} \tag{2.18}$$

for some $\boldsymbol{\gamma}$. By equations (2.14) and (2.15), the solutions for $\boldsymbol{\gamma}$ and $\boldsymbol{d}$ are

$$\boldsymbol{\gamma} = (\boldsymbol{F}_2^{'}\boldsymbol{M}\boldsymbol{F}_2)^{-1}\boldsymbol{F}_2^{'}\boldsymbol{y}, \tag{2.19}$$

$$\boldsymbol{d} = (\boldsymbol{F}_1'\boldsymbol{T})^{-1}\boldsymbol{F}_1^{'}(\boldsymbol{y} - \boldsymbol{M}\boldsymbol{c}), \tag{2.20}$$

where $\boldsymbol{M} = \boldsymbol{\Sigma} + \lambda\boldsymbol{I}$.

So far the smoothing parameter λ has been fixed. Good choice of λ is crucial to the performance of the spline estimates (Wahba, 1990). Several methods have been proposed including cross-validation (CV), generalized cross-validation (GCV), generalized maximum likelihood (GML) and unbiased risk (UBR) methods (Gu, 2002).

2.5 Bayesian interpretation for Reproducing Kernel Solution to Smoothing Spline

As shown first by Kimeldorf and Wahba (1971), the solution to the penalized likelihood equation (2.3) is equivalent to the Bayes estimate to a certain Bayes model with a limiting Gaussian prior.

Example 2.5

Consider the classical one-way ANOVA model with independent observations $y_i \sim N(\eta(x_i), \delta_0), i = 1, \ldots, n$, where $x_i \in \{1, \ldots, K\}$. Set $\boldsymbol{\eta} = \alpha \mathbf{1} + \boldsymbol{\eta}_1$ with independent priors $\alpha \sim N(0, \tau^2)$ for the mean and $\boldsymbol{\eta}_1 \sim N(0, \frac{\delta_0}{\lambda}(\boldsymbol{I} - \mathbf{1}\mathbf{1}'/K))$. Note that $\boldsymbol{\eta}_1'\mathbf{1} = 0$ almost surely and that $\overline{\eta} = \sum_{x=1}^{K} \eta(x)/K = \alpha$. The posterior mean of $\boldsymbol{\eta}$ is given by the minimizer of

$$\frac{1}{\delta_0}\sum_{i=1}^{n}(y_i - \eta(x_i))^2 + \frac{1}{\tau^2}\overline{\eta}^2 + \frac{\lambda}{\delta_0}\sum_{x=1}^{K}(\eta(x) - \overline{\eta})^2. \tag{2.21}$$

Letting $\tau \to \infty$ implies α has a flat prior.

Example 2.6

Consider $\eta(x) = \eta_0(x) + \eta_1(x)$ on $[0, 1]$, with $\eta_0(x)$ and $\eta_1(x)$ having independent Gaussian priors with mean 0 and covariance matrices,

$$\begin{aligned} E[\eta_0(x)\eta_0(y)] &= \tau^2 R_0(x, y) = \tau^2 \sum_{\nu=0}^{m-1} \frac{x^\nu}{\nu!}\frac{y^\nu}{\nu!}, \\ E[\eta_1(x)\eta_1(y)] = bR_1(x, y) &= b\int_0^1 \frac{(x-u)_+^{m-1}}{(m-1)!}\frac{(y-u)_+^{m-1}}{(m-1)!}du. \end{aligned}$$

Observing $y_i \sim N(\eta(x_i), \delta_0), \quad i = 1, \ldots, n$, the joint distribution of $\boldsymbol{y}$ and $\eta(x)$ is normal with mean zero and covariance matrix

$$\begin{pmatrix} b\boldsymbol{\Sigma} + \tau^2 \boldsymbol{T}\boldsymbol{T}' + \delta_0 \boldsymbol{I} & b\boldsymbol{\xi} + \tau^2 \boldsymbol{T}\boldsymbol{\phi} \\ b\boldsymbol{\xi}' + \tau^2 \boldsymbol{\phi}'\boldsymbol{T}' & bR_1(x, x) + \tau^2 \boldsymbol{\phi}'\boldsymbol{\phi} \end{pmatrix}, \tag{2.22}$$

where $\boldsymbol{\Sigma}$ is $n \times n$ with the (i, j)th entry $R_1(x_i, x_j)$, $\boldsymbol{T}$ is $n \times M$ with the (i, ν)th entry $x_i^{\nu-1}/(\nu - 1)!$, $\boldsymbol{\xi}$ is $n \times 1$ with the ith entry $R_1(x_i, x)$, and $\boldsymbol{\phi}$ is $M \times 1$ with the νth entry $x^{\nu-1}/(\nu-1)!$. Using a standard result from the multivariate

normal distribution, the posterior mean of $\eta(x)$ given $\boldsymbol{y}$ is

$$\begin{aligned} E[\eta(x) \mid \boldsymbol{y}] &= (b\boldsymbol{\xi}' + \tau^2\boldsymbol{\phi}'\boldsymbol{T}')(b\boldsymbol{\Sigma} + \tau^2\boldsymbol{T}\boldsymbol{T}' + \delta_0\boldsymbol{I})^{-1}\boldsymbol{y} \\ &= \boldsymbol{\xi}'(\boldsymbol{\Sigma} + \rho\boldsymbol{T}\boldsymbol{T}' + \lambda\boldsymbol{I})^{-1}\boldsymbol{y} + \boldsymbol{\phi}'\rho\boldsymbol{T}'(\boldsymbol{\Sigma} + \rho\boldsymbol{T}\boldsymbol{T}' + \lambda\boldsymbol{I})^{-1}\boldsymbol{y}, \end{aligned}$$

where $\rho = \tau^2/b$ and $b = \delta_0/\lambda$. Letting $\rho \to \infty$, it can be shown that the posterior mean $E[\eta(x) \mid \boldsymbol{y}]$ is of the form $\boldsymbol{\xi}'\boldsymbol{c} + \boldsymbol{\phi}'\boldsymbol{d}$, with the coefficients given by

$$\begin{aligned} \boldsymbol{c} &= (\boldsymbol{M}^{-1} - \boldsymbol{M}^{-1}\boldsymbol{T}(\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1})\boldsymbol{y}, \\ \boldsymbol{d} &= (\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{M}^{-1}\boldsymbol{y}, \end{aligned}$$

where $\boldsymbol{M} = \boldsymbol{\Sigma} + \lambda\boldsymbol{I}$ (Wahba, 1990).

These two cases demonstrate that the limiting posterior mean for the Bayesian problem is the same as the smoothing spline solution (2.18), (2.19) and (2.20). The setting for those two cases is based on equation (2.12). The penalty term is $\lambda\boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{c}$, which implies the prior for $\boldsymbol{c}$ has a Gaussian distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}^-/\lambda$. The inverse covariance matrices (the precision matrices) are proportional to the reproducing kernel R_1.

2.6 Bayesian approach for Reproducing Kernel Solution to Smoothing Spline

Now an alternative for the prior distribution is proposed. Instead of working with a precision matrix derived from a reproducing kernel, after certain

transformations we have the covariance matrix for the prior itself derived from the reproducing kernel. The setting for the posterior distribution stays the same. But there are some advantages to this transformation. This transformation has better properties for Bayesian computation. And the reproducing kernel as the covariance matrix in the prior distribution provides flexibility in decomposing the covariance components in the smoothing spline ANOVA, which is the focus in a later section.

Now set $\boldsymbol{\eta} = \boldsymbol{\Sigma c}$, so

$$\boldsymbol{\eta}'\boldsymbol{\Sigma}^{-}\boldsymbol{\eta} = \boldsymbol{c}'\boldsymbol{\Sigma}\boldsymbol{\Sigma}^{-}\boldsymbol{\Sigma c} = \boldsymbol{c}'\boldsymbol{\Sigma c}. \tag{2.23}$$

Maximizing (2.12) is equivalent to minimizing

$$-\frac{1}{2\delta_0}\|\boldsymbol{y} - \boldsymbol{T d} - \boldsymbol{\eta}\|^2 - \frac{\lambda}{2\delta_0}\boldsymbol{\eta}'\boldsymbol{\Sigma}^{-}\boldsymbol{\eta}, \tag{2.24}$$

which implies the corresponding prior distribution for $\boldsymbol{\eta}$ is

$$\boldsymbol{\eta} \mid \delta_0, \lambda \sim N_n\left(\mathbf{0}, \frac{\delta_0}{\lambda}\boldsymbol{\Sigma}\right). \tag{2.25}$$

Assign the prior distribution for $\boldsymbol{d}$ as

$$\boldsymbol{d} \mid \tau \sim N_M(\mathbf{0}, \tau\mathbf{I_M}). \tag{2.26}$$

In order to improve the MCMC in Bayesian computation, we will be working in the coordinates obtained by transforming the data using $\boldsymbol{F}$, the eigenvectors

from the spectral decomposition of $\boldsymbol{TT}'$. Set $\boldsymbol{v} = \boldsymbol{Td} + \boldsymbol{\eta}$ and $\boldsymbol{u} = \boldsymbol{F}'\boldsymbol{v}$. Then

$$\boldsymbol{v} = \boldsymbol{Td} + \boldsymbol{\eta} \mid \tau, \delta_0, \lambda \sim N_n \left(\mathbf{0}, \tau \mathbf{T}\mathbf{T}' + \frac{\delta_0}{\lambda} \boldsymbol{\Sigma} \right) \tag{2.27}$$

and

$$\boldsymbol{u} = \boldsymbol{F}'\boldsymbol{v} \mid \tau, \delta_0, \lambda \sim N_n \left(\mathbf{0}, \mathbf{F}' \left(\tau \mathbf{T}\mathbf{T}' + \frac{\delta_0}{\lambda} \boldsymbol{\Sigma} \right) \mathbf{F} \right). \tag{2.28}$$

To work in the orthogonal coordinates, also transform $\boldsymbol{y}$, so

$$\boldsymbol{y} \mid \boldsymbol{v}, \delta_0 \sim \mathrm{N}_n(\boldsymbol{v}, \delta_0 \boldsymbol{I}_n), \tag{2.29}$$

$$\boldsymbol{F}'\boldsymbol{y} \mid \boldsymbol{u}, \delta_0 \sim \mathrm{N}_n(\boldsymbol{u}, \delta_0 \boldsymbol{I}_n). \tag{2.30}$$

The joint distribution of $\boldsymbol{u}$ and $\boldsymbol{F}'\boldsymbol{y}$ is

$$\begin{pmatrix} \boldsymbol{u} \\ \boldsymbol{F}'\boldsymbol{y} \end{pmatrix} \sim N \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \boldsymbol{F}' \left(\tau \boldsymbol{T}\boldsymbol{T}' + \frac{\delta_0}{\lambda} \boldsymbol{\Sigma} \right) \boldsymbol{F} & \boldsymbol{F}' \left(\tau \boldsymbol{T}\boldsymbol{T}' + \frac{\delta_0}{\lambda} \boldsymbol{\Sigma} \right) \boldsymbol{F} \\ \boldsymbol{F}' \left(\tau \boldsymbol{T}\boldsymbol{T}' + \frac{\delta_0}{\lambda} \boldsymbol{\Sigma} \right) \boldsymbol{F} & \delta_0 \boldsymbol{I}_{n \times n} + \boldsymbol{F}' \left(\tau \boldsymbol{T}\boldsymbol{T}' + \frac{\delta_0}{\lambda} \boldsymbol{\Sigma} \right) \boldsymbol{F} \end{pmatrix} \right)$$

Using a standard result on multivariate normal distribution (e.g. Johnson and Wichern, 1998), the conditional distribution for $\boldsymbol{u} \mid \delta_0, \lambda, \tau$ is

$$\boldsymbol{u} \mid \boldsymbol{y}, \delta_0, \lambda, \tau \sim N(\boldsymbol{B}_\tau^{-1} \boldsymbol{F}'\boldsymbol{y}, \delta_0 \boldsymbol{B}_\tau^{-1}), \tag{2.31}$$

where $\boldsymbol{B}_\tau = \boldsymbol{I}_n + (\boldsymbol{F}'(\frac{\tau}{\delta_0} \boldsymbol{T}\boldsymbol{T}' + \frac{1}{\lambda} \boldsymbol{\Sigma}) \boldsymbol{F})^{-1}$.

The spectral decomposition showed $\boldsymbol{TT}' = \boldsymbol{F\Lambda F}'$ and $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{I}_n$, so

$$\boldsymbol{F}'(\boldsymbol{TT}')\boldsymbol{F} = \boldsymbol{F}'(\boldsymbol{F\Lambda F}')\boldsymbol{F} = \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_1 & \mathbf{0}_{M\times(n-M)} \\ \mathbf{0}_{(n-M)\times M} & \mathbf{0}_{(n-M)\times(n-M)} \end{pmatrix}, \quad (2.32)$$

where $\boldsymbol{\Lambda}_1$ is an $M \times M$ diagonal matrix. Also

$$\boldsymbol{F}'\boldsymbol{\Sigma F} = \begin{pmatrix} \boldsymbol{F}_1' \\ \boldsymbol{F}_2' \end{pmatrix} \boldsymbol{\Sigma}(\boldsymbol{F}_1, \boldsymbol{F}_2) = \begin{pmatrix} \boldsymbol{F}_1'\boldsymbol{\Sigma F}_1 & \boldsymbol{F}_1'\boldsymbol{\Sigma F}_2 \\ \boldsymbol{F}_2'\boldsymbol{\Sigma F}_1 & \boldsymbol{F}_2'\boldsymbol{\Sigma F}_2 \end{pmatrix}. \quad (2.33)$$

We have

$$\left(\boldsymbol{F}'\left(\frac{\tau}{\delta_0}\boldsymbol{TT}'\right)\boldsymbol{F} + \frac{1}{\lambda}\boldsymbol{F}'\boldsymbol{\Sigma F}\right)^{-1} = \begin{pmatrix} \frac{\tau}{\delta_0}\boldsymbol{\Lambda}_1 + \frac{1}{\lambda}\boldsymbol{F}_1'\boldsymbol{\Sigma F}_1 & \frac{1}{\lambda}\boldsymbol{F}_1'\boldsymbol{\Sigma F}_2 \\ \frac{1}{\lambda}\boldsymbol{F}_2'\boldsymbol{\Sigma F}_1 & \frac{1}{\lambda}\boldsymbol{F}_2'\boldsymbol{\Sigma F}_2 \end{pmatrix}^{-1} \quad (2.34)$$

Applying the well-known result (e.g. Horn and Johnson, 1985) for the inverse of a block matrix and letting $\tau \to \infty$, which leads to a noninformative prior distribution on $\boldsymbol{d}$,

$$\begin{pmatrix} \frac{\tau}{\delta_0}\boldsymbol{\Lambda}_1 + \frac{1}{\lambda}\boldsymbol{F}_1'\boldsymbol{\Sigma F}_1 & \frac{1}{\lambda}\boldsymbol{F}_1'\boldsymbol{\Sigma F}_2 \\ \frac{1}{\lambda}\boldsymbol{F}_2'\boldsymbol{\Sigma F}_1 & \frac{1}{\lambda}\boldsymbol{F}_2'\boldsymbol{\Sigma F}_2 \end{pmatrix}^{-1} \to \begin{pmatrix} \mathbf{0}_{M\times M} & \mathbf{0}_{M\times(n-M)} \\ \mathbf{0}_{(n-M)\times M} & \lambda(\boldsymbol{F}_2'\boldsymbol{\Sigma F}_2)^{-1} \end{pmatrix} \quad (2.35)$$

From (2.31), the limiting conditional distribution for $\boldsymbol{u} \mid \delta_0, \lambda$ is

$$\boldsymbol{u} \mid \boldsymbol{y}, \delta_0, \lambda \sim N(\boldsymbol{B}^{-1}\boldsymbol{F}'\boldsymbol{y}, \delta_0\boldsymbol{B}^{-1}), \quad (2.36)$$

where

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{I}_M & 0 \\ 0 & \boldsymbol{I}_{(n-M)\times(n-M)} + \lambda(\boldsymbol{F}_2'\boldsymbol{\Sigma}\boldsymbol{F}_2)^{-1} \end{pmatrix} \tag{2.37}$$

Note that letting $\tau \to \infty$ in the prior distribution of $\boldsymbol{d}$ is the same as putting a flat prior on $\boldsymbol{d}$. Now we have to transform $\boldsymbol{u}$ back to the original scale $\boldsymbol{v}$, so

$$\boldsymbol{v} \mid \boldsymbol{y}, \delta_0, \lambda \sim N(\boldsymbol{F}\boldsymbol{B}^{-1}\boldsymbol{F}'\boldsymbol{y}, \delta_0\boldsymbol{F}\boldsymbol{B}^{-1}\boldsymbol{F}) \tag{2.38}$$

2.7 Smoothing Spline ANOVA Models

Smoothing spline ANOVA is a function estimate based on an ANOVA type decomposition of the unknown mean function f. Before we go in to detail, we start this section with a two-way ANOVA example.

Example 2.7

Consider $\alpha \in \mathcal{X}_1 = \{1, \ldots, K_1\}$ and $\beta \in \mathcal{X}_2 = \{1, \ldots, K_2\}$. The usual model is

$$E(y_{jk}) = \mu + \alpha_j + \beta_k + \gamma_{jk}, \quad j = 1, \ldots, K_1, \quad k = 1, \ldots, K_2, \tag{2.39}$$

with side conditions $\sum_{j=1}^{K_1} \alpha_j = \sum_{k=1}^{K_2} \beta_k = \sum_{j=1}^{K_1} \gamma_{jk} = \sum_{k=1}^{K_2} \gamma_{jk} = 0$. Define the averaging operators as $A_1 f = \sum_{j=1}^{K_1} f(j,k)/J = \overline{f}_{\cdot k}$ and $A_2 f = \sum_{k=1}^{K_2} f(j,K)/K = \overline{f}_{j\cdot}$. The side conditions imply $A_1(I - A_1) = A_2(I - A_2) = 0$. The effects of each component in the two-way ANOVA model are defined

in terms of the averaging operators applied to f to obtain μ, α, β and γ by

$$\begin{aligned} A_1A_2f &= A_1\overline{f}_{k\cdot} = \overline{f}_{\cdot\cdot}, \\ (I-A_1)A_2f &= (I-A_1)\overline{f}_{j\cdot} = \overline{f}_{j\cdot} - \overline{f}_{\cdot\cdot}, \\ A_1(I-A_2)f &= A_1f - A_1A_2f = \overline{f}_{\cdot k} - \overline{f}_{\cdot\cdot}, \\ (I-A_1)(I-A_2)f &= (I-A_1-A_2+A_1A_2)f = f - \overline{f}_{j\cdot} - \overline{f}_{\cdot k} - \overline{f}_{\cdot\cdot}, \end{aligned}$$

where A_1A_2f represents the constant, $(I-A_1)A_2f$ represents the main effect of α, $A_1(I-A_2)f$ represents the main effect of β, and $(I-A_1)(I-A_2)f$ represents the interaction effect between α and β. The averaging operators, A_1 and A_2, decompose f into four parts, and each part is in a subspace which is orthogonal to the others.

This idea of ordinary ANOVA decomposition has been borrowed and applied to the decomposition of functions in a Hilbert $\mathcal{H}$ space. Suppose one has reproducing kernel Hilbert spaces $\mathcal{H}_{(\gamma)}$ on domains $\mathcal{X}_\gamma$, $\gamma = 1, \ldots, \Gamma$, respectively. Further, assume that each space $\mathcal{H}_{(\gamma)}$ has a one-way ANOVA decomposition built in via the tensor sum decompositions $\mathcal{H}_{(\gamma)} = \mathcal{H}_{0(\gamma)} \oplus \mathcal{H}_{1(\gamma)}$, where $\mathcal{H}_{0(\gamma)} = \{f : f \propto 1\}$ has a reproducing kernel $R_{0(\gamma)} \propto 1$ and $\mathcal{H}_{1(\gamma)}$ has a reproducing kernel $R_{1(\gamma)}$ satisfying side conditions $A_\gamma R_{1(\gamma)}(x_{(\gamma)}, \cdot) = 0$, for all $x_{(\gamma)} \in \mathcal{X}_\gamma$, where the A_γ are the averaging operators defining the one-way ANOVA decompositions on $\mathcal{X}_\gamma$. The tensor product space $\mathcal{H} = \otimes_{\gamma=1}^{\Gamma}\mathcal{H}_{(\gamma)}$ has a tensor sum decomposition

$$\mathcal{H} = \otimes_{\gamma=1}^{\Gamma}(\mathcal{H}_{0(\gamma)} \oplus \mathcal{H}_{1(\gamma)}). \tag{2.40}$$

If Γ=2, then

$$\begin{aligned}\mathcal{H} &= \otimes_{\gamma=1}^{2}(\mathcal{H}_{0(\gamma)} \oplus \mathcal{H}_{1(\gamma)}) \\ &= (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{0(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{0(x_2)}) \\ &\quad \oplus (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{1(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{1(x_2)}), \end{aligned} \tag{2.41}$$

where $\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{0(x_2)}$ is the space of constants, $\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{0(x_2)}$ is space of main effects in x_1, $\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{1(x_2)}$ is space of main effects in x_2, and $\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{1(x_2)}$ is the space of interaction effects between x_1 and x_2.

With reproducing kernel Hilbert space $\mathcal{H}$ with two discrete variables in $x_1 \in \mathcal{X}_1 = \{1, \ldots, K_1\}$ and $x_2 \in \mathcal{X}_2 = \{1, \ldots, K_2\}$, Example 2.7 can be reinterpreted. Set $\mathcal{H} = \{\boldsymbol{\mu} : \boldsymbol{\mu} \in \mathbb{R}^{K_1 \times K_2}\}$, where $\boldsymbol{u} = (u_{11}, \ldots, u_{1K_2}, \ldots, u_{K_1 1}, \ldots, u_{K_1 K_2})'$. The representer ξ_{ij} in $\mathcal{H}$ is $\boldsymbol{e}_{ij} = (0, \ldots, 0, 1, 0, \ldots, 0)'$, the ijth unit vector, such that

$$\langle \xi_{ij}, \boldsymbol{\mu} \rangle = \langle \boldsymbol{e}_{ij}.\boldsymbol{\mu} \rangle = \mu_{ij}. \tag{2.42}$$

This defines the reproducing kernel

$$R((i,j),(k,l)) = \langle \boldsymbol{e}_{ij}, \boldsymbol{e}_{kl} \rangle = I_{\{(i,j)=(k,l)\}}. \tag{2.43}$$

The null space is

$$\mathcal{H}_0 = \{\boldsymbol{\mu} : \mu_{ij} = c, \forall i, j, i.e. \;\; \boldsymbol{\mu} = c\mathbf{1}_{K_1 K_2}\}. \tag{2.44}$$

Since

$$\mu = \langle \frac{1}{K_1 K_2} \mathbf{1}_{K_1 K_2}, \boldsymbol{\mu} \rangle = \langle \xi_{ij}, \boldsymbol{u} \rangle, \tag{2.45}$$

the reproducing kernel is

$$R_0((i,j),(k,l)) = \langle \xi_{ij}, \xi_{ij} \rangle = \langle \frac{1}{K_1 K_2} \mathbf{1}_{K_1 K_2}, \frac{1}{K_1 K_2} \mathbf{1}_{K_1 K_2} \rangle = \frac{1}{K_1 K_2} \mathbf{1}_{K_1 K_2} \tag{2.46}$$

where $\mathbf{1}_k = (1, \ldots, 1)'$ is with k terms of 1 in a vector.

Consider the main effect at level $x_1 \in \mathcal{X}_1$ with reproducing kernel Hilbert space is $\mathcal{H}_{1(x_1)}$. The main effects for x_1 are defined to be

$$\begin{aligned} \overline{\mu}_{i\cdot} - \overline{\mu}_{\cdot\cdot} &= \boldsymbol{e}'_{ij} \boldsymbol{\mu} - \frac{1}{K_1 K_2} \mathbf{1}'_{K_1 K_2} \boldsymbol{\mu} \\ &= \langle \boldsymbol{e}_{ij} - \frac{1}{K_1 K_2} \mathbf{1}_{K_1 K_2}, \boldsymbol{\mu} \rangle = \langle \xi_{ij}, \boldsymbol{\mu} \rangle, \end{aligned} \tag{2.47}$$

where $\boldsymbol{e}_{ij} = (0, \ldots, 0, 1/K_2, \ldots, 1/K_2, 0, \ldots, 0)'$, the ith block with elements $1/K_2$. The representer is $\xi_{ij} = \boldsymbol{e}_{ij} - \frac{1}{K_1 K_2} \mathbf{1}_{K_1 K_2}$, which is

$$\xi_{ij}(k,l) = \begin{cases} \frac{1}{K_2} - \frac{1}{K_1 K_2} & if\ i = k \\ -\frac{1}{K_1 K_2} & if\ i \neq k \end{cases}.$$

The reproducing kernel is defined as

$$\begin{aligned} R_{1(x_1)}((i,j),(k,l)) &= \langle \xi_{(ij)}, \xi_{(kl)} \rangle \\ &= \begin{cases} \frac{1}{K_2} - \frac{1}{K_1 K_2} & if\ i = k \\ -\frac{1}{K_1 K_2} & if\ i \neq k \end{cases} \end{aligned}$$

$$= \frac{1}{K_2} I_{\{i=k\}} - \frac{1}{K_1 K_2} = \frac{1}{K_2}\left(I_{\{i=k\}} - \frac{1}{K_1}\right). \quad (2.48)$$

Thus the tensor product space for $\mathcal{H}_{(x_1)} = \{\boldsymbol{\mu} \in \mathbb{R}^{K_1 \times K_2} : \mu_{ij} = \overline{\mu}_{i\cdot}, \forall i, j\}$ can be expressed as

$$\mathcal{H}_{(x_1)} = \mathcal{H}_{0(x_1)} \oplus \mathcal{H}_{1(x_1)}. \quad (2.49)$$

The reproducing kernel under subspace $\mathcal{H}_{0(x_1)}$ is R_0 in equation (2.46), and the reproducing kernel for subspace $\mathcal{H}_{1(x_1)}$ is $R_{1(x_1)}$ in equation (2.48).

With the same approach applied to x_2, the tensor product space for $\mathcal{H}_{(x_2)} = \{\boldsymbol{\mu} \in \mathbb{R}^{K_1 \times K_2} : \mu_{ij} = \overline{\mu}_{\cdot j}, \forall i, j\}$ can be expressed as

$$\mathcal{H}_{(x_2)} = \mathcal{H}_{0(x_2)} \oplus \mathcal{H}_{1(x_2)}. \quad (2.50)$$

The reproducing kernels under subspace $\mathcal{H}_{0(x_2)}$ is R_0 in equation (2.46) and $\mathcal{H}_{1(x_2)}$ is

$$R_{1(x_2)}((i,j),(k,l)) = \frac{1}{K_1}\left(I_{\{j=l\}} - \frac{1}{K_2}\right). \quad (2.51)$$

Thus the tensor product space can be expressed as

$$\begin{aligned}
\mathcal{H} &= (\mathcal{H}_{0(x_1)} \oplus \mathcal{H}_{1(x_1)}) \otimes (\mathcal{H}_{0(x_2)} \oplus \mathcal{H}_{1(x_2)}) \\
&= (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{0(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{0(x_2)}) \oplus (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{1(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{1(x_2)}) \\
&= \mathcal{H}_0 \oplus \mathcal{H}_{(x_1)} \oplus \mathcal{H}_{(x_2)} \oplus \mathcal{H}_{(x_1,x_2)}, \quad (2.52)
\end{aligned}$$

where $\mathcal{H}_0$ is the space of constants, $\mathcal{H}_{(x_1)}$ is the space of main effects for

x_1, $\mathcal{H}_{(x_2)}$ is the space of main effects for x_2 and $\mathcal{H}_{(x_1,x_2)}$ is the space of the interaction effects. The corresponding reproducing kernels for each subspace are $R_0((i,j),(k,l))$ in equation (2.46), $R_{1(x_1)}((i,j),(k,l))$ in equation (2.48), $R_{1(x_2)}((i,j),(k,l))$ in equation (2.51) and

$$\begin{aligned} R_{1(x_1,x_2)}((i,j),(k,l)) &= R((i,j),(k,l)) - R_0((i,j),(k,l)) - R_{1(x_1)}((i,j),(k,l)) \\ &- R_{1(x_2)}((i,j),(k,l)). \end{aligned} \tag{2.53}$$

Example 2.8

Now consider two variables, $x_1 \in \mathcal{X}_1 = \{1,\ldots,K_1\}$ and $x_2 \in \mathcal{X}_2 = [0,1]$. Consider reproducing kernel Hilbert spaces $\mathcal{H}_{(x_1)} = \mathcal{H}_{0(x_1)} \oplus \mathcal{H}_{1(x_1)}$ on domain $\mathcal{X}_1$ and $\mathcal{H}_{(x_2)} = \mathcal{H}_{0(x_2)} \oplus \mathcal{H}_{1(x_2)}$ on domain $\mathcal{X}_2$ with the linear spline. Analogous to Example 2.7, the reproducing kernels for the discrete variable x_1 under subspaces $\mathcal{H}_{0(x_1)} = \{\mu \mathbf{1}_{K_1} : \mu \in \mathbb{R}\}$ and $\mathcal{H}_{1(x_1)} = \{\boldsymbol{\mu} \in \mathbb{R}^{K_1} \; s.t. \; \langle \boldsymbol{\mu}, \mathbf{1}_{K_1} \rangle = 0, \; i.e. \; \sum_{i=1}^{K_1} \mu_i = 0\}$ are

$$\begin{aligned} R_{0(x_1)}(i,j) &= \frac{1}{K_1}, \\ R_{1(x_1)}(i,j) &= I_{\{i=j\}} - \frac{1}{K_1}. \end{aligned} \tag{2.54}$$

The reproducing kernel Hilbert spaces for x_2 are

$$\begin{aligned} \mathcal{H}_{0(x_2)} &= \{f : f' = 0\}, \\ \mathcal{H}_{1(x_2)} &= \{f : f(0) = 0, \int_0^1 (f')^2 \, dx < \infty\}. \end{aligned} \tag{2.55}$$

Refer to equation (2.6) and (2.7). When $m = 1$, the reproducing kernels for

subspaces $\mathcal{H}_{0(x_2)}$ and $\mathcal{H}_{1(x_2)}$ are

$$\begin{aligned} R_{0(x_2)}(x,y) &= 1, \\ R_{1(x_2)}(x,y) &= \int_0^1 (x-u)_+(y-u)_+\,du = x \wedge y. \end{aligned} \tag{2.56}$$

The tensor product space is the same as Example 2.7,

$$\begin{aligned} \mathcal{H} &= (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{0(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{0(x_2)}) \oplus (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{1(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{1(x_2)}) \\ &= \mathcal{H}_0 \oplus \mathcal{H}_{(x_1)} \oplus \mathcal{H}_{(x_2)} \oplus \mathcal{H}_{(x_1,x_2)}. \end{aligned}$$

The reproducing kernels for each of the subspaces are

$$\begin{aligned} \mathcal{H}_0 &: R_{0(x_1)}(i,x)R_{0(x_2)}(j,y) = \frac{1}{K_1}, \\ \mathcal{H}_{(x_1)} &: R_{1(x_1)}(i,x)R_{0(x_2)}(j,y) = I_{\{i=j\}} - \frac{1}{K_1}, \\ \mathcal{H}_{(x_2)} &: R_{0(x_1)}(i,x)R_{1(x_2)}(j,y) = (\frac{1}{K_1})(x \wedge y), \\ \mathcal{H}_{(x_1,x_2)} &: R_{1(x_1)}(i,x)R_{1(x_2)}(j,y) = (I_{\{i=j\}} - \frac{1}{K_1})(x \wedge y). \end{aligned}$$

Example 2.9

Another case to consider is with the same two spaces in $x_1 \in \mathcal{X}_1 = \{1, \ldots, K_1\}$ and $x_2 \in \mathcal{X}_2 = [0,1]$ but with cubic spline smoothing on x_2. The reproducing kernel Hilbert space for $x_1 \in \mathcal{X}_1$ stays the same with $\mathcal{H}_{(x_1)} = \mathcal{H}_{0(x_1)} \oplus \mathcal{H}_{1(x_1)}$. Now the reproducing kernel Hilbert space for $x_2 \in \mathcal{X}_2$ with cubic spline is $\mathcal{H}_{(x_2)} = \mathcal{H}_{00(x_2)} \oplus \mathcal{H}_{01(x_2)} \oplus \mathcal{H}_{11(x_2)}$. With $m = 2$, equation (2.6)

showed

$$\begin{aligned} R_{0(x_2)}(x,y) &= R_{00(x_2)}(x,y) + R_{01(x_2)}(x,y) = 1 + xy, \\ R_{11(x_2)}(x,y) &= \int_0^1 (x-u)_+(y-u)_+ \, du, = (x \wedge y)^2(3(x \vee y) - (x \wedge y))/6. \end{aligned}$$

The reproducing kernel $R_{0(x_2)}(x,y)$ is decomposed into $R_{00(x_2)}(x,y)$ plus $R_{01(x_2)}(x,y)$ for the corresponding reproducing kernel Hilbert spaces $\mathcal{H}_{00(x_2)}$ and $\mathcal{H}_{01(x_2)}$. The space $\mathcal{H}_{00(x_2)} = \{f : f(x_2) = c\}$ is the space of constants, $\mathcal{H}_{01(x_2)} = \{f : f(x_2) = bx_2\}$ is the space of linear effects, and $\mathcal{H}_{11(x_2)} = \{f : f(0) = f'(0) = 0, \ \int_0^1 (f'')^2 \, dx < \infty\}$ is the space of smooth effects.

Thus the corresponding reproducing kernels under each subspace are

$$\begin{aligned} R_{0(x_1)}(i,j) &= \frac{1}{K_1}, \\ R_{1(x_1)}(i,j) &= I_{\{i=j\}} - \frac{1}{K_1}, \\ R_{00(x_2)}(x,y) &= 1, \\ R_{01(x_2)}(x,y) &= xy, \\ R_{11(x_2)}(x,y) &= (x \wedge y)^2(3(x \vee y) - (x \wedge y))/6. \end{aligned} \tag{2.57}$$

We construct a tensor product space with six tensor sum terms,

$$\begin{aligned} \mathcal{H} &= (\mathcal{H}_{0(x_1)} \oplus \mathcal{H}_{1(x_1)}) \otimes (\mathcal{H}_{00(x_2)} \oplus \mathcal{H}_{01(x_2)} \oplus \mathcal{H}_{11(x_2)}) \\ &= (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{00(x_2)}) \oplus (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{01(x_2)}) \oplus (\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{11(x_2)}) \\ &\quad \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{00(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{01(x_2)}) \oplus (\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{11(x_2)}) \end{aligned} \tag{2.58}$$

where $\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{00(x_2)}$ is the space of constants, $\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{01(x_2)}$ is the one

dimensional space of a linear effect in x_2, $\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{11(x_2)}$ is the space of smooth effects in x_2, $\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{00(x_2)}$ is the space of level effects in x_1, $\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{01(x_2)}$ is the space of interaction effects between level effects in x_1 and linear effects in x_2, and $\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{11(x_2)}$ is the space of interaction effects between level effects in x_1 and smooth effects in x_2. The corresponding reproducing kernels for each subspace are

$$\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{00(x_2)} \quad : \quad R_{0(x_1)}(i,x)R_{00(x_2)}(j,y) = \frac{1}{K_1}, \tag{2.59}$$

$$\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{01(x_2)} \quad : \quad R_{0(x_1)}(i,x)R_{01(x_2)}(j,y) = (\frac{1}{K_1})(xy), \tag{2.60}$$

$$\mathcal{H}_{0(x_1)} \otimes \mathcal{H}_{11(x_2)} \quad : \quad R_{0(x_1)}(i,x)R_{11(x_2)}(j,y) = \left(I_{\{i=j\}} - \frac{1}{K_1}\right) \times ((x \wedge y)^2(3(x \vee y) - (x \wedge y))/6), \tag{2.61}$$

$$\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{00(x_2)} \quad : \quad R_{1(x_1)}(i,x)R_{00(x_2)}(j,y) = \left(I_{\{i=j\}} - \frac{1}{K_1}\right), \tag{2.62}$$

$$\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{01(x_2)} \quad : \quad R_{1(x_1)}(i,x)R_{01(x_2)}(j,y) = \left(I_{\{i=j\}} - \frac{1}{K_1}\right)(xy), \tag{2.63}$$

$$\mathcal{H}_{1(x_1)} \otimes \mathcal{H}_{11(x_2)} \quad : \quad R_{1(x_1)}(i,x)R_{11(x_2)}(j,y) = \left(I_{\{i=j\}} - \frac{1}{K_1}\right) \times (x \wedge y)^2(3(x \vee y) - (x \wedge y))/6), \tag{2.64}$$

respectively.

For each of the cases discussed above, the reproducing kernel Hilbert space has decomposition as $\mathcal{H} = \mathcal{H}_0 \oplus \sum_{k=1}^{p} \mathcal{H}_k$, where H_0 is the space that is not penalized, and each $\mathcal{H}_k$ is a reproducing kernel Hilbert space with reproducing kernel R_k. The SSANOVA model is constructed based on $\mathcal{H}_0 \oplus \sum_{k=1}^{p} \mathcal{H}_k$. The estimate of f is the minimizer of

$$\sum_{i=1}^{n} \{y_i - L_i f\}^2 + \lambda \sum_{k=1}^{p} \theta_k^{-1} \|P_k f\|^2, \tag{2.65}$$

where P_k is the orthogonal projection operator of f onto $\mathcal{H}_k$ in $\mathcal{H}$. Let $\xi_{ki}(t) = L_i R_k(t, \cdot)$ and $\boldsymbol{\Sigma}_k = \{\langle \xi_{ki}, \xi_{kj} \rangle\}_{n \times n}$. The solution to equation (2.65) is

$$\begin{aligned} \widehat{f}(t) &= \sum_{i=1}^{M} d_i \phi_i(t) + \sum_{j=1}^{n} c_j (\sum_{k=1}^{p} \theta_k \xi_{kj}(t)) \\ \widehat{\boldsymbol{f}} &= \boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c}, \end{aligned} \tag{2.66}$$

where $\boldsymbol{c}$ and $\boldsymbol{d}$ are solutions to equation (2.7) with $\boldsymbol{\Sigma}$ replaced by $\sum_{k=1}^{p} \theta_k \boldsymbol{\Sigma}_k$. Smoothing parameters $\lambda/\theta_1, \ldots, \lambda/\theta_p$ can be estimated similarly using GCV, GML, and UBR methods. Wang (1997) has developed an R function "*ssr*" to fit SSANOVA models. Gu (2002) has an algorithm to solve for $\boldsymbol{c}$, $\boldsymbol{d}$, and $\lambda/\theta_1, \ldots, \lambda/\theta_p$ with GCV, GML, and UBR methods simultaneously.

We propose a Bayesian approach in the next chapter to estimate the parameters $\boldsymbol{c}$ and $\boldsymbol{d}$ and smoothing parameters simultaneously through MCMC.

Chapter 3

Bayesian Smoothing Spline ANOVA

With the Bayesian interpretation for the reproducing kernel solution to smoothing splines, it's natural to consider a fully Bayesian approach. In this chapter, we discuss a fully Bayesian approach to smoothing splines and SSANOVA.

3.1 Fully Bayesian approach to Smoothing Splines

In Section 2.5, we discussed a Bayesian interpretation for the reproducing kernel solution to smoothing splines following Kimeldorf and Wahba (1971). The Bayesian interpretation is based on equation (2.12), which implies the prior for $\boldsymbol{c}$ has a Gaussian distribution with mean zero and covariance matrix $\delta_0/\lambda\boldsymbol{\Sigma}^-$. To implement a Bayesian approach, the computation could be intensive since each MCMC step requires the precision of $\boldsymbol{\Sigma}$. The alternative we propose in Section 2.6 transforms the model so the prior has a Gaussian distribution with covariance matrix proportional to $\boldsymbol{\Sigma}$. This not only improves MCMC computation but facilitates the construction of priors in the Bayesian approach to SSANOVA. This alternative provides the conditional distribution

for $\boldsymbol{u}$ in equation (2.36). So the solution for the smoothing spline is $\boldsymbol{v} = \boldsymbol{F}\boldsymbol{u}$ given δ_0 and λ.

3.1.1 Bayesian hierarchical model

To implement a fully Bayesian approach in a hierarchical model, we require priors for δ_0 and λ. Since we took $\tau \to \infty$, we applied the result of equation (2.35). The prior distribution for $\boldsymbol{u} = (\boldsymbol{u}_1', \boldsymbol{u}_2')'$ in equation (2.28) is now

$$\boldsymbol{u}_2 \mid \delta_0, \lambda \sim N_{(n-M)}\left(\mathbf{0}, \frac{\delta_0}{\lambda}(\mathbf{F_2'\Sigma F_2})\right). \tag{3.1}$$

Following many authors seeking objective priors for regression models, we will use the Jeffrey's (invariance) prior for δ_0. We will follow White (2006) and Liang et al. (2008) and use the Pareto for λ. Thus we have

$$\begin{aligned} [\delta_0] &\propto \frac{1}{\delta_0}, \quad \delta_0 > 0, \\ [\lambda \mid a_0] &= \frac{a_0}{(a_0+\lambda)^2}, \quad \lambda \geq 0. \end{aligned} \tag{3.2}$$

To sample from the Pareto distribution efficiently, we use a hierarchical structure as in White (2006). With

$$\begin{aligned} \lambda \mid \phi &\sim Exp(\phi), \\ \phi \mid a_0 &\sim Exp(a_0), \end{aligned} \tag{3.3}$$

the marginal for λ is the Pareto distribution (3.2). Since $\boldsymbol{F}_2'\boldsymbol{\Sigma}\boldsymbol{F}_2$ may be singular, we may need a full rank parameterization. Let $\boldsymbol{F}_2'\boldsymbol{\Sigma}\boldsymbol{F}_2 = \boldsymbol{Q}\boldsymbol{D}\boldsymbol{Q}'$ be the spectral decomposition, where $\boldsymbol{Q}$ is the matrix of eigenvectors corresponding

to the nonzero eigenvalues and $\boldsymbol{D}$ is the diagonal matrix with the nonzero eigenvalues in the diagonals. Letting $\boldsymbol{u}_2 = \boldsymbol{Q}\boldsymbol{v}$, then

$$\boldsymbol{v} \mid \delta_0, \lambda \sim N\left(\mathbf{0}, \frac{\delta_0}{\lambda}\mathbf{D}\right). \tag{3.4}$$

The full conditionals for each of the parameters are

$$\begin{aligned}
\boldsymbol{d} \mid \delta_0 &\sim N((\boldsymbol{T}'\boldsymbol{F}_1\boldsymbol{F}_1'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{F}_1\boldsymbol{F}_1'\boldsymbol{y}, \delta_0(\boldsymbol{T}'\boldsymbol{F}_1\boldsymbol{F}_1'\boldsymbol{T})^{-1}), \\
\boldsymbol{v} \mid \lambda, \delta_0 &\sim N((\boldsymbol{I}_r + \lambda\boldsymbol{D}^{-1})^{-1}\boldsymbol{Q}'\boldsymbol{F}_2'\boldsymbol{y}, \delta_0(\boldsymbol{I}_r + \lambda\boldsymbol{D}^{-1})), \\
\lambda \mid \boldsymbol{v}, \delta_0, \phi &\sim Gamma\left(\frac{r}{2} + 1, \frac{2\delta_0}{\boldsymbol{v}'\boldsymbol{D}^{-1}\boldsymbol{v} + 2\delta_0\phi}\right), \\
\phi \mid \lambda, a_0 &\sim Gamma\left(2, \frac{1}{\lambda + a_0}\right), \\
\delta_0 \mid \lambda &\sim Inverse\ Gamma\left(\frac{n+r}{2}, \frac{(\boldsymbol{F}_2'\boldsymbol{y} - \boldsymbol{u}_2)^2}{2} + \frac{\lambda\boldsymbol{v}'\boldsymbol{D}^{-1}\boldsymbol{v}}{2}\right)
\end{aligned} \tag{3.5}$$

where r is the rank of $\boldsymbol{D}$. We demonstrate this model by a simulated example first, and then apply the model to a dataset from a manufacturing production line.

3.1.2 Effective Degrees of Freedom in Smoothing Spline and Prior for λ

Equation (2.38) showed the conditional distribution of $\boldsymbol{v}$ in a Bayesian smoothing spline model. The Bayes estimate for $\boldsymbol{v}$ given λ is the smoothing spline estimate,

$$\boldsymbol{v} = \widehat{\boldsymbol{y}} = \boldsymbol{F}\boldsymbol{B}^{-1}\boldsymbol{F}'\boldsymbol{y}, \tag{3.6}$$

where

$$\boldsymbol{B} = \begin{pmatrix} \boldsymbol{I}_M & 0 \\ 0 & \boldsymbol{I}_{(n-M)} + \lambda(\boldsymbol{F}_2^{'}\boldsymbol{\Sigma}\boldsymbol{F}_2)^{-1} \end{pmatrix}. \tag{3.7}$$

Thus the matrix $\boldsymbol{F}\boldsymbol{B}^{-1}\boldsymbol{F}'$ is known as the smoother matrix. Following Hastie et al. (2001), the complexity of a model can be described by the "effective number of parameters," which is defined to be the trace of the smoother matrix. The smoother matrix $\boldsymbol{S}(\lambda)$ is

$$\boldsymbol{S}(\lambda) = \boldsymbol{F}\boldsymbol{B}^{-1}\boldsymbol{F}' = (\boldsymbol{F}_1, \boldsymbol{F}_2) \begin{pmatrix} \boldsymbol{I}_M & 0 \\ 0 & (\boldsymbol{I}_{(n-M)} + \lambda(\boldsymbol{F}_2^{'}\boldsymbol{\Sigma}\boldsymbol{F}_2)^{-1})^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{F}_1' \\ \boldsymbol{F}_2' \end{pmatrix} \tag{3.8}$$

Thus the trace of this smoother matrix is

$$\begin{aligned} d(\lambda) = tr(\boldsymbol{S}(\lambda)) &= tr\left[\boldsymbol{F}_1\boldsymbol{F}_1' + \boldsymbol{F}_2(\boldsymbol{I}_{(n-M)} + \lambda(\boldsymbol{F}_2^{'}\boldsymbol{\Sigma}\boldsymbol{F}_2)^{-1})^{-1}\boldsymbol{F}_2'\right] \\ &= M + \sum_{i=1}^{r} \frac{1}{1+\lambda d_i^{-1}}. \end{aligned} \tag{3.9}$$

This equation shows how λ is related to the complexity of the model. The trace of the smoother matrix $d(\lambda)$ defines the effective degrees of freedom of a smoothing spline. This very useful tool allows us a more intuitive way to specify a prior for λ. In the next two examples we discuss how does the effective degrees of freedom facilitate us to select the prior information on λ. Following White (2006), since $d(\lambda)$ is a monotone function of λ, the median of the prior distribution of $d(\lambda)$ is $d(\lambda_m)$, where λ_m is the median of the prior on λ. For the Pareto distribution (3.2), the median is a_0. Thus the median for the prior on $d(\lambda)$ is $d(a_0)$. White (2006) suggested choosing a_0 by trial and error to have

a desirable prior degrees of freedom.

3.1.3 Simulated Example

Consider a simple simulation from a balanced design. The independent variables include a discrete variable (x_1) with two levels and a continuous variable (x_2). Let $\boldsymbol{e} = (1, \ldots, m)'/m$, $\mathbf{1}_1 = (1, 1)'$ and $\mathbf{1}_2 = (1, \ldots, 1)'_{m\times 1}$ where $m = 10$. Take $\boldsymbol{x}_1 = (1, 2)' \otimes \mathbf{1}_2$, $\boldsymbol{x}_2 = \boldsymbol{e} \otimes \mathbf{1}_1$ and $y_i = 1 + 2I_{\{x_{1i}=1\}} + 3\sin(2\pi x_{2i} - \pi) + \varepsilon_i$ with $\delta_0 = 1$. The sample size is $n = 20$. The model proposed is a smoothing spline function

$$f(x_2) \quad = \quad \mu + \beta x_2 + s_2(x_2), \tag{3.10}$$

where μ is the constant term, β is the coefficient for the linear effect in x_2 and $s_2(x_2)$ is the smooth effect of x_2.

After assigning a value for the parameter a_0, all the parameters can be sampled by Gibbs sampler in MCMC from their full conditionals by (3.5). Then the Bayes estimates for each parameter are their posterior means.

The parameter a_0 is the median of the Pareto distribution for the prior distribution of λ. Even though the Pareto distribution has no mean or variance, the choice of median is still informative (refer to Section 3.1.2 for details). So what is a reasonable choice for a_0? Following Hastie et al. (2001), the complexity of a model can be described by the "effective number of parameters," which is defined to be the trace of the smoother matrix as defined in equation (3.9). Based on the smoother matrix, effective number of parameters are 7.9, 7.2, 4.8 and 2.4 when λ is set to 0.000001, 0.00001, 0.0001 and 0.001 respec-

tively. We chose a_0 to be 0.0001, which provided about 5 effective degrees of freedom. Note that includes 1 degree of freedom in the linear term plus 5 degrees of freedom from the smooth term to give 6 degrees of freedom for the fit. This fit should be able to catch the curve effect in the data. To observe the smoothing effect of λ, we tried different values a_0 =0.000001, 0.00001, 0.0001 and 0.001 in the prior distribution for λ. The Bayes estimates for λ are 0.0006, 0.00062, 0.001 and 0.005, which corresponds to equivalent degrees of freedom of 2.9, 2.8, 2.2 and 1.3 in (3.8) respectively. Figures 3.1 shows the MCMC trace plots for samples of the $\log(\lambda)$ from the models with a_0=0.001, a_0=0.0001, a_0=0.00001 and a_0=0.000001. Convergence is rapid in all cases. Figure 3.2 shows the fits of the models with a_0=0.001, a_0=0.0001, a_0=0.00001, a_0=0.000001 and the true function. The fits for a_0=0.00001 and a_0=0.000001 are almost identical. The fit is quite robust to choice of a_0. With the a_0 ranges from 0.001 to 0.000001, those fits are similar. The results suggest that the prior has little influence in estimating λ.

3.1.4 Manufacturing Example

The data used to illustrate this model come from the monitoring process for a production line stamping parts for circuit breakers. Regular measurements were taken on the metal parts and measured in inches. The data consists of 474 measurements of the metal parts along with the corresponding date and time when the measurements took place and the operators who took the measurements. There were 18 operators. The date and time variable has been converted into Julian date, which transforms the variable into a continuous scale. In this analysis, we used only nonrepeated measurements. We will use

the full dataset in later analysis.

The model proposed is again the smoothing spline function

$$f(x_2) \quad = \quad \mu + \beta x_2 + s_2(x_2), \tag{3.11}$$

where μ is the constant term, β is the coefficient for the linear effect in x_2 and $s_2(x_2)$ represents the smooth effect of x_2.

Based on the smoother matrix (3.8), the effective number of parameters are 50.7, 30.4, 17.4 and 9.6 when λ is set to $a_0 = 1 \times 10^{-8}$, $a_0 = 1 \times 10^{-7}$, $a_0 = 1 \times 10^{-6}$ and $a_0 = 1 \times 10^{-5}$ respectively. So a_0, the median of the Pareto distribution, was chosen to be $a_0 = 1 \times 10^{-6}$, which provided about 17.4 effective degrees of freedom. (Again note that this includes 1 degree of freedom in the linear term plus 17.4 degrees of freedom from the smooth term, leading to 18.4 degrees of freedom for the median of the prior distribution.) The Bayes estimate for λ is 1.77×10^{-5}, which corresponds to 8.2 effective degrees of freedom. To observe the smoothing effect of λ, we select different values $a_0 = 1 \times 10^{-8}, 1 \times 10^{-7}$ and 1×10^{-5} in the prior distribution for λ. The Bayes estimates for λ are 1.70×10^{-5}, 1.63×10^{-5} and 4.76×10^{-5}, which corresponds to equivalent degrees of freedom of 8.3, 8.4 and 6.2. Figures 3.3 shows the MCMC trace plots for samples of $log(\lambda)$ from the models with $a_0 = 1 \times 10^{-5}$, $a_0 = 1 \times 10^{-6}$, $a_0 = 1 \times 10^{-7}$ and $a_0 = 1 \times 10^{-8}$. Convergence is rapid in all cases. Figure 3.4 shows the fits of the models with $a_0 = 1 \times 10^{-5}$, $a_0 = 1 \times 10^{-6}$, $a_0 = 1 \times 10^{-7}$ and $a_0 = 1 \times 10^{-8}$. The fits for $a_0 = 1 \times 10^{-6}$, $a_0 = 1 \times 10^{-7}$ and $a_0 = 1 \times 10^{-8}$ are almost identical. With a_0 ranges from 1×10^{-5} to 1×10^{-8}, the fits are similar. The results suggest that the fit is

not sensitive to the choice of λ when the right range of degrees of freedom is chosen.

This section discussed Bayesian smoothing splines with fully Bayesian estimates. With the posterior distribution, other statistical inference can be done for the fitted model such as credible sets. While these quantities can be computed with GCV, GML and UBR, the uncertainty due to estimating δ_0 and λ are not included in the total inference.

3.2 Fully Bayesian approach in SSANOVA

In Section 2.7 we discussed the frequentist approach to SSANOVA based on Gu (2002). In this section, we will extend the work to a fully Bayesian approach. As mentioned in the previous section, the alternative proposed to improve the MCMC computation will also facilitate the construction of SSANOVA in the Bayesian approach. So we continue to adopt this alternative approach.

The advantage of Bayesian SSANOVA over Bayesian smoothing splines is the added flexibility of testing fixed effects and interaction effects. After decomposing the estimated function into several subspaces, we can test each component in the model and come out with the most parsimonious model.

3.2.1 Bayesian hierarchical SSANOVA model

Based on the procedure discussed in Section 2.7, consider a variable x in domain $\mathcal{X}$ and a Hilbert space $\mathcal{H}$ of functions on $\mathcal{X}$. A tensor sum decomposition

of the reproducing kernel Hilbert space $\mathcal{H}$ has the form

$$\mathcal{H} = \oplus_{k=0}^{p} \mathcal{H}_k \tag{3.12}$$

with inner product

$$\langle f, g \rangle = \sum_{k=0}^{p} \langle f, g \rangle_k = \sum_{k=0}^{p} \langle f_k, g_k \rangle_k \tag{3.13}$$

and a reproducing kernel

$$R(x, y) = \sum_{k=0}^{p} R_k(x, y), \tag{3.14}$$

where $\langle f, g \rangle_k$ is an inner product in $\mathcal{H}_k$ with corresponding reproducing kernel R_k, f_k is the projection of f in $\mathcal{H}_k$, and $\mathcal{H}_0$ is a finite dimensional space with no penalty.

As discussed in Examples 2.7, 2.8 and 2.9, consider two variables $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$ in an experiment with n observations. Suppose $\mathcal{H}_{(x_1)}$ consists of functions on $\mathcal{X}_1$ and $\mathcal{H}_{(x_2)}$ consists of functions on $\mathcal{X}_2$, and assume the reproducing kernel Hilbert space is $\mathcal{H} = \mathcal{H}_{(x_1)} \otimes \mathcal{H}_{(x_2)}$. As in equation (3.12), suppose $\mathcal{H}_{(x_1)}$ and $\mathcal{H}_{(x_2)}$ have tensor sum decompositions

$$\mathcal{H}_{(x_1)} = \oplus_{i=0}^{p_1} \mathcal{H}_{i(x_1)},$$

$$\mathcal{H}_{(x_2)} = \oplus_{j=0}^{p_2} \mathcal{H}_{j(x_2)}.$$

Thus $\mathcal{H} = \mathcal{H}_{(x_1)} \otimes \mathcal{H}_{(x_2)} = \oplus_p \mathcal{H}_p$, where p is the number of all the pairwise combinations of $i = 0, \ldots, p_1$ and $j = 0, \ldots, p_2$. Each orthogonal space $\mathcal{H}_p$

is constructed by taking the product of subspaces $\mathcal{H}_{i(x_1)}$ and $\mathcal{H}_{j(x_2)}$ of $\mathcal{H}_{(x_1)}$ and $\mathcal{H}_{(x_2)}$ respectively. The subspace $\mathcal{H}_{i(x_1)}$ is from the ith tensor sum decomposition of $\mathcal{H}_{(x_1)}$, and $H_{j(x_2)}$ is from the jth tensor sum decomposition of $H_{(x_2)}$. Moreover, the corresponding reproducing kernel R_p is the product of reproducing kernels under $\mathcal{H}_{i(x_1)}$ and $\mathcal{H}_{j(x_2)}$.

After the model is selected, the reproducing kernel Hilbert space can be rewritten as $\mathcal{H} = \mathcal{H}_0 \oplus \sum_{k=1}^{p} \mathcal{H}_k$, and the reproducing kernel R_k in $\mathcal{H}_k$ is the product of $R_{i(x_1)}$ in $\mathcal{H}_{(x_1)}$ and $R_{j(x_2)}$ in $\mathcal{H}_{(x_2)}$. Define

$$\boldsymbol{\Sigma} = \sum_{k=1}^{p} \boldsymbol{\Sigma}_k, \tag{3.15}$$

where $\boldsymbol{\Sigma}_k$ is the $n \times n$ matrix with the (i, j)th entry $R_k((x_{1i}, x_{2i}), (x_{1j}, x_{2j}))$.

To view the effect of each $\boldsymbol{\Sigma}_k$, we assigned different weights, θ_k, to each term and assume

$$\boldsymbol{\Sigma} = \sum_{k=1}^{p} \theta_k \boldsymbol{\Sigma}_k. \tag{3.16}$$

Under a fully Bayesian approach, those weights will be estimated through the smoothing parameters $\lambda_k = \lambda/\theta_k$. Instead of estimating the θ_k directly, we will get the Bayes estimates for each λ_k, which controls the smoothing effect under each $\boldsymbol{\Sigma}_k$.

The minimizer for equation (2.65) now is

$$\widehat{f}(t) = \boldsymbol{T}\boldsymbol{d} + \boldsymbol{\Sigma}\boldsymbol{c} = \boldsymbol{T}\boldsymbol{d} + \left(\sum_{k=1}^{p} \boldsymbol{\Sigma}_k \right) \boldsymbol{c}. \tag{3.17}$$

As the alternative approach considered in equation (2.24), the prior distribu-

tion for $\boldsymbol{\eta}$ is

$$\boldsymbol{\eta} \mid \delta_0, \lambda \sim N_n \left(\mathbf{0}, \frac{\delta_0}{\lambda} \boldsymbol{\Sigma} \right). \tag{3.18}$$

To apply the SSANOVA decomposition, refer to the setup for $\boldsymbol{u}_2$ in equation (3.24), which has a diffuse prior in $\mathcal{H}_0$. The covariance matrix for $\boldsymbol{u}_2$ is $\boldsymbol{F}_2^{'} \boldsymbol{\Sigma} \boldsymbol{F}_2$. We decompose this covariance matrix according to the reproducing kernel under each subspace,

$$\boldsymbol{F}_2^{'} \boldsymbol{\Sigma} \boldsymbol{F}_2 = \sum_{k=1}^{p} \theta_k \boldsymbol{F}_2^{'} \boldsymbol{\Sigma}_k \boldsymbol{F}_2. \tag{3.19}$$

So each covariance matrix $\theta_k \boldsymbol{F}_2^{'} \boldsymbol{\Sigma}_k \boldsymbol{F}_2$ corresponds to a prior distribution $\boldsymbol{w}_k$ for each subspace k from the penalty term. Let

$$\boldsymbol{u}_2 \;\; = \;\; \sum_{k=1}^{p} \boldsymbol{w}_k.$$

If the priors on the $\boldsymbol{w}_k$ are independent, the prior on $\boldsymbol{w}_k$ satisfies

$$\boldsymbol{w}_k \;\; \sim \;\; N \left(0, \lambda_k \boldsymbol{F}_2^{'} \boldsymbol{\Sigma}_k \boldsymbol{F}_2 \right), \;\; k = 1, \ldots, p,$$

where $\lambda_k = \lambda / \theta_k, \; k = 1, \ldots, p$. Then $\boldsymbol{u}_2$ has the required prior distribution (3.24). To implement a fully Bayesian hierarchical model, we need to assign prior distribution to the smoothing parameter $\lambda_k = \lambda / \theta_k$.

We now take the prior distributions for δ_0 and λ_k to be

$$[\delta_0] \;\; \propto \;\; \frac{1}{\delta_0}, \;\; \delta_0 > 0, \tag{3.20}$$

$$[\lambda_k \mid a_k] \;=\; \frac{a_k}{(a_k+\lambda_k)^2}, \quad \lambda_k \geq 0 \quad , k = 1 \quad , \ldots, p, \tag{3.21}$$

where the Pareto distribution will be sampled based on the hierarchical structure in White (2006). Since the covariance matrix $\boldsymbol{F}_2^{'}\boldsymbol{\Sigma}_k\boldsymbol{F}_2$ might be singular, again take the spectral decomposition

$$\boldsymbol{F}_2^{'}\boldsymbol{\Sigma}_k\boldsymbol{F}_2 = \boldsymbol{Q}_k\boldsymbol{D}_k\boldsymbol{Q}_k' \quad k = 1, \ldots, p, \tag{3.22}$$

where $\boldsymbol{Q}_k$ is the matrix of eigenvectors corresponding to the nonzero eigenvalues at $\boldsymbol{F}_2^{'}\boldsymbol{\Sigma}_k\boldsymbol{F}_2$ and $\boldsymbol{D}_k$ is the diagonal matrix with the nonzero eigenvalues on the diagonal.

Set $\boldsymbol{w}_k = \boldsymbol{Q}_k\boldsymbol{v}_k$. Then the prior on $\boldsymbol{v}_k$ and $\boldsymbol{u}_2$ is given by

$$\boldsymbol{v}_k \mid \delta_0, \lambda_k \;\sim\; N\left(\mathbf{0}, \frac{\delta_0}{\lambda_k}\mathbf{D}_k\right) \quad k = 1, \ldots, p, \tag{3.23}$$

$$\boldsymbol{u}_2 \;=\; \sum_{k=1}^{p} \boldsymbol{Q}_k\boldsymbol{v}_k. \tag{3.24}$$

The full conditionals for each of the parameters are easily calculated:

$$\boldsymbol{d} \mid \delta_0 \;\sim\; N((\boldsymbol{T}'\boldsymbol{F}_1\boldsymbol{F}_1'\boldsymbol{T})^{-1}\boldsymbol{T}'\boldsymbol{F}_1\boldsymbol{F}_1'\boldsymbol{y}, \delta_0(\boldsymbol{T}'\boldsymbol{F}_1\boldsymbol{F}_1'\boldsymbol{T})^{-1}), \tag{3.25}$$

$$\begin{aligned}\boldsymbol{v}_k \mid \boldsymbol{v}_{-k}, \lambda_k, \delta_0 \;\sim\; & N\Big((\boldsymbol{I}_{r_k} + \lambda_k\boldsymbol{D}_k^{-1})^{-1}\boldsymbol{Q}_k'\Big(\boldsymbol{F}_2^{'}\boldsymbol{y} - \sum_{i=1, i\neq k}^{p} \boldsymbol{Q}_i\boldsymbol{v}_i\Big), \\ & \delta_0(\boldsymbol{I}_{r_k} + \lambda_k\boldsymbol{D}_k^{-1})\Big), \quad k = 1, \ldots, p, \end{aligned} \tag{3.26}$$

$$\begin{aligned}\lambda_k \mid \boldsymbol{v}_k, \delta_0, \phi_k \;\sim\; & \text{Gamma}\left(\frac{r_k}{2} + 1, \frac{2\delta_0}{\boldsymbol{v}_k'\boldsymbol{D}_k^{-1}\boldsymbol{v}_k + 2\delta_0\phi_k}\right), \\ & k = 1, \ldots, p, \end{aligned} \tag{3.27}$$

$$\phi_k \mid \lambda_k, a_k \sim \text{Gamma}\left(2, \frac{1}{\lambda_k + a_k}\right), \quad k = 1, \ldots, p, \tag{3.28}$$

$$\delta_0 \mid \lambda_1, \ldots, \lambda_p, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_p \sim \text{Inverse Gamma}\left(\frac{n + \sum_{k=1}^{p} r_k}{2}, \frac{\|\boldsymbol{F}_2'\boldsymbol{y} - \boldsymbol{u}_2\|^2}{2} + \sum_{k=1}^{p} \frac{\lambda_k \boldsymbol{v}_k' \boldsymbol{D}_k^{-1} \boldsymbol{v}_k}{2}\right), \tag{3.29}$$

where $r_1, \ldots, r_p$ are the ranks of the nonzero eigenvalue matrices $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_p$, $a_1, \ldots, a_p$ are the parameters for the prior distributions of $\lambda_1, \ldots, \lambda_p$, and $\boldsymbol{v}_{-k} = \{\boldsymbol{v}_j, j \neq k\}$, the collection of all $\boldsymbol{v}_j$ for $j \neq k$.

The posterior distributions for each of the parameters can be simulated based on the full conditionals. Since the decomposition of the reproducing kernel Hilbert space provides orthogonal subspaces that span main effects and interaction effects, we will test the components under each subspace for different effects. The model selection criteria for hypothesis testing is based on Bayes factors.

3.2.2 Priors for one-way ANOVA compared to smoothing spline ANOVA

Consider $X = \{1, \ldots, K\}$ corresponding to a one-way ANOVA model. The balanced one-way ANOVA hierarchical model is

$$y_{ij} = \mu + \alpha_i + \varepsilon_{ij}, \quad i = 1, \ldots, K, \quad j = 1, \ldots, r, \tag{3.30}$$

$$\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2), \tag{3.31}$$

where r is the number of replicates. The customary prior on α_i is

$$\alpha_i \overset{iid}{\sim} N(0, \sigma_\alpha^2), \quad i = 1, \ldots, K. \tag{3.32}$$

In smoothing spline ANOVA for the identifiability of $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)'$, the prior for $\boldsymbol{\alpha}$ is restricted to the complement of the null space. The null space is

$$\mathcal{H}_0 = \{\boldsymbol{\alpha} = c(1, \ldots, 1)', -\infty < c < \infty\}, \tag{3.33}$$

with the projection matrix

$$\boldsymbol{P}_0 = \mathbf{1}(\mathbf{1}'\mathbf{1})^{-1}\mathbf{1}' = \frac{1}{K}\mathbf{1}\mathbf{1}'. \tag{3.34}$$

Let $\mathcal{H}_1$ be the complement of the null space with projection matrix

$$\boldsymbol{P}_1 = \boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}'. \tag{3.35}$$

If we take the projection of $\boldsymbol{\alpha}$ onto $\mathcal{H}_1$, the resulting prior is

$$\boldsymbol{P}_1\boldsymbol{\alpha} \sim N(0, \boldsymbol{P}_1\sigma_\alpha^2\boldsymbol{I}\boldsymbol{P}_1') = N(0, \sigma_\alpha^2\boldsymbol{P}_1). \tag{3.36}$$

Now represent the balanced one-way ANOVA hierarchical model (3.30) in matrices,

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \tag{3.37}$$

where $\boldsymbol{y}$ is the $n \times 1$ vector of observed responses, μ is the overall mean, $\boldsymbol{X} = \boldsymbol{I}_k \otimes \mathbf{1}_r$ is an $n \times K$ design matrix, $\boldsymbol{\alpha}$ is a $K \times 1$ vector for factor level effects and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector for random errors. Since the prior for $\boldsymbol{\alpha}$ is

restricted to the complement of null space, the prior is

$$\boldsymbol{\alpha} \sim N\left(0, \sigma_\alpha^2\left(\boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}'\right)\right) = N(0, \sigma_\alpha^2 \boldsymbol{P}_1). \tag{3.38}$$

In keeping with the SSANOVA development, we replace σ_α^2 by δ_0/λ. Then the prior on $\boldsymbol{X}\boldsymbol{\alpha}$ is

$$\boldsymbol{X}\boldsymbol{\alpha} \sim N\left(0, \frac{\delta_0}{\lambda}\boldsymbol{X}\left(\boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}'\right)\boldsymbol{X}'\right) = N\left(0, \frac{\delta_0}{\lambda} r\boldsymbol{\Sigma}\right), \tag{3.39}$$

is the same as the prior distribution for $\boldsymbol{\eta}$ in equation (3.18) with covariance matrix $r\boldsymbol{\Sigma}$ for the discrete variable in SSANOVA.

Take $\widetilde{\boldsymbol{P}}$ to be a $K \times (K-1)$ matrix whose columns are orthonormal and $\widetilde{\boldsymbol{P}} \perp \mathbf{1}_K$. Then $\boldsymbol{P}_1 = \widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{P}}'$, $\widetilde{\boldsymbol{P}}'\widetilde{\boldsymbol{P}} = \boldsymbol{I}_{(k-1)}$, and $\boldsymbol{\alpha} \sim N\left(0, \frac{\delta_0}{\lambda}\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{P}}'\right)$.

Let $\boldsymbol{\beta} = \widetilde{\boldsymbol{P}}'\boldsymbol{\alpha}$, so

$$\begin{aligned} \boldsymbol{\beta} = \widetilde{\boldsymbol{P}}'\boldsymbol{\alpha} &\sim N\left(0, \frac{\delta_0}{\lambda}\widetilde{\boldsymbol{P}}'\widetilde{\boldsymbol{P}}\widetilde{\boldsymbol{P}}'\widetilde{\boldsymbol{P}}\right) \\ &\sim N\left(0, \frac{\delta_0}{\lambda}\boldsymbol{I}_{(K-1)}\right), \end{aligned}$$

where $\boldsymbol{\beta}$ is a $(K-1) \times 1$ vector. Then $\boldsymbol{\beta}$ has a proper normal distribution with a non-singular covariance matrix. Thus we can write

$$\boldsymbol{y} = \mathbf{1}\mu + \widetilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{3.40}$$

where $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\widetilde{\boldsymbol{P}}$. With this transformation between $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, the priors for both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are equivalent but in different dimensions. This is the customary prior for the coefficients in a full rank linear model.

Since $\widetilde{\boldsymbol{X}}$ has full rank, consider the Zellner's g-prior (Zellner, 1986) for $\boldsymbol{\beta}$:

$$\begin{aligned} \boldsymbol{\beta} &\sim N(0, g\delta_0(\widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}})^{-1}) \\ &\sim N(0, g\delta_0(\widetilde{\boldsymbol{P}}'\boldsymbol{X}'\boldsymbol{X}\widetilde{\boldsymbol{P}})^{-1}) \\ &\sim N\left(0, \frac{g\delta_0}{r}\boldsymbol{I}_{(K-1)}\right). \end{aligned} \tag{3.41}$$

Without loss of generality, assume that $\boldsymbol{X} = \boldsymbol{I}_K \otimes \mathbf{1}_r$ so $\boldsymbol{X}'\boldsymbol{X} = r\boldsymbol{I}_K$.

The Zellner-Siow prior (Zellner and Siow, 1980) on $\boldsymbol{\beta}$ is (3.41) with

$$g \sim \text{Inverse Gamma}\left(\frac{1}{2}, \frac{n}{2}\right). \tag{3.42}$$

When the prior on smoothing parameter λ is

$$\lambda \sim \text{Gamma}\left(\frac{1}{2}, \frac{K}{2}\right), \tag{3.43}$$

i.e.,

$$\begin{aligned} K\lambda &\sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right), \\ &\sim \boldsymbol{\chi}^2_1, \end{aligned} \tag{3.44}$$

then the prior on the discrete term $\boldsymbol{\eta}$ in the SSANOVA is exactly the Zellner-Siow prior. In a balanced design, the Zellner's g-prior is depends on the number of factor levels K.

We take a variant prior with $K = 1$ for λ so

$$\lambda \sim \text{Gamma}\left(\frac{1}{2}, \frac{1}{2}\right) \tag{3.45}$$

for level effects in a discrete variable in SSANOVA. For smooth effect in a continuous variable, the prior for λ is

$$\lambda \sim \text{Gamma}\left(\frac{1}{2}, \frac{b}{2}\right), \tag{3.46}$$

where the scale parameter b is determined by the effective degrees of freedom.

3.2.3 Prediction at new points

The posterior means $\overline{\boldsymbol{d}}$ and $\overline{\boldsymbol{v}}_k$ are computed as the mean of the MCMC samples from equation (3.25) and (3.26) separately.

The estimate at the observed points is

$$\widehat{y} = \boldsymbol{F}\overline{\boldsymbol{u}} = \boldsymbol{F}\boldsymbol{F}'\overline{\boldsymbol{v}} = \boldsymbol{F}\boldsymbol{F}'\left(\boldsymbol{T}\overline{\boldsymbol{d}} + \sum_{k=1}^{p}\overline{\boldsymbol{\eta}}_k\right),$$

where $\overline{\boldsymbol{u}} = (\overline{\boldsymbol{u}}_1', \overline{\boldsymbol{u}}_2')'$,

$$\begin{aligned}\overline{\boldsymbol{u}}_1 &= \boldsymbol{F}_1'\boldsymbol{T}\overline{\boldsymbol{d}} \\ \overline{\boldsymbol{u}}_2 &= \sum_{k=1}^{p}\overline{\boldsymbol{w}}_k = \sum_{k=1}^{p}\boldsymbol{Q}_k\overline{\boldsymbol{v}}_k.\end{aligned}$$

Consider predicting at new points $s_1, \ldots, s_t$. The conditional mean for $\widetilde{\boldsymbol{y}}$ at the new points is

$$\widetilde{y} = \widetilde{\boldsymbol{T}}\overline{\boldsymbol{d}} + \sum_{k=1}^{p}\widetilde{\boldsymbol{\eta}}_k, \tag{3.47}$$

where $\widetilde{\boldsymbol{T}}$ spans the null space for the new points and $\widetilde{\boldsymbol{\eta}}_k = (\widetilde{\boldsymbol{\eta}}_k(s_1), \ldots, \widetilde{\boldsymbol{\eta}}_k(s_t))'$ is the conditional mean of $\boldsymbol{\eta}_k$ given the data at the new points. The covariance

matrix between the observed and new points under each reproducing kernel Hilbert space $\mathcal{H}_k$ is

$$\boldsymbol{C}_k = \text{Cov}(\widetilde{\boldsymbol{\eta}}_k, \boldsymbol{\eta}_k) = [R_k(s_i, x_j)]_{t\times n}. \tag{3.48}$$

Define the reproducing kernel for the new points as

$$\widetilde{\boldsymbol{\Sigma}}_k = [R_k(s_i, s_j)]_{t\times t}. \tag{3.49}$$

Then the prior distribution for $\widetilde{\boldsymbol{\eta}}_k$ is

$$\widetilde{\boldsymbol{\eta}}_k \mid \delta_0, \lambda \sim N\left(0, \frac{\delta_0}{\lambda}\widetilde{\boldsymbol{\Sigma}}_k\right). \tag{3.50}$$

The joint distribution of $\widetilde{\boldsymbol{\eta}}_k$ and $\boldsymbol{w}_k$ is

$$\begin{pmatrix} \widetilde{\boldsymbol{\eta}}_k \\ \boldsymbol{w}_k \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \frac{\delta_0}{\lambda}\widetilde{\boldsymbol{\Sigma}}_k & \frac{\delta_0}{\lambda}\boldsymbol{C}_k\boldsymbol{F}_2 \\ \frac{\delta_0}{\lambda}\boldsymbol{F}_2'\boldsymbol{C}_k' & \frac{\delta_0}{\lambda}\boldsymbol{F}_2'\boldsymbol{\Sigma}_k\boldsymbol{F}_2 \end{pmatrix}\right)$$

Using a standard result on multivariate normal distributions, the conditional distribution of $\widetilde{\boldsymbol{\eta}}_k \mid \boldsymbol{w}_k$ is

$$\begin{aligned} \widetilde{\boldsymbol{\eta}}_k \mid \boldsymbol{w}_k, \delta_0, \lambda \;\sim\; & N(\boldsymbol{C}_k\boldsymbol{F}_2(\boldsymbol{F}_2'\boldsymbol{\Sigma}_k\boldsymbol{F}_2)^{-1}\boldsymbol{Q}_k\boldsymbol{v}_k, \\ & \frac{\delta_0}{\lambda}(\widetilde{\boldsymbol{\Sigma}}_k - \boldsymbol{C}_k\boldsymbol{F}_2(\boldsymbol{F}_2'\boldsymbol{\Sigma}_k\boldsymbol{F}_2)^{-1}\boldsymbol{F}_2'\boldsymbol{C}_k')). \end{aligned} \tag{3.51}$$

To predict at the new points, samples from the conditional distribution $\widetilde{\boldsymbol{\eta}}_k \mid \boldsymbol{w}_k, \delta_0, \lambda$ must be projected onto the complement of the null space for the new points.

The projection matrix in the null space for the new points is

$$\boldsymbol{P}_{N0} = \widetilde{\boldsymbol{T}}(\widetilde{\boldsymbol{T}}'\widetilde{\boldsymbol{T}})^{-1}\widetilde{\boldsymbol{T}}'. \tag{3.52}$$

Then the projection matrix onto the complement of null space is

$$\boldsymbol{P}_{N1} = \boldsymbol{I}_t - \widetilde{\boldsymbol{T}}(\widetilde{\boldsymbol{T}}'\widetilde{\boldsymbol{T}})^{-1}\widetilde{\boldsymbol{T}}'. \tag{3.53}$$

Finally, the estimate of the new points is

$$\begin{aligned} \boldsymbol{P}_{N1}\widetilde{\boldsymbol{\eta}}_k \mid \boldsymbol{w}_k \quad \sim \quad & N(\boldsymbol{P}_{N1}\boldsymbol{C}_k\boldsymbol{F}_2(\boldsymbol{F}_2'\boldsymbol{\Sigma}_k\boldsymbol{F}_2)^{-1}\boldsymbol{Q}_k\boldsymbol{v}_k, \\ & \boldsymbol{P}_{N1}\frac{\delta_0}{\lambda}(\widetilde{\boldsymbol{\Sigma}}_k - \boldsymbol{C}_k\boldsymbol{F}_2(\boldsymbol{F}_2'\boldsymbol{\Sigma}_k\boldsymbol{F}_2)^{-1}\boldsymbol{F}_2'\boldsymbol{C}_k')\boldsymbol{P}_{N1}'). \end{aligned} \tag{3.54}$$

3.2.4 Bayes Factor in Bayesian approach

A well-known and widely adopted model selection criteria in Bayesian approach is the Bayes factor. Kass and Raftery (1995), Albert and Chib (1997) and Bayarri and Garcia-Donato (2007) have discussed the definition, computation issues and the choice of priors for Bayes factors. The Bayes factor is the ratio between the posterior odds and prior odds of two models. Let p_k denote the prior distribution for model k, where $k = 1, 2, p_k \geq 0$, and $p_1 + p_2 = 1$. For the two competing models, suppose model 2 has parameters $\boldsymbol{\omega}_{M2}$ and model 1 nested within model 2 has parameters $\boldsymbol{\omega}_{M1}$. Suppose the prior distribution for model 1 is $p_1(\boldsymbol{\omega}_{M1})$ and that for model 2 is $p_2(\boldsymbol{\omega}_{M2}) = p_1(\boldsymbol{\omega}_{M1})p_a(\boldsymbol{\gamma})$, where $\boldsymbol{\gamma}$ contains the parameters in model 2 but not in model 1. Note that the parameters that appear in both model 2 and model 1 have the same prior distribution

under both models. Also, $p_a(\boldsymbol{\gamma})$ must be a proper prior to have a well-defined Bayes factor. The marginal likelihood function of y under model M_k is

$$p(y \mid M_k) = \int_{\boldsymbol{\omega}_{M_k}} f_k(y \mid \boldsymbol{\omega}_{M_k}) p_k(\boldsymbol{\omega}_{M_k}) d\boldsymbol{\omega}_{M_k}, \quad k = 1, 2,$$

and the posterior probability of $M_k \mid y$ is

$$p(M_k \mid y) = \frac{p(y \mid M_k) p_k}{\sum_{k=1}^{2} p(y \mid M_k) p_k}.$$

The Bayes factor for comparing model 2 with model 1 is defined to be

$$BF_{21} = \frac{p(M_2 \mid y)/p(M_1 \mid y)}{p_1/p_2} = \frac{p(y \mid M_2)}{p(y \mid M_1)}.$$

Since sometimes it is impossible to integrate out all the parameters, we use bridge sampling (Meng and Wong, 1996) to estimate the Bayes factor if we can integrate out $\boldsymbol{\gamma}$. Assume we have output from MCMC simulations under both models. Let $\boldsymbol{\omega}_{M_k}$ be the sequence of common parameters generated under model k, $k = 1, 2$, and let $l_{1i} = \frac{q_2(\widetilde{\boldsymbol{\omega}}_{M1})}{q_1(\widetilde{\boldsymbol{\omega}}_{M1})}$, $l_{2i} = \frac{q_2(\widetilde{\boldsymbol{\omega}}_{M2})}{q_1(\widetilde{\boldsymbol{\omega}}_{M2})}$, where q_k is the product of the likelihood and the marginal prior density of $\boldsymbol{\omega}_{Mk}$ under model k. The Meng and Wong (1996) algorithm is iterative. At the $(j+1)$th iteration, compute

$$\widehat{BF_{21}}^{j+1} = \frac{\frac{1}{n1} \sum_{i=1}^{n1} \frac{l_{1i}}{d_1 l_{1i} + d_2 \widehat{BF_{21}}^{j}}}{\frac{1}{n2} \sum_{i=1}^{n2} \frac{1}{d_1 l_{2i} + d_2 \widehat{BF_{21}}^{j}}}, \tag{3.55}$$

where $d_1 = 1 - d_2 = \frac{n_1}{n_1 + n_2}$, and n_k is the number of random samples generated from the full conditionals through Gibbs sampling after a suitable burn-in

period for model k, $k = 1, 2$. Convergence of $\widehat{BF_{21}}^{j}$ is generally rapid. Larger values of BF_{21} provide increasing evidence to support model 2 and smaller values of BF_{21} support model 1.

Kass and Raftery (1995) provided an interpretation of BF_{21}. They suggest that when $1 < BF_{21} < 3.2$, the evidence against model 1 is negligible. When $3.2 < BF_{21} < 10$, the evidence against model 1 is substantial. When $10 < BF_{21} < 100$, the evidence against model 1 is strong. When $BF_{21} > 100$, the evidence against model 1 is decisive.

Example 3.1

Consider two variables, $x_1 \in \mathcal{X}_1 = \{1, \ldots, K_1\}$ and $x_2 \in \mathcal{X}_2 = [0, 1]$. Suppose we want to model the x_1 effect using an one-way ANOVA effects model and the x_2 effect using a cubic spline. The tensor product decomposition for the Hilbert space $\mathcal{H}$ is listed in equation (2.58) and the reproducing kernel under each subspace is in equation (2.57). Recall that $\mathcal{H}_{0\langle x_1\rangle} \otimes \mathcal{H}_{00\langle x_2\rangle}$ is the space of constants with reproducing kernels $R_{0\langle x_1\rangle} R_{00\langle x_2\rangle}$; $\mathcal{H}_{0\langle x_1\rangle} \otimes \mathcal{H}_{01\langle x_2\rangle}$ is the one dimensional space of a linear effect in x_2 with reproducing kernel $R_{0\langle x_1\rangle} R_{01\langle x_2\rangle}$; $\mathcal{H}_{0\langle x_1\rangle} \otimes \mathcal{H}_{11\langle x_2\rangle}$ is the space of smooth effect in x_2 with reproducing kernel $R_{0\langle x_1\rangle} R_{11\langle x_2\rangle}$; $\mathcal{H}_{1\langle x_1\rangle} \otimes \mathcal{H}_{00\langle x_2\rangle}$ is the space of level effects in x_1 with reproducing kernel $R_{1\langle x_1\rangle} R_{00\langle x_2\rangle}$. $\mathcal{H}_{1\langle x_1\rangle} \otimes \mathcal{H}_{01\langle x_2\rangle}$ is the space of interaction effects between level effects in x_1 and linear effect in x_2 with reproducing kernel $R_{1\langle x_1\rangle} R_{01\langle x_2\rangle}$. Finally, $\mathcal{H}_{1\langle x_1\rangle} \otimes \mathcal{H}_{11\langle x_2\rangle}$ is the space of interaction effects between level effects in x_1 and smooth effect in x_2 with reproducing kernel $R_{1\langle x_1\rangle} R_{11\langle x_2\rangle}$. Refer to Equations (2.59), (2.60), (2.61), (2.62), (2.63) and (2.64) for the corresponding reproducing kernel under each subspace.

A fully Bayesian approach was implemented with the prior distribution

assigned for δ_0 and $\lambda_1, \ldots, \lambda_4$ from equations (3.20) and (3.21). The prior distribution on each term corresponding to the penalty on its subspace is given by equation (3.23) with $k = 1, 2, 3, 4$. The corresponding full conditionals for $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_4$, $\lambda_1, \ldots, \lambda_4$, $\phi_1, \ldots, \phi_4$ and δ_0 are from equations (3.26), (3.27), (3.28) and (3.29).

The tensor product decomposition decomposes the Hilbert space $\mathcal{H}$ into six subspaces as discussed in the beginning of this example. The spaces $\mathcal{H}_{0\langle x_1\rangle} \otimes \mathcal{H}_{00\langle x_2\rangle}$ and $\mathcal{H}_{0\langle x_1\rangle} \otimes \mathcal{H}_{01\langle x_2\rangle}$ are finite dimensional spaces containing functions that are not going to be penalized. We will only smooth the functions from the other four subspaces.

In this model, we are interested in testing the effects of the estimated functions from each subspace. We set up a series of partially nested models of interest and use the Bayes factor as the model selection criteria to test for the effects. The partially nested models are:

$$\begin{aligned}
\text{Model } 0 : f(x_1, x_2) &= \mu + \beta x_2 + s_1(x_1) + s_2(x_2) \\
&\quad + l_{12}(x_1, x_2) + s_{22}(x_1, x_2), && (3.56)\\
\text{Model } 1 : f(x_1, x_2) &= \mu + \beta x_2 + s_1(x_1) + s_2(x_2) + l_{12}(x_1, x_2), && (3.57)\\
\text{Model } 2 : f(x_1, x_2) &= \mu + \beta x_2 + s_1(x_1) + s_2(x_2), && (3.58)\\
\text{Model } 3 : f(x_1, x_2) &= \mu + \beta x_2 + s_1(x_1), && (3.59)\\
\text{Model } 4 : f(x_1, x_2) &= \mu + \beta x_2 + s_2(x_2), && (3.60)
\end{aligned}$$

where μ is the constant term, β is the coefficient for the linear effect in x_2, s_1 represents the level effects in x_1, s_2 represents the smooth effect in x_2, l_{12} represents the interaction effects between the level effects in x_1 and linear

effects in x_2, and s_{22} represents the interaction effects between the level effects in x_1 and the smooth effect in x_2.

To test if there is evidence to conclude the level effects from x_1 interact with the smooth effect from x_2, we will have model 0 compete with model 1. The parameters for model 0 are $\boldsymbol{\omega}_{M0} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_4, \lambda_1, \ldots, \lambda_4, \delta_0\}$ and for model 1 are $\boldsymbol{\omega}_{M1} = \{\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3, \lambda_1, \lambda_2, \lambda_3, \delta_0\}$. Since it is intractable to compute the marginal distribution of $\boldsymbol{y}$ for both model 0 and model 1, we will use bridge sampling to compute the Bayes factor after integrating out the parameters $\boldsymbol{v}_4$ and λ_4.

The product of likelihood and priors under model 0 is

$$\begin{aligned} lp(M_0) &= (2\pi\delta_0)^{-\frac{(n-2)}{2}} e^{-\frac{1}{2\delta_0}\|\boldsymbol{F}_2'\boldsymbol{y} - \sum_{i=1}^4 \boldsymbol{Q}_i\boldsymbol{v}_i\|^2} \times \frac{1}{\delta_0} \times \\ &\quad \prod_{i=1}^4 \left((2\pi)^{-\frac{r_i}{2}} \left| \frac{\delta_0}{\lambda_i}\boldsymbol{D}_i \right|^{-\frac{1}{2}} e^{-\frac{\lambda_i}{2\delta_0}\boldsymbol{v}_i'\boldsymbol{D}_i\boldsymbol{v}_i} \right) \times \prod_{i=1}^4 \left(\frac{a_i}{(a_i+\lambda_i)^2} \right) \end{aligned} \quad (3.61)$$

The product of likelihood and priors under model 1 is

$$\begin{aligned} lp(M_1) &= (2\pi\delta_0)^{-\frac{(n-2)}{2}} e^{-\frac{1}{2\delta_0}\|\boldsymbol{F}_2'\boldsymbol{y} - \sum_{i=1}^3 \boldsymbol{Q}_i\boldsymbol{v}_i\|^2} \times \frac{1}{\delta_0} \times \\ &\quad \prod_{i=1}^3 \left((2\pi)^{-\frac{r_i}{2}} \left| \frac{\delta_0}{\lambda_i}\boldsymbol{D}_i \right|^{-\frac{1}{2}} e^{-\frac{\lambda_i}{2\delta_0}\boldsymbol{v}_i'\boldsymbol{D}_i^{-1}\boldsymbol{v}_i} \right) \times \prod_{i=1}^3 \left(\frac{a_i}{(a_i+\lambda_i)^2} \right) \end{aligned} \quad (3.62)$$

After integrating out $\boldsymbol{v}_4$ analytically, the term λ_4 will be integrated out numerically under model 0. Then

$$\begin{aligned} q_0 &= \int_{\mathbb{R}}\int_{\mathbb{R}} lp(M_0)\, d\boldsymbol{v}_4 d\lambda_4 \\ &= \int_{\mathbb{R}} (2\pi\delta_0)^{-\frac{(n-2)}{2}} \left|2\pi\boldsymbol{A}^{-1}\delta_0\right|^{\frac{1}{2}} e^{-\frac{1}{2\delta_0}\|\boldsymbol{F}_2'\boldsymbol{y} - \sum_{i=1}^3 \boldsymbol{Q}_i\boldsymbol{v}_i\|^2} \times \\ &\quad e^{\frac{1}{2\delta_0}(\boldsymbol{F}_2'\boldsymbol{y} - \sum_{i=1}^3 \boldsymbol{Q}_i\boldsymbol{v}_i)'\boldsymbol{Q}_4\boldsymbol{A}^{-1}\boldsymbol{Q}_4'(\boldsymbol{F}_2'\boldsymbol{y} - \sum_{i=1}^3 \boldsymbol{Q}_i\boldsymbol{v}_i)} \times \end{aligned}$$

$$\prod_{i=1}^{3}\left((2\pi)^{-\frac{r_i}{2}}\left|\frac{\delta_0}{\lambda_i}\boldsymbol{D}_i\right|^{-\frac{1}{2}}exp^{-\frac{\lambda_i}{2\delta_0}\boldsymbol{v}_i'\boldsymbol{D}_i^{-1}\boldsymbol{v}_i}\right)\times$$
$$\prod_{i=1}^{3}\left(\frac{a_i}{(a_i+\lambda_i)^2}\right)\times\frac{1}{\delta_0}\left|\frac{\delta_0}{\lambda_4}\boldsymbol{D}_4\right|^{-\frac{1}{2}}d\lambda_4, \tag{3.63}$$

where $\boldsymbol{A}=\boldsymbol{I}_{r_4}+\lambda_4\boldsymbol{D}_4^{-1}$. Finally,

$$\begin{aligned}\frac{q_0}{q_1} &= \int_{\mathbb{R}}|2\pi\boldsymbol{A}^{-1}\delta_0|^{\frac{1}{2}}e^{\frac{1}{2\delta_0}(\boldsymbol{F}_2'\boldsymbol{y}-\sum_{i=1}^{3}\boldsymbol{Q}_i\boldsymbol{v}_i)'\boldsymbol{Q}_4\boldsymbol{A}^{-1}\boldsymbol{Q}_4'(\boldsymbol{F}_2'\boldsymbol{y}-\sum_{i=1}^{3}\boldsymbol{Q}_i\boldsymbol{v}_i)}\times,\\ &\quad\frac{a_4}{(a_4+\lambda_4)^2}\left|\frac{\delta_0}{\lambda_4}\boldsymbol{D}_4\right|^{-\frac{1}{2}}(2\pi)^{-\frac{r_4}{2}}d\lambda_4\end{aligned} \tag{3.64}$$

We use the random samples generated from the full conditionals and apply equation (3.55) to compute the Bayes factor BF_{01} and test the significance of $\boldsymbol{v}_4$, the interaction effect between the level effects in x_1 and the smooth effect in x_2. A similar approach can be done to calculate BF_{12} to test the interaction effect between the level effects in x_1 and the linear effect in x_2, BF_{23} to test the smooth effect in x_2, and BF_{24} to test the level effects in x_1. By definition, $BF_{02}=BF_{01}\times BF_{12}$ provided the statistic to test for interaction effects. To illustrate the application of this model selection method, the simulated and manufacturing examples in Sections 3.1.3 and 3.1.4 will be revisited and the results will be compared.

3.2.5 Simulated Example in Bayesian SSANOVA

Consider again the example in Section 3.1.3. Instead of fitting the data with a smoothing spline, we will fit the data with the five models (3.56), (3.57), (3.58), (3.59) and (3.60) proposed in Example 3.1.

The effect of each component is tested by model selection. The Bayes

factors for each pair of models is listed in Table 3.1. The hyper-parameters $a_1,\ a_2,\ a_3$ and a_4 for the Pareto priors on λ was set to be $a_1 = a_2 = a_3 = a_4 = 0.01$, which corresponding to effective degrees of freedom 1, 1, 1 and 2 for each term in model (3.56). The results showed $BF_{01} = 0.438$ for testing the effect of s_{12} was inconclusive, $BF_{12} = 0.192$ favored model 2 and concluded l_{12} is insignificant, $BF_{23} = 883$ strongly favored model 2 and concluded s_2 is significant and $BF_{24} = 75.7$ favored model 2 and concluded s_1 is significant. As shown in Figure 3.5, model 0, model 1 and model 2 provided almost the same fit as confirmed in the model selection, while model 2 is the most parsimonious model.

Table 3.1: Bayes factors in simulated example

BF_{01}	BF_{12}	BF_{23}	BF_{24}
0.438	0.192	883	75.7

3.2.6 Manufacturing Example in Bayesian SSANOVA

Now we revisit the manufacturing example in Section 3.1.4. Since there were 18 operators, we do not consider the interaction effects between x_1 and x_2. We fit the data with three partially nested models so we can inspect the linear and smooth effects individually.

$$\begin{aligned}
\text{Model 2} : f(x_1, x_2) &= \mu + \beta x_2 + s_1(x_1) + s_2(x_2), \\
\text{Model 3} : f(x_1, x_2) &= \mu + \beta x_2 + s_1(x_1), \\
\text{Model 4} : f(x_1, x_2) &= \mu + \beta x_2 + s_2(x_2),
\end{aligned}$$

where μ is the constant term, β is the coefficient for the linear effect in x_2, s_1 represents the level effects in x_1, and s_2 represents the smooth effect in x_2.

The effect of each component is tested by model selection. The Bayes factors for each pair of models is listed in Table 3.2. The hyper-parameters a_1 and a_2 for the Pareto priors of the λ are $a_1 = a_2 = 0.01$, which corresponding to effective degrees of freedom 16.9 and 0.9. Note that $BF_{23} = 4.04$ favored model 2 and $BF_{24} = 4.38 \times 10^{-3}$ strongly favored model 4. We conclude that the level effects in x_1 are insignificant since we favored model 4 over 2. The smooth effect in x_2 is significant since we favored model 2 over model 3. As shown in Figure 3.6, model 2 and model 4 provide almost the same fit as confirmed in the model selection, while model 4 is the most parsimonious model.

Table 3.2: Bayes factors in manufacturing example

BF_{23}	BF_{24}
4.04	4.38×10^{-3}

3.2.7 Alternative Bayes Factor Computation

As discussed in Section 3.2.4, bridge sampling provides one way to compute Bayes factors for fully Bayesian smoothing spline ANOVA models. So far, bridge sampling has been used to test the smooth effect, the level effects and the interaction effects among independent variables. In addition to testing for these effects, it is sometimes interesting to test for variables in the null space. In particular, when testing for the effect of a continuous variable, one wants to simultaneously test for the linear and smooth effects.

Consider p independent discrete variables $x_1, \ldots, x_p$, q independent contin-

uous variables $x_{p+1}, \ldots, x_{p+q}$ and the response variable y. The model proposed to fit the data is a smoothing spline function

$$\begin{aligned} \text{Model 2} : f(x_1, \ldots, x_{p+q}) &= \mu + \sum_{i=p+1}^{p+q} \beta_i x_i + \sum_{i=1}^{p} l_i(x_i) + \sum_{i=p+1}^{p+q} s_i(x_i) + \\ & \sum_{j>i} f_{ij}(x_i, x_j), \end{aligned}$$

where μ is the constant term, $\beta_{p+1}, \ldots, \beta_{p+q}$ are the coefficients for the linear effects in $x_{p+1}, \ldots, x_{p+q}$, $l_1, \ldots, l_p$ represent the level effects for discrete variables $x_1, \ldots, x_p$, $s_{p+1}, \ldots, s_{p+q}$ represent the smooth effects for continuous variables $x_{p+1}, \ldots, x_{p+q}$ and f_{ij} represent the interaction effects between x_i and x_j for both linear and nonlinear effects.

Suppose we are interested in testing the significance of some independent variables, smooth effects or interaction effects such as $\{l_{s_1}, \ldots, l_{p_1}, \beta_{t_1}, \ldots, \beta_{q_1}, s_{t_1}, \ldots, s_{q_1}, f_{s_1,t_1}, \ldots, f_{p_1,q_1}\}$, where $1 \leq s_1, p_1 \leq p$, $p+1 \leq t_1$ and $q_1 \leq p+q$. The corresponding reduced model for testing is

$$\begin{aligned} \text{Model 1} : f(x_1, \ldots, x_{p+q}) &= \mu + \sum_{i=p+1, i\neq(t_1,\ldots,q_1)}^{p+q} \beta_i x_i + \sum_{i=1, i\neq(s_1,\ldots,p_1)}^{p} l_i(x_i) \\ &+ \sum_{i=p+1, i\neq(t_1,\ldots,q_1)}^{p+q} s_i(x_i) \\ &+ \sum_{j>i, i\neq(s_1,\ldots,p_1), j\neq(t_1,\ldots,q_1)} f_{ij}(x_i, x_j) \end{aligned}$$

With Zellner-Siow priors assigned for $(\beta_{t_1}, \ldots, \beta_{q_1})$, $(l_{s_1}, \ldots, l_{p_1})$ and the interaction effects between level effects as discussed in Section 3.2.2 and the scaled χ^2_1 priors assigned for $(s_{t_1}, \ldots, s_{q_1})$ and the interaction effects between the linear effects and nonlinear effects, we can consolidate the terms receiving

the same type of priors. Let p_2 represent the number of terms receiving Zellner-Siow priors and q_2 represent the number of terms receiving scaled χ^2_1 priors in model 2, and let p_3 represent the number of terms receiving Zellner-Siow priors and q_3 represent the number of terms receiving scaled χ^2_1 priors in model 1. Both model 2 and model 1 can be rewritten as

$$\begin{aligned}
\text{Model 2} : f(x_1, \ldots, x_{p+q}) &= \mu + \sum_{i=p+1}^{p+q} \beta_i x_i + \sum_{i=1}^{p_2} l_i(x_i) + \sum_{i=1}^{q_2} s_i(x_i) \\
\text{Model 1} : f(x_1, \ldots, x_{p+q}) &= \mu + \sum_{i=p+1, i\neq(t_1,\ldots,q_1)}^{p+q} \beta_i x_i + \sum_{i=1}^{p_3} l_i(x_i) + \sum_{i=1}^{q_3} s_i(x_i),
\end{aligned}$$

where l_i represents the terms received Zellner-Siow priors, s_i represents the terms received scaled χ^2_1 priors, $p_2 \geq p_3$ and $q_2 \geq q_3$.

The parameters in model 2 are $\boldsymbol{\omega}_{M2} = \{\beta_{p+1}, \ldots, \beta_{p+q}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p_2+q_2}, \lambda_1, \ldots, \lambda_{p_2+q_2}, g, \delta_0\}$. The parameters in model 1 are $\boldsymbol{\omega}_{M1} = \{\beta_{p+1}, \ldots, \beta_{t_1}, \beta_{q_1}, \ldots, \beta_{p+q}, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_{p_3+q_3}, \lambda_1, \ldots, \lambda_{p_3+q_3}, \delta_0\}$. One way to implement bridge sampling for computing the Bayes factor BF_{21} is to integrate out those parameters that are in model 2 but not in model 1 from the product of likelihood and priors under model 2. This is intractable.

An alternative has been proposed to accommodate the computation of Bayes factor BF_{21}. Instead of integrating out those parameters that are in model 2 but not in model 1 to match the parameter spaces for both model 2 and model 1, those parameters will be included in the MCMC steps under model 1 by sampling from the full conditionals derived from the priors those parameters received under model 2. The product of likelihood and priors under

model 2 is

$$
\begin{aligned}
lp(M_2) &= (2\pi\delta_0)^{-n/2} e^{-\frac{1}{2\delta_0}\|\boldsymbol{F}_1'\boldsymbol{y}-\boldsymbol{F}_1'\boldsymbol{T}\boldsymbol{d}\|^2} e^{-\frac{1}{2\delta_0}\|\boldsymbol{F}_2'\boldsymbol{y}-\sum_{i=1}^{p_2+q_2}\boldsymbol{Q}_i\boldsymbol{v}_i\|^2} \\
&\quad \times \prod_{i=1}^{q_2}(2\pi)^{-r_i/2}\delta_0^{-r_i/2}\lambda_i^{r_i/2}|\boldsymbol{D}_i|^{-\frac{1}{2}} e^{-\frac{\lambda_i}{2\delta_0}\boldsymbol{v}_i'\boldsymbol{D}_i^{-1}\boldsymbol{v}_i} \prod_{i=1}^{q_2}\frac{b_i^{1/2}}{\Gamma(1/2)}\lambda_i^{-1/2}e^{-\lambda_i b_i} \\
&\quad \times \prod_{i=1}^{p_2}(2\pi)^{-r_i/2}\delta_0^{-r_i/2}\lambda_i^{r_i/2}(\frac{n}{k_i})^{r_i/2} e^{-\frac{\lambda_i n}{2\delta_0 k_i}\boldsymbol{v}_i'\boldsymbol{v}_i} \prod_{i=1}^{p_2}\frac{(1/2)^{1/2}}{\Gamma(1/2)}\lambda_i^{-1/2}e^{-\lambda_i/2} \\
&\quad \times (2\pi g\delta_0(\boldsymbol{T}_2'\boldsymbol{T}_2)^{-1})^{-1/2} e^{-\frac{1}{2g\delta_0(\boldsymbol{T}_2'\boldsymbol{T}_2)^{-1}}\beta^2} \times \\
&\quad \frac{(n/2)^{1/2}}{\Gamma(1/2)} g^{-3/2} e^{-\frac{n}{2g}} \times \frac{1}{\delta_0}, \qquad (3.65)
\end{aligned}
$$

where $\boldsymbol{T} = [\boldsymbol{x}_{p+1}, \ldots, \boldsymbol{x}_{p+q}]$, $\boldsymbol{T}_1 = [\boldsymbol{x}_{p+1}, \ldots, \boldsymbol{x}_{t_1-1}, \boldsymbol{x}_{q_1+1}, \ldots, \boldsymbol{x}_{p+q}]$, $\boldsymbol{T}_2 = [\boldsymbol{x}_{t_1}, \ldots, \boldsymbol{x}_{q_1}]$, $\boldsymbol{\beta} = (\beta_{t_1}, \ldots, \beta_{q_1})'$, b_i is the scale parameter for χ_1^2 and g is the hyperparameter for the Zellner-Siow prior. The full conditionals for $\boldsymbol{d}_1 = (\beta_{p+1}, \ldots, \beta_{t_1-1}, \beta_{q_1+1}, \ldots, \beta_{p+q})'$, $\boldsymbol{v}_i$, λ_i and δ_0 refer to equations (3.25), (3.26), (3.27) and (3.29) in Section 3.2.1.

The full conditionals for $\boldsymbol{\beta}$ and g are

$$
\begin{aligned}
\boldsymbol{\beta} \mid \delta_0, g &\sim N((\boldsymbol{T}_2'\boldsymbol{F}_{1b}\boldsymbol{F}_{1b}'\boldsymbol{T}_2 + (g\boldsymbol{T}_2'\boldsymbol{T}_2)^{-1})^{-1}(\boldsymbol{T}_2'\boldsymbol{F}_{1b}\boldsymbol{F}_{1b}'\boldsymbol{y}), \\
&\qquad \delta_0(\boldsymbol{T}_2'\boldsymbol{F}_{1b}\boldsymbol{F}_{1b}'\boldsymbol{T}_2 + (g\boldsymbol{T}_2'\boldsymbol{T}_2)^{-1})^{-1}), \qquad (3.66) \\
g \mid \boldsymbol{\beta}, \delta_0 &\sim \text{Inverse Gamma}\left(1, \frac{\boldsymbol{T}_2'\boldsymbol{T}_2\boldsymbol{\beta}^2}{2\delta_0} + \frac{n}{2}\right), \qquad (3.67)
\end{aligned}
$$

where $\boldsymbol{F}_1 = [\boldsymbol{F}_{1a}, \boldsymbol{F}_{1b}]$, $\boldsymbol{F}_1$ is the $n \times q$ matrix of vectors spanning the column space of $\boldsymbol{T}$ and $\boldsymbol{F}_{1b}$ has dimension $n \times (q_1 - t_1)$.

The product of likelihood and priors under model 1 is

$$
lp(M_1) = (2\pi\delta_0)^{-n/2} e^{-\frac{1}{2\delta_0}\|\widetilde{\boldsymbol{F}}_1'\boldsymbol{y}-\widetilde{\boldsymbol{F}}_1'\boldsymbol{T}_1\boldsymbol{d}_1\|^2} e^{-\frac{1}{2\delta_0}\|\widetilde{\boldsymbol{F}}_2'\boldsymbol{y}-\sum_{i=1}^{p_3+q_3}\boldsymbol{Q}_i\boldsymbol{v}_i\|^2}
$$

$$\times \prod_{i=1}^{q_3}(2\pi)^{-r_i/2}\delta_0^{-r_i/2}\lambda_i^{r_i/2}|\boldsymbol{D}_i|^{-\frac{1}{2}}e^{-\frac{\lambda_i}{2\delta_0}\boldsymbol{v}_i'\boldsymbol{D}_i^{-1}\boldsymbol{v}_i}\prod_{i=1}^{q_3}\frac{b_i^{1/2}}{\Gamma(1/2)}\lambda_i^{-1/2}e^{-\lambda_i b_i}$$
$$\times \prod_{i=1}^{p_3}(2\pi)^{-r_i/2}\delta_0^{-r_i/2}\lambda_i^{r_i/2}(\frac{n}{k_i})^{r_i/2}e^{-\frac{\lambda_i n}{2\delta_0 k_i}\boldsymbol{v}_i'\boldsymbol{v}_i}\prod_{i=1}^{p_3}\frac{(1/2)^{1/2}}{\Gamma(1/2)}\lambda_i^{-1/2}e^{-\lambda_i/2}$$
$$\times(2\pi g\delta_0(\boldsymbol{T}_2'\boldsymbol{T}_2)^{-1})^{-1/2}e^{-\frac{1}{2g\delta_0(\boldsymbol{T}_2'\boldsymbol{T}_2)^{-1}}\beta^2}\times$$
$$\frac{(n/2)^{1/2}}{\Gamma(1/2)}g^{-3/2}e^{-\frac{n}{2g}}\times\frac{1}{\delta_0}, \tag{3.68}$$

where $\widetilde{\boldsymbol{F}}=[\widetilde{\boldsymbol{F}}_1,\widetilde{\boldsymbol{F}}_2]$, $\widetilde{\boldsymbol{F}}_1$ is the $n\times(q-q_1+t_1)$ matrix of vectors spanning the column space of $\boldsymbol{T}_1$.

Take the ratio of $lp(M_2)$ to $lp(M_1)$,

$$\frac{q_0}{q_1}=\frac{e^{-\frac{1}{2\delta_0}\|\boldsymbol{F}_1'\boldsymbol{y}-\boldsymbol{F}_1'\boldsymbol{T}\boldsymbol{d}\|^2}e^{-\frac{1}{2\delta_0}\|\boldsymbol{F}_2'\boldsymbol{y}-\sum_{i=1}^{p_2+q_2}\boldsymbol{Q}_i\boldsymbol{v}_i\|^2}}{e^{-\frac{1}{2\delta_0}\|\widetilde{\boldsymbol{F}}_1'\boldsymbol{y}-\widetilde{\boldsymbol{F}}_1'\boldsymbol{T}_1\boldsymbol{d}_1\|^2}e^{-\frac{1}{2\delta_0}\|\widetilde{\boldsymbol{F}}_2'\boldsymbol{y}-\sum_{i=1}^{p_3+q_3}\boldsymbol{Q}_i\boldsymbol{v}_i\|^2}}. \tag{3.69}$$

Utilize the random samples generated from the full conditionals under both model 2 and model 1, then apply equation (3.55) to compute the Bayes factor BF_{21} to test the significance of $\{l_{s_1},\ldots,l_{p_1},\beta_{t_1},\ldots,\beta_{q_1},s_{t_1},\ldots,s_{q_1},f_{s_1,t_1},\ldots,f_{p_1,q_1}\}$.

In order to adapt the Bayes factor for testing the linear effects in continuous variables, the linear terms must have informative priors such as Zellner-Siow priors, which is different from the estimation and prediction situation where a flat prior in the null spaces constructed by the linear effects of continuous variables is favored. So if the goal of the study is parameter estimation or prediction, then the flat prior in the null space is chosen. However, if the goal is testing the linear effects in continuous variables, then the informative priors are needed.

Two simulated examples demonstrate this alternative approach in comput-

ing Bayes factors.

Example 3.2

Consider simulated data from a balanced design with two independent continuous variables x_1 and x_2 and one response variable y. Take $x_{1i} = (i-1)/(m-1)$, $x_{2j} = (j-1)/(m-1)$ and $y_{ij} = 1 + 3\sin(2\pi x_{1i} - \pi) + \varepsilon_{ij}$ with $\delta_0 = 1$, where $i = 1, \ldots, m,\ j = 1, \ldots, m$. We vectorized this model as follows. Let $\boldsymbol{e} = (0, \ldots, (m-1))'/(m-1)$ and $\mathbf{1} = (1, \ldots, 1)'_{m\times 1}$ with $m = 30$. Then let $\boldsymbol{x}_1 = \boldsymbol{e} \otimes \mathbf{1}$ and $\boldsymbol{x}_2 = \mathbf{1} \otimes \boldsymbol{e}$, where $\otimes$ is the Kronecker product. Let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})'$ and $\boldsymbol{y} = (\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_m)'$. The sample size is $n = 900$.

Suppose one wants to test if x_2 is a significant variable to predict y. The models proposed for testing are

$$\begin{aligned} \text{Model 2} : f(x_1, x_2) &= \mu + \beta_1 x_1 + \beta_2 x_2 + s_1(x_1) + s_2(x_2), \qquad (3.70) \\ \text{Model 1} : f(x_1, x_2) &= \mu + \beta_1 x_1 + s_1(x_1), \end{aligned}$$

where μ is the constant term, β_1 is the coefficient for the linear effect in x_1, β_2 is the coefficient for the linear effect in x_2, s_1 represents the smooth effect in x_1 and s_2 represents the smooth effect in x_2.

To test for the effect of x_2, it is required to test $f(x_2) = \beta_2 x_2 + s_2(x_2)$. The Zellner-Siow prior was used for β_2, the coefficient for the linear effect in x_2. The scaled χ^2_1 priors were applied to test smooth effects, both s_1 and s_2. The scale parameters in χ^2_1 priors were selected to be 0.01 for both s_1 and s_2, which provided 5.25 effective degrees of freedom for each of the smooth effects, s_1 and s_2. Figure 3.7 shows the MCMC trace plots for parameters u_{10}, λ_1, λ_2 and δ_0 under model 2. Convergence is rapid in all cases.

Followed the procedure of the alternative approach for computing the Bayes factor, $BF_{21} = 8.24 \times 10^{-88}$, providing conclusive evidence that x_2 contributes no information in predicting y. This conclusion agrees with the true model.

The second simulated dataset is also a balanced design with two independent variables, one discrete (x_1) and one continuous (x_2), and response variable y. Take $x_{1i} = i$, $x_{2j} = (j-1)/(m-1)$ and $y_{ij} = 1 + 2I_{\{x_{1i}=1\}} + 3\sin(2\pi x_{2j} - \pi) + \varepsilon_{ij}$ with $\delta_0 = 4$, where $i = 1, 2, 3, 4,\ j = 1, \ldots, m$. We vectorized this model as follows. Let $\boldsymbol{e} = (0, \ldots, (m-1))'/(m-1)$, $\mathbf{1}_1 = (1, \ldots, 1)'_{4\times 1}$ and $\mathbf{1}_2 = (1, \ldots, 1)'_{100\times 1}$, with $m = 100$, and let $\boldsymbol{x}_1 = (1, 2, 3, 4)' \otimes \mathbf{1}_2$ and $\boldsymbol{x}_2 = \mathbf{1}_1 \otimes \boldsymbol{e}$. Finally, let $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{im})'$ and $\boldsymbol{y} = (\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_4)'$. The sample size is $n = 400$.

Again the goal is to identify if x_2 is a significant variable to predict y, and the models proposed for testing are

$$\text{Model 2} : f(x_1, x_2) = \mu + \beta_2 x_2 + l_1(x_1) + s_2(x_2), \tag{3.71}$$

$$\text{Model 1} : f(x_1, x_2) = \mu + l_1(x_1), \tag{3.72}$$

where μ is the constant term, β_2 is the coefficient for the linear effect in x_2, l_1 represents the level effects in x_1 and s_2 represents the smooth effect in x_2.

To test for the effect of x_2, it is required to test $f(x_2) = \beta_2 x_2 + s_2(x_2)$. The Zellner-Siow priors were used for β_2 and the level effects in x_1. The scaled χ^2_1 prior was applied to test the smooth effects in x_2. The scale parameter in χ^2_1 prior was selected to be 0.001, which provided 5.33 effective degrees of freedom for the smooth effect, s_2. Figure 3.8 shows the MCMC trace plots for parameters u_{10}, λ_1, λ_2 and δ_0 under model 2. Convergence is rapid in all

cases.

The result for testing $f(x_2) = \beta_2 x_2 + s_2(x_2)$ is $BF_{21} = 1.70 \times 10^{130}$, providing conclusive evidence that x_2 contributes information in predicting y. This again agrees with the true model.

Example 3.3

We revisit the manufacturing dataset in Section 3.2.6 using the alternative Bayes factor to test the effect of *time*. The models proposed for testing are equations (3.71) and (3.72), where x_1 represents the variable *operator* and x_2 represents the variable *time*.

To test for the effect of *time*, it is required to test $f(time) = \beta_2 time + s_2(time)$. The same set of priors for model 2 in equation (3.71) were used. The scale parameter in the χ^2_1 prior was selected to be 0.0001, which provided 5.01 effective degrees of freedom for the smooth effect, s_2. Figure 3.9 shows the MCMC trace plots for parameters u_{10}, λ_1, λ_2 and δ_0 under model 2. Convergence is rapid in all cases.

The corresponding $BF_{21} = 3.44 \times 10^{14}$ provides conclusive evidence that *time* contributes information in predicting y.

3.2.8 Simulated Example 1

Consider a simulated dataset with three independent variables, a discrete variable (x_1) with four levels, a second discrete variable (x_2) with twenty levels and one continuous variable (x_3), along with response variable y. Take $(x_{11}, x_{12}, x_{13}, x_{14})' = (a, b, c, d)$, $x_{2j} = j$, $x_{3k} = (k-1)/(m-1)$ and $y_{ijk} = 1 + 2I_{\{x_{1i}=a\}} + 3\sin(2\pi x_{3k} - \pi) + \varepsilon_{ijk}$ with $\delta_0 = 1$, where $i = 1, 2, 3, 4$, $j = 1, \ldots, 20$, $k = 1, \ldots, m$. We vectorized this model as follows. Let $\boldsymbol{e} =$

$(0, \ldots, (m-1))'/(m-1)$, $\mathbf{1}_1 = (1, \ldots, 1)'_{100\times 1}$ and $\mathbf{1}_2 = (1, \ldots, 1)'_{20\times 1}$, where $m = 20$ then $\boldsymbol{x}_1 = (a, b, c, d)' \otimes \mathbf{1}_1$, $\boldsymbol{x}_2 = (1, \ldots, 20)' \otimes \mathbf{1}_2$ and $\boldsymbol{x}_3 = \boldsymbol{e} \otimes \mathbf{1}_2$. Let $\boldsymbol{y}_{ij} = (y_{ij1}, \ldots, y_{ijm})'$, and $\boldsymbol{y}_i = (\boldsymbol{y}'_{i1}, \ldots, \boldsymbol{y}'_{i,20})'$ and $\boldsymbol{y} = (\boldsymbol{y}'_1, \ldots, \boldsymbol{y}'_4)'$. The sample size is $n = 400$.

The model proposed to analyze the effect of each covariate is a smoothing spline function

$$\begin{aligned} \text{Model 7} : f(x_1, x_2, x_3) &= \mu + \beta_3 x_3 + s_3(x_3) + l_1(x_1) + l_2(x_2) + l_{13}(x_1, x_3) \\ &\quad + ls_{13}(x_1, x_3) + l_{23}(x_2, x_3) + ls_{23}(x_2, x_3), \end{aligned} \tag{3.73}$$

where μ is the constant term, β_3 is the coefficient for the linear effect in x_3, s_3 represents the smooth effect in x_3, l_1 represents the level effects in x_1, l_2 represents the level effects in x_2, l_{13} represents the interaction effects between the level effects in x_1 and linear effects in x_3, ls_{13} represents the interaction effects between the level effects in x_1 and smooth effect in x_3, l_{23} represents the interaction effects between the level effects in x_2 and linear effects in x_3 and ls_{23} represents the interaction effects between the level effects in x_2 and smooth effect in x_3.

To test for the significance of each term in the model 7, a series of partially nested models of interest is proposed:

$$\begin{aligned} \text{Model 6} : f(x_1, x_2, x_3) &= \mu + \beta_3 x_3 + s_3(x_3) + l_1(x_1) + l_2(x_2) + l_{13}(x_1, x_3) \\ &\quad + ls_{13}(x_1, x_3) + l_{23}(x_2, x_3), \end{aligned} \tag{3.74}$$

$$\text{Model 5} : f(x_1, x_2, x_3) = \mu + \beta_3 x_3 + s_3(x_3) + l_1(x_1) + l_2(x_2) + l_{13}(x_1, x_3)$$

$$+ls_{13}(x_1, x_3), \tag{3.75}$$

$$\text{Model 4} : f(x_1, x_2, x_3) = \mu + \beta_3 x_3 + s_3(x_3) + l_1(x_1) + l_2(x_2)$$

$$+l_{13}(x_1, x_3), \tag{3.76}$$

$$\text{Model 3} : f(x_1, x_2, x_3) = \mu + \beta_3 x_3 + s_3(x_3) + l_1(x_1) + l_2(x_2), \tag{3.77}$$

$$\text{Model 2} : f(x_1, x_3) = \mu + \beta_3 x_3 + s_3(x_3) + l_1(x_1), \tag{3.78}$$

$$\text{Model 1} : f(x_3) = \mu + \beta_3 x_3 + s_3(x_3), \tag{3.79}$$

$$\text{Model 0} : f(x_1, x_3) = \mu + \beta_3 x_3 + l_1(x_1). \tag{3.80}$$

The Zellner-Siow prior was used for β_3, the coefficient for the linear effect in x_3. Two different types of priors, Pareto and scaled χ^2_1, were applied to test the smooth effects. The scale parameters in the Pareto priors or those for the χ^2_1 priors were selected by the effective degrees of freedom method as discussed in Section 3.1.2. The values of λ corresponding to reasonable prior effective degrees of freedom for each term in model 7 are listed in Table 3.3. Figure 3.10 shows the MCMC trace plots for parameters u_{10}, λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, λ_7 and δ_0 under model 7 with scaled χ^2_1 priors. Convergence is rapid in all cases. These were used in the computation of the Bayes factors. The Bayes factors under both the Pareto and scaled χ^2_1 priors for model comparison and the specific terms to be tested are listed in Table 3.4.

Based on Table 3.4, both the Pareto and scaled χ^2_1 priors have come up with the same conclusion that there is insufficient evidence to conclude that $ls_{23}(x_2, x_3)$, $l_{23}(x_2, x_3)$, $ls_{13}(x_1, x_3)$ and $l_{13}(x_1, x_3)$ are significant and sufficient evidence to conclude that $l_2(x_2)$, $l_1(x_1)$ and $s_3(x_3)$ are significant in predicting y. The most parsimonious model to predict y is model 3. Figure 3.11 shows

the model 7 and model 3 provided almost the same fit and is a more suitable fit than those in model 1 and model 0 as confirmed in the model selection. When constructing the response variable y, the x_2 is not involved. But the hypothesis testing concluded that $l_2(x_2)$ is significant. This is caused by the confounding effect between x_1 and x_2. There are 5 levels of x_2 for each level of x_1 in this data. Since the x_1 is significant, so is x_2.

Table 3.3: The λs for each term in Model 7 giving the desirable effective degrees of freedom.

	s_3	l_{13}	ls_{13}	l_{23}	ls_{23}
λ	0.0001	0.1	0.1	10	1
Effective *df*	4.44	2.83	2.781	2.775	5.92

Table 3.4: The Bayes factors for testing each term in model 7 adapted both the scaled χ^2_1 and the Pareto priors in all the terms except $l_2(x_2)$ and $l_1(x_1)$, which received Zellner-Siow priors.

	BF_{76}	BF_{65}	BF_{54}	BF_{43}
scaled χ^2_1	0.202	0.371	0.357	0.121
Pareto	0.351	0.539	0.479	0.244
Terms	$ls_{23}(x_2, x_3)$	$l_{23}(x_2, x_3)$	$ls_{13}(x_1, x_3)$	$l_{13}(x_1, x_3)$
	BF_{32}	BF_{21}	BF_{20}	
scaled χ^2_1	1.0×10^{70}	1.4×10^{104}	∞	
Pareto	1.0×10^{70}	2.5×10^{104}	∞	
Terms	$l_2(x_2)^*$	$l_1(x_1)^*$	$s_3(x_3)$	

3.2.9 Potassium Measurement on Dogs

In this section, a dataset from Wang and Ke (2004) is revisited using fully Bayesian SSANOVA models. Thirty-six dogs were assigned to four groups: control, extrinsic cardiac denervation three weeks prior to coronary occlusion, extrinsic cardiac denervation immediately prior to coronary occlusion, and bilateral thoratic sympathectomy and stellectomy three weeks prior to coronary

occlusion. Coronary sinus potassium concentrations were measured on each dog every two minutes from 1 to 13 minutes after occlusion. The goal is to identify if the variables *group*, *dog* and *time* are useful to predict the *potassium* concentrations.

In order to adapt the model setting in Section 3.2.8, x_1 represents the variable *group*, x_2 represents the variable *dog*, x_3 represents the variable *time* and y represents the response variable *potassium*. The model proposed to fit the data is model 7 in equation (3.73). To test the significance of each term, a series of models in equations (3.74), (3.75), (3.76), (3.77), (3.78), (3.79) and (3.80) are proposed with the same priors, linear/level effects with Zellner-Siow priors and smooth effects with Pareto or scaled χ^2_1 priors, chosen in the Section 3.2.8. The scale parameter in the Pareto prior or that for the scaled χ^2_1 prior were selected by effective degrees of freedom. The values of λ corresponding to desirable effective degrees of freedom for each term in model 7 are listed in Table 3.5. Figure 3.12 shows the MCMC trace plots for parameters u_{10}, λ_1, λ_2, λ_3, λ_4, $log(\lambda_5)$, $log(\lambda_6)$, $log(\lambda_7)$ and δ_0 under model 7 with scaled χ^2_1 priors. Convergence is rapid in all cases. The Bayes factors under both Pareto and scaled χ^2_1 priors for model comparison and the specific terms to be tested are listed in Table 3.6.

Based on Table 3.6, under either the Pareto or the scaled χ^2_1 priors, we conclude that there is sufficient evidence that $ls_{23}(dog, time)$, $l_{23}(dog, time)$, $l_{13}(dog, time)$, $l_2(dog)$, $l_1(group)$ and $s_3(time)$ are significant in predicting *potassium*. However BF_{54} didn't provide sufficient evidence to conclude if $ls_{13}(group, time)$ is significant or not. Figure 3.13 shows the fits from model 7, model 3, model 1 and model 0, while the most suitable fit to the dataset is

model 7 as confirmed in the model selection.

In the study conducted by Wang and Ke (2004), model 7, model 6 and model 5 in equations (3.73), (3.74) and (3.75) were fit to the data. Through the AIC criteria, model 7 was judged best. The Bayes factor analysis suggests that the $ls_{13}(group, time)$ term could be omitted. The focus of their study is more in model fitting and parameter estimation rather than hypothesis testing. The flexibility of hypothesis testing is the advantage of this fully Bayesain SSANOVA approach.

Table 3.5: The λs for each term in Model 7 giving the desirable effective degrees of freedom.

	$s_3(time)$	$l_{13}(group, time)$	$ls_{13}(group, time)$
λ	0.0001	0.1	0.1
Effective *df*	3.09	2.59	2.01
	$l_{23}(dog, time)$	$ls_{23}(dog, time)$	
λ	10	1	
Effective *df*	2.08	5.13	

Table 3.6: The Bayes factors for testing each term in model 7 adapted both scaled χ^2_1 and Pareto priors in all the terms except $l_2(dog)$ and $l_1(group)$, which received Zellner-Siow priors.

	BF_{76}	BF_{65}	BF_{54}	BF_{43}
scaled χ^2_1	954	1.58×10^4	1.23	4.56×10^5
Pareto	116	3.85×10^3	1.27	3.86×10^5
Terms	$ls_{23}(dog, time)$	$l_{23}(dog, time)$	$ls_{13}(group, time)$	$l_{13}(group, time)$
	BF_{32}	BF_{21}	BF_{20}	
scaled χ^2_1	9×10^{63}	4.3×10^{51}	8.08×10^{63}	
Pareto	6×10^{62}	3.8×10^{51}	8.07×10^{63}	
Terms	$l_2(dog)^*$	$l_1(group)^*$	$s_3(time)$	

Chapter 4

Fully Bayesian SSANOVA for Binary response variables

In Chapter 3, the fully Bayesian SSANOVA model is for Gaussian response variables. In this chapter it will be extended to binary response variables.

4.1 Binary response variable

Let $\boldsymbol{y}$ denotes the vector with components $y_i,\ i = 1, 2, \ldots, n$, where y_i is the observed response with only two possible outcomes. Take y_i to be independent Bernoulli random variables with probability $p_i = p(y_i = 1) = H(\boldsymbol{x}_i'\boldsymbol{\beta})$, where $\boldsymbol{x}_i' = (x_{i1}, \ldots, x_{ip})$, the data vector for the ith case, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ are the parameters to be estimated and H is a cumulative distribution function for a continuous random variable. The inverse, H^{-1}, is also known as the link function. There are many choices for link function. One popular choice is $H = \Phi$, the cumulative distribution function of a standard normal distribution, leading to the probit model. Thus the probability function of $\boldsymbol{y}$ is

$$p(\boldsymbol{y} \mid \boldsymbol{\beta}) \quad = \quad \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{(1-y_i)}$$

$$= \prod_{i=1}^{n} \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})^{y_i}(1-\Phi(\boldsymbol{x}_i'\boldsymbol{\beta}))^{(1-y_i)}$$

As in Hastie and Tibshirani (1999), Gu (2002) and Wang and Ke (2004), this generalized linear model setup can be extended to a generalized additive model with the probability

$$p_i = p(y_i = 1) = H\left(\mu + \sum_{j=1}^{k} f_j(x_{ij})\right), \tag{4.1}$$

where μ is the constant term and the f_j represent linear effects, smooth effects or interaction effects.

The SSANOVA approach has been applied to generalized linear models by Gu (2002). In this section, we extend the fully Bayesian SSANOVA models of Section 3.2.1 to the case of probit regression. The solution for fully Bayesian SSANOVA model discussed in Section 3.2.1 was for a Gaussian response variable. To extend the work for binary response variables, one solution is through data augmentation. Followed Albert and Chib (1993), introduce n independent latent variables $\boldsymbol{z} = (z_1, \ldots, z_n)$, where the z_i are independent $N((\boldsymbol{T}\boldsymbol{d} + (\sum_{k=1}^{p} \boldsymbol{\Sigma}_k)\,\boldsymbol{c})_i, 1)$, and $y_i = 1$ if $z_i > 0$ and $y_j = 0$ if $z_j < 0$. With the solution for the fully Bayesian SSANOVA model discussed in Section 3.2.1, the distribution of the latent variables $\boldsymbol{z}$ is

$$\begin{aligned} \begin{pmatrix} \boldsymbol{F}_1'\boldsymbol{z} \\ \boldsymbol{F}_2'\boldsymbol{z} \end{pmatrix} \mid \boldsymbol{d}, \boldsymbol{u}_2, \lambda &\sim N\left(\begin{pmatrix} \boldsymbol{F}_1'\boldsymbol{T}\boldsymbol{d} \\ \boldsymbol{u}_2 \end{pmatrix}, \boldsymbol{I}_n\right), \\ \boldsymbol{z} \mid \boldsymbol{d}, \boldsymbol{u}_2, \lambda &\sim N(\boldsymbol{T}\boldsymbol{d} + \boldsymbol{F}_2\boldsymbol{u}_2, \boldsymbol{I}_n), \end{aligned} \tag{4.2}$$

where the variance is set to be 1 for identifiability.

The prior for $\boldsymbol{d}$ is assigned to be flat. The prior distribution for $\boldsymbol{u}_2$ is suggested as equations (3.23) and (3.24). As discussed by Sun et al. (2001), the posterior distribution for $\boldsymbol{z}$ is proper. For the linear effects, Zellner-Siow priors in equation (3.44) are suggested as discussed in Section 3.2.2. For smooth effects, the scaled χ^2_1 priors are suggested as discussed in Section 3.2.2. The scale for the scaled χ^2_1 is selected by the effective degrees of freedom discussed in Section 3.1.2.

Given the data y_i, the full conditional truncated normal distribution for the latent variables $\boldsymbol{z}$ given $\boldsymbol{y}$ satisfies

$$\begin{aligned} p(\boldsymbol{z} \mid \boldsymbol{y}, \boldsymbol{d}, \boldsymbol{u}_2, \lambda) \quad \propto \quad & \prod_{i=1}^{n} \{1_{(\boldsymbol{z}_i>0)} 1_{(\boldsymbol{y}_i=1)} + 1_{(\boldsymbol{z}_i<0)} 1_{(\boldsymbol{y}_i=0)}\} \\ & \times (2\pi)^{-n/2} e^{\{-\frac{1}{2}(\boldsymbol{z}-\boldsymbol{T}\boldsymbol{d}-\boldsymbol{F}_2\boldsymbol{u}_2)^2\}}, \end{aligned}$$

where $1_{(x\in A)}$ is the indicator function, $1_{(x\in A)} = 1$ if $x \in A$ and 0 otherwise. This can be simplified as

$$y_i = 1, \quad z_i \mid \boldsymbol{d}, \boldsymbol{u}_2, \lambda \quad \sim \quad N_+((\boldsymbol{T}\boldsymbol{d} + \boldsymbol{F}_2\boldsymbol{u}_2)_i, 1), \tag{4.3}$$

$$y_i = 0, \quad z_i \mid \boldsymbol{d}, \boldsymbol{u}_2, \lambda \quad \sim \quad N_-((\boldsymbol{T}\boldsymbol{d} + \boldsymbol{F}_2\boldsymbol{u}_2)_i, 1), \tag{4.4}$$

where N_+ is the positive normal distribution restricted to $(0, \infty)$ and N_- is the negative normal distribution restricted to $(-\infty, 0)$.

Followed the algorithm in Devroye (1986), sampling from the full conditionals in equations (4.3) and (4.4) can be accomplished as follows:

$$\boldsymbol{y}_i = 1, \ \boldsymbol{z}_i > 0: \qquad \boldsymbol{z}_i = (\boldsymbol{T}\boldsymbol{d} + \boldsymbol{F}_2\boldsymbol{u}_2)_i + \Phi^{-1}(1 - u\Phi((\boldsymbol{T}\boldsymbol{d} + \boldsymbol{F}_2\boldsymbol{u}_2)_i)),$$

$$\boldsymbol{y}_i = 0,\ \boldsymbol{z}_i < 0: \qquad \boldsymbol{z}_i = (\boldsymbol{T}\boldsymbol{d} + \boldsymbol{F}_2\boldsymbol{u}_2)_i + \Phi^{-1}(u\Phi(-(\boldsymbol{T}\boldsymbol{d} + \boldsymbol{F}_2\boldsymbol{u}_2)_i)),$$

where u is a random draw from the uniform distribution on $[0, 1]$.

The full conditionals for each of the parameters are similar to equations (3.25), (3.26), (3.27) except that $\boldsymbol{z}$ is involved, scaled χ^2_1 priors have been assigned for smooth effects, and $\delta_0 = 1$:

$$\begin{aligned}
\boldsymbol{d} \mid \boldsymbol{v}_k &\sim N\Big((\boldsymbol{T}'\boldsymbol{T})^{-1}\boldsymbol{T}'\Big(\boldsymbol{z} - \boldsymbol{F}_2\sum_{i=1}^{p}\boldsymbol{Q}_i\boldsymbol{v}_i\Big), \Big(\boldsymbol{T}'\boldsymbol{T}\Big)^{-1}\Big), && (4.5)\\
\boldsymbol{v}_k \mid \boldsymbol{v}_{-k}, \lambda_k, &\sim N\Big((\boldsymbol{I}_{r_k} + \lambda_k\boldsymbol{D}_k^{-1})^{-1}\boldsymbol{Q}_k'\boldsymbol{F}_2'\Big(\boldsymbol{z} - \boldsymbol{T}\boldsymbol{d} - \boldsymbol{F}_2\Big(\sum_{i=1, i\neq k}^{p}\boldsymbol{Q}_i\boldsymbol{v}_i\Big),\\
&\quad (\boldsymbol{I}_{r_k} + \lambda_k\boldsymbol{D}_k^{-1})^{-1}\Big), \quad k = 1, \ldots, p, && (4.6)\\
\lambda_k \mid \boldsymbol{v}_k &\sim \text{Gamma}\Big(\frac{r_k + 1}{2}, \Big(\frac{2}{\boldsymbol{v}_k'\boldsymbol{D}_k^{-1}\boldsymbol{v}_k + 2\beta_k}\Big)\Big), && (4.7)\\
&\quad k = 1, \ldots, p,
\end{aligned}$$

where $r_1, \ldots, r_p$ are the ranks of the nonzero eigenvalue matrices $\boldsymbol{D}_1, \ldots, \boldsymbol{D}_p$, $\beta_1, \ldots, \beta_p$ are the scale parameters in the prior distribution for each smoothing parameter $\lambda_1, \ldots, \lambda_p$, and $\boldsymbol{v}_{-k}$ denotes the set of vectors $\{\boldsymbol{v}_j, j \neq k\}$. When the covariates are discrete, the full conditionals for each of the parameters stay the same except in equation (4.6), where the diagonal elements in $\boldsymbol{D}_k^{-1}$ are replaced by n/K and in equation (4.7), where β_k is replaced by $1/2$.

The posterior distributions for each of the parameters can be simulated based on the full conditionals.

4.1.1 Simulated Example 2

Consider a simulated dataset with two continuous independent variables. The sample size is $n = 500$, and x_{1i} and x_{2i} are independent samples from the uniform distribution on $[0, 1]$. The binary response variable y_i is 0 if $0.25x_{1i} + 1.5\sin(2\pi x_{2i} - \pi) + \varepsilon_i$ is negative and y_i is 1 if $0.25x_{1i} + 1.5\sin(2\pi x_{2i} - \pi) + \varepsilon_i$ is positive. The variance $\delta_0 = 1$. The model proposed for this study is a smoothing spline function

$$\begin{aligned} \text{Model 6} : f(x_1, x_2) &= \mu + \beta_1 x_1 + \beta_2 x_2 + s_1(x_1) + s_2(x_2) + ls_{12}(x_1, x_2) \\ &\quad + sl_{12}(x_1, x_2) + s_{12}(x_1, x_2), \end{aligned} \tag{4.8}$$

where μ is the constant term, β_1 is the coefficient for the linear effect in x_1, β_2 is the coefficient for the linear effect in x_2, s_1 represents the smooth effect in x_1, s_2 represents the smooth effect in x_2, ls_{12} represents the interaction effects between the level effects in x_1 and smooth effect in x_2, sl_{12} represents the interaction effects between the smooth effect in x_1 and level effects in x_2 and s_{12} represents the interaction effects between the smooth effect in x_1 and smooth effect in x_2. To test for the significance of each term in model 6, a series of partially nested models of interest is proposed:

$$\begin{aligned} \text{Model 5} : f(x_1, x_2) &= \mu + \beta_1 x_1 + \beta_2 x_2 + s_1(x_1) + s_2(x_2) + ls_{12}(x_1, x_2) \\ &\quad + sl_{12}(x_1, x_2), \end{aligned} \tag{4.9}$$

$$\begin{aligned} \text{Model 4} : f(x_1, x_2) &= \mu + \beta_1 x_1 + \beta_2 x_2 + s_1(x_1) + s_2(x_2) + \\ &\quad ls_{12}(x_1, x_2), \end{aligned} \tag{4.10}$$

$$\text{Model 3} : f(x_1, x_2) = \mu + \beta_1 x_1 + \beta_2 x_2 + s_1(x_1) + s_2(x_2), \tag{4.11}$$

$$\text{Model 2} : f(x_1, x_2) = \mu + \beta_1 x_1 + \beta_2 x_2 + s_1(x_1), \tag{4.12}$$

$$\text{Model 1} : f(x_1, x_2) = \mu + \beta_1 x_1 + \beta_2 x_2 + s_2(x_2). \tag{4.13}$$

Two different types of priors, Pareto and scaled χ^2_1, were applied to test the smooth effects. The scale parameters in the Pareto priors or those for the χ^2_1 priors were selected by effective degrees of freedom. The values of λ corresponding to desirable effective degrees of freedom for each term in model 6 are listed in Table 4.1. Figure 4.1 shows the MCMC trace plots for parameters u_{10}, $log(\lambda_1)$, $log(\lambda_2)$, $log(\lambda_3)$, $log(\lambda_4)$ and $log(\lambda_5)$ under model 6 with scaled χ^2_1 priors, which have converged in all cases. These were used in the computation of Bayes factors. The Bayes factors under both the Pareto and scaled χ^2_1 priors for model comparison and the specific term to be tested are listed in Table 4.2.

Based on Table 4.2, under either the Pareto or the scaled χ^2_1 priors, we conclude that there is insufficient evidence that $s_{12}(x_1, x_2)$, $sl_{12}(x_1, x_2)$, $ls_{12}(x_1, x_2)$ and $s_1(x_1)$ are significant and sufficient evidence that $s_2(x_2)$ is significant in predicting y. The most parsimonious model to predict y is model 1, which agrees with the true model. Figure 4.3 shows the fits of components in model 6 with 95% credible sets, and Figure 4.2 shows the fits of each component in model 6 and prediction by model 6, which help us to visualize the effects.

Table 4.1: The λs for each term in Model 6 giving the desirable effective degrees of freedom.

	$s_1(x_1)$	$s_2(x_2)$	$ls_{12}(x_1, x_2)$	$sl_{12}(x_1, x_2)$	$s_{12}(x_1, x_2)$
λ	0.005	0.005	0.005	0.005	0.005
Effective *df*	5.27	5.25	4.57	4.43	5.57

Table 4.2: The Bayes factors for testing each term in model 6 adapted both scaled χ^2_1 and Pareto priors.

	BF_{65}	BF_{54}	BF_{43}	BF_{32}	BF_{31}
scaled χ^2_1	0.471	0.345	0.125	3.02×10^7	0.232
Pareto	0.347	0.272	0.156	2.15×10^7	0.260
Terms	$s_{12}(x_1, x_2)$	$sl_{12}(x_1, x_2)$	$ls_{12}(x_1, x_2)$	$s_2(x_2)$	$s_1(x_1)$

4.1.2 Wisconsin Epidemiological Study of Diabetic Retinopathy

A dataset from Wahba et al. (1995) is revisited using a fully Bayesian binary response SSANOVA model. The Wisconsin Epidemiological Study of Diabetic Retinopathy (WESDR) is an epidemiological study of a cohort of diabetic patients receiving their medical care in an 11-county area in Southern Wisconsin. Detailed descriptions of the data can be found in Klein and DeMets (1988). A number of medical and demographical variables were collected for this study. We analyze the subgroup of the younger onset population, consisting of 669 subjects with no or non-proliferative retinopathy at the start, and no missing data from the variables we studied. The goal of this study was to examine how the progression of diabetic retinopathy at the first follow-up depends on the following continuous covariates: *dur* (duration of diabetes at baseline), *gly* (glycosylated hemoglobin, a measure of hyperglycemia), and *bmi* (body mass index = weight in kg/(height in m)^2).

A smoothing spline function is proposed to study the effect of each covariate.

$$\begin{aligned} \text{Model 13} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_1(gly) \\ &\quad + s_2(dur) + s_3(bmi) + sl_{12}(gly, dur) \\ &\quad + ls_{12}(gly, dur) + s_{12}(gly, dur), \qquad (4.14) \end{aligned}$$

where μ is the constant term, β_1 is the coefficient for the linear effect in *gly*, β_2 is the coefficient for the linear effect in *dur*, β_3 is the coefficient for the linear effect in *bmi*, s_1 represents the smooth effect in *gly*, s_2 represents the smooth effect in *dur*, s_3 represents the smooth effect in *bmi*, sl_{12} represents the interaction effects between the smooth effect in *gly* and linear effects in *dur*, ls_{12} represents the interaction effects between the linear effects in *gly* and smooth effect in *dur*, s_{12} represents the interaction effects between the smooth effect in *gly* and smooth effect in *dur*.

To test for the significance of each term in model 13, a series of partially nested models of interest is proposed. The partially nested models are:

$$\begin{aligned}
\text{Model 12} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_1(gly) \\
&\quad + s_2(dur) + s_3(bmi) + sl_{12}(gly, dur) \\
&\quad + ls_{12}(gly, dur), \qquad (4.15) \\
\text{Model 11} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_1(gly) \\
&\quad + s_2(dur) + s_3(bmi) \\
&\quad + sl_{12}(gly, dur), \qquad (4.16) \\
\text{Model 10} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_1(gly) \\
&\quad + s_2(dur) + s_3(bmi), \qquad (4.17) \\
\text{Model 9} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_1(gly) \\
&\quad + s_2(dur), \qquad (4.18) \\
\text{Model 8} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi \\
&\quad + s_1(gly), \qquad (4.19) \\
\text{Model 7} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi
\end{aligned}$$

$$+s_2(dur). \tag{4.20}$$

The scaled χ^2_1 priors have been applied to test the smooth effects. The scale parameters for the χ^2_1 priors were selected by effective degrees of freedom. The values of λ corresponding to desirable effective degrees of freedom for each term in model 13 are listed in Table 4.3. Figure 4.4 shows the MCMC trace plots for parameters $log(\lambda_1)$, $log(\lambda_2)$, $log(\lambda_3)$, $log(\lambda_4)$, $log(\lambda_5)$ and $log(\lambda_6)$ under model 13, which have converged in all cases. These were used in the computation of Bayes factors. The Bayes factors with scaled χ^2_1 priors for model comparison and the specific term to be tested are listed in Table 4.4.

Based on Table 4.4, there is insufficient evidence to conclude that $s_{12}(gly, dur)$, $ls_{12}(gly, dur)$, $sl_{12}(gly, dur)$ and $s_1(gly)$ are significant and sufficient evidence to conclude that $s_3(bmi)$ and $s_2(dur)$ are significant in predicting progression of diabetic retinopathy. The most parsimonious model to predict y is

$$\begin{aligned} f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_2(dur) \\ &\quad + s_3(bmi). \end{aligned} \tag{4.21}$$

Wahba et al. (1995) analyzed this dataset and concluded the model to be

$$\begin{aligned} \text{Model 6} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_2(dur) \\ &\quad + s_3(bmi) + sl_{32}(bmi, dur) + ls_{32}(bmi, dur) \\ &\quad + s_{32}(bmi, dur). \end{aligned} \tag{4.22}$$

Table 4.3: The λs for each term in Model 13 giving the desirable effective degrees of freedom.

	$s_1(gly)$	$s_3(bmi)$	$s_2(dur)$
λ	0.01	0.01	0.01
Effective df	3.88	3.66	3.37
	$sl_{12}(gly, dur)$	$ls_{12}(gly, dur)$	$s_{12}(gly, dur)$
λ	0.001	0.001	0.001
Effective df	3.73	5.11	4.41

Table 4.4: The Bayes factors for testing each term in model 13 with scaled χ^2_1 priors.

	$BF_{13,12}$	$BF_{12,11}$	$BF_{11,10}$
scaled χ^2_1	0.450	0.349	0.068
Terms	$s_{12}(gly, dur)$	$ls_{12}(gly, dur)$	$sl_{12}(gly, dur)$
	$BF_{10,9}$	BF_{98}	BF_{97}
scaled χ^2_1	10.3	525	0.145
Terms	$s_2(dur)$	$s_3(bmi)$	$s_1(gly)$

Wahba et al. (1995) excluded $s_1(gly)$ from the model, agreeing with our conclusion based on BF_{97}. However, they claimed that the interaction effects between bmi and dur, $f(bmi, dur) = sl_{32}(bmi, dur) + ls_{32}(bmi, dur) + s_{32}(bmi, dur)$, were not negligible by the evidence of examining the size of the fitted $f(bmi, dur)$ term, along with cross sections of its confidence intervals, suggesting that the components were not negligible in a practical sense.

To verify the significance of $f(bmi, dur)$, another set of partially nested models of interest was proposed based on model 6 in equation (4.22) to test the interaction effects between "dur" and "bmi":

$$\begin{aligned} \text{Model 5} : f(gly, dur, bmi) &= \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_2(dur) \\ &\quad + s_3(bmi) + sl_{32}(bmi, dur) \\ &\quad + ls_{32}(bmi, dur), \end{aligned} \tag{4.23}$$

$$\text{Model 4} : f(gly, dur, bmi) = \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_2(dur) + s_3(bmi) + sl_{32}(bmi, dur), \quad (4.24)$$

$$\text{Model 3} : f(gly, dur, bmi) = \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_2(dur) + s_3(bmi), \quad (4.25)$$

$$\text{Model 2} : f(gly, dur, bmi) = \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_2(dur), \quad (4.26)$$

$$\text{Model 1} : f(gly, dur, bmi) = \mu + \beta_1 gly + \beta_2 dur + \beta_3 bmi + s_3(bmi). \quad (4.27)$$

Scaled χ^2_1 priors were applied to test each smooth effect. The scale parameters for the χ^2_1 priors were selected by effective degrees of freedom. The values of λ corresponding to desirable effective degrees of freedom for each term in model 6 are listed in Table 4.5. Figure 4.5 shows the MCMC trace plots for parameters u_{10}, $log(\lambda_1)$, $log(\lambda_2)$, $log(\lambda_3)$, $log(\lambda_4)$ and $log(\lambda_5)$ under model 6, which have converged in all cases. These were used in the computation of the Bayes factors. The Bayes factors with scaled χ^2_1 priors for model comparison and the specific terms to be tested are listed in Table 4.6.

Based on Table 4.6, there is insufficient evidence to conclude that $sl_{32}(bmi, dur)$, $ls_{32}(bmi, dur)$ and $s_{32}(bmi, dur)$ are significant and sufficient evidence to conclude that $s_3(bmi)$ and $s_2(dur)$ are significant in predicting progression of diabetic retinopathy. Figure 4.6 shows the fits of $s_3(bmi)$ and $s_2(dur)$ with 95% credible sets in Model 3 and the fit by model 3, which helps us to visualize the effects. Based on the definition of Bayes factor, the statistic to test $f(bmi, dur)$ is $BF_{63} = BF_{65} \times BF_{54} \times BF_{43} = 0.011$. This

is fairly strong evidence that the effect of $f(bmi, dur)$ is insignificant, which disagrees with the conclusion by Wahba et al. (1995). Thus we conclude the most parsimonious model is model 3. This fully Baysesian SSANOVA method has provided a more powerful tool for hypothesis testing in the interaction effects than the frequentist approach.

Table 4.5: The λs for each term in Model 6 giving the desirable effective degrees of freedom.

	$s_2(dur)$	$s_3(bmi)$	$sl_{32}(bmi, dur)$
λ	0.01	0.01	0.001
Effective *df*	3.66	3.37	3.20
	$ls_{32}(bmi, dur)$	$s_{32}(bmi, dur)$	
λ	0.001	0.0005	
Effective *df*	4.07	3.87	

Table 4.6: The Bayes factors for testing each term in model 6 with scaled χ^2_1 priors.

	BF_{65}	BF_{54}	BF_{43}
scaled χ^2_1	0.609	0.703	0.340
Terms	$s_{32}(bmi, dur)$	$ls_{32}(bmi, dur)$	$sl_{32}(bmi, dur)$
	BF_{32}	BF_{31}	
scaled χ^2_1	10.1	48.9	
Terms	$s_3(bmi)$	$s_2(dur)$	

Chapter 5

Comments and Future Work

Previous research in smoothing spline ANOVA models has focused on statistical inference in estimation and prediction while the needs of hypothesis testing for model selection have emerged. This study adapts a Bayesian approach to smoothing spline ANOVA models. These fully Bayesian smoothing spline ANOVA models provide flexibility for hypothesis testing and better statistical inference on parameters of interest since the posterior distribution for those parameters are available.

Smoothing spline ANOVA models have a tensor sum decomposition of inner product spaces to ensure that the estimated functions are from the orthogonal subspaces. This nice property facilitates Bayesian computation with better mixing in the MCMC steps, which provided more efficient estimation. However, computation in fully Bayesian smoothing spline ANOVA models is still intensive. To promote fully Bayesian SSANOVA models, we would like to improve the computation performance. Nychka (2000) suggests selecting a subset of the full bases derived from locations that are widely separated, which will give a good approximation to the full expansion. This is an opportunity for us to cut down the computation but still obtain a good approximation.

An alternative to model selection in some contexts is Bayesian model averaging (see, e.g. Clyde (1999)). One advantage is that Bayesian model averaging can be applied with improper priors, so it's not necessary to use different priors for testing and estimation. We plan to explore model averaging methods in Bayesian SSANOVA models.

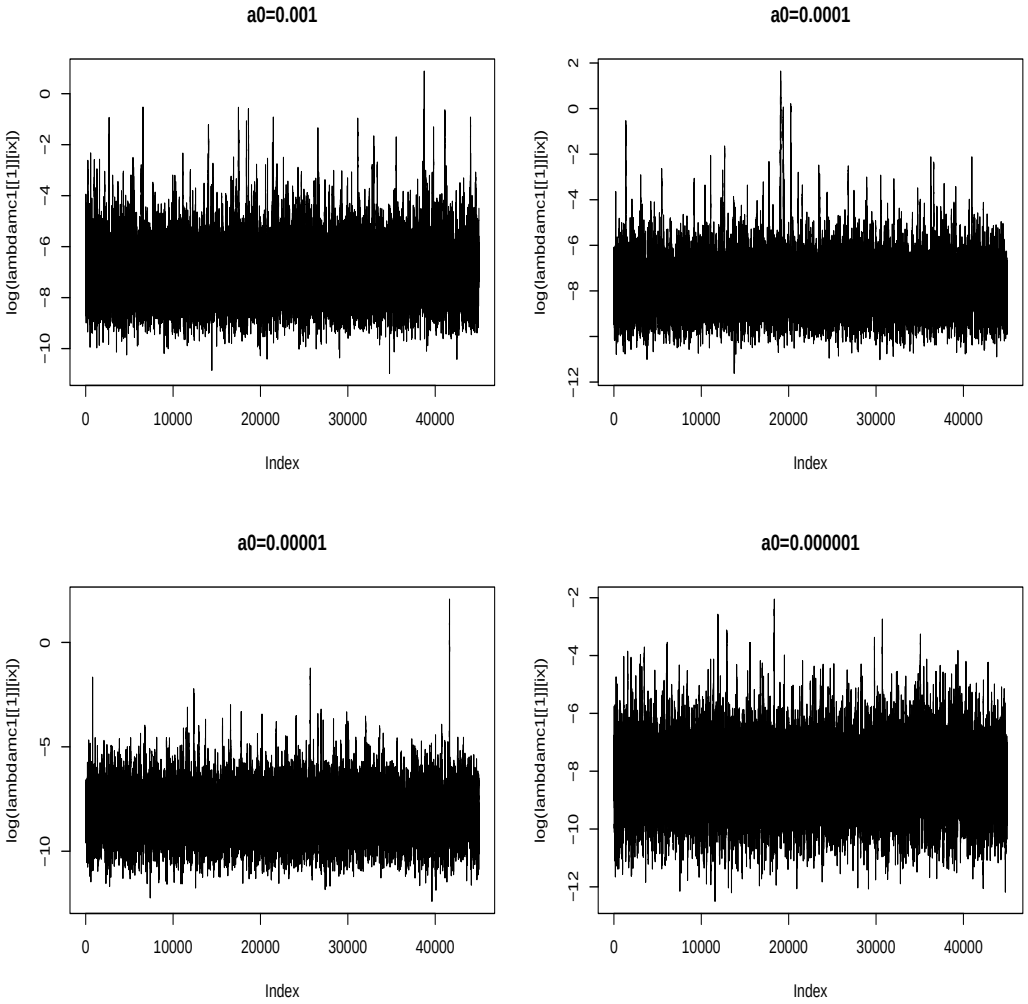

Figure 3.1: MCMC trace plots for samples of the $\log(\lambda)$ from the models with a_0=0.001, a_0=0.0001, a_0=0.00001 and a_0=0.000001 for the simulated example in Section 3.1.3. This is based on 50,000 iterations with 5,000 iterations for burnin.

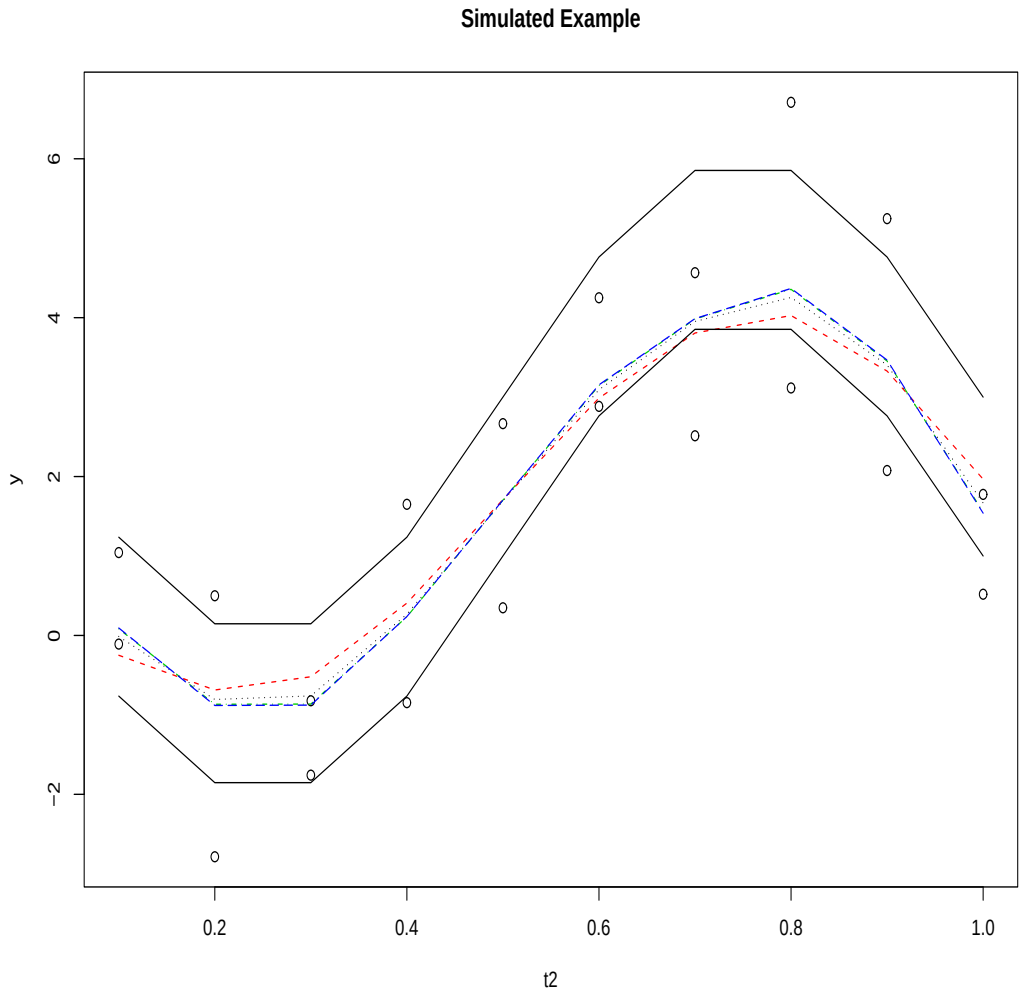

Figure 3.2: The estimated $f(x_2)$ of the models with a_0=0.001 (dashed line), a_0=0.0001 (dotted line), a_0=0.00001 (dotdash line) and a_0=0.000001 (longdash line). The fits for a_0=0.00001 and a_0=0.000001 are overlapped. The solid line represents the true function. This is for the simulated example in Section 3.1.3 and based on 50,000 iterations with 5,000 iterations for burnin.

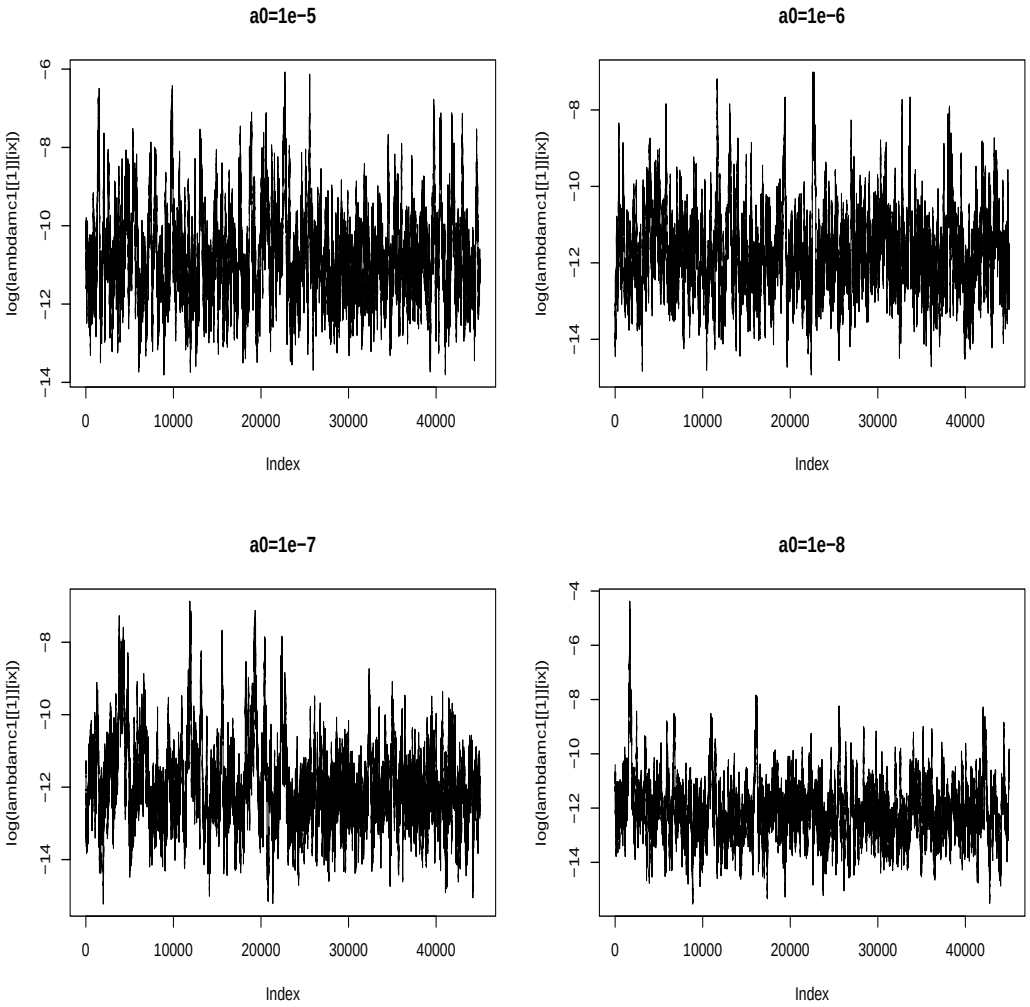

Figure 3.3: MCMC trace plots for samples of the $\log(\lambda)$ from the models with $a_0 = 1 \times 10^{-5}$, $a_0 = 1 \times 10^{-6}$, $a_0 = 1 \times 10^{-7}$ and $a_0 = 1 \times 10^{-8}$ for the manufacturing example in Section 3.1.4. This is based on 50,000 iterations with 5,000 iterations for burnin.

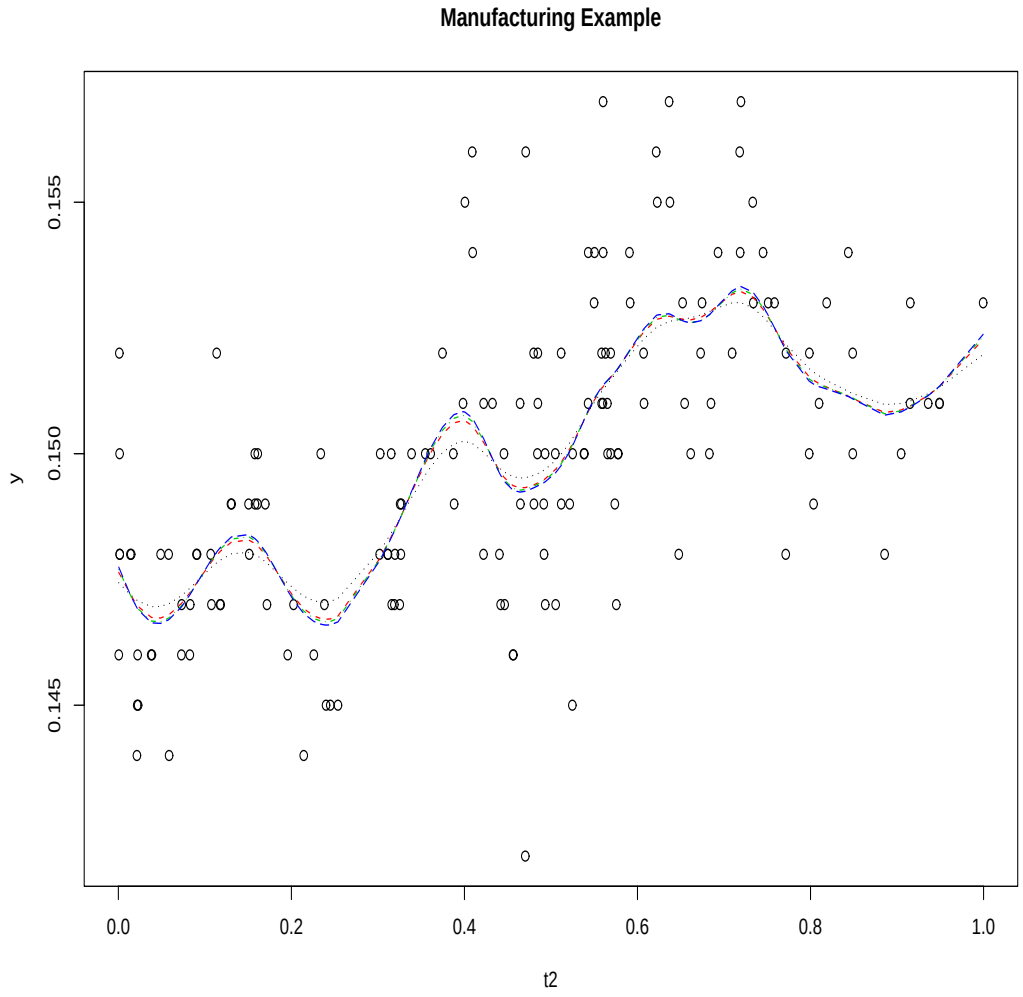

Figure 3.4: The estimated $f(x_2)$ of the models with $a_0 = 1 \times 10^{-5}$ (dotted line), $a_0 = 1 \times 10^{-6}$ (dashed line), $a_0 = 1 \times 10^{-7}$ (dotdash line) and $a_0 = 1 \times 10^{-8}$ (longdash line). The fits for $a_0 = 1 \times 10^{-6}$, $a_0 = 1 \times 10^{-7}$ and $a_0 = 1 \times 10^{-8}$ are almost overlapped. This is for the manufacturing example in Section 3.1.4 and based on 50,000 iterations with 5,000 iterations for burnin.

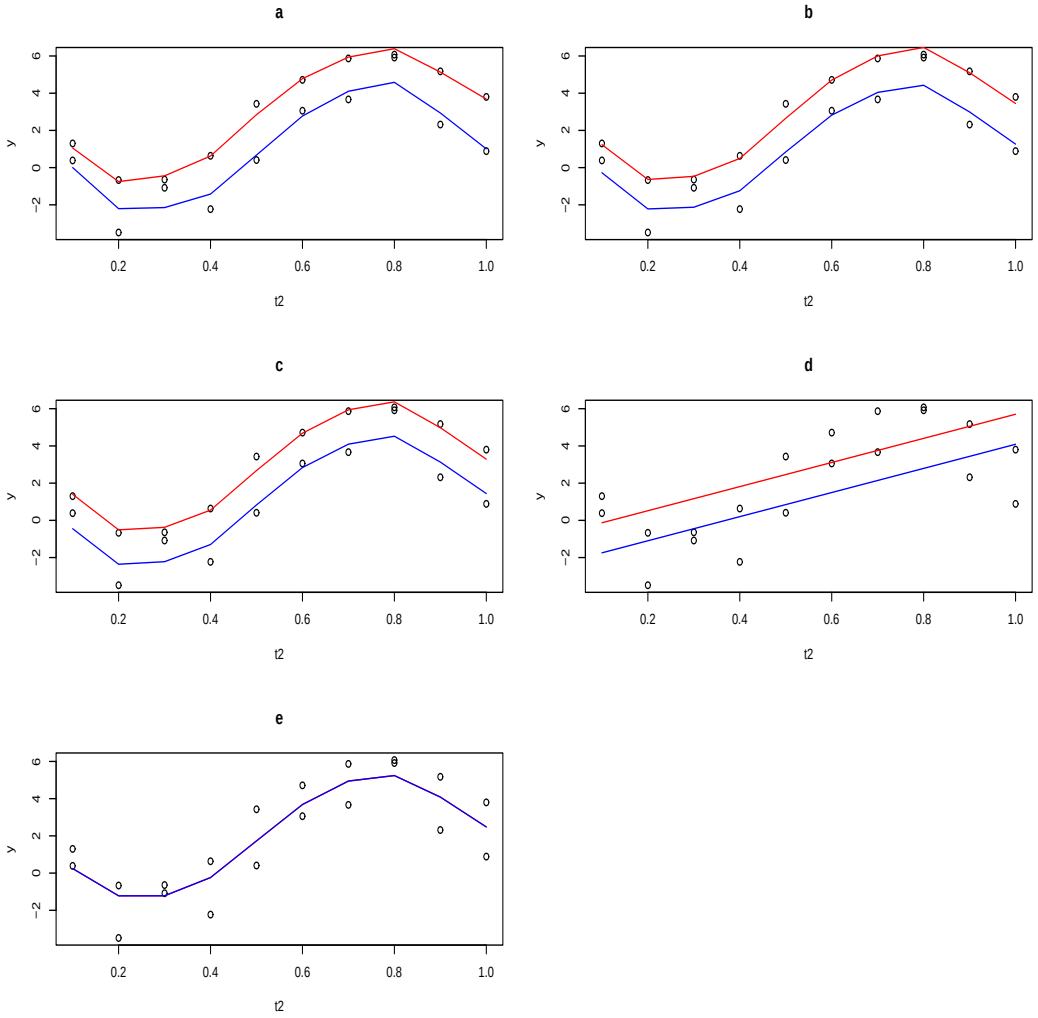

Figure 3.5: The estimated f of the model 0, model 1, model 2, model 3 and model 4 (corresponds to panels a, b, c, d and e) for the simulated example in Section 3.2.5 when a_0=0.001. This is based on 30,000 iterations with 5,000 iterations for burnin.

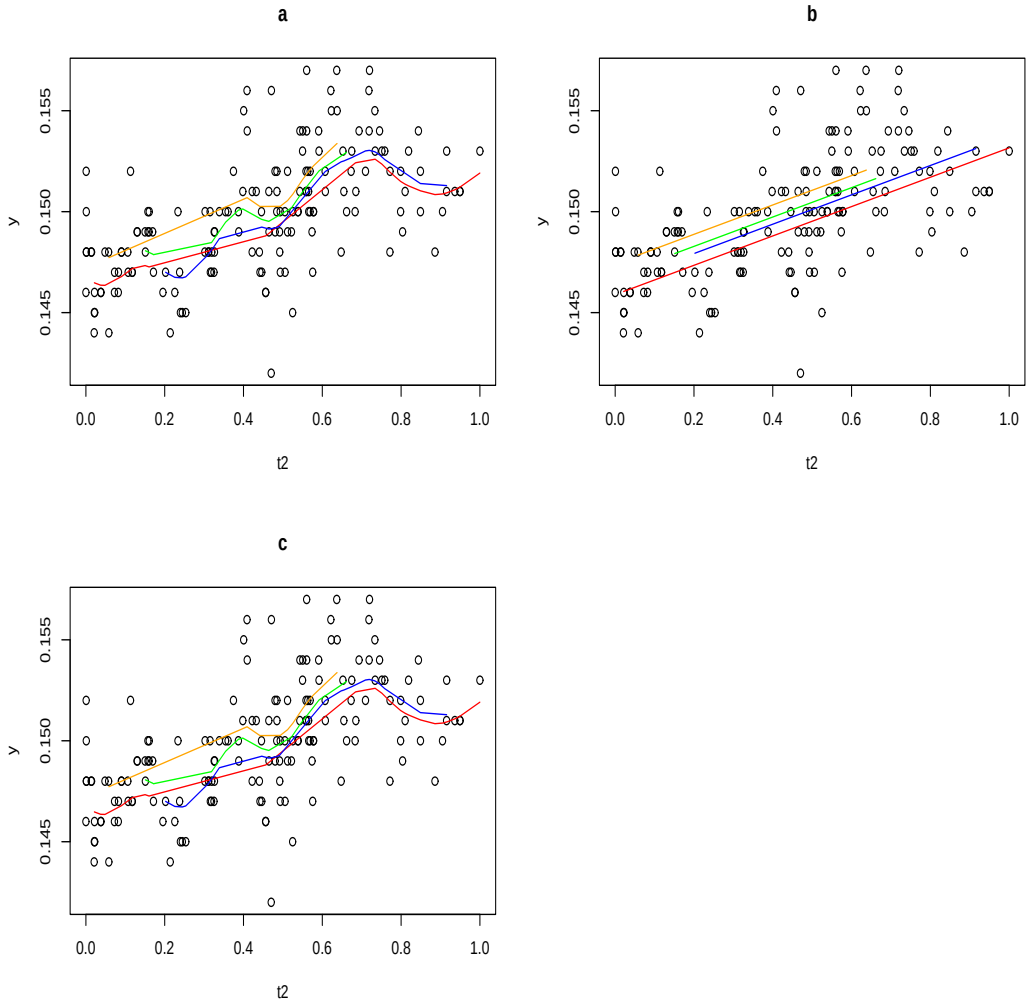

Figure 3.6: The estimated f of the model 2, model 3 and model 4 (corresponds to panels a, b and c) for the manufacturing example in Section 3.2.6 when a_0=0.001. This is based on 30,000 iterations with 5,000 iterations for burnin.

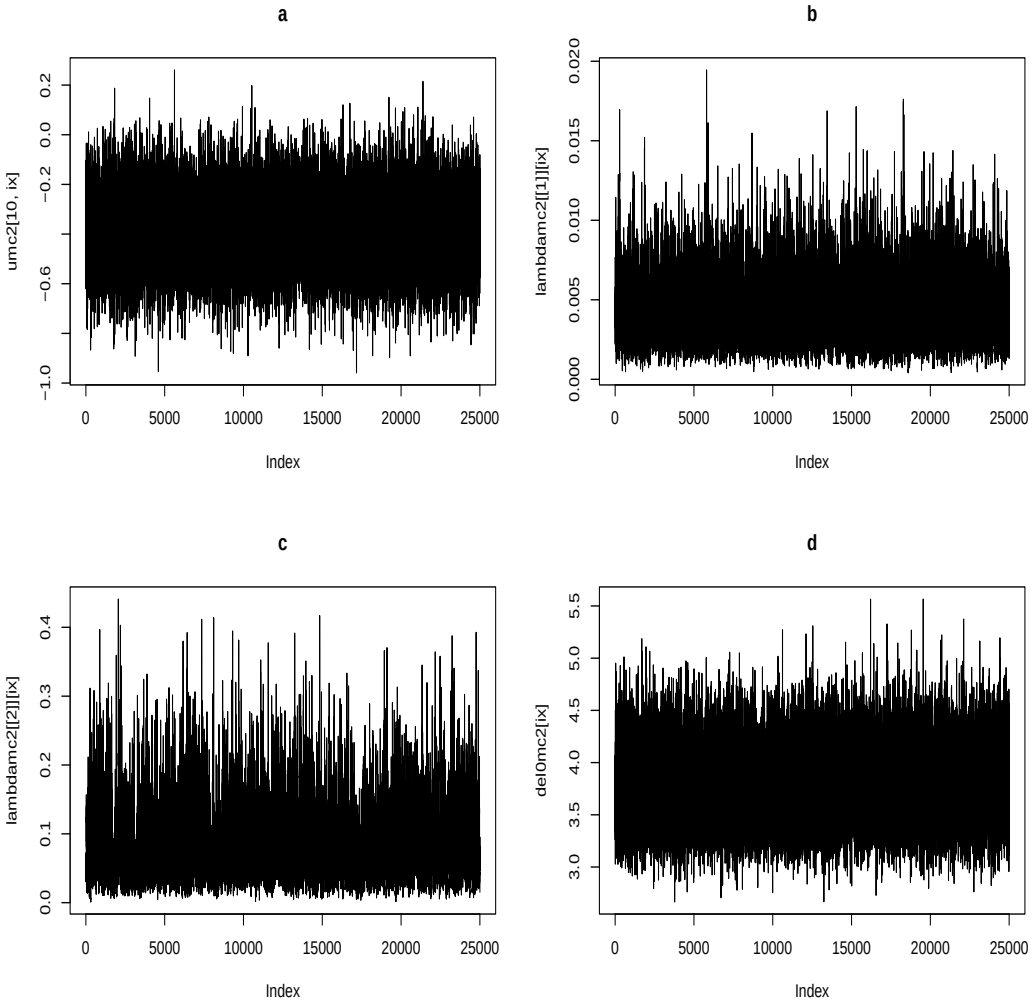

Figure 3.7: MCMC trace plots for samples of the u_{10}, the 10th component of $\boldsymbol{u}$, λ_1, λ_2 and δ_0 (corresponds to panels a, b, c and d) under model 2 in equation (3.70) for the Example 3.2 in Section 3.2.7. This is based on 30,000 iterations with 5,000 iterations for burnin.

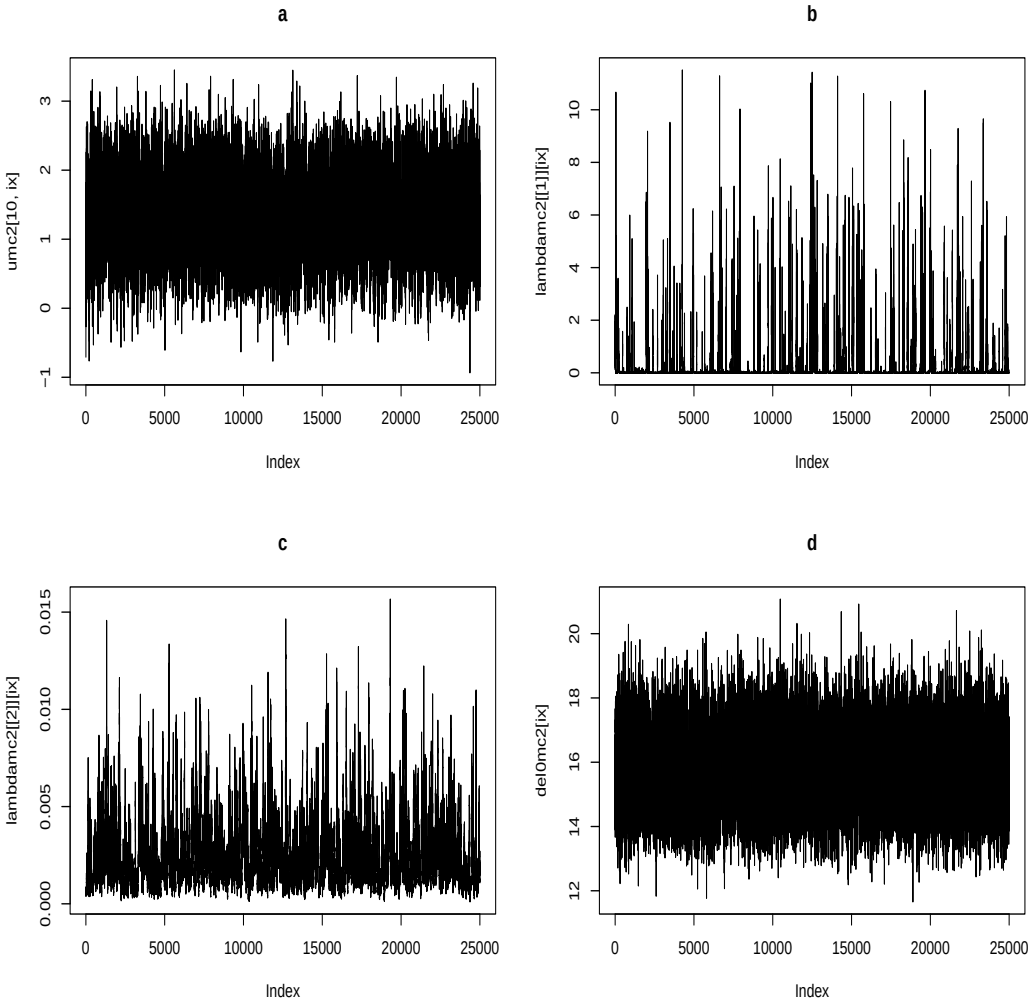

Figure 3.8: MCMC trace plots for samples of the u_{10}, the 10th component of $\boldsymbol{u}$, λ_1, λ_2 and δ_0 (corresponds to panels a, b, c and d) under model 2 in equation (3.71) for the Example 3.2 in Section 3.2.7. This is based on 30,000 iterations with 5,000 iterations for burnin.

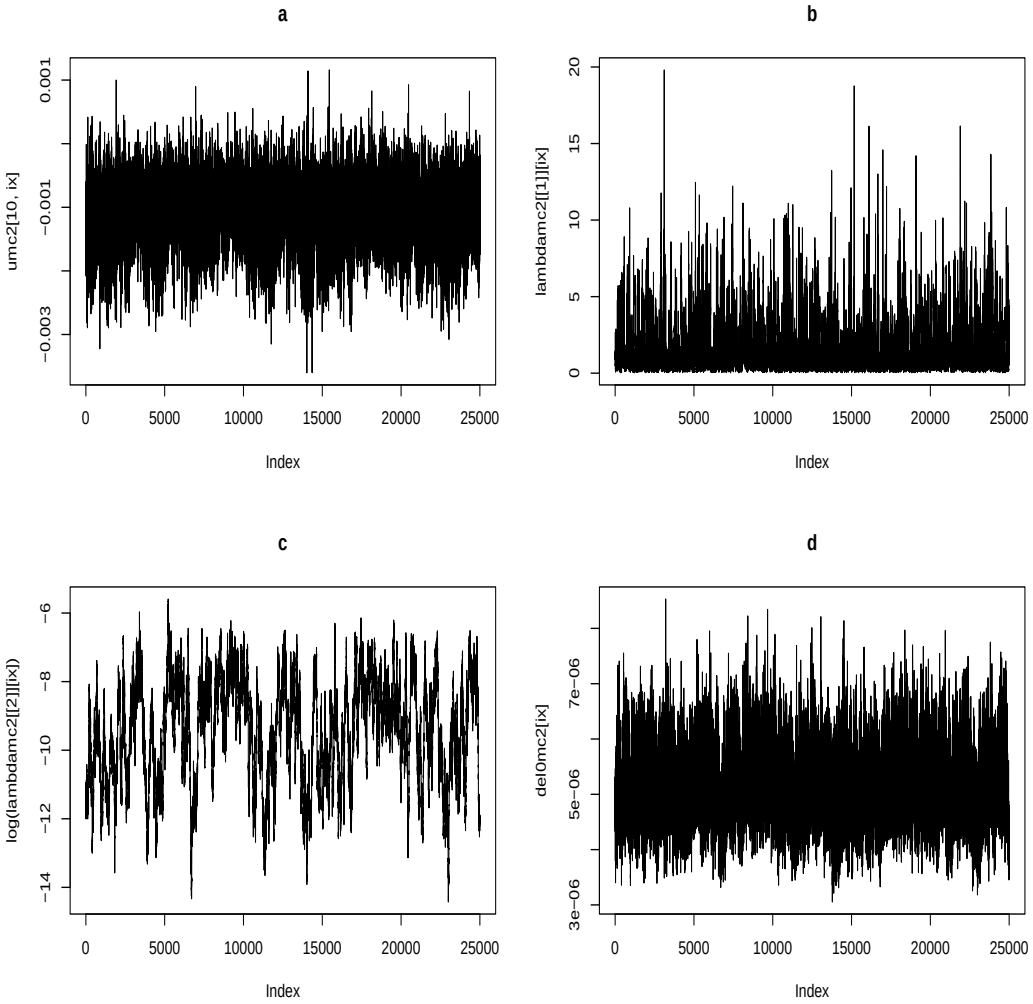

Figure 3.9: MCMC trace plots for samples of the u_{10}, the 10th component of $\boldsymbol{u}$, λ_1, λ_2 and δ_0 (corresponds to panels a, b, c and d) under model 2 for manufacturing example in Section 3.2.7. This is based on 30,000 iterations with 5,000 iterations for burnin.

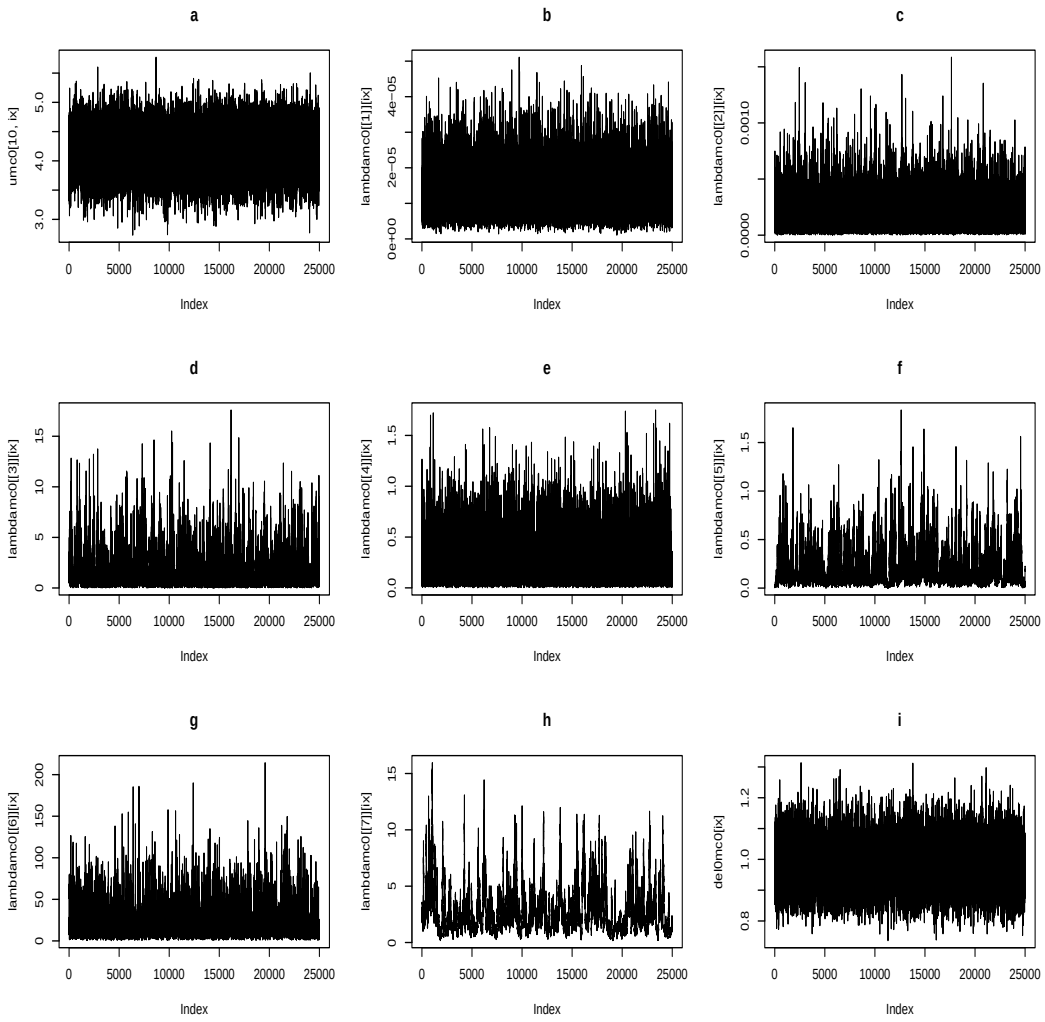

Figure 3.10: MCMC trace plots for samples of the u_{10}, the 10th component of $\boldsymbol{u}$, λ_1, λ_2, λ_3, λ_4, λ_5, λ_6, λ_7 and δ_0 (corresponds to panels a, b, c, d, e, f, g, h and i) under model 7 in equation (3.73) with scaled χ^2_1 priors for the Simulated Example 1 in Section 3.2.8. This is based on 30,000 iterations with 5,000 iterations for burnin.

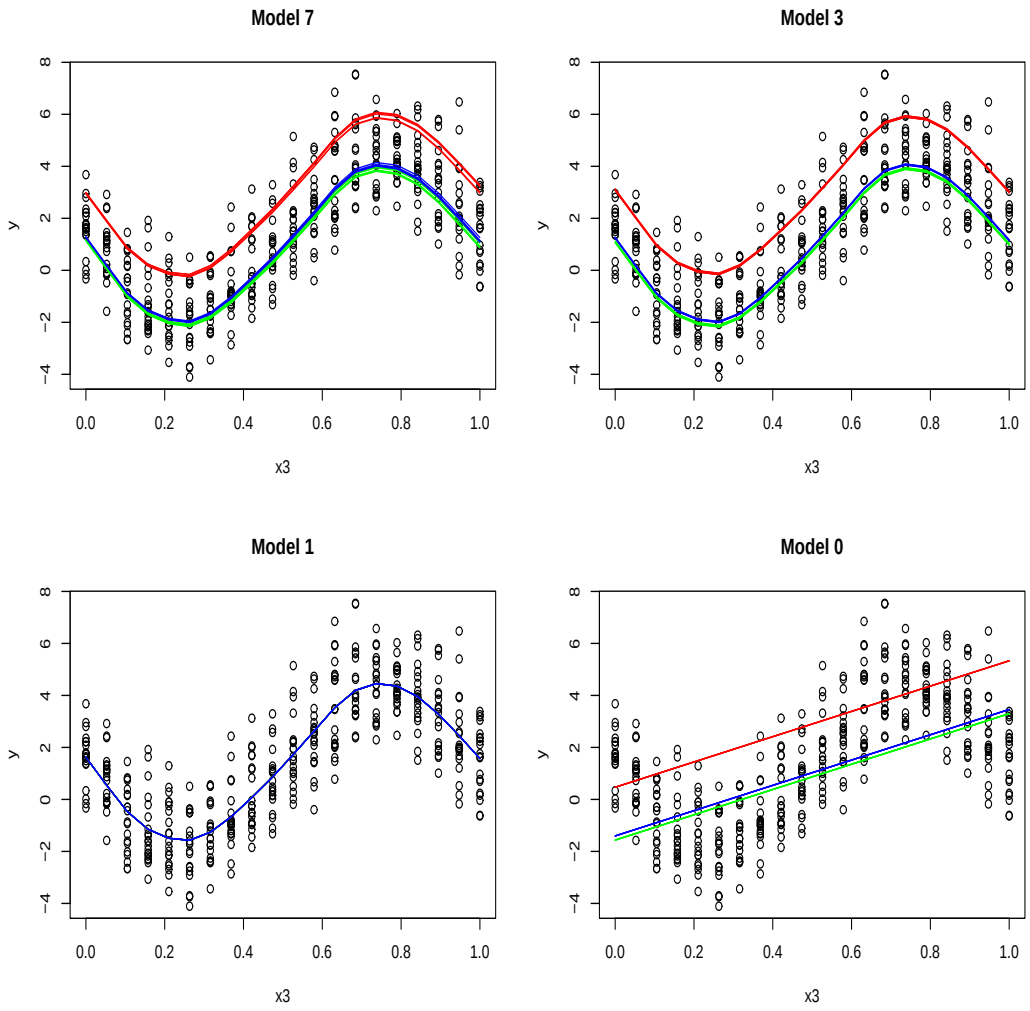

Figure 3.11: The estimated f of the model 7, model 3, model 1 and model 0 for the Simulated Example 1 in Section 3.2.8. This is based on 30,000 iterations with 5,000 iterations for burnin.

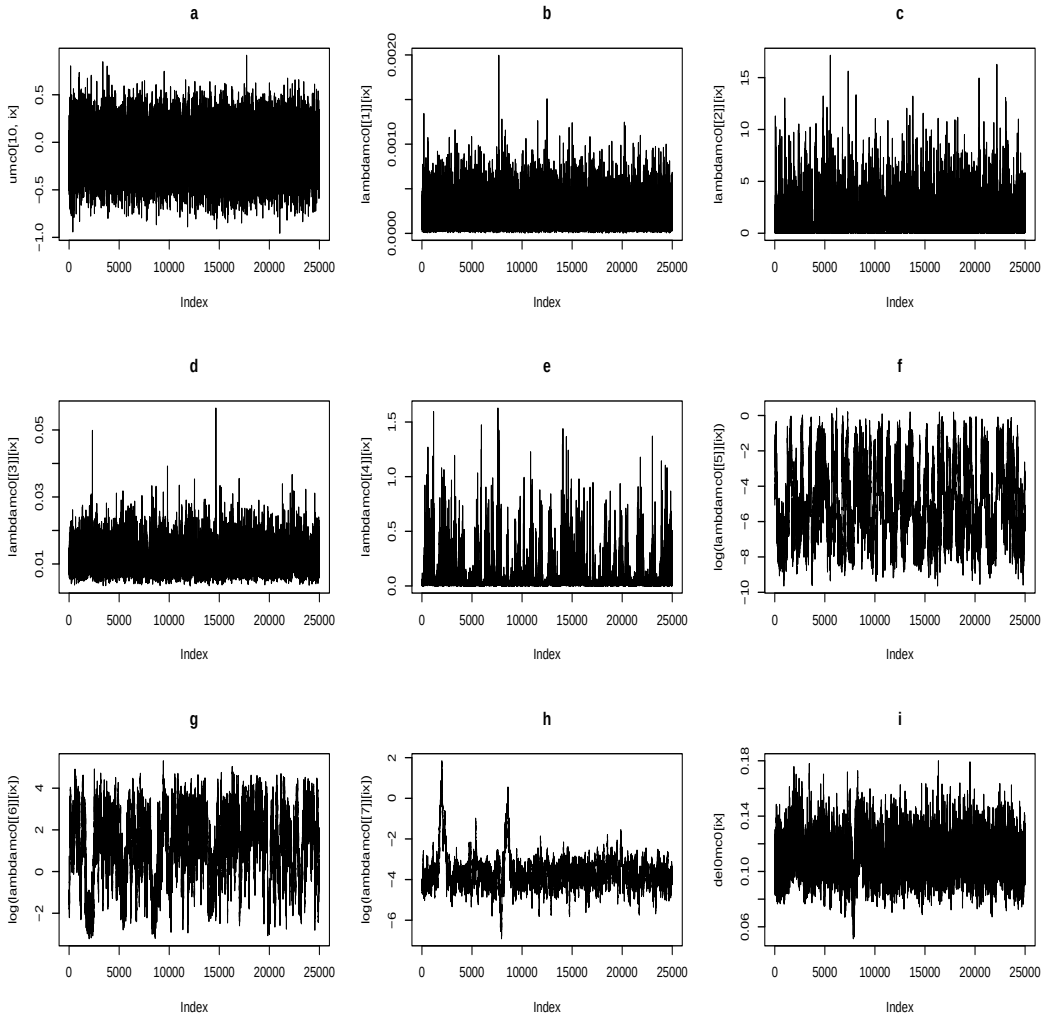

Figure 3.12: MCMC trace plots for samples of the u_{10}, the 10th component of $\boldsymbol{u}$, λ_1, λ_2, λ_3, λ_4, $log(\lambda_5)$, $log(\lambda_6)$, $log(\lambda_7)$ and δ_0 (corresponds to panels a, b, c, d, e, f, g, h and i) under model 7 with scaled χ^2_1 priors for the Dogs Example in Section 3.2.9. This is based on 30,000 iterations with 5,000 iterations for burnin.

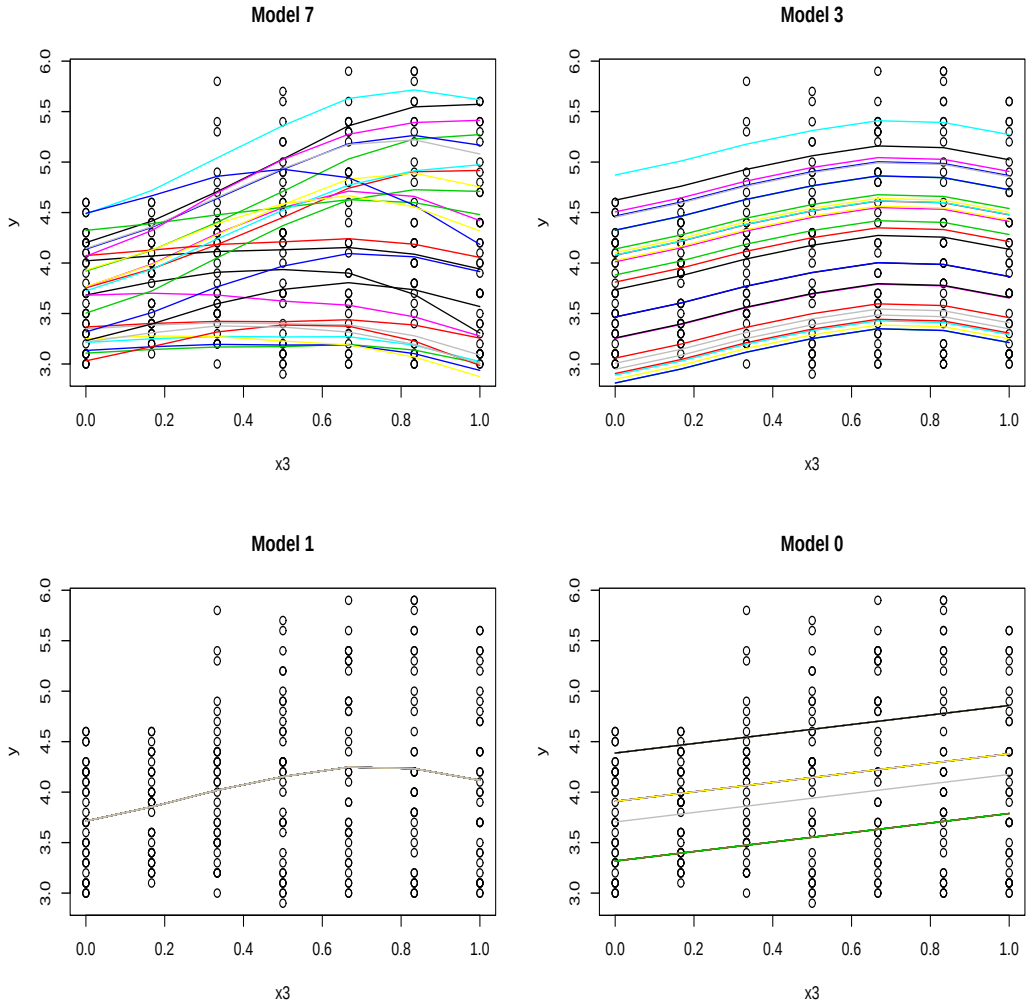

Figure 3.13: The estimated f of the model 7, model 3, model 1 and model 0 for the Dogs Example in Section 3.2.9. This is based on 30,000 iterations with 5,000 iterations for burnin.

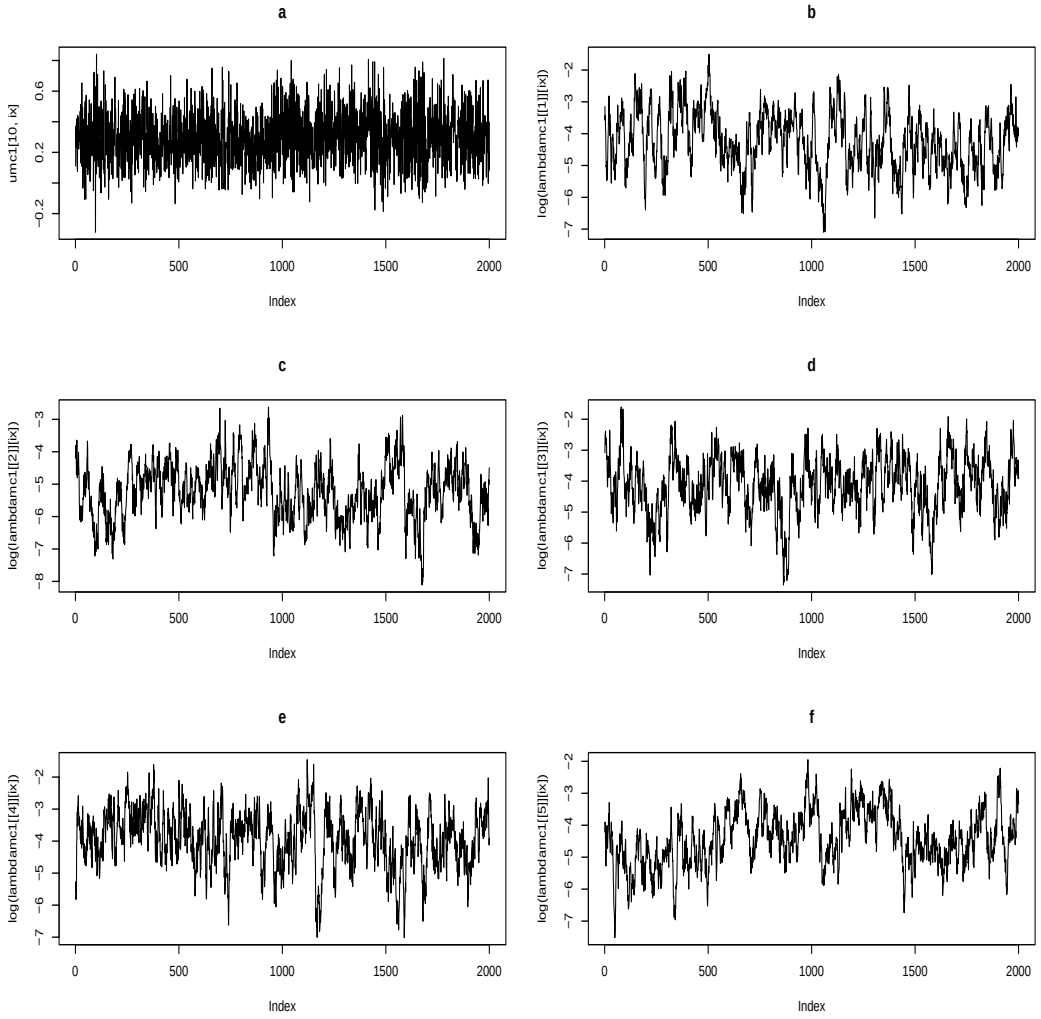

Figure 4.1: MCMC trace plots for samples of the u_{10}, the 10th component of $\boldsymbol{u}$, $log(\lambda_1)$, $log(\lambda_2)$, $log(\lambda_3)$, $log(\lambda_4)$ and $log(\lambda_5)$ (corresponds to panels a, b, c, d, e and f) under model 6 in equation (4.8) with scaled χ^2_1 priors for the Simulated Example 2 in Section 4.1.1. Those samples are the every 10th of the MCMC samples from the 30,000 iterations after 10,000 iterations for burnin.

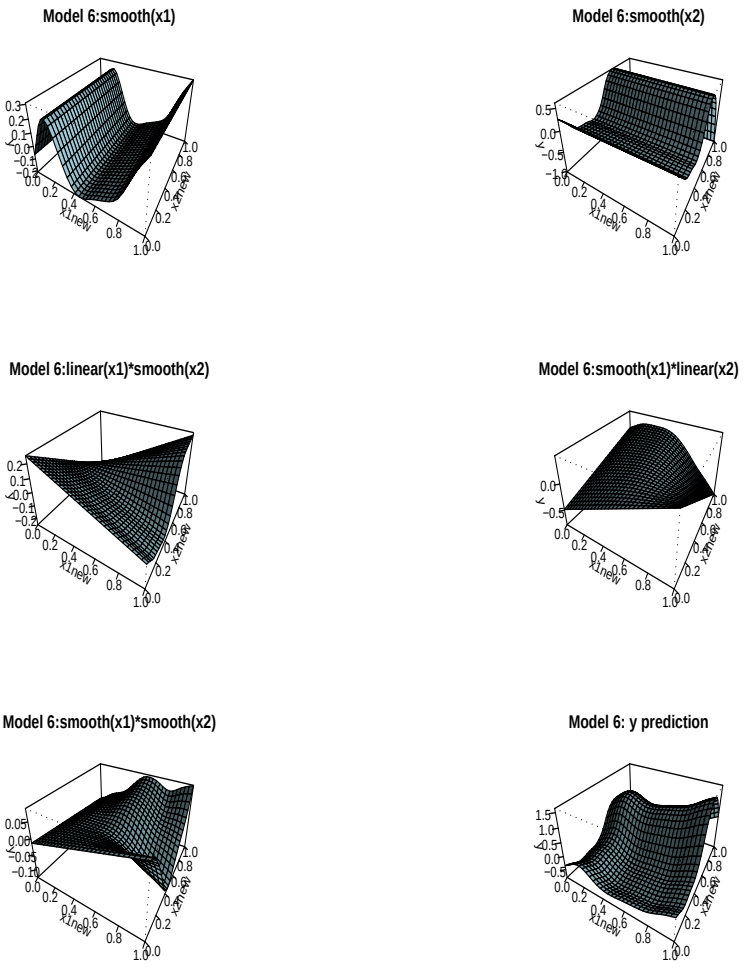

Figure 4.2: The estimate of each component in Model 6, $s_1(x_1)$, $s_2(x_2)$, $ls_{12}(x_1, x_2)$, $sl_{12}(x_1, x_2)$, $s_{12}(x_1, x_2)$ and the fit by model 6 in equation (4.8) with scaled χ^2_1 priors for the Simulated Example 2 in Section 4.1.1. Those samples are the every 10th of the MCMC samples from the 30,000 iterations after 10,000 iterations for burnin.

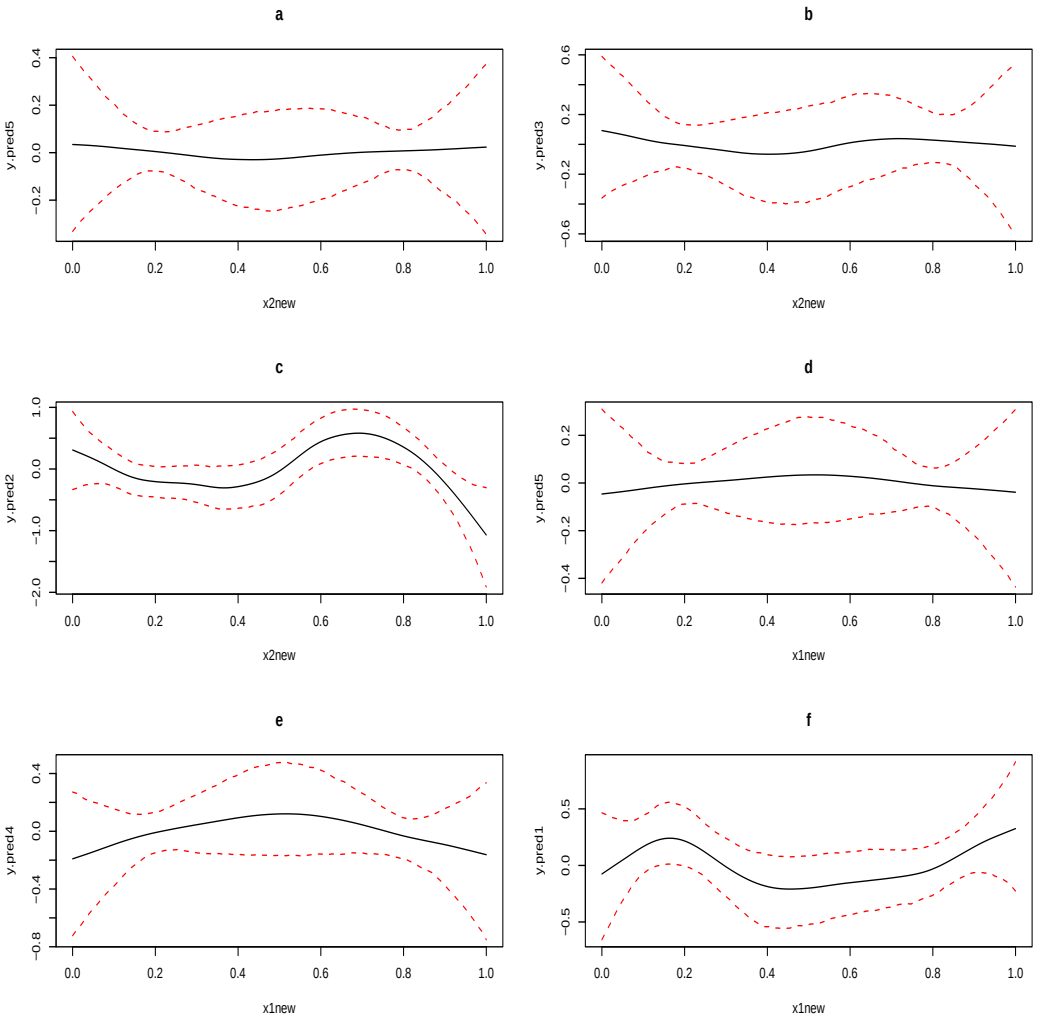

Figure 4.3: The estimates of $s_{12}(0.5, x_2)$, $ls_{12}(0.5, x_2)$, $s_2(x_2)$, $s_{12}(x_1, 0.5)$, $sl_{12}(x_1, 0.5)$ and $s_1(x_1)$ with 95% credible sets in Model 6 with scaled χ^2_1 priors for the Simulated Example 2 in Section 4.1.1(corresponds to panels a, b, c, d, e and f). Those samples are the every 10th of the MCMC samples from the 30,000 iterations after 10,000 iterations for burnin.

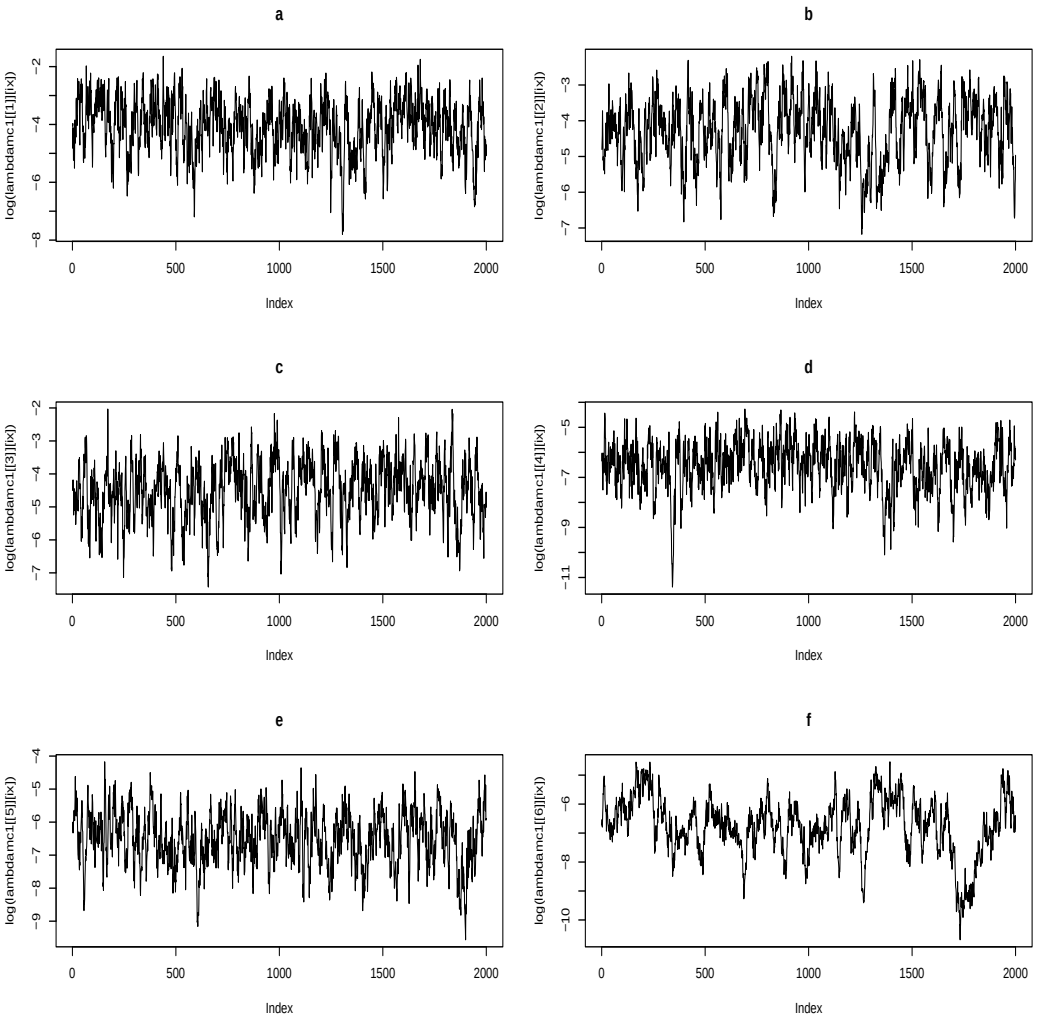

Figure 4.4: MCMC trace plots for samples of the $log(\lambda_1)$, $log(\lambda_2)$, $log(\lambda_3)$, $log(\lambda_4)$, $log(\lambda_5)$ and $log(\lambda_6)$ (corresponds to panels a, b, c, d, e and f) under model 13 in equation (4.14) with scaled χ^2_1 priors for the Diabetic Retinopathy Example in Section 4.1.2. Those samples are the every 10th of the MCMC samples from the 30,000 iterations after 10,000 iterations for burnin.

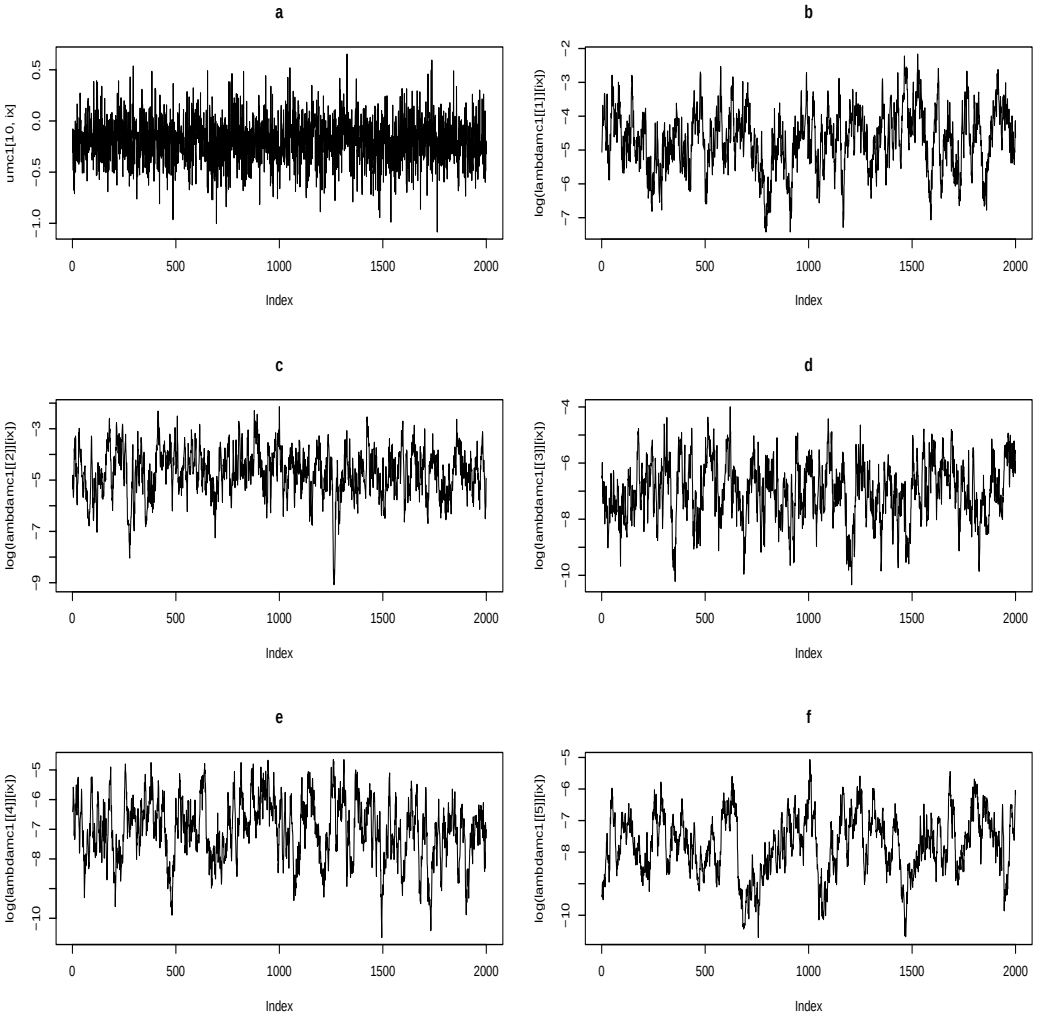

Figure 4.5: MCMC trace plots for samples of the u_{10}, the 10th component of $\boldsymbol{u}$, $log(\lambda_1)$, $log(\lambda_2)$, $log(\lambda_3)$, $log(\lambda_4)$, and $log(\lambda_5)$ (corresponds to panels a, b, c, d, e and f) under model 6 in equation (4.22) with scaled χ^2_1 priors for the Diabetic Retinopathy Example in Section 4.1.2. Those samples are the every 10th of the MCMC samples from the 30,000 iterations after 10,000 iterations for burnin.

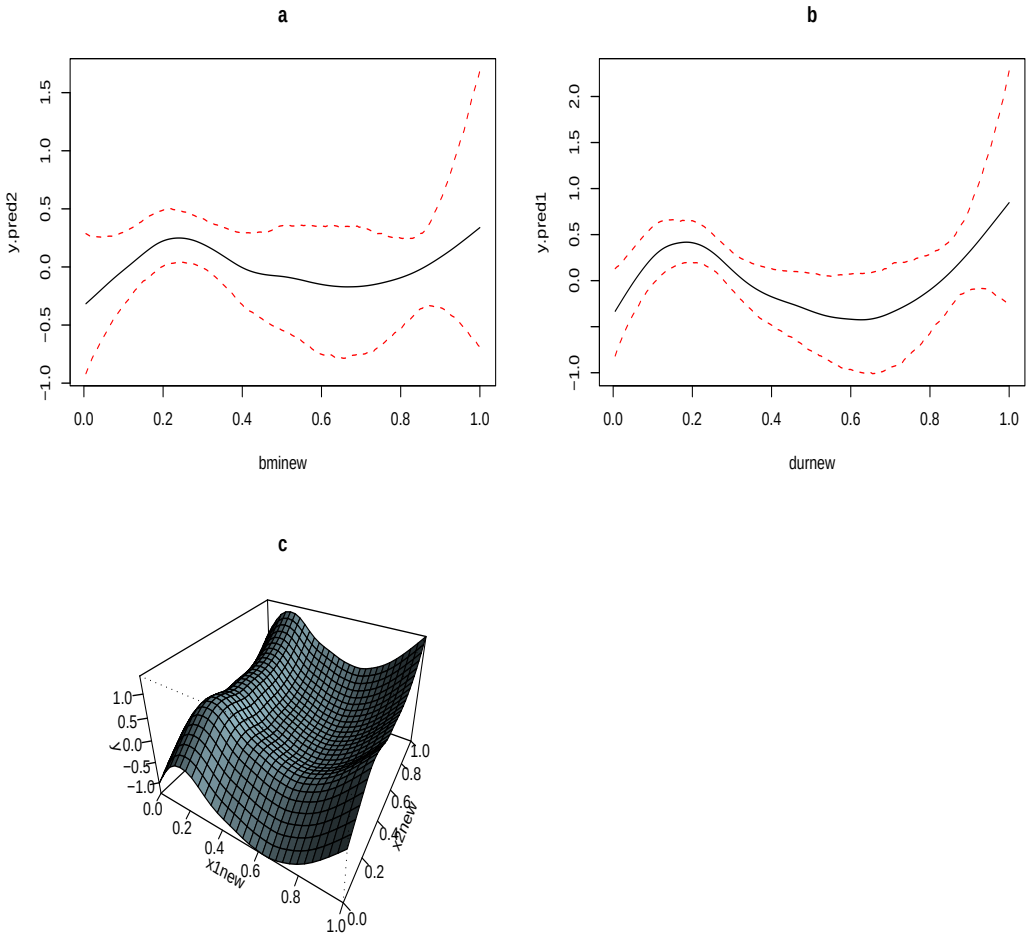

Figure 4.6: The estimates of $s_3(bmi)$ and $s_2(dur)$ with 95% credible sets in Model 3 and the fit by model 3 in equation (4.25)(corresponds to panels a, b and c) with scaled χ^2_1 priors for the Diabetic Retinopathy Example in Section 4.1.2. Those samples are the every 10th of the MCMC samples from the 30,000 iterations after 10,000 iterations for burnin.

Bibliography

Albert, J. and S. Chib (1997). Bayesian tests and model diagnostics in conditionally independent hierarchical models. *Journal of the American Statistical Association 92*, 916–925.

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

Bayarri, M. and G. Garcia-Donato (2007). Extending conventional priors for testing general hypotheses in linear models. *Biometrika*, asm014.

Breiman, L., J. H. Friedman, O. R.A., and C. J. Stone (1984). *Classification and regression trees*. Wadsworth, Belmont, CA.

Buja, A., T. Hastie, and R. Tibshirani (1989). Linear smoothers and additive models (C/R: P510-555). *The Annals of Statistics 17*, 453–510.

Clyde, M. A. (1999). Bayesian model averaging and model search strategies. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. Smith (Eds.), *Bayesian Statistics 6 – Proceedings of the Sixth Valencia International Meeting*, pp. 157–185. Clarendon Press [Oxford University Press].

Devroye, L. (1986). *Non-uniform random variate generation*. Springer-Verlag Inc.

Dey, D. e., P. e. Muller, P. e. Mueller, P. e. Müller, and D. e. Sinha (1998). *Practical Nonparametric and Semiparametric Bayesian Statistics*. Springer-Verlag Inc.

Eubank, R. L. (1988). *Spline Smoothing and Nonparametric Regression*. Marcel Dekker Inc.

Friedman, J. H. (1991). Reply to comments on "Multivariate adaptive regression splines". *The Annals of Statistics 19*, 123–141.

Friedman, J. H. and B. W. Silverman (1989). Flexible parsimonious smoothing and additive modeling (C/R: P23-39). *Technometrics 31*, 3–21.

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

Gu, C. (1992). Diagnostics for nonparametric regression models with additive terms. *Journal of the American Statistical Association 87*, 1051–1058.

Gu, C. (2002). *Smoothing Spline ANOVA Models*. Springer-Verlag Inc.

Hastie, T. (1989). Comments on "Flexible parsimonious smoothing and additive modeling". *Technometrics 31*, 23–29.

Hastie, T. and R. Tibshirani (1999). *Generalized Additive Models*. Chapman & Hall Ltd.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction: with 200 Full-color Illustrations*. Springer-Verlag Inc.

Horn, R. A. and C. R. Johnson (1985). *Matrix Analysis*. Cambridge University Press.

Johnson, R. A. and D. W. Wichern (1998). *Applied Multivariate Statistical Analysis*. Prentice-Hall Inc.

Karcher, P. and Y. Wang (2001). Generalized nonparametric mixed effects models. *Journal of Computational and Graphical Statistics 10*(4), 641–655.

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

Kimeldorf, G. S. and G. Wahba (1971). Some results on tchebycheffian spline functions. *Journal of Mathematical Analysis and Application 33*, 82–94.

Klein, R., K. B. E. K. M. S. E. D. M. D. and D. L. DeMets (1988). Glycosylated hemoglobin predicts the incidence and progression of diabetic retinopathy. *Journal of the American Medical Association 260*, 2864–2871.

Kleinman, K. P. and J. G. Ibrahim (1998). A semi-parametric Bayesian approach to generalized linear mixed models. *Statistics in Medicine 17*, 2579–2596.

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g priors for bayesian variable selection. *Journal of the American Statistical Association 103*(481), 410–423.

Luo, Z. (1998). Backfitting in smoothing spline ANOVA. *The Annals of Statistics 26*(5), 1733–1759.

Luo, Z. and G. Wahba (1997). Hybrid adaptive splines. *Journal of the American Statistical Association 92*, 107–116.

Meng, X.-L. and W. H. Wong (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica 6*, 831–860.

Murray, F. J. and J. Von Neumann (1936). On rings of operators. *The Annals of Mathematics 37*, 116–229.

Nychka, D. W. (2000). Spatial-process estimates as smoothers. In M. G. a. Schimek (Ed.), *Smoothing and regression: approaches, computation, and application*, pp. 393–424. John Wiley & Sons.

Ruppert, D., M. Wand, and R. Carroll (2003). *Semiparametric Regression*. Cambridge University Press, Cambridge.

Smith, M. and R. Kohn (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics 75*, 317–343.

Smith, M., C.-M. Wong, and R. Kohn (1998). Additive nonparametric regression with autocorrelated errors. *Journal of the Royal Statistical Society 60*, 311–331.

Speckman, P. L. and D. Sun (2003). Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika 90*(2), 289–302.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics 13*, 689–705.

Sun, D., R. K. Tsutakawa, and Z. He (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica 11*(1), 77–95.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM [Society for Industrial and Applied Mathematics].

Wahba, G., Y. Wang, C. Gu, R. Klein, and B. Klein (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy : the 1994 Neyman memorial lecture. *The Annals of Statistics 23*(6), 1865–1895.

Wang, Y. (1997). GRKPACK: Fitting smoothing spline ANOVA models for exponential families. *Communications in Statistics: Simulation and Computation 26*, 765–782.

Wang, Y. and C. Ke (2004). Assist: A suite of s functions implementing spline smoothing techniques.

White, G. A. (2006). Bayesian semiparametric spatial and joint spatial temporal smoothing. Ph.D. dissertation, University of Missouri Columbia, Department of Statistics.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de*

Finetti, pp. 233–243. Elsevier/North-Holland [Elsevier Science Publishing Co., New York; North-Holland Publishing Co., Amsterdam].

Zellner, A. and A. Siow (1980). Posterior odds ratios for selected regression hypotheses. In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith (Eds.), *Bayesian Statistics: Proceedings of the First International Meeting held in Valencia (Spain)*, pp. 585–603. University of Valencia.

Zhang, H. H. and Y. Lin (2006). Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica 16*(3), 1021–1041.

Zhang, H. H., G. Wahba, Y. Lin, M. Voelker, M. Ferris, R. Klein, and B. Klein (2004). Variable selection and model building via likelihood basis pursuit. *Journal of the American Statistical Association 99*(467), 659–672.

VITA

Chin-I Cheng was born in Taoyuan, Taiwan R.O.C. She received her Bachelor degree of Science in Applied Mathematics from the Chung Yuan Christian University, Taiwan R.O.C. in 1994. In 1997 she earned her Master degree of Science in Statistics from the University of Arkansas in Fayetteville, AR. After gaining few years' work experience, she came to the United States again in 2004 for pursuing her doctoral degree. She is expecting to receive her Doctoral degree of Philosophy in Statistics from the University of Missouri-Columbia, Columbia, MO in July 2009. She will join the Department of Mathematics at Central Michigan University, Mt. Pleasant, MI as an assistant professor in August 2009. Her research interests are in the area of Bayesian statistics, nonparametric and semiparametric models, Bayesian smoothing spline ANOVA. (Last updated: July 2009).