DEPARTMENT OF STATISTICS
University of Wisconsin
1210 West Dayton St.
Madison, WI 53706

TECHNICAL REPORT NO. 928
July 21, 1994

Smoothing Spline Analysis of Variance of Data From
Exponential Families

by
Yuedong Wang

\[1\text{Supported by the National Science Foundation under Grant DMS-9121003 and the National Eye Institute under Grant R01 EY09946.}\]
SMOOTHING SPLINE ANALYSIS OF VARIANCE OF DATA FROM EXPONENTIAL FAMILIES

by
Yuedong Wang

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

Doctor of Philosophy
(Statistics)

at the
UNIVERSITY OF WISCONSIN-MADISON
1994
Acknowledgements

I would like to express my sincerest gratitude to my advisor, Professor Grace Wahba, for her invaluable advice and continuing encouragement during my years as a graduate student.

Appreciation is extended to Professors Douglas Bates, Rick Chappell, Wei-Yin Loh and Brian Yandell for their service on my final examination committee, especially to Professors Douglas Bates and Brian Yandell for their careful reading of this thesis, their valuable comments and many discussions. I would like to thank Professor Douglas Bates and Professor Wei-Yin Loh for their invaluable work in setting up and maintaining the computing resources. I would like to thank Professor Chong Gu at the Department of Statistics of Purdue University for his computer programs, advice and discussions. I would like to thank Ronald Klein, MD and Barbara Klein, MD for providing the WESDR data.

Finally, special thanks go to my parents, my sister and my wife Yan Li for their love and support.

This research is supported by the National Eye Institute-NIH under Grant R01 EY09946.
Abstract

We consider the penalized likelihood method with smoothing spline ANOVA for estimating nonparametric regression functions given data from exponential families. Generalized cross-validation or unbiased risk estimation is used to empirically assess the amount of smoothing at each Newton-Raphson iteration. We construct Bayesian confidence intervals and bootstrap confidence intervals for the estimates. Several simulations are conducted to examine the performance of the algorithm, to compare the two methods for choosing smoothing parameters and to assess the performance of confidence intervals. The method is applied to binary response data from the Wisconsin Epidemiological Study of Diabetic Retinopathy to estimate the risk of incidence, progression and progression to proliferative diabetic retinopathy, given several potential risk factors at the start of the study. We also consider the penalized conditional likelihood method with smoothing spline ANOVA for estimating nonparametric log odds functions given data from matched case-control studies. We apply this method to a case-control study of breast cancer.

Key words and phrases: Bayesian confidence intervals; bootstrap confidence intervals; cross-validation; delta algorithm; GCV estimate; generalized linear model; jackknife; matched case-control data; Newton-Raphson iteration; penalized likelihood method; PSA method; smoothing parameter; smoothing splines; smoothing spline ANOVA’s; symmetrized Kullback-Leibler discrepancy; unbiased risk estimate.
# Contents

Acknowledgements ....................................................... ii
Abstract .................................................................... iii

1 Introduction .......................................................... 1
  1.1 Generalized Linear Models ................................. 1
  1.2 Penalized Likelihood Method ............................... 1
  1.3 Smoothing Spline Analysis of Variance .................. 2
  1.4 PSA Method .................................................. 4

2 Choosing the Smoothing Parameters ......................... 7
  2.1 GCV and UBR Methods ................................... 7
  2.2 The PSA Algorithm ...................................... 12
  2.3 Simulations ................................................ 12

3 Confidence Intervals .............................................. 31
  3.1 Introduction ............................................... 31
  3.2 Confidence Intervals for Smoothing Splines ........... 32
    3.2.1 Smoothing Splines .................................. 32
    3.2.2 Bayesian Confidence Intervals ................... 33
    3.2.3 Bootstrap Confidence Intervals ................. 35
    3.2.4 Simulations ......................................... 38
    3.2.5 Sample Size Needed to Distinguish Two Curves . 45
  3.3 Componentwise Confidence Intervals for SS-ANOVA .... 51
    3.3.1 Bayesian Confidence Intervals for Components ... 51
    3.3.2 Bootstrap Confidence Intervals for Components ... 52
    3.3.3 Simulations ........................................ 52
  3.4 Confidence Intervals for Univariate Penalized Likelihood Regression of Non-Gaussian Data .............. 57
    3.4.1 Approximate Bayesian Confidence Intervals ....... 57
    3.4.2 Bootstrap Confidence Intervals ................... 58
    3.4.3 Simulation Study ................................... 58
Chapter 1

Introduction

1.1 Generalized Linear Models

Nelder and Wedderburn (1972) introduced a collection of statistical regression models known as generalized linear models (GLIM's) for the analysis of data from exponential families (McCullagh and Nelder, 1989). The applications of GLIM's to practical data analysis has greatly expanded with the subsequent development of theory and computer software. As the popularity of these methods has increased, so has the need for more sophisticated model building and diagnostic checking techniques.

Suppose data have the form \((y_i, x_i), \ i = 1, 2, \cdots, n\), where \(y_i\) are independent observations. The distribution function of \(y_i\) is from an exponential family with density function

\[
g(y_i; f_i, \phi) = \exp((y_i h(f_i) - b(f_i))/a(\phi) + c(y_i, \phi)), \tag{1.1.1}
\]

where \(f_i = f(x_i)\) is the parameter of interest and \(h(f_i)\) is a monotone transformation of \(f_i\) known as the canonical parameter. \(\phi\) is an unknown scale parameter. Assume \(x_i \in X\), where \(X\) is a measurable space. The \(x_i\)'s are often called covariates.

The log likelihood given \(y_i, x_i\) is

\[
l_i(f_i) = \log g(y_i; f_i, \phi) = (y_i h(f_i) - b(f_i))/a(\phi) + c(y_i, \phi). \tag{1.1.2}
\]

The purpose is to investigate the global relationship between \(f\) and \(x\). GLIM assumes that \(f\) is a linear function of \(x\) when \(X\) is a linear space.

1.2 Penalized Likelihood Method

To achieve greater flexibility, many authors proposed nonparametric regression models to relax the rigid linear constraint. In particular, the penalized likelihood smoothing
spline regression (O'Sullivan, 1983; Green and Yandell, 1985; O’Sullivan, Yandell and Raynor, 1986; Gu, 1990) assumes only that $f$ is smooth but imposes a roughness penalty $J(f)$, say, to discourage arbitrariness of the estimator. More precisely, it assumes that $f \in \mathcal{H}$, where $\mathcal{H}$ is a reproducing kernel Hilbert space. A reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions on $X$ in which the evaluation functional is continuous (Aronszajn, 1950). Suppose $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, where $\mathcal{H}_0$ is finite dimensional (the “parametric” part, usually polynomials), and $\mathcal{H}_1$ (the “smooth” part) is the orthocomplement of $\mathcal{H}_0$ in $\mathcal{H}$. Let $J(f) = \| P_1 f \|^2$, where $P_1$ is the orthogonal projection operator in $\mathcal{H}$ onto $\mathcal{H}_1$. The penalized likelihood method estimates $f$ by solving the following variational problem

$$
\min_{f \in \mathcal{H}} \left\{ -\sum_{i=1}^{n} l_i(f_i) + \lambda \| P_1 f \|^2 \right\}, \quad (1.2.1)
$$

where the first part is the negative log likelihood. It measures the goodness of fit. $\lambda > 0$ is called a smoothing parameter. It controls the trade-off between goodness of fit and roughness of the estimator. We discuss how to choose the smoothing parameter in the next chapter.

### 1.3 Smoothing Spline Analysis of Variance

The smoothing spline analysis of variance (SS ANOVA) model with Gaussian noise assumes

$$
y_i = f(x_1(i), \cdots, x_d(i)) + \epsilon_i, \quad i = 1, \cdots, n, \quad (1.3.1)
$$

where $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(0, \sigma^2 I_{n \times n})$, $\sigma^2$ unknown. $x_k \in X^{(k)}$, where $X^{(k)}$ is a measurable space, $k = 1, \cdots, d$. Denote $X = X^{(1)} \otimes \cdots \otimes X^{(d)}$ and an element in $X$ as $x = (x_1, \cdots, x_d)$. Let $d\mu_j$ be a probability measure on $X^{(j)}$ and let $\mathcal{H}^{(j)}$ be a RKHS of functions on $X^{(j)}$ with

$$
\int_{X^{(j)}} f_j(x_j) d\mu_j = 0 \quad (1.3.2)
$$

for $f_j(x_j) \in \mathcal{H}^{(j)}$. Let $\{1^{(j)}\}$ be the one dimensional space of constant functions on $X^{(j)}$. From (1.3.2), $\{1^{(j)}\}$ is orthogonal to $\mathcal{H}^{(j)}$ under the inner product

$$
< a(x_j), b(x_j) > = \int_{X^{(j)}} a(x_j) b(x_j) d\mu_j. \quad (1.3.3)
$$

Consider the RKHS

$$
\mathcal{G} = \prod_{j=1}^{d} (\{1^{(j)}\} \oplus \{\mathcal{H}^{(j)}\}). \quad (1.3.4)
$$
We can expand $\mathcal{G}$ as

$$
\mathcal{G} = [1] \oplus \sum_j \mathcal{H}^{(j)} \oplus \sum_{j<k} [\mathcal{H}^{(j)} \otimes \mathcal{H}^{(k)}] \oplus \cdots,
$$

where $[1]$ denotes the constant functions on $X$. With some abuse of notation, we are omitting factors of the form $[1^{(j)}]$ whenever it multiplies a term of a different form. An element $f_j$ in $\mathcal{H}^{(j)}$ is called a main effect, an element $f_{jk}$ in $\mathcal{H}^{(j)} \otimes \mathcal{H}^{(k)}$ is called a two-factor interaction, and so on. Suppose $\mathcal{H}^{(j)}$ can be further decomposed: $\mathcal{H}^{(j)} = \mathcal{H}^{(j)}_{\pi} \oplus \mathcal{H}^{(j)}_s$, where $\mathcal{H}^{(j)}_{\pi}$ is finite dimensional subspace, $\mathcal{H}^{(j)}_s$ is the orthocomplement of $\mathcal{H}^{(j)}_{\pi} = \mathcal{H}^{(j)}$. We will later let the penalty $J_j(f_j)$ on the roughness of $f_j$ equal $\|P^{(j)}_s f_j\|_{\mathcal{H}^{(j)}}^2$, where $P^{(j)}_s$ is the orthogonal projection operator in $\mathcal{H}^{(j)}$ onto $\mathcal{H}^{(j)}_s$. Thus no penalty is imposed to elements in $\mathcal{H}^{(j)}_{\pi}$ and $\{1^{(j)}\}$. We can decompose the two factor interaction space $\mathcal{H}^{(j)} \otimes \mathcal{H}^{(k)}$ as the sum of four orthogonal subspaces:

$$
\mathcal{H}^{(j)} \otimes \mathcal{H}^{(k)} = \mathcal{H}^{(j)}_{\pi} \otimes \mathcal{H}^{(k)}_{\pi} \oplus \mathcal{H}^{(j)}_{\pi} \otimes \mathcal{H}^{(k)}_s \oplus \mathcal{H}^{(j)}_s \otimes \mathcal{H}^{(k)}_{\pi} \oplus \mathcal{H}^{(j)}_s \otimes \mathcal{H}^{(k)}_s.
$$

We refer to (1.3.6), (1.3.7), (1.3.8) and (1.3.9) as parametric $\otimes$ parametric $(\pi, \pi)$, parametric $\otimes$ smooth $(\pi, s)$, smooth $\otimes$ parametric $(s, \pi)$ and smooth $\otimes$ smooth $(s, s)$ components. Substituting into (1.3.5), we have

$$
\mathcal{G} = [1] \oplus [\oplus_{j=1}^d \mathcal{H}^{(j)} \oplus (\oplus_{j=1}^d \mathcal{H}^{(j)}_s)] \oplus [(\oplus_{j<k} (\mathcal{H}^{(j)}_{\pi} \otimes \mathcal{H}^{(k)}_{\pi})) \oplus (\oplus_{j<k} (\mathcal{H}^{(j)}_{\pi} \otimes \mathcal{H}^{(k)}_s)) \oplus \cdots.
$$

Consider the special case that $X^{(j)} = [0, 1]$ and $d\mu_j$ is the Lebesque measure on $[0,1]$, $j = 1, \cdots, d$. Let the component space on $X^{(j)}$ be the RKHS

$$
W^{m}_2 = \{ f : f^{(v)} \text{ absolutely continuous, } v = 0, \cdots, m-1, \int_0^1 f^{(m)}(x)^2 dx < \infty \}
$$

with norm

$$
\|f\|^2 = \sum_{v=0}^{m-1} (\int_0^1 f^{(v)}(x) dx)^2 + \int_0^1 f^{(m)}(x)^2 dx.
$$

We can decompose $W^{m}_2 = N \oplus P_{m-1} \oplus S_m$, where $N$ is the space of constants, with the square norm $(\int_0^1 f(x) dx)^2$; $P_{m-1}$ is the space of polynomials of degree less than $m$ which integrate to zero, with the square norm $\sum_{v=1}^{m-1} (\int_0^1 f^{(v)}(x) dx)^2$; and $S_m$ is the space of functions with square integrable $m$th derivatives satisfying $\int_0^1 f^{(v)}(x) dx = 0$, $v = 0, \cdots, m-1$, with the square norm $\int_0^1 f^{(m)}(x)^2 dx$. So the polynomial spline penalty
functional $J(f) = \int_0^1 (f^{(m)}(x))^2 \, dx$ is the norm of the projection of $f$ in $W_2^m$ onto $S_m$. The space $P_0$ vanishes when $m = 1$.

Similar to the usual ANOVA, the model space is a subspace $\mathcal{M}$ of $\mathcal{G}$. By deleting some higher-order interactions, we can get less flexible, but more “estimable” models. We can also select a model from simple to complex by including the constants, or the constants and some of the main effects (this case is often called an additive model, see Hastie and Tibshirani (1990)), or the constants plus some of the main effects and two-factor interactions, and so on. When a model is chosen, we can regroup and write the model space as

$$\mathcal{M} = \mathcal{H}^0 \oplus \sum_{j=1}^q \mathcal{H}^j,$$  \hspace{1cm} (1.3.13)

where $\mathcal{H}^0$ is a finite dimensional space containing functions which are not going to be penalized, and the $\mathcal{H}^j$'s are orthogonal subspaces of the form $\mathcal{H}^{(j)}_\pi$ or composite spaces of tensor products of two or more spaces of the form $\mathcal{H}^{(j)}_\pi$ and $\mathcal{H}^{(k)}_\pi$, and so on. The norms on the composite $\mathcal{H}^j$ are the tensor product norms induced by the norms on the component subspaces, $\|f\|^2 = \|P_0 f\|^2 + \sum_{j=1}^q \|P_j f\|^2$, where $P_j$ is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^j$. The smoothing spline ANOVA estimate of $f$ is the solution to the following variational problem

$$\min_{f \in \mathcal{M}} \left\{ \sum_{i=1}^n (y_i - f(x_i))^2 + n \sum_{j=1}^q \lambda_j \|P_j f\|^2 \right\},$$  \hspace{1cm} (1.3.14)

where $x_i = (x_{1i}, \ldots, x_{di})$. The first term in (1.3.14) is the sum of squared residuals. It measures the goodness of fit. The second part is the penalty on roughness of the estimate. The $\lambda_j$'s are smoothing parameters controlling the trade-off between goodness of fit and roughness. These smoothing parameters can be estimated from data by the generalized cross validation method or by the unbiased risk method (see Wahba (1990), Gu (1989) and next chapter for details).

1.4 PSA Method

We propose a method which combines results from Penalized likelihood estimation, Smoothing splines and Analysis of variance. We call this a PSA method and an estimate using such a method is a PSA estimate.

We assume that the data are from an exponential family and we have chosen a model space $\mathcal{M}$ with the form (1.3.13). As a direct generalization of (1.2.1) to multivariate functions and a direct generalization of (1.3.14) to non-Gaussian data, a PSA estimate is the solution to the following variational problem:

$$\min_{f \in \mathcal{M}} \left\{ -\sum_{i=1}^n l_i(f_i) + \frac{n}{2} \sum_{j=1}^q \lambda_j \|P_j f\|^2 \right\},$$  \hspace{1cm} (1.4.1)
where $f_i = f(x_i)$. The first part in (1.4.1) is the negative log likelihood. It measures the goodness of fit. In the second part, $P_j$ is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}^j$ and $\|P_j f\|^2$ is a quadratic roughness penalty. The $\lambda_j$'s are a set of smoothing parameters controlling the trade-off between goodness of fit and roughness of the estimate. We will discuss how to objectively choose smoothing parameters in the next chapter. Writing $\lambda_j = \lambda / \theta_j$, (1.4.1) becomes

$$
\min_{f \in \mathcal{M}} \left\{ - \sum_{i=1}^n l_i(f_i) + \frac{n}{2} \lambda \| P_* f \|_{\Theta}^2 \right\},
$$

where $P_* = \sum_{j=1}^q P_j$ is the orthogonal projection in $\mathcal{M}$ onto $\mathcal{H}_* = \sum_{j=1}^q \mathcal{H}^j$ and

$$
\| f \|_{\Theta}^2 = \| P_* f \|^2 + \sum_{j=1}^q \theta_j^{-1} \| P_j f \|^2,
$$

is a modified norm indexed by $\Theta = (\theta_1, \cdots, \theta_q)$. We denote by $R_j$ the reproducing kernel (RK) (Aronszajn, 1950) for $\mathcal{H}^j$ under the original norm. It can be shown that $\theta_j R_j$ is the RK under the norm $\| \cdot \|_{\Theta}$. Thus the RK for $\sum_{j=1}^q \mathcal{H}^j$ under $\| \cdot \|_{\Theta}$ is

$$
R_{\Theta} = \sum_{j=1}^q \theta_j R_j.
$$

Since the RK of the tensor product space is the product of the RK's of the component spaces (Aronszajn, 1950), the computation of the $R_j$'s is straightforward. For example, the RK corresponding to the space $\mathcal{H}_k^{(j)} \otimes \mathcal{H}_k^{(k)}$ is

$$
R_{\mathcal{H}_k^{(j)} \otimes \mathcal{H}_k^{(k)}}(x_j(j_1), x_j(j_2)) R_{\mathcal{H}_k^{(k)}}(x_{k}(k_1), x_{k}(k_2)),
$$

where $x_{v}(v)$ denotes the $v$th coordinate of the $v$th design point.

For the special case that $X^{(j)} = [0, 1]$ and $d\mu_j$ is the Lebesgue measure on $[0,1]$, $j = 1, \cdots, d$, the RK for subspaces $N_j$, $P_{m-1}$ and $S_m$ are $R_N(x, z) = 1$, $R_{P_{m-1}}(x, z) = \sum_{v=1}^{m-1} k_v(x)k_v(z)$ and $R_{S_m}(x, z) = k_m(x)k_m(z) + (-1)^{m-1}k_{2m}([x-z])$ respectively, where $k_v(\cdot) = B_v(\cdot)/v!$ and $B_v(\cdot)$ is the $v$-th Bernoulli polynomials defined by $B_0(x) \equiv 1$, $B'_1(x) = vB_{v-1}(x)$ and $\int_0^1 B_v(x)dx = 0$. [x] denotes the fractional part of x.

The solution to (1.4.2) has the form (Wahba, 1990; O'Sullivan et al., 1986)

$$
f_{\lambda, \Theta}(x) = \sum_{v=1}^M d_v \phi_v(x) + \sum_{i=1}^n \sum_{j=1}^q \theta_j R_j(x_i, x) = \phi(x)^T d + \xi(x)^T c,
$$

where $\{\phi_v\}_{v=1}^M$ is a set of basis functions of $\mathcal{H}^n$, $\phi^T(x) = (\phi_1(x), \cdots, \phi_M(x))$, $\xi^T(x) = (R_{\Theta}(x_1, x), \cdots, R_{\Theta}(x_n, x))$. $c_{n \times 1}$ and $d_{M \times 1}$ are vectors of coefficients to be estimated. Substituting (1.4.4) into (1.4.2), we can estimate $c$ and $d$ by minimizing

$$
I(c, d) = - \sum_{i=1}^n l_i(\phi^T(x_i) d + \xi^T(x_i)c) + \frac{n}{2} \lambda c^T Q_{\Theta} c,
$$

(1.4.5)
where \( Q_\Theta \) is an \( n \times n \) matrix with \( Q_\Theta(i,j) = \langle R_\Theta(x_i, x), R_\Theta(x_j, x) \rangle = R_\Theta(x_i, x_j) \). Since \( l_i \)'s are not quadratic, (1.4.5) cannot be solved directly. But if all \( l_i(f_i) \)'s are strictly concave, we can use a Newton-Raphson procedure to compute \( c \) and \( d \) for fixed \( \lambda \) and \( \Theta \). Let \( u_i = -dl_i/df_i, \ u^T = (u_1, \cdots, u_n), w_i = -d^2l_i/df_i^2, \ W = \text{diag}(w_1, \cdots, w_n), \) and \( S = (\phi(x_1), \cdots, \phi(x_n))^T \). Then

\[
\begin{align*}
\frac{\partial I}{\partial c} &= Q_\Theta u + n\lambda Q_\Theta c, \\
\frac{\partial I}{\partial d} &= S^T u, \\
\frac{\partial^2 I}{\partial c \partial c^T} &= Q_\Theta W Q_\Theta + n\lambda Q_\Theta, \\
\frac{\partial^2 I}{\partial c \partial d^T} &= Q_\Theta W S, \\
\frac{\partial^2 I}{\partial d \partial d^T} &= S^T W S. 
\end{align*}
\]

(1.4.6) (1.4.7) (1.4.8) (1.4.9) (1.4.10)

The Newton-Raphson iteration satisfies the linear system

\[
\begin{pmatrix}
Q_\Theta W Q_\Theta + n\lambda Q_\Theta & Q_\Theta W S \\
S^T W Q_\Theta & S^T W S
\end{pmatrix}
\begin{pmatrix}
c - c_- \\
d - d_-
\end{pmatrix}
= 
\begin{pmatrix}
-Q_\Theta u_- - n\lambda Q_\Theta c_- \\
-S^T u_-
\end{pmatrix}
\]

(1.4.11)

where the subscript minus indicates quantities evaluated at the previous Newton-Raphson iteration. Similar to Gu (1990), \( f = Sd + Q_\Theta c \) is always unique as long as \( S \) is of full column rank. Thus all we need is a solution of (1.4.11). If \( Q_\Theta \) is nonsingular, (1.4.11) is equivalent to the system

\[
\begin{align*}
(W_\cdot Q_\Theta + n\lambda I)c + W_\cdot Sd &= W_\cdot f_\cdot - u_-, \\
S^T c &= 0.
\end{align*}
\]

(1.4.12)

If \( Q_\Theta \) is singular, any solution to (1.4.12) is also a solution to (1.4.11). Let \( \tilde{Q}_\Theta = W_{\cdot}^{-1/2} Q_\Theta W_{\cdot}^{1/2}, \ \tilde{c} = W_{\cdot}^{-1/2} c, \ \tilde{S} = W_{\cdot}^{1/2} S, \ \tilde{d} = d, \) and \( \tilde{y} = W_{\cdot}^{-1/2}(W_\cdot f_\cdot - u_-); \) (1.4.12) is simplified to

\[
\begin{align*}
(\tilde{Q}_\Theta + n\lambda I)\tilde{c} + \tilde{S} \tilde{d} &= \tilde{y}, \\
\tilde{S}^T \tilde{c} &= 0.
\end{align*}
\]

(1.4.13)
Chapter 2

Choosing the Smoothing Parameters

2.1 GCV and UBR Methods

In Chapter 1, the smoothing parameters $\lambda_j = \lambda/\theta_j$ are fixed. As all $\lambda_j \to 0$, $f$ follows the data and is very wiggly. It then has small bias but large variance. As all $\lambda_j \to \infty$, $f$ is forced in the null space $H^0$, which is a parametric fit. It then has large bias but small variance. As the $\lambda_j$'s vary, we have a family of models. Therefore choosing appropriate smoothing parameters is crucial for effectively estimating the true function from data by fitting smoothing spline models. Choosing the $\lambda_j$'s is equivalent to choosing $\lambda$ and $\Theta = (\theta_1, \ldots, \theta_q)$ after imposing an identifiability constraint to $\lambda$ and $\Theta$. We call $\lambda$ the main smoothing parameter and $\Theta$ the subsidiary smoothing parameters.

Reconsider the system (1.4.12). This is a one step of Newton-Raphson procedure. Suppose we have data

$$\tilde{\bar{y}}_i = f_i + \epsilon_i, \quad i = 1, \ldots, n,$$

(2.1.1)

where the $\epsilon_i$'s are independent to each other and $\text{Var}(\epsilon_i) = \psi_i^{-1}\sigma^2$. The weighted least square SS ANOVA estimate of $f$ based on these data is the minimizer of:

$$\min_{f \in \mathcal{M}} \left\{ \sum_{i=1}^{n} \psi_i (\tilde{\bar{y}}_i - f_i)^2 + n \sum_{j=1}^{q} \lambda_j \|P_j f\|^2 \right\}. \quad (2.1.2)$$

Let $\Psi = \text{diag}(\psi_1, \ldots, \psi_n)$, $\tilde{\bar{y}} = (\tilde{\bar{y}}_1, \ldots, \tilde{\bar{y}}_n)^T$. The solution of (2.1.2) has the form (1.4.4), where $c$ and $d$ are solutions of the system

$$(\Psi Q_\theta + n\lambda I)c + \Psi S d = \Psi \tilde{\bar{y}},$$

$$S^T c = 0. \quad (2.1.3)$$
Comparing (1.4.12) with (2.1.3), it is clear that the solution of (1.4.12) gives the minimizer of the following problem:

$$
\min_{f \in \mathcal{M}} \left\{ \sum_{i=1}^{n} w_i (\tilde{y}_i - f_i)^2 + n \sum_{j=1}^{q} \lambda_j \|P_j f\|^2 \right\},
$$

(2.1.4)

where \(\tilde{y}_i = f_i - u_i / w_i\). Notice that both \(\tilde{y}_i\)'s and \(w_i\)'s change with iterations. They are called the pseudo-data (or working values and working weights). Thus the Newton-Raphson procedure iteratively reformulates the problem to model \(f_i\)'s on pseudo-data by the reweighted least squares SS-ANOVA. However, the \(\tilde{y}_i\)'s do not exactly have the form (2.1.1). The following lemma indicates that they approximately have the form (2.1.1) if \(f\) is the canonical parameter and \(f_-\) is not far from \(f\). This lemma is an extension of Lemma 3.1 of Gu (1990) from binary data to all data from exponential families.

Lemma 2.1 Suppose that \(h(f) = f, \phi \) is known, \(b \in C^2 \) and \(b''\) is uniformly bounded away from 0. If \(|f_i - f_i| = o(1)\) uniformly in \(i\), then

$$
\tilde{y}_i = f_i + \epsilon_i + o_p(1),
$$

where \(\epsilon_i\) has mean 0 and variance \(w_i^{-1}\).

[Proof] For convenience of notation, we drop the subscript \(i\). Since \(h(f) = f\), we have \(E(y) = b'(f), \text{Var}(y) = b''(f)a(\phi), u = (b'(f) - y)/a(\phi), w = b''(f)/a(\phi), \text{E}(u/w) = 0\) and \(\text{Var}(u/w) = w^{-1}\). Let

$$
\zeta = f_- - u_- / w_- - (f - u/w) = f_- - f - \left( \frac{b'(f_-) - y}{b''(f_-)} - \frac{b'(f) - y}{b''(f)} \right).
$$

Then

$$
E(\zeta) = f_- - f - \frac{b'(f_-) - b'(f)}{b''(f_-)} = \left( 1 - \frac{b''(f_+)}{b''(f_-)} \right) (f_- - f) = o(f_- - f),
$$

$$
\text{Var}(\zeta) = \left( \frac{1}{b''(f_-)} - \frac{1}{b''(f)} \right)^2 b''(f)a(\phi) = o(1) \text{Var}(\frac{u}{w}),
$$

So

$$
\tilde{y} = f_- - \frac{u_-}{w_-} = f - \frac{u}{w} + \zeta = f + \epsilon + o_p(1),
$$

where \(\epsilon = -u/w\) has mean 0 and variance \(w^{-1}\).

From the above discussion, we can use well known methods to select smoothing parameters at each step of the Newton-Raphson procedure. Two of the commonly recognized data-driven methods for choosing smoothing parameters are the generalized
cross validation (GCV) and the unbiased risk (UBR) methods (Wahba, 1990). The GCV method estimates smoothing parameters by minimizing the GCV score

$$V(\lambda, \Theta) = \frac{1/n \| (I - A(\lambda, \Theta)) W_1^{1/2} \tilde{y} \|^2}{\left[ (1/n) tr(I - A(\lambda, \Theta)) \right]^2},$$  \hspace{1cm} (2.1.5)

where \( \tilde{y} = (\tilde{y}_1, \ldots, \tilde{y}_n)^T \), \( A(\lambda, \Theta) \) satisfies

$$(w_1^{1/2} f_{\lambda, \phi}(x_1), \ldots, w_n^{1/2} f_{\lambda, \phi}(x_n))^T = A(\lambda, \Theta)(w_1^{1/2} \tilde{y}_1, \ldots, w_n^{1/2} \tilde{y}_n)^T,$$  \hspace{1cm} (2.1.6)

and the \( f_{\lambda, \phi}(x_i) \)'s are computed as the solution of (1.4.13).

The UBR method estimates smoothing parameters by minimizing the following unbiased risk estimate

$$\tilde{U}(\lambda, \Theta) = \frac{1}{n} \| (I - A(\lambda, \Theta)) W_1^{1/2} \tilde{y} \|^2 + 2 \tilde{\sigma}^2 tr A(\lambda, \Theta),$$  \hspace{1cm} (2.1.7)

where \( \tilde{\sigma}^2 \) is an estimate of the dispersion parameter \( \sigma^2 \). We will give the form of \( \tilde{\sigma}^2 \) later.

It is well known that for a given \( W_1 \), the expectations of \( V(\lambda, \Theta) \) and \( \tilde{U}(\lambda, \Theta) \) are proxies of the expectation of the integrated squared error \( R(\lambda, \Theta) = \int_X (f_{\lambda, \phi}(x) - f(x))^2 d\mu(x) \), where \( \mu(x) \) is the probability measure of experiment design on \( X \) (Wahba, Gu, Wang, and Chappell, 1993). Hence the minimizers of \( V(\lambda, \Theta) \) and \( \tilde{U}(\lambda, \Theta) \) are good estimates of the minimizers of \( R(\lambda, \Theta) \) (Craven and Wahba, 1979; Li, 1985; Li, 1986).

As argued by Gu (1992a), GCV scores and unbiased risk estimates from different data are not comparable. So we should not choose smoothing parameters based on different pseudo-data. Instead, since the fit from the latest Newton-Raphson iteration represents the “best” information available about the true function, we should use them as much as possible. The ultimate goal of our analysis is not to get any specific smoothing parameters, but rather to get an estimate \( \hat{\gamma} \) of the density function \( g \) such that \( L(g, \hat{\gamma}) \) is a minimum for some loss function \( L \) by tuning these smoothing parameters along with the pseudo-data. The integrated squared error is one such loss function. Another loss function is the Kullback-Leibler discrepancy or the symmetric Kullback-Leibler discrepancy, which is a more sensible loss function for density functions. The Kullback-Leibler discrepancy between two density functions \( g \) and \( g_1 \) at the design points is defined as

$$KL(g, g_1) = \frac{1}{n} \sum_{i=1}^n E_i \ln(g_i/g_{1i}),$$

where \( g_i = g(\cdot; f_i, \phi) \) and \( g_{1i} = g(\cdot; f_{1i}, \phi) \). The symmetric Kullback-Leibler discrepancy is defined as

$$SKL(g, g_1) = KL(g, g_1) + KL(g_1, g).$$

KL and SKL discrepancies are proxies of each other. In the following discussion, similar to that in Gu (1992a), we provide evidence that \( \tilde{U}(\lambda, \Theta) \) is a proxy of \( SKL(g_1, g_1, \phi) \), where \( g_{1, \phi} \) is calculated from the solution of (1.4.13). In addition, we develop an estimate of \( \sigma^2 \).

Suppose \( \phi \) is known. Note that \( E(y_i) = b'(f_i)/h'(f_i) \), \( u_i = (b'(f_i) - y_i h'(f_i))/\alpha(\phi) \), \( w_i = (b''(f_i) - y_i h''(f_i))/\alpha(\phi) \). At a particular step of the Newton-Raphson iteration, denote the pseudo-data as

$$\tilde{y}_i = f_i - u_i/w_i = f_i + \delta_i,$$  \hspace{1cm} (2.1.8)
where
\[ \delta_i = -u_{i-}/w_{i-} = (y_i h'(f_{i-}) - b'(f_{i-}))/w_{i-}a(\phi). \]  
(2.1.9)

Again we use the subscript minus to indicate quantities evaluated at the previous Newton-Raphson iteration. Then

\[
\begin{align*}
E_{g_i} \ln \frac{g_i}{g_{\lambda, \Theta}(x_i)} &= E_{f_i} \left[ \frac{y_i h(f_i) - b(f_i)}{a(\phi)} - c(y_i, \phi) - \frac{y_i h(f_{\lambda, \Theta}(x_i)) - b(f_{\lambda, \Theta}(x_i))}{a(\phi)} + c(y_i, \phi) \right] \\
&= \frac{h'(f_i)(h(f_i) - h(f_{\lambda, \Theta}(x_i))) - [b(f_i) - b(f_{\lambda, \Theta}(x_i))]}{a(\phi)}.
\end{align*}
\]

Similarly,

\[
E_{g_{\lambda, \Theta}(x_i)} \ln \frac{g_{\lambda, \Theta}(x_i)}{g_i} = \frac{\frac{\partial (f_{\lambda, \Theta}(x_i))}{\partial h(f_i)} [h(f_{\lambda, \Theta}(x_i)) - h(f_i)] - [b(f_{\lambda, \Theta}(x_i)) - b(f_i)]}{a(\phi)}.
\]

We have

\[
SKL(g_i, g_{\lambda, \Theta})_{proxy} = \frac{1}{n} \sum_{i=1}^{n} \left[ \frac{b'(f_{i-})}{h'(f_{\lambda, \Theta}(x_i))} \right] [h(f_i) - h(f_{\lambda, \Theta}(x_i))]
\]

\[
\approx \frac{1}{n} \sum_{i=1}^{n} \frac{b''(f_{i-})h'(f_{i-}) - b'(f_{i-})h''(f_{i-})h'(f_{i-})(f_i - f_{\lambda, \Theta}(x_i))}{[h'(f_{i-})]^2}
\]

\[
\approx \frac{1}{n} \sum_{i=1}^{n} [b''(f_{i-}) - y_i h''(f_{i-})](f_i - f_{\lambda, \Theta}(x_i))^2
\]

\[
_{proxy} = \frac{1}{n} \sum_{i=1}^{n} w_{i-}(f_i - f_{\lambda, \Theta}(x_i))^2
\]

\[
_{proxy} = \frac{1}{n} \sum_{i=1}^{n} w_{i-}(f_{\lambda, \Theta}(x_i) - f_{i-})^2 + \frac{2}{n} \sum_{i=1}^{n} w_{i-} f_{\lambda, \Theta}(x_i)(f_{i-} - f_i).
\]

Let \( \delta = (\delta_1, \cdots, \delta_n) \), \( f_- = (f_{i-}, \cdots, f_{n-}) \). From (2.1.6), it is easy to show that

\[
\sum_{i=1}^{n} w_{i-}(f_{\lambda, \Theta}(x_i) - f_{i-})^2 = \|(I - A(\lambda, \Theta))W_{\frac{1}{2}}f_- - A(\lambda, \Theta)W_{\frac{1}{2}}\delta\|^2.
\]

Let \( a_{i,j} \) be the \((i,j)\)th entry of \( A(\lambda, \Theta) \). Since

\[
f_{\lambda, \Theta}(x_i) = w_{i-}^{-\frac{1}{2}} \sum_{j=1}^{n} a_{i,j} w_{j-}^{-\frac{1}{2}} y_j = w_{i-}^{-\frac{1}{2}} \sum_{j=1}^{n} a_{i,j} (w_{j-}^{-\frac{1}{2}} f_{j-} + w_{j-}^{-\frac{1}{2}} \delta_j),
\]

\[
w_{i-}(f_{i-} - f_i) \approx h'(f_{i-}) \left[ \frac{b'(f_{i-})}{h'(f_{i-})} - \frac{b'(f_i)}{h'(f_i)} \right]/a(\phi),
\]
it follows that
\[ \sum_{i=1}^{n} w_i f_{i,0}(x_i)(f_i - f_i) \]
\[ \approx \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} \left( w_j f_j + w_j \delta_j \right) \frac{h'(f_j) [y'(f_j) - y_i h'(f_i)]}{w_i h'(f_i)} \]
\[ \approx \sum_{i=1}^{n} \sum_{j=1}^{n} a_{i,j} w_j f_j w_i \frac{[y'(f_i) - y_i h'(f_i)]}{w_i a(\phi)} \]
\[ + \sum_{i=1}^{n} a_{i,i} w_i \delta_i^2 \]
\[ = -(W^{1/2} f)^T A(\lambda, \Theta) (W^{1/2} \delta) - (W^{1/2} \delta)^T A(\lambda, \Theta) (W^{1/2} \delta) + \sum_{i=1}^{n} a_{i,i} w_i \delta_i^2, \]

where we used the following approximations:
\[ \sum_{i=1}^{n} a_{i,j} h'(f_j) \left( \frac{b'(f_j)}{h'(f_j)} - \frac{b'(f_i)}{h'(f_i)} \right) w_i \approx \sum_{i=1}^{n} a_{i,j} w_i \frac{[b'(f_i) - y_i h'(f_i)]}{w_i a(\phi)}, \]
\[ \delta_j \sum_{i=1}^{n} a_{i,j} h'(f_i) \left( \frac{b'(f_i)}{h'(f_i)} - \frac{b'(f_j)}{h'(f_j)} \right) w_i \approx \delta_j \sum_{i \neq j} a_{i,j} w_i \frac{[b'(f_i) - y_i h'(f_i)]}{w_i a(\phi)}. \]

Let
\[ \hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} w_i \delta_i^2 \]

be the weighted average of residuals, which is the usual estimate of the dispersion parameter (McCullagh and Nelder, 1989). \[ \sum_{i=1}^{n} a_{i,i} w_i \delta_i^2 \] may be further approximated by \[ \sigma^2 \sum_{i=1}^{n} a_{i,i} = \sigma^2 tr A(\lambda, \Theta) \] when \[ a_{i,i} \] are roughly balanced, which is generally true (Nychka, 1988). Putting things together and adding \(2(W^{1/2} f)^T (W^{1/2} \delta) + (W^{1/2} \delta)^T (W^{1/2} \delta)\), which does not depend on \( \lambda \) and \( \Theta \), we have

\[ SKL(f, f_{0,0}) \approx \tilde{U}(\lambda, \Theta) \]  

with \( \sigma^2 \) estimated by \( \hat{\sigma}^2 \). If the dispersion parameter is known to be 1, such as in the case of Bernoulli data and Poisson data, a better estimate is

\[ U(\lambda, \Theta) = \frac{1}{n} \| (I - A(\lambda, \Theta)) W^{1/2} \tilde{y} \|^2 + \frac{2}{n} tr A(\lambda, \Theta). \]

We choose \( \lambda \) and \( \Theta \) based on minimizing \( V(\lambda, \Theta) \) or \( U(\lambda, \Theta) \) (\( \tilde{U}(\lambda, \Theta) \) for the case of possible over or under dispersion) for each Newton-Raphson iteration and call them options V and U (U^-) respectively.
2.2 The PSA Algorithm

A generic code RKPACK (Gu, 1989; Gu and Wahba, 1991) is available to solve (1.4.13) and estimate $\lambda$ and $\Theta$ via GCV (option V) or the UBR method at the same time. When using the UBR method, we can either specify $\sigma^2 = 1$ (option U) or estimate $\sigma^2$ (option $U^-$). This suggests the following algorithm:

**PSA Algorithm.** Given the matrices $S$, $Q_j$'s, the response vector $y$ and the starting vector $f_0$:

1. Compute $u_-$ and $W_-$. Compute the transformations $\tilde{S}$, $\tilde{Q}_j = W_+^{1/2}Q_jW_-^{1/2}$ and $\tilde{y}$;

2. Call RKPACK with inputs $\tilde{S}$, $\tilde{Q}_1$ and $\tilde{y}$. That is, solve (1.4.13) and choose $\lambda$ and $\Theta$ by GCV (option V) or the UBR method (option U or option $U^-$);

3. Compute the new $f$. Stop if the algorithm converges under some criteria (for example, the relative weighted mean square error $\sum_{i=1}^n w_i - (f_i - f_i^-)/(1 + |f_i|)^2 / \sum_{i=1}^n w_i$ is used in our programs); otherwise go to step 1.

The starting value $f_0$ may be a constant function, a GLIM fit or some other estimates. Since changing $\lambda$ and $\Theta$ at each iteration means modifying the problem successively, convergence is not guaranteed. Nevertheless, the algorithm converges most of the time.

2.3 Simulations

In this section, we conduct 6 simulations to

1. Evaluate the performance of the PSA method;

2. Compare the performance of the GCV method and the UBR method for choosing smoothing parameters;

3. Assess the effectiveness of the GCV method and the UBR method. We compare the SKL's of estimates with smoothing parameters chosen by GCV or UBR with the SKL's of estimates with optimal smoothing parameters.

In the first 4 simulations, binary data are generated based on an additive model

$$\text{logit}[P(Y = 1| x_1, x_2)] = f(x_1, x_2) = C + f_1(x_1) + f_2(x_2), \quad 0 \leq x_1, x_2 \leq 1,$$

where $C$ is a constant, $f_1(x_1), f_2(x_2)$ are smooth functions which integrate to zero over $[0,1]$. Two different underlying functions $f$ are used:

**Case I** \hspace{1cm} $C = 0$, \hspace{0.5cm} $f_1(x_1) = 2x_1 - 1$, \hspace{0.5cm} $f_2(x_2) = x_2 - 0.5$;
Case II \[ C = 0, \]
\[ f_1(x_1) = -8(x_1 - 0.5)^2 + 8(x_1 - 0.5)^3 I_{\{x_1 > 0.5\}} - s_1, \]
\[ f_2(x_2) = (2.5x_2 - 1)^3 - s_2, \]

where \( I_{\{x_1 > 0.5\}} = 1 \) if \( x_1 > 0.5 \) and \( I_{\{x_1 > 0.5\}} = 0 \) otherwise. \( s_1 \) and \( s_2 \) are constants such that the integrals of \( f_1 \) and \( f_2 \) over \([0,1] \) equal zero. For each case, the two main effects and the 3-dimensional probability surface are plotted in Figure 2.3.1. Two different sample sizes are used: \( n = 200 \) and \( n = 400 \). Hence we have \( 2 \times 2 = 4 \) simulations. We generated the design points \( x_i \) from an uniform distribution on \([0,1]^2 \) and generated binary responses using the underlying function. Designs and responses are generated for 100 replications of each of these 4 simulations. Notice that we have a random design. That is, the designs are different for different replications. We use the Fortran routines \texttt{uni} and \texttt{rmor} of the Core Mathematics Library (Cmlib) from the National Bureau of Standards to generate random numbers in all the simulations. The symmetrized Kullback-Leibler discrepancy \( \text{SKL} = \frac{1}{n} \sum_{i=1}^{n} (p_i - \hat{p}_i) \times (f_i - \hat{f}_i) \) measures the performance of an estimate, where \( p = e^i / (1 + e^i) \) is the probability of \( Y = 1 \). In one replication of Case I with \( n=200 \), RKPACK failed to find a reasonable descent direction and the estimation was poor. We excluded this replication in the following plots.

First, we evaluate the performance of the PSA method. The UBR method (option U) is used to choose the smoothing parameters since it works better than option U of UBR and GCV (see below). We select the 5th, 25th, 50th, 75th and 95th best estimates ordered by SKL. Their main effects estimates are plotted in Figure 2.3.2 for Case I and Figure 2.3.3 for Case II. Their estimates of the probability functions are plotted in Figure 2.3.4 for Case I and \( n=200 \), Figure 2.3.5 for Case I and \( n=400 \), Figure 2.3.6 for Case II and \( n=200 \), and Figure 2.3.7 for Case II and \( n=400 \). We conclude from these plots that the PSA method can capture the shape of a underlying model most of the time even for small sample sizes. But it needs a moderate large sample size to capture the shapes of components.

To compare the GCV method and the option U of UBR with the option U of UBR method for choosing smoothing parameters, we repeat the same simulations with option U of U and V. As shown in Figure 2.3.8, the SKL’s with option U are uniformly smaller than the SKL’s with option U, and the SKL’s with option U are smaller than the SKL’s with option V most of the time. The differences become smaller when the sample size increases. So the option U of UBR is recommended if there is no over- or under-dispersion.

To assess the performance of the GCV method and the UBR method, for a fixed underlying function and a fixed sample, we use the simplex method to search for the minimum SKL as \( \lambda_1 \) and \( \lambda_2 \) vary. That is, for 3 fixed pairs of \( \lambda_1 \) and \( \lambda_2 \), we used RKPACK to calculate the estimates with fixed smoothing parameters. The SKL’s for each pair of \( \lambda_1 \) and \( \lambda_2 \) can be calculated since we know the true function. A new pair of \( \lambda_1 \) and \( \lambda_2 \) is found by the simplex method. This procedure is iterated until convergence.
Figure 2.3.1: Main effects and the probability functions of Case I and Case II.
Figure 2.3.2: Estimates of the main effects of Case I. Solid lines are the true main effects. Five dashed lines in each graph are main effect estimates: 1, 2, 3, 4 and 5 are the 5th, 25th, 50th, 75th, 95th best estimates ordered by SKL.
Figure 2.3.3: Estimates of the main effects of Case II. Solid lines are the true main effects. Five dashed lines in each graph are main effect estimates: 1, 2, 3, 4 and 5 are the 5th, 25th, 50th, 75th, 95th best estimates ordered by SKL.
Figure 2.3.4: The true probability function and their estimates for Case I with n=200.
Figure 2.3.5: The true probability function and their estimates for Case I with $n=400$. 
Figure 2.3.6: The true probability function and their estimates for Case II with \(n=200\).
Figure 2.3.7: The true probability function and their estimates for Case II with n=400.
The SKL reaches a minimum at the converged point. Comparing the SKL’s using the GCV method and the UBR method with the optimal SKL’s, figure 2.3.9 shows that both GCV and UBR work well.

In the next two simulations, let $X = X^{(1)} \otimes X^{(2)} \otimes X^{(3)} = [0,1]^3$, $f(x) = C + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2)$, where $C = -1$, $f_1(x_1) = e^{\alpha x_1}/100 - (e^\beta - 1)/600$, $f_2(x_2) = 10^5[\beta_{2,1}^2(1-x_2) + \beta_{2,7}] + 10^5[\beta_{2,11}^2(1-x_2)^2 - \beta_{2,7}]$, and $f_{12}(x_1, x_2) = 0.5 \cos(2\pi(x_1-x_2))$, where $\beta_{p,q}$ is the Beta function: $\beta_{p,q}(x) = \frac{\Gamma(p+q)}{\Gamma(p)\Gamma(q)} x^{p-1}(1-x)^{q-1}$, $0 \leq x \leq 1$. We call this underlying function Case III. The true main effects, interaction and probability are shown in Figure 2.3.10. Two sample sizes are used: n=200 and n=400. We fit with a model having three main effects and one two factor interaction: $f(x) = C + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2) + f_5(x_3)$. So we have 6 smoothing parameters.

We use fixed designs in these two simulations. That is, we first generate $n$ design points $x_i$'s from the uniform distribution on $[0,1]^3$. We then use this design to generate 100 replications of responses. The algorithm fails in some replications due to the following reasons:

1. The RKPACK subroutine fails to converge within the specified number of iterations (15 is used in our simulations). We denote this situation as info = -4;

2. RKPACK fails to find a reasonable descent direction. We denote this situation as info = -5;

3. The PSA algorithm fails to converge within the specified number of iterations (30 is used in our simulations). We denote this situation as info = -6;

4. The PSA algorithm stops since there are some $w$'s equals to zero. We denote this situation as info = -7.

The numbers of replications out of 100 replications which belong to one of the above 4 cases are listed in Table 2.3.1. From this table, we can see that the sample size of 200 is too small for 6 smoothing parameters without assuming the dispersion parameter $\sigma^2 = 1$ (option V and option U^-). The sample size of 200 is big enough for Option U. The sample size of 400 is big enough for all options.

Table 2.3.1: Number of replications out of 100 total that the algorithm fails.

<table>
<thead>
<tr>
<th>Info</th>
<th>n = 200</th>
<th>n = 400</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-4</td>
<td>-5</td>
</tr>
<tr>
<td>option U</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>option V</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>option U^-</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Obviously, option U works better than the two other options since it fails in fewer replications. The estimation is fine even when the algorithm fails to converge within the
Figure 2.3.8: Comparison of option U with option V and option U⁻ by SKL.
Figure 2.3.9: Comparison of option U and V with the optimum by SKL.
Figure 2.3.10: Main effects, interaction and the probability functions of Case III.
specified steps with option U. Figure 2.3.11 plots the SKL’s of option U against SKL’s of option V and option U* excluding the failed replications. Option U works better than two other options in these replications. If the dispersion parameter is unknown, option V should be used and the dispersion parameter can be estimated.

We select the 5th, 25th, 50th, 75th and 95th best estimates with option U. Their main effect estimates are plotted in Figure 2.3.12. The estimates of the interactions for n=200 are plotted in Figure 2.3.13. The estimates of the interactions for n=400 are plotted in Figure 2.3.14. The estimates of the probability function at $x_3 = 0.5$ are plotted in Figure 2.3.15 for $n = 200$ and in Figure 2.3.16 for $n = 400$. From these plots, we conclude that sample size $n = 200$ is too small. Only the 5th percentile fit catches the shape of the probability function. $n = 400$ is big enough for the estimates to catch the shape of the true probability function. The main effects estimates are also good. But the estimates of the interaction are bad for both sample size $n = 200$ and $n = 400$. This is because the signal of the interaction is weak. We need a bigger sample size to catch the shape of this interaction.
Figure 2.3.12: Estimates of the main effects of Case III. Solid lines are the true main effects. Five dashed lines in each graph are main effect estimates: 1, 2, 3, 4 and 5 are the 5th, 25th, 50th, 75th, 95th best estimates ordered by SKL.
Figure 2.3.13: The true interaction function and some estimates for Case III with n=200.
Figure 2.3.14: The true interaction function and some estimates for Case III with n=400.
Figure 2.3.15: The true probability function and some estimates for Case III at $x_3 = 0.5$ with $n=200$. 
Figure 2.3.16: The true probability function and some estimates for Case III at $x_3 = 0.5$ with $n=400$. 
Chapter 3

Confidence Intervals

3.1 Introduction

Smoothing splines and SS ANOVAs have been used successfully in a broad range of applications requiring flexible nonparametric regression models. It is highly desirable to have interpretable confidence intervals for these estimates for various reasons; for example, to decide whether a spline estimate is more suitable than a particular parametric regression estimate. A parametric regression model may be considered unsuitable if its estimate is outside of a large portion of the confidence intervals of a smoothing spline estimate.

In this chapter, we include the Gaussian distribution as a special case of exponential family. When we say non-Gaussian data, we mean data from exponential families other than Gaussian.

One way to construct confidence intervals for nonparametric estimates is via the bootstrap. Dikta (1990) constructed pointwise bootstrap confidence intervals for a smoothed nearest neighbor estimate. Härdle and Bowman (1988) and Härdle and Marron (1991) used the bootstrap to construct pointwise and simultaneous confidence intervals for a kernel estimate. Kooperberg, Stone and Truong (1993) constructed bootstrap confidence intervals for a regression spline estimate of a hazard function. Wahba (1990) suggested the use of an estimate-based bootstrap to construct confidence intervals for a smoothing spline. Meier and Nychka (1993) used bootstrap confidence intervals for spline estimates to obtain the properties of a statistic to test the equality of two rate equations. However, no reported research to date has studied the properties and performance of bootstrap confidence intervals for a smoothing spline.

Another way to construct confidence intervals for a smoothing spline is to use the correspondence between a smoothing spline and a Bayes estimate (Wahba, 1978). The so-called Bayesian confidence intervals were proposed in Wahba (1983) for a smoothing spline, where their frequentist properties were discussed. Gu and Wahba (1993) extended Bayesian confidence intervals to the components of an SS ANOVA, and Gu
(1992b) extended them to penalized likelihood smoothing spline estimates of non-
Gaussian data. It is well established that these Bayesian confidence intervals have
the average coverage probability property (we define this property later), as opposed to
a pointwise property (Nychka, 1988). They have performed well in a number of simul-
ations. See also Abramovich and Steinberg (1993), who generalized the Bayesian intervals
to the case of a variable smoothing parameter. In this chapter, we extend Bayesian con-
fidence intervals to components of PSA estimates of non-Gaussian data. We prove that
these Bayesian confidence intervals approximately have the average coverage probability
property.

As far as we know, direct comparisons between smoothing spline bootstrap con-
fidence intervals and Bayesian confidence intervals have not yet been done. In this
chapter, we construct bootstrap confidence intervals for smoothing splines and provide
some evidence that the bootstrap confidence intervals have an average coverage prob-
ability across the function being estimated, as opposed to a pointwise property. We
also propose bootstrap confidence intervals for SS ANOVAs and spline estimates of
non-Gaussian data, which appears to be new.

In Section 3.2, we review Bayesian confidence intervals and bootstrap confidence
intervals for smoothing splines with Gaussian data. We show evidence supporting the
average coverage probability property of bootstrap confidence intervals. Six variations
of bootstrap confidence intervals are considered. We run several simulations to find the
best bootstrap confidence intervals and compare them to Bayesian confidence intervals.
We give an estimate of the sample size needed to distinguish two curves if Bayesian
confidence intervals are used. The parallel comparisons for SS ANOVA are given in
Section 3.3. In Section 3.4, we run a simulation to compare the performance between
Bayesian confidence intervals and bootstrap confidence intervals for a penalized likel-
hood smoothing spline estimate based on binary data. We have found that the best
variations of the bootstrap intervals behave similar to the Bayesian intervals. Bootstrap
intervals have the advantage that they are easy to explain and appear to work better
than Bayesian intervals for small sample sizes with Gaussian data. The disadvantage
of bootstrap intervals is that they are computer intensive. In Section 3.5, we introduce
Bayesian confidence intervals for components of PSA estimates of non-Gaussian data
and run a simulation to assess their performance.

3.2 Confidence Intervals for Smoothing Splines

3.2.1 Smoothing Splines

Consider the model

$$y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \quad x_i \in [0, 1],$$  \hspace{1cm} (3.2.1)
where $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \sim N(0, \sigma^2 I_{n \times n})$, $\sigma^2$ unknown and $f \in W_2^m$. The smoothing spline $\hat{f}_\lambda$ is the minimizer of

$$\frac{1}{n} \sum_{i=1}^{n} (y_i - f(t_i))^2 + \lambda \int_0^1 (f^{(m)}(t))^2 dt$$  \hspace{1cm} (3.2.2)$$

over $f \in W_2^m$. The smoothing parameter $\lambda$ controls the trade-off between the goodness of fit and the roughness. When $\lambda$ is fixed, $\hat{f}_\lambda = (\hat{f}_\lambda(x_1), \cdots, \hat{f}_\lambda(x_n))^T$ is a linear function of $y = (y_1, \cdots, y_n)^T$: $\hat{f}_\lambda = A(\lambda)y$, where $A(\lambda)$ is the so-called "hat", or influence, matrix. $\lambda$ can be selected by a data-based procedure such as generalized cross validation (GCV) or unbiased risk estimation (UBR) (see Wahba, 1990). The GCV estimate of $\lambda$ is the minimizer of the GCV function

$$V(\lambda) = \frac{1}{n} ||(I - A(\lambda))y||^2 / \left[ \frac{1}{n} tr(I - A(\lambda)) \right]^2.$$  \hspace{1cm} (3.2.3)$$

The UBR estimate of $\lambda$ is the minimizer of

$$U(\lambda) = \frac{1}{n} ||(I - A(\lambda))y||^2 + 2 \sigma^2 / tr A(\lambda),$$  \hspace{1cm} (3.2.4)$$

assuming that $\sigma^2$ is known. $\sigma^2$ could be replaced by its estimate. Denote by $\hat{\lambda}$ an estimate of $\lambda$ by one of these procedures. Denote by $\hat{f}_\lambda$ the solution of (3.2.2) with $\lambda = \hat{\lambda}$.

### 3.2.2 Bayesian Confidence Intervals

Suppose that $f$ in (3.2.2) is a sample path from the Gaussian process

$$F_\xi(x) = \sum_{j=1}^{m} \frac{\tau_j x^{j-1}}{(j-1)!} + b^j \int_0^x \frac{(x - s)^{m-1}}{(m-1)!} dW(s),$$  \hspace{1cm} (3.2.5)$$

where $W(\cdot)$ is a standard Weiner process and $\tau = (\tau_1, \cdots, \tau_m)^T \sim N(0, \xi I_{m \times m})$. Wahba (1978) showed that with $n\lambda = \sigma^2 / b$,

$$\hat{f}_\lambda(x) = \lim_{\xi \to \infty} E(F_\xi(x)|y), \quad \sigma^2 A(\lambda) = \lim_{\xi \to \infty} \text{Cov}(F_\xi|y),$$  \hspace{1cm} (3.2.6)$$

where $F_\xi = (F_\xi(x_1), \cdots, F_\xi(x_n))^T$.

This connection between a smoothing spline and the posterior mean and variance led Wahba (1983) to propose the $(1 - \alpha)100\%$ Bayesian confidence intervals for $\{f(x_i)\}_{i=1}^{n}$ as

$$\hat{f}_\lambda(x_i) \pm z_{\alpha/2} \sqrt{\sigma^2 A_{ii}}, \quad i = 1, \cdots, n.$$  \hspace{1cm} (3.2.7)$$
where $z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the standard Normal distribution. $\hat{\sigma}^2 = \|I - A(\lambda)x\|^2/\text{tr}(I - A(\lambda))$ is an estimate of $\sigma^2$. Both simulations (Wahba, 1983) and theory (Nychka, 1988; Nychka, 1990) suggested that these Bayesian confidence intervals have good frequency properties for $f \in W_2^2$ provided $\lambda$ is a good estimate of the $\lambda$ which minimizes the predictive mean square error. The intervals must be interpreted “across the function”, rather than pointwise. More precisely, assume $f \in W_2^2$, instead of a realization of the stochastic process (3.2.5). Let $\tau_n$ be a point randomly selected from $\{x_i\}_{i=1,n}$. The average coverage probability (ACP) is defined as

$$ACP = \frac{1}{n} \sum_{i=1}^{n} P(f(x_i) \in C(\alpha, x_i)) = P(f(\tau_n) \in C(\alpha, \tau_n)). \quad (3.2.8)$$

Let $T_n(\lambda) = \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_\lambda(x_i) - f(x_i))^2$ be the average squared error. Let $\lambda^0$ be the value that minimizes $ET_n(\lambda)$. Let $b(x) = Ef_{\lambda^0}(x) - f(x)$ and $v(x) = \hat{f}_{\lambda^0}(x) - E\hat{f}_{\lambda^0}(x)$, the bias term and the variation term of the estimate $\hat{f}_{\lambda^0}(x)$ respectively. Set $b = b(\tau_n)$, $v = v(\tau_n)$ and $U = (b + v)/(ET_n(\lambda^0))^{1/2}$. Nychka argued that the distribution of $U$ is close to a standard normal distribution since it is the convolution of two random variables, one normal and the other with a variance that is small relative to the normal component.

We only consider $\{x_i\}_{i=1,n}$ as fixed design points. Denote by $E_n$ the empirical distribution for $\{x_i\}_{i=1,n}$. Assume $\sup_{u \in [0,1]} |E_n - u| = O(\frac{1}{n})$.

**Assumption 1:** $\hat{\lambda}$ is the minimizer of the GCV function $V(\lambda)$ over the interval $[\lambda_n, \infty)$, where $\lambda_n \sim n^{-d/m/5}$.

**Assumption 2:** $f$ is such that for some $\gamma > 0$, $\frac{1}{n} \sum_{i=1}^{n} (Ef_{\lambda}(x_i) - f(x_i))^2 = \gamma \lambda^2 (1 + o(1))$ uniformly for $\lambda \in [\lambda_n, \infty)$.

**Lemma 3.1 (Nychka 1988)** Suppose $\hat{T}_n$ is a consistent estimator of $ET_n(\lambda^0)$. Let $C(\alpha, t) = \hat{f}_{\lambda}(x) \pm z_{\alpha/2} \sqrt{\hat{T}_n}$. Then under Assumptions 1 and 2,

$$\frac{1}{n} \sum_{i=1}^{n} P(f(x_i) \in C(\alpha, x_i)) - P(|U| \leq z_{\alpha/2}) \rightarrow 0$$

uniformly in $\alpha$ as $n \rightarrow \infty$.

Nychka also proved that

$$\frac{\hat{\sigma}^2 \text{tr}A(\hat{\lambda})/n}{ET_n(\lambda^0)} \xrightarrow{p} \frac{8m^2}{(2m - 1)(4m + 1)} \text{ as } n \rightarrow \infty. \quad (3.2.9)$$

For $m = 2$, the right hand side of (3.2.9) equals 32/27. Thus for large sample size, confidence intervals with $\hat{T}_n$ replaced by $\hat{\sigma}^2 \text{tr}A(\hat{\lambda})/n$ should have ACP close to or a little bit over the nominal coverage. Bayesian confidence intervals actually use the individual diagonal elements of $A(\lambda)$ instead of the average $\text{tr}A(\hat{\lambda})/n$. It is reasonable since most of the diagonal elements are essentially the same (Nychka, 1988).
3.2.3 Bootstrap Confidence Intervals

The following bootstrap method is described in Wahba (1990). Suppose \( \{x_i\}_{i=1,n} \) are fixed design points. Let \( \hat{f}_\lambda \) and \( \hat{\sigma}^2 \) be the estimates of \( f \) and \( \sigma^2 \) from the data. Pretending that \( \hat{f}_\lambda \) is the “true” \( f \), generate a bootstrap sample

\[
y_i^* = \hat{f}_\lambda(x_i) + \epsilon_i^*, \quad i = 1, \cdots, n,
\]

where \( \epsilon^* = (\epsilon_1^*, \cdots, \epsilon_n^*)^T \sim N(\mathbf{0}, \hat{\sigma}^2 I_{n \times n}) \). Then find the smoothing spline estimate \( \hat{f}_\lambda^* \) based on the bootstrap sample. Let \( f^*(x_i) \) be the random variable defined by the bootstrap fit at \( x_i \) given the “true” function \( \hat{f}_\lambda \), variance \( \hat{\sigma}^2 \) and design \( \{x_i\}_{i=1,n} \). Repeat this process \( B \) times. So at each point \( x_i \), we have \( B \) bootstrap estimates of \( \hat{f}_\lambda(x_i) \), which are \( B \) realizations of \( f^*(x_i) \). For each fixed \( x_i \), we use six methods to construct a bootstrap confidence interval for \( f(x_i) \):

(A) Percentile-\( t \) interval (denoted by T-I). Similar to a Student’s \( t \) statistic, consider \( D_i = (\hat{f}_\lambda(x_i) - f(x_i))/s_i \), where \( s_i \) is an appropriate scale parameter. It is called a pivotal statistic since it is independent of the nuisance parameter \( \sigma \) in certain parametric models. We expect it to reduce the dependence on \( \sigma \) in our case. Denote by \( D_i^* \) the bootstrap estimate of \( D_i \); that is, \( D_i^* = (\hat{f}_\lambda^*(x_i) - \hat{f}_\lambda(x_i))/s_i^* \). Let \( \delta_{\alpha/2} \) and \( \delta_{1-\alpha/2} \) be the lower and upper \( \alpha/2 \) points of the empirical distribution of \( D_i^* \). The \( (1-\alpha)100\% \) T bootstrap confidence interval is \( (\hat{f}_\lambda(x_i) - \delta_{1-\alpha/2}s_i, \hat{f}_\lambda(x_i) - \delta_{\alpha/2}s_i) \). The standard deviation of \( \hat{f}_\lambda(x_i) - f(x_i) \) generally equals a constant times \( \sigma \). Setting \( s_i = \hat{\sigma} \), we have the T-I bootstrap confidence intervals:

\[
(\hat{f}_\lambda(x_i) - \delta_{1-\alpha/2}\hat{\sigma}, \hat{f}_\lambda(x_i) - \delta_{\alpha/2}\hat{\sigma}).
\]  

(B) Another percentile-\( t \) interval (denoted by T-II). From the Bayesian model, the exact standard deviation of \( \hat{f}_\lambda(x_i) - f(x_i) \) equals \( \sqrt{\hat{\sigma}^2 a_{i,i}} \). Setting \( s_i^* = \sqrt{\hat{\sigma}^2 a_{i,i}} \) and \( s_i = \sqrt{\sigma^2 a_{i,i}} \) in (A), we have new estimates \( \delta_{\alpha/2}, \delta_{1-\alpha/2} \) and the T-II bootstrap confidence intervals:

\[
(\hat{f}_\lambda(x_i) - \delta_{1-\alpha/2}\sqrt{\hat{\sigma}^2 a_{i,i}}, \hat{f}_\lambda(x_i) - \delta_{\alpha/2}\sqrt{\hat{\sigma}^2 a_{i,i}}).
\]

(C) Normal interval (denoted by Nor). Let \( T_i = (\hat{f}_\lambda(x_i) - f(x_i))^2 \) be the squared error at \( x_i \). Let \( T_i^* \) be the bootstrap estimate of \( T_i \). The \( (1-\alpha)100\% \) normal bootstrap confidence interval is

\[
(\hat{f}_\lambda(x_i) - z_{\alpha/2}\sqrt{T_i^*}, \hat{f}_\lambda(x_i) + z_{\alpha/2}\sqrt{T_i^*}).
\]  

We use the individual squared error estimate instead of the average squared error because we want the length of a confidence interval to depend on the distribution of the design points. Generally, the confidence intervals are narrower in a neighbourhood with more data.
(D) Percentile interval (denoted by Per) (Efron (1982)). Let $f^*_L(x_i)$ and $f^*_U(x_i)$ be the lower and upper $\alpha/2$ points of the empirical distribution of $f^*(x_i)$. The $(1-\alpha)100\%$ confidence interval is

$$
(f^*_L(x_i), f^*_U(x_i)).
$$

(E) Pivotal method (denoted by Piv) (Efron (1981)). Let $\rho_{\alpha/2}$ and $\rho_{1-\alpha/2}$ be the lower and upper $\alpha/2$ points of the empirical distribution of $f^*(x_i) - \hat{f}_\lambda(x_i)$. Then $\rho_{\alpha/2} = f^*_L(x_i) - \hat{f}_\lambda(x_i)$, $\rho_{1-\alpha/2} = f^*_U(x_i) - \hat{f}_\lambda(x_i)$. If the empirical distribution of $f^*(x_i) - \hat{f}_\lambda(x_i)$ approximates the distribution of $\hat{f}_\lambda(x_i) - f(x_i)$, then $P(\rho_{\alpha/2} < \hat{f}_\lambda(x_i) - f(x_i) < \rho_{1-\alpha/2}) \approx 1-\alpha$. The $(1-\alpha)\%$ pivotal confidence interval for $f(x_i)$ is

$$
(2\hat{f}_\lambda(x_i) - f^*_L(x_i), 2\hat{f}_\lambda(x_i) - f^*_U(x_i)).
$$

(F) Bias corrected percentile interval (denoted by BC) (Efron (1982)). Suppose there exists an increasing function $r$ such that

$$
\eta = r(f(x_i)), \quad \hat{\eta} = r(\hat{f}_\lambda(x_i)), \quad \hat{\eta}^* = r(f^*(x_i)),
$$

$$
\hat{\eta} - \eta \sim N(-a, 1), \quad \hat{\eta}^* - \hat{\eta} \sim N(-a, 1),
$$

for some constant $a$. The $(1-\alpha)100\%$ bias corrected confidence interval is

$$
(G_n^{-1}[\phi(2a - z_{\alpha/2})], G_n^{-1}[\phi(2a + z_{\alpha/2})]),
$$

where $G_n^*$ is the empirical distribution of $f^*(x_i)$, $\phi$ is the density function of a standard normal distribution, and $a = \phi^{-1}(G^*(\hat{f}_\lambda(x_i)))$.

The properties of the bootstrap confidence intervals can be studied by rewriting the expected average squared error as

$$
E_f T_n(\lambda^0) = \frac{1}{n} \sum_{i=1}^{n} E_f(\hat{f}_\lambda^0(x_i)f - f(x_i))^2,
$$

where $\hat{f}_\lambda^0(\cdot|f)$ is the smoothing spline estimate of the true function $f$ with the smoothing parameter equals $\lambda^0$. The bootstrap method replaces $f$ by $\hat{f}_\lambda$:

$$
E_f T_n(\lambda^0) \approx E_{\hat{f}_\lambda} T_n = \frac{1}{n} \sum_{i=1}^{n} E_{\hat{f}_\lambda}(\hat{f}_\lambda^* (x_i)|\hat{f}_\lambda) - \hat{f}_\lambda(x_i))^2,
$$

where $\hat{f}_\lambda^* (\cdot|\hat{f}_\lambda)$ is the smoothing spline estimate of $\hat{f}_\lambda$ for a bootstrap sample $y^*$ when $\hat{f}_\lambda$ is the true function. We can estimate the right hand side of (3.2.18) since the “true” function $\hat{f}_\lambda$ and variance $\hat{\sigma}^2$ are known. One way is to repeatedly generate $B$ bootstrap samples from this true model and fit smoothing splines with $\lambda^0$ chosen by GCV or UBR. The average squared error of these $B$ repetitions could be used as an estimate of the expected average squared error. The following theorem proves that for fixed sample size such an estimate is consistent for the right hand side of (3.2.18).
Theorem 3.1 Denote the $B$ bootstrap samples by\
\[ y^{*j} = \hat{f}_\lambda + \epsilon^{*j}, \quad j = 1, \ldots, B. \]

Let $\hat{f}^{*j}_{\lambda_0}$ be the smoothing spline fit for the $j$th bootstrap sample with the smoothing parameter equals $\lambda_0$. Then for fixed $n$,
\[ \frac{1}{B} \sum_{j=1}^{B} \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_{\lambda_0}^{*j}(x_i) - \hat{f}_\lambda(x_i))^2 \xrightarrow{a.s.} E \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_{\lambda_0}^{*}(x_i) - \hat{f}_\lambda(x_i))^2, \quad B \to \infty. \]

[Proof] For simplicity of notations, we use $f$ instead $\hat{f}_\lambda$ as the true function, $\sigma^2$ instead of $\hat{\sigma}^2$ as the true variance. Write
\[ \hat{f}_{\lambda_0}^{*} - f = (A(\lambda_0) - I)f + A(\lambda_0)\epsilon^*. \]

Then
\[ \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_{\lambda_0}^{*}(x_i) - f(x_i))^2 = \frac{1}{n} (\hat{f}_{\lambda_0}^{*} - f)^T (\hat{f}_{\lambda_0}^{*} - f) = \frac{1}{n} [f^T (A(\lambda_0) - I)^2 f + 2f^T (A(\lambda_0) - I)A(\lambda_0)\epsilon^* + \epsilon^{*T} A^2(\lambda_0)\epsilon^*]. \]

Therefore
\[ E \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_{\lambda_0}^{*}(x_i) - f(x_i))^2 = \frac{1}{n} [f^T (A(\lambda_0) - I)^2 f + \sigma^2 tr A^2(\lambda_0)]. \]

Similarly, we have
\[ \frac{1}{B} \sum_{j=1}^{B} \frac{1}{n} \sum_{i=1}^{n} (\hat{f}_{\lambda_0}^{*j}(x_i) - f(x_i))^2 \]
\[ = \frac{1}{B} \sum_{j=1}^{B} \frac{1}{n} [f^T (A(\lambda_0) - I)^2 f + 2f^T (A(\lambda_0) - I)A(\lambda_0)\epsilon^{*j} + (\epsilon^{*j})^T A^2(\lambda_0)\epsilon^{*j}]
\[ = \frac{1}{n} f^T (A(\lambda_0) - I)^2 f + 2f^T (A(\lambda_0) - I)A(\lambda_0) \frac{1}{B} \sum_{j=1}^{B} \epsilon^{*j} +
\[ \frac{1}{B} \sum_{j=1}^{B} (\epsilon^{*j})^T A^2(\lambda_0) \epsilon^{*j}
\[ \xrightarrow{a.s.} \frac{1}{n} f^T (A(\lambda_0) - I)^2 f + \sigma^2 tr A^2(\lambda_0), \quad B \to \infty. \]

Thus the bootstrap method tries to get an estimate $\hat{\epsilon}_n^*$ of $ET_n(\lambda_0)$ directly. From Lemma 1, the bootstrap confidence intervals $C(\alpha, t) = \hat{f}_\lambda(x) \pm z_{\alpha/2} \sqrt{\hat{\epsilon}_n^*}$ should have the
ACP property rather than a pointwise coverage property. The Nor bootstrap intervals use individual squared error estimates at each data point instead of $T_n^*$. They should behave similarly to the intervals using $T_n^*$ when pointwise squared errors are not too different from each other. So we expect the normal bootstrap confidence intervals to have the ACP property rather than the pointwise property. Actually, all these bootstrap confidence intervals seem to have the ACP property, based on simulations presented in the next section.

As pointed out by many authors, the bootstrap bias $E(\hat{f}_{\hat{T}}(x) - \hat{f}(x)|\hat{T})$ generally underestimates the true bias $E(\hat{f}_T(x) - f(x)|f)$, particular at bump points. Hall (1990) suggests using a bootstrap resample of smaller size (say $n_1$) than the original sample for kernel estimates. He showed that for second-order kernels, the optimal choice of $n_1$ is of order $n^{1/2}$. It is hard to get a good estimate of $n_1$ in practice. Furthermore, a bootstrap sample of size $n_1$ may give a very bad smoothing spline estimate. Dikta (1990) and Härdle and Marron (1991) suggested using an undersmoothed estimate to generate the bootstrap samples. They proved that after the right scaling, for a kernel estimate $\hat{f}_\lambda$ with $\lambda$ as the optimal bandwidth, $\hat{f}_{\hat{T}}(x) - \hat{f}_\lambda(x)$ and $\hat{f}_\lambda(x) - f(x)$ have the same limiting distributions as $n \to \infty$, if $\lambda$ tends to zero at a rate slower than $\hat{\lambda}$. Again, it is difficult to get an estimate of $\hat{\lambda}$ in practice. The optimal $\lambda$ depends on some order of the derivative of $f$. Also, the performance for finite samples may not be satisfactory, which is shown in their simulations. Here we do not intend to construct pointwise confidence intervals. Instead, we only need a decent estimate of $ET(\lambda^0)$. Without trying to estimate the correct bias, the bootstrap estimates of mean squared errors prove satisfactory in our simulations.

### 3.2.4 Simulations

In this section, we use some simulations to

1. study the performance of 6 kinds of bootstrap confidence intervals and find out which are better;

2. show the ACP property of bootstrap confidence intervals;

3. compare the performance of bootstrap confidence intervals with the Bayesian confidence intervals.

The experimental design is the same as in Wahba (1983). Three functions are used:

- **Case 1**
  \[
  f(x) = \frac{1}{3} \beta_{10,5}(x) + \frac{1}{3} \beta_{7,7}(x) + \frac{1}{3} \beta_{8,10}(x),
  \]

- **Case 2**
  \[
  f(x) = \frac{6}{10} \beta_{30,17}(x) + \frac{4}{10} \beta_{8,11}(x),
  \]

- **Case 3**
  \[
  f(x) = \frac{1}{3} \beta_{20,5}(x) + \frac{1}{3} \beta_{12,12}(x) + \frac{1}{3} \beta_{7,30}(x),
  \]
where $\beta_{p,q}$ is the Beta function.

Case 1, Case 2 and Case 3 have 1, 2 and 3 bumps respectively (see Figure 3.2.1 next). They reflect an increasingly complex “truth”. The experiment consists of $3 \times 3 \times 5 = 45$ combinations of Case 1, 2 or 3, $n = 32, 64$ or 128 and $\sigma = 0.0125, 0.025, 0.05, 0.1$ or 0.2. In all cases, $x_i = i/n$. Data are generated for 100 replications of each of these 45 combinations. Spline fits are calculated using RKPACK (Gu, 1989). Percentiles of a standard normal distribution are calculated using the CDLIB, developed by Dr. B. W. Brown and Dr. J. Lovato and obtained from statlib.

We use GCV to select $\lambda$ in all simulations in this section. It has been noted that for a small sample size, there is a positive probability that GCV selects $\lambda = 0$, especially for small $\sigma^2$. In practice, if $\sigma$ is known only to within a few orders of magnitude, these extreme cases can be readily identified. The numbers of replications out of 100 simulation replications which have ratio $\hat{\sigma}/\sigma < 0.001$ are listed in Table 3.2.1. Examination of the simulation output reveals that we have the same numbers if we count cases with ratios smaller than 0.1, that is, there are no cases with ratio between 0.1 and 0.001. The number decreases as the sample size increases, as $\sigma$ increases or as the number of bumps decreases. All these cases have their $\lambda$ smaller than $-14$ in the $\log_{10}$ scale, while others have $\lambda$ between $-9$ to $-4$ on $\log_{10}$ scale. We do not impose any limitation on the range of $\lambda$ since we do not want to assume any specific prior knowledge. Instead, we can easily identify these “bad” (interpolation) cases if we know $\sigma$ within 3 orders of magnitude, which is generally true in practice. After identifying the “bad” cases, one can refit by choosing $\lambda$ in a limited range. For our simulations, we simply dropped these cases since they indicate failure of GCV rather than failure of confidence intervals and they are “correctable”. That is, all summary statistics of coverages are based on the remaining cases. Actually, for $\sigma$ as small as .0125 and .025 here, the confidence intervals are visually so close to the estimate that they are hard to see. Confidence intervals are unlikely to be very interesting in these cases. We decided to include these cases since we want to know when these confidence intervals fail.

Table 3.2.1: Number of replications out of 100 total that have ratio $\hat{\sigma}/\sigma < 0.001$.

<table>
<thead>
<tr>
<th>Case</th>
<th>$n = 128$</th>
<th>$n = 64$</th>
<th>$n = 32$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>$\sigma = 0.0125$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma = 0.025$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma = 0.05$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma = 0.1$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$\sigma = 0.2$</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

For all bootstrap confidence intervals, $B = 500$. Using a similar argument to the one above, there is a positive probability that the bootstrap sample fit selects an extreme
Table 3.2.2: Mean coverages of 95% bootstrap confidence intervals.

<table>
<thead>
<tr>
<th>Case</th>
<th>n = 128</th>
<th>n = 64</th>
<th>n = 32</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>σ = 0.125</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.935</td>
<td>.925</td>
<td>.932</td>
</tr>
<tr>
<td>T-II</td>
<td>.925</td>
<td>.918</td>
<td>.923</td>
</tr>
<tr>
<td>Nor</td>
<td>.958</td>
<td>.946</td>
<td>.950</td>
</tr>
<tr>
<td>Per</td>
<td>.940</td>
<td>.927</td>
<td>.930</td>
</tr>
<tr>
<td>Piv</td>
<td>.924</td>
<td>.905</td>
<td>.913</td>
</tr>
<tr>
<td>BC</td>
<td>.919</td>
<td>.897</td>
<td>.905</td>
</tr>
<tr>
<td>σ = 0.025</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.938</td>
<td>.927</td>
<td>.933</td>
</tr>
<tr>
<td>T-II</td>
<td>.929</td>
<td>.919</td>
<td>.925</td>
</tr>
<tr>
<td>Nor</td>
<td>.960</td>
<td>.949</td>
<td>.952</td>
</tr>
<tr>
<td>Per</td>
<td>.940</td>
<td>.924</td>
<td>.931</td>
</tr>
<tr>
<td>Piv</td>
<td>.930</td>
<td>.910</td>
<td>.916</td>
</tr>
<tr>
<td>BC</td>
<td>.923</td>
<td>.903</td>
<td>.909</td>
</tr>
<tr>
<td>σ = 0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.939</td>
<td>.929</td>
<td>.932</td>
</tr>
<tr>
<td>T-II</td>
<td>.928</td>
<td>.920</td>
<td>.924</td>
</tr>
<tr>
<td>Nor</td>
<td>.957</td>
<td>.949</td>
<td>.954</td>
</tr>
<tr>
<td>Per</td>
<td>.940</td>
<td>.923</td>
<td>.930</td>
</tr>
<tr>
<td>Piv</td>
<td>.932</td>
<td>.914</td>
<td>.918</td>
</tr>
<tr>
<td>BC</td>
<td>.923</td>
<td>.908</td>
<td>.909</td>
</tr>
<tr>
<td>σ = 0.1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.938</td>
<td>.932</td>
<td>.929</td>
</tr>
<tr>
<td>T-II</td>
<td>.926</td>
<td>.923</td>
<td>.923</td>
</tr>
<tr>
<td>Nor</td>
<td>.956</td>
<td>.952</td>
<td>.955</td>
</tr>
<tr>
<td>Per</td>
<td>.942</td>
<td>.924</td>
<td>.929</td>
</tr>
<tr>
<td>Piv</td>
<td>.932</td>
<td>.922</td>
<td>.919</td>
</tr>
<tr>
<td>BC</td>
<td>.922</td>
<td>.914</td>
<td>.915</td>
</tr>
<tr>
<td>σ = 0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.937</td>
<td>.935</td>
<td>.930</td>
</tr>
<tr>
<td>T-II</td>
<td>.928</td>
<td>.926</td>
<td>.925</td>
</tr>
<tr>
<td>Nor</td>
<td>.959</td>
<td>.954</td>
<td>.952</td>
</tr>
<tr>
<td>Per</td>
<td>.944</td>
<td>.923</td>
<td>.925</td>
</tr>
<tr>
<td>Piv</td>
<td>.930</td>
<td>.926</td>
<td>.920</td>
</tr>
<tr>
<td>BC</td>
<td>.921</td>
<td>.916</td>
<td>.912</td>
</tr>
</tbody>
</table>
Table 3.2.3: Standard deviations of coverages of 95% bootstrap confidence intervals.

<table>
<thead>
<tr>
<th>Case</th>
<th>n = 128</th>
<th>n = 64</th>
<th>n = 32</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td><strong>σ = .0125</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.050</td>
<td>.038</td>
<td>.038</td>
</tr>
<tr>
<td>T-II</td>
<td>.052</td>
<td>.038</td>
<td>.039</td>
</tr>
<tr>
<td>Nor</td>
<td>.039</td>
<td>.041</td>
<td>.040</td>
</tr>
<tr>
<td>Per</td>
<td>.045</td>
<td>.043</td>
<td>.044</td>
</tr>
<tr>
<td>Piv</td>
<td>.055</td>
<td>.050</td>
<td>.049</td>
</tr>
<tr>
<td>BC</td>
<td>.058</td>
<td>.054</td>
<td>.054</td>
</tr>
<tr>
<td><strong>σ = .025</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.051</td>
<td>.042</td>
<td>.039</td>
</tr>
<tr>
<td>T-II</td>
<td>.054</td>
<td>.042</td>
<td>.041</td>
</tr>
<tr>
<td>Nor</td>
<td>.041</td>
<td>.039</td>
<td>.039</td>
</tr>
<tr>
<td>Per</td>
<td>.048</td>
<td>.046</td>
<td>.047</td>
</tr>
<tr>
<td>Piv</td>
<td>.055</td>
<td>.051</td>
<td>.051</td>
</tr>
<tr>
<td>BC</td>
<td>.060</td>
<td>.056</td>
<td>.054</td>
</tr>
<tr>
<td><strong>σ = .05</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.051</td>
<td>.046</td>
<td>.044</td>
</tr>
<tr>
<td>T-II</td>
<td>.057</td>
<td>.047</td>
<td>.044</td>
</tr>
<tr>
<td>Nor</td>
<td>.044</td>
<td>.042</td>
<td>.040</td>
</tr>
<tr>
<td>Per</td>
<td>.051</td>
<td>.047</td>
<td>.049</td>
</tr>
<tr>
<td>Piv</td>
<td>.057</td>
<td>.056</td>
<td>.053</td>
</tr>
<tr>
<td>BC</td>
<td>.065</td>
<td>.058</td>
<td>.059</td>
</tr>
<tr>
<td><strong>σ = .1</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.058</td>
<td>.051</td>
<td>.050</td>
</tr>
<tr>
<td>T-II</td>
<td>.062</td>
<td>.051</td>
<td>.049</td>
</tr>
<tr>
<td>Nor</td>
<td>.047</td>
<td>.043</td>
<td>.042</td>
</tr>
<tr>
<td>Per</td>
<td>.055</td>
<td>.048</td>
<td>.055</td>
</tr>
<tr>
<td>Piv</td>
<td>.062</td>
<td>.054</td>
<td>.053</td>
</tr>
<tr>
<td>BC</td>
<td>.071</td>
<td>.056</td>
<td>.054</td>
</tr>
<tr>
<td><strong>σ = .2</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>T-I</td>
<td>.064</td>
<td>.053</td>
<td>.053</td>
</tr>
<tr>
<td>T-II</td>
<td>.067</td>
<td>.054</td>
<td>.053</td>
</tr>
<tr>
<td>Nor</td>
<td>.047</td>
<td>.046</td>
<td>.046</td>
</tr>
<tr>
<td>Per</td>
<td>.059</td>
<td>.050</td>
<td>.060</td>
</tr>
<tr>
<td>Piv</td>
<td>.068</td>
<td>.057</td>
<td>.056</td>
</tr>
<tr>
<td>BC</td>
<td>.076</td>
<td>.062</td>
<td>.062</td>
</tr>
</tbody>
</table>
Table 3.2.4: Mean coverage and their standard deviations (inside the parentheses) of 95% Bayesian confidence intervals.

<table>
<thead>
<tr>
<th>Case</th>
<th>n = 128</th>
<th></th>
<th>n = 64</th>
<th></th>
<th>n = 32</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>σ = .0125</td>
<td>.967</td>
<td>.956</td>
<td>.960</td>
<td>.955</td>
<td>.930</td>
<td>.924</td>
</tr>
<tr>
<td></td>
<td>(.032)</td>
<td>(.040)</td>
<td>(.039)</td>
<td>(.076)</td>
<td>(.088)</td>
<td>(.091)</td>
</tr>
<tr>
<td>σ = 0.025</td>
<td>.966</td>
<td>.958</td>
<td>.961</td>
<td>.961</td>
<td>.944</td>
<td>.940</td>
</tr>
<tr>
<td></td>
<td>(.036)</td>
<td>(.041)</td>
<td>(.040)</td>
<td>(.050)</td>
<td>(.076)</td>
<td>(.081)</td>
</tr>
<tr>
<td>σ = 0.05</td>
<td>.963</td>
<td>.959</td>
<td>.961</td>
<td>.956</td>
<td>.947</td>
<td>.943</td>
</tr>
<tr>
<td></td>
<td>(.039)</td>
<td>(.044)</td>
<td>(.041)</td>
<td>(.055)</td>
<td>(.076)</td>
<td>(.077)</td>
</tr>
<tr>
<td>σ = 0.1</td>
<td>.962</td>
<td>.963</td>
<td>.963</td>
<td>.952</td>
<td>.948</td>
<td>.945</td>
</tr>
<tr>
<td></td>
<td>(.043)</td>
<td>(.039)</td>
<td>(.034)</td>
<td>(.058)</td>
<td>(.076)</td>
<td>(.076)</td>
</tr>
<tr>
<td>σ = 0.2</td>
<td>.961</td>
<td>.962</td>
<td>.960</td>
<td>.938</td>
<td>.953</td>
<td>.947</td>
</tr>
<tr>
<td></td>
<td>(.048)</td>
<td>(.041)</td>
<td>(.039)</td>
<td>(.073)</td>
<td>(.059)</td>
<td>(.053)</td>
</tr>
</tbody>
</table>
\[ \lambda^* = 0. \] Since we know the true function and variance when bootstrapping, we certainly can identify these "bad" cases and should drop them. Therefore these "bad" repetitions are dropped when we calculate the bootstrap confidence intervals. This is not a limitation on bootstrap confidence intervals, but rather a subtle point about which one needs to be careful when constructing bootstrap confidence intervals. Our results here and in Wahba and Wang (1993) suggest that this phenomena of "bad" cases is rare in practice for \( n \) much bigger than 100.

In each case, the number of data points at which the confidence intervals cover the true values are recorded. These numbers are then divided by the corresponding sample sizes to form the coverage percentage of the intervals on the design points. Table 3.2.2 and 3.2.3 list mean coverages and their standard deviations of 95% bootstrap confidence intervals. For almost all cases, T-I, T-II, Nor and Per bootstrap confidence intervals are better than Piv and BC bootstrap confidence intervals. T-I intervals work better than T-II's. Nor intervals are a little bit better than Per's. T-I intervals are better than Nor intervals in small sample size cases but this is reversed in large sample cases. The average coverages are much improved when the sample size increases from 32 to 64 and improved a little bit when the sample size increases from 64 to 128.

In the remainder of this section, when we mention bootstrap confidence intervals, we mean either T-I or Nor bootstrap confidence intervals. To compare with the Bayesian confidence intervals, we use the same data to construct Bayesian confidence intervals. The mean coverages and standard deviations of 95% confidence intervals are listed in Table 3.2.4. Comparing Tables 3.2.2, 3.2.3 and 3.2.4, we see that for \( n = 32 \), bootstrap confidence intervals have better average coverages and smaller standard deviations than Bayesian intervals. For \( n = 64 \), bootstrap confidence intervals and Bayesian confidence intervals are about the same. For \( n = 128 \), Bayesian confidence intervals have average coverages a little bit over the nominal value, while bootstrap confidence intervals have average coverage a little bit under the nominal value.

For each repetition of the experiment, we calculated the true MSE, a bootstrap estimate of MSE (denoted by MSE*) and the estimate of \( \sigma \) (denoted by \( \hat{\sigma} \)). We then get the ratios: MSE*/MSE and \( \hat{\sigma}/\sigma \). The average ratios and their standard deviations are listed in Table 3.2.5. Notice that \( \hat{\sigma} \) underestimates \( \sigma \) on average, which agrees with Carter and Eagleson (1992). Thus the bootstrap samples have smaller variation than they should. This causes the average coverages of bootstrap confidence intervals to be a little bit smaller than the nominal value. On the other hand, underestimation of \( \hat{\sigma}^2 \) does help the performance of Bayesian confidence intervals since \( \hat{\sigma}^2 tr.A(\lambda)/n \) overestimates \( ET_n(\lambda^0) \) by a factor of \( 8m^2/(2m-1)(4m+1) \). Carter and Eagleson (1992), who studied the same examples used here, found that for these functions, a better choice to the estimate of \( \sigma^2 \) is \( y^T(I - A(\lambda))^2 y/ tr[I - A(\lambda)]^2 \). We do not know to what extent these results concerning \( \hat{\sigma}^2 \) are example-dependent, but we would expect that such a choice of \( \hat{\sigma}^2 \) would make bootstrap confidence intervals work better relative to the Bayesian intervals in the present experiments. Notice also that even though the bootstrap bias is generally smaller than true bias, MSE* overestimates MSE on average, especially for
Table 3.2.5: Means and standard deviations (inside the parentheses) of the ratios of bootstrap MSE and true MSE and ratios of estimated $\sigma$ and true $\sigma$.

<table>
<thead>
<tr>
<th></th>
<th>n = 128</th>
<th></th>
<th>n = 64</th>
<th></th>
<th>n = 32</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MSE*/MSE</td>
<td>$\tilde{\sigma}/\sigma$</td>
<td>MSE*/MSE</td>
<td>$\tilde{\sigma}/\sigma$</td>
<td>MSE*/MSE</td>
</tr>
<tr>
<td>$\sigma = .0125$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>1.196(.434)</td>
<td>.979(.072)</td>
<td>1.154(.457)</td>
<td>.944(.144)</td>
<td>1.018(.604)</td>
</tr>
<tr>
<td>Case 2</td>
<td>.989(.288)</td>
<td>.964(.086)</td>
<td>.991(.415)</td>
<td>.989(.168)</td>
<td>.333(.245)</td>
</tr>
<tr>
<td>Case 3</td>
<td>1.060(.284)</td>
<td>.954(.085)</td>
<td>.956(.395)</td>
<td>.859(.162)</td>
<td>—</td>
</tr>
<tr>
<td>$\sigma = .025$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>1.231(.502)</td>
<td>.983(.071)</td>
<td>1.210(.502)</td>
<td>.964(.121)</td>
<td>1.109(.645)</td>
</tr>
<tr>
<td>Case 2</td>
<td>1.047(.351)</td>
<td>.969(.084)</td>
<td>1.071(.425)</td>
<td>.922(.153)</td>
<td>.647(.237)</td>
</tr>
<tr>
<td>Case 3</td>
<td>1.091(.316)</td>
<td>.963(.084)</td>
<td>1.042(.404)</td>
<td>.904(.154)</td>
<td>.356(.306)</td>
</tr>
<tr>
<td>$\sigma = .05$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>1.266(.587)</td>
<td>.985(.071)</td>
<td>1.238(.547)</td>
<td>.967(.120)</td>
<td>1.209(.747)</td>
</tr>
<tr>
<td>Case 2</td>
<td>1.106(.417)</td>
<td>.972(.084)</td>
<td>1.120(.456)</td>
<td>.935(.149)</td>
<td>.897(.612)</td>
</tr>
<tr>
<td>Case 3</td>
<td>1.121(.355)</td>
<td>.969(.084)</td>
<td>1.093(.433)</td>
<td>.928(.152)</td>
<td>.716(.478)</td>
</tr>
<tr>
<td>$\sigma = .1$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>1.301(.687)</td>
<td>.987(.070)</td>
<td>1.268(.588)</td>
<td>.971(.119)</td>
<td>1.286(.891)</td>
</tr>
<tr>
<td>Case 2</td>
<td>1.166(.473)</td>
<td>.980(.071)</td>
<td>1.151(.482)</td>
<td>.945(.144)</td>
<td>1.066(.677)</td>
</tr>
<tr>
<td>Case 3</td>
<td>1.152(.388)</td>
<td>.978(.072)</td>
<td>1.124(.470)</td>
<td>.942(.148)</td>
<td>.933(.549)</td>
</tr>
<tr>
<td>$\sigma = .2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Case 1</td>
<td>1.326(.735)</td>
<td>.988(.070)</td>
<td>1.236(.636)</td>
<td>.972(.118)</td>
<td>1.256(1.000)</td>
</tr>
<tr>
<td>Case 2</td>
<td>1.203(.537)</td>
<td>.983(.072)</td>
<td>1.182(.518)</td>
<td>.960(.127)</td>
<td>1.149(.715)</td>
</tr>
<tr>
<td>Case 3</td>
<td>1.167(.437)</td>
<td>.982(.072)</td>
<td>1.158(.518)</td>
<td>.964(.124)</td>
<td>1.023(.543)</td>
</tr>
</tbody>
</table>
large sample sizes. The variances of \( \text{MSE}^*/\text{MSE} \)'s are quite big.

We visually inspected many of the plotted intervals and pointwise coverages. They all give a similar visual impression. Therefore, we just plot some “typical” cases. In Figure 3.2.1, we plot both T-I bootstrap confidence intervals and Bayesian confidence intervals when \( \sigma = 0.2 \) for some selected functions and sample sizes. These cases are actually the first replicates in simulations. Some pointwise coverages are plotted in Figure 3.2.2 (T-I and Bayesian) and 3.2.3 (Per and Nor). Pointwise coverages of the Per bootstrap and Bayesian confidence intervals vs \( |\hat{f}| \) are plotted in Figure 3.2.4. We also plot the pointwise coverages of bootstrap confidence intervals vs the pointwise coverages of Bayesian confidence intervals in Figure 3.2.5. From these plots, it is obvious that the pointwise coverages of that bootstrap confidence intervals are similar to that of the Bayesian confidence intervals. That is, the pointwise coverage is smaller than the nominal value at high curvature points, particularly for Per intervals. These plots support the argument that the bootstrap confidence intervals have the ACP property.

### 3.2.5 Sample Size Needed to Distinguish Two Curves

Suppose we have two different curves \( f \) and \( g \) defined on \([0, 1]\), with some data generated from curve \( f \) plus Gaussian noise. That is, the data have the form (3.2.1). We can construct Bayesian confidence intervals for \( f \) such as \( \hat{f}_\lambda(x_i) \pm z_{\alpha/2} \sqrt{\sigma^2 a_{i,i}} \), \( i = 1, \cdots, n \).

Here we consider the design as random and let \( \kappa \) be a random design point with density function \( \epsilon \). Assume \( \epsilon \) is strictly positive on \([0, 1]\) and \( \epsilon \in C^\infty[0, 1] \). The following assumption is a variation of Assumption 1 with \( \lambda_n \sim n^{-4m/5} \) replaced by \( \lambda_n \sim n^{-2m/5} \log(n)^m \).

**Assumption 1**: \( \lambda \) is the minimizer of the GCV function \( V(\lambda) \) over the interval \( [\lambda_n, \infty) \), where \( \lambda_n \sim n^{-2m/5} \log(n)^m \).

Nychka (1990) proved that if Assumption 1' and Assumption 2 hold, then \( \hat{\sigma}^2 \to_p \sigma^2 \) and \( tr A(\lambda)/n \to 0 \) as \( n \to \infty \). So the average Bayesian interval length approaches to zero. Thus more and more points of \( g \) lie outside of the confidence intervals as the sample size increases. The question is: what is the smallest sample size such that if we randomly choose a point \( \kappa \) according to the distribution of design, the probability of \( g(\kappa) \) is inside the 100(1 - \( \alpha \))% Bayesian confidence interval at this point is less than \( \alpha \)?

Define \( d(f, g) = |f(\kappa) - g(\kappa)| \). Denote by \( D \) the cumulative distribution function of \( d(f, g) \) and let \( d_\alpha \) be the \( \alpha \) percentile of \( D \). Assume that \( A_{i,i} = tr A(\lambda)/n \), \( i = 1, \cdots, n \) (the following formulas are approximately true when the diagonal elements of \( A(\lambda) \) are almost equal). Then

\[
P(|f(\kappa) - g(\kappa)| > z_{\alpha/2} \sqrt{tr A(\lambda)/n}) \geq P(|f(\kappa) - g(\kappa)| > d_\alpha) = 1 - \alpha
\]
Figure 3.2.1: Display of the 95% confidence intervals for $\sigma = 0.2$. Solid lines: true function; dashed lines: confidence intervals.
Figure 3.2.2: Stars are pointwise coverages of 95% confidence intervals when $\sigma = 0.1$ and $n=128$. Dashed curves are the magnitude of $|f|$. 
Figure 3.2.3: Stars are pointwise coverages of 95% confidence intervals when $\sigma = 0.1$ and $n=128$. Dashed curves are the magnitude of $|f|$. 
Figure 3.2.4: Plots of pointwise coverages of 95% confidence intervals vs $|\hat{f}|$ when $\sigma = 0.1$ and $n=128$. 
Figure 3.2.5: Plots of pointwise coverages of 95% bootstrap confidence intervals vs pointwise coverages of 95% Bayesian confidence intervals when $\sigma = 0.1$ and $n=128$. n, p and t in these plots indicate Nor, Per and T-I.
as long as \( z_{a/2} \sigma \sqrt{\text{tr}(\hat{A}(\lambda)) / n} \leq d_{\alpha} \). Therefore the sample size we need is

\[
n = \left( \frac{z_{a/2} \sigma}{d_{\alpha}} \right)^2 \text{tr}(\hat{A}(\lambda)).
\] (3.2.19)

The quantity \( \text{tr}(A(\lambda)) \) is sometimes called the degrees of freedom for signal. It goes to infinity as \( n \) goes to infinity if \( f \notin \mathcal{H}_0 \). For a special case, suppose Assumptions 1' and 2 hold. From Lemma 3.1 of Nychka (1990),

\[
\text{tr}(A(\lambda)) = a l_1 \lambda^{-\frac{1}{m}} (1 + o(1)), \quad \hat{\lambda} = \left( \frac{a l_2 \sigma^2}{n \gamma 4m} \right)^{\frac{1}{2m+4}} (1 + o_p(1)),
\] (3.2.20)

where \( a = \pi / \int_0^1 (e(x))^{\frac{1}{m}} dx \) and \( l_k = \int_0^\infty \frac{dx}{(1+x^{2m})^k}, \ k = 1, 2 \). Replacing them in (3.2.19), we have

\[
n \approx \left[ \left( \frac{z_{a/2} \sigma}{d_{\alpha}} \right)^2 a l_1 \left( \frac{a l_2 \sigma^2}{n \gamma 4m} \right)^{-\frac{1}{2m+4}} \right]^{-\frac{1}{4m+4}}.
\] (3.2.21)

In some cases (cf. Wahba (1983)), \( \gamma = \int_0^1 (f^{(2m)}(x))^2 dx \). If \( f, g, e \) and \( \sigma \) are known, \( d_{\alpha} \) could be calculated or estimated through Monte Carlo simulation. In practice, we do not know \( f, g, e \) and \( \sigma \). Some pilot study or prior knowledge are necessary for the design in order to estimate sample size adequately.

### 3.3 Componentwise Confidence Intervals for SS-ANOVA

#### 3.3.1 Bayesian Confidence Intervals for Components

An SS ANOVA estimate \( f_\lambda \) is the solution to (1.4.1) when the data are Gaussian. The minimization also gives estimates of components in the model space such as main effects and two-factor interactions. We would like to construct confidence intervals not only for the overall function \( f \), but also for all components in the model. These confidence intervals may, for example, be used as an aid in eliminating unimportant components, or deciding whether certain features are "real".

Suppose that \( f \) is a sample path from the Gaussian process

\[
F_\xi(x) = \sum_{k=1}^M \tau_k \phi_k(x) + b^\frac{1}{2} \sum_{j=1}^q \sqrt{\tau_j} Z_j(x),
\] (3.3.1)

where \( \tau = (\tau_1, \cdots, \tau_M)^T \sim N(0, \xi I_{M \times M}), M = \dim(\mathcal{H}^0), \phi_1, \cdots, \phi_M \) span \( \mathcal{H}^0 \), the \( Z_j \)'s are independent, zero mean Gaussian stochastic processes, independent of \( \tau \), with
\[ E[Z_{ij}(z)Z_{j}(x)] = R_{ij}(z,x), \] where \( R_{ij}(z,x) \) is the reproducing kernel of \( H^1 \). Gu and Wahba (1993) showed that with \( n\lambda = \sigma^2/b \), \( \hat{f}_{\lambda}(x) = \lim_{t \to \infty} E(F_t(x)|y) \), and that this relation holds with \( \hat{f}_{\lambda} \) and \( f \) replaced by \( P^T\hat{f}_{\lambda} \) and \( P^Tf \). They also obtained the posterior variances for each component. Then they constructed Bayesian confidence intervals for each component in an manner analogous to the univariate case.

### 3.3.2 Bootstrap Confidence Intervals for Components

As in the univariate smoothing spline case, a componentwise bootstrap tries to estimate the distribution of an SS ANOVA estimate directly. We first get estimates of \( f \), the components of \( f \) and \( \sigma^2 \). Then we generate \( B \) bootstrap samples and fit each with an SS ANOVA model and collect its components. We can calculate bootstrap confidence intervals for each component separately as in the univariate case. We construct T-I (simply denoted as T), Nor, Per, Piv and BC intervals in this case. We do not construct T-II intervals since they are inferior to T-I's.

### 3.3.3 Simulations

We use the same example function and the same model space as in Gu and Wahba (1993). Let \( \mathbf{X} = X^{(1)} \otimes X^{(2)} \otimes X^{(3)} = [0,1]^3 \), \( f(x) = C + f_1(x_1) + f_2(x_2) + f_{12}(x_{12}) \), where \( C = 5 \), \( f_1(x_1) = e^{3x_1} - (e^3 - 1)/3 \), \( f_2(x_2) = 10^6[x_2^3(1-x_2)^6 - \beta_{12,7}] + 10^4[x_2^3(1-x_2)^{10} - \beta_{4,11}] \), and \( f_{12}(x_1, x_2) = 5 \cos(2\pi(x_1-x_2)) \), where \( \beta_{p,q} \) is the Beta function.

We fit with a model having three main effects and one two-factor interaction: \( f(x) = C + f_1(x_1) + f_2(x_2) + f_{12}(x_1, x_2) + f_3(x_3) \).

We only run simulations for \( n = 100 \) (\( n = 200 \) was too computer intensive) with three levels of \( \sigma \) (1, 3 and 10). The GCV method was used to choose smoothing parameters for all simulations. \( B = 100 \) for all simulations. 100 trials are conducted for each experiment, and data for the 95% confidence intervals are collected. In each case, the number of data points at which the confidence intervals cover the true values of \( f, f_1, f_2, f_{12} \) and \( f_3 \) are recorded. These numbers are then divided by the sample size to form the coverage percentage of the intervals on design points. We summarize these coverage percentages using box-plots (Figure 3.3.1, 3.3.2 and 3.3.3). We only plot the box-plots for T, Nor, Per and Piv intervals since they are uniformly better than the BC intervals. As in the smoothing spline case, the GCV criterion selects \( \lambda \approx 0 \) (nearly interpolates the data) in one of the 100 \( \sigma = 1 \) trials, in one of the \( \sigma = 3 \) trials and in two of the \( \sigma = 10 \) trials. Again, these cases can be readily detected by examining estimates of \( \sigma^2 \) which are orders of magnitude smaller than the true values. We exclude these four cases.

From these box-plots, we see that T, Nor, Per and Piv intervals work well. The good performance of Nor intervals suggests that Lemma 3.1 might be true for componentwise Nor intervals. Comparing with box-plots of Gu and Wahba (1993) in the top row of their
Figure 3.3.1: Coverage percentages of 95% bootstrap intervals when $\sigma = 1$. Plusses: sample means; dotted lines: nominal coverage.
Figure 3.3.2: Coverage percentages of 95% bootstrap intervals when $\sigma = 3$. Plusses: sample means; dotted lines: nominal coverage.
Figure 3.3.3: Coverage percentages of 95% bootstrap intervals when $\sigma = 10$. Plusses: sample means; dotted lines: nominal coverage.
Figure 3.3.4: Display of the 95% intervals in a "typical" n=100 and σ = 3 trial. Solid lines: true function; dotted lines: SS ANOVA fit; dashed lines: confidence intervals.
Figure 1, we can see that bootstrap confidence intervals have somewhat better mean coverages and smaller variability than the Bayesian confidence intervals. A point worth noting is that Bayesian confidence intervals for $f_\beta$ are actually simultaneous confidence intervals. This is not true for bootstrap confidence intervals.

We visually inspected many of the plotted intervals and (with the above noted four exceptions) they all look similar. A “typical” trial for $\sigma = 3$ is plotted in Figure 3.3.4. We can see that bootstrap confidence intervals are not very smooth. This is because $B = 100$ is not big enough. We expect that with $B \geq 500$, the bootstrap confidence intervals will look smoother.

### 3.4 Confidence Intervals for Univariate Penalized Likelihood Regression of Non-Gaussian Data

#### 3.4.1 Approximate Bayesian Confidence Intervals

Let $f_\lambda$ be the solution of (1.2.1). Again, suppose $f$ is a sample path from the Gaussian process

$$F_{\xi}(x) = \sum_{k=1}^{M} \tau_k \phi_k(x) + b^\delta Z(x), \tag{3.4.1}$$

where $\tau = (\tau_1, \ldots, \tau_M)^T \sim N(0, \xi I_{M \times M})$, $\phi_1, \ldots, \phi_M$ span $\mathcal{H}^0$, $Z(x)$ is a zero mean Gaussian process and is independent of $\tau$, with $EZ(x)Z(x) = R(z, t)$, where $R(z, t)$ is the reproducing kernel of $\mathcal{H}^0$. Gu (1992b) set $n \lambda = \sigma^2/b$, and obtained the approximate posterior distribution of $f$ given $y$ as Gaussian with $\hat{f}_\lambda(x) \approx \lim_{\xi \to \infty} \mathbb{E}(F_{\xi}(x) | y)$. He found the posterior covariance $\lim_{\xi \to \infty} \text{Cov}(F_{\xi} | y)$, in terms of the relevant “hat” or influence matrix for the problem, and the Hessian of the log likelihood with respect to $f$, evaluated at the fixed point of the Newton-Raphson iteration for the minimizer of (1.2.1). See Gu (1992b) for details.

We now prove that these Bayesian confidence intervals approximately have the ACP property. Consider the model: $y_i = f(x_i) + \epsilon_i$, $i = 1, \ldots, n$. Let $f_i = f(x_i)$, $f = (f_1, \ldots, f_n)^T$, $y = (y_1, \ldots, y_n)^T$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$. Suppose that $\text{E} \epsilon = 0$ and $\text{Cov}(\epsilon) = \sigma^2 W^{-1}$, where $W$ is known and $W = \text{diag}(w_1, \ldots, w_n)$. We make the following transformations: $\tilde{y} = W^{1/2} y$, $\tilde{f} = W^{1/2} f$, $\tilde{\epsilon} = W^{1/2} \epsilon$. Let $\tilde{f}_{\lambda}$ be the smoothing spline estimate of $\tilde{f}$ using the data $\tilde{y}$, where $\lambda$ is chosen by GCV. Let $A(\tilde{\lambda})$ be the “hat” matrix, that is, $\tilde{f}_{\lambda} = A(\tilde{\lambda}) \tilde{y}$. Then $\tilde{f}_{\lambda} = W^{-1/2} \tilde{f}_{\lambda}$ is a fit of the original function. Define $\mathcal{U}$ as section 3.2.2 for the transformed problem. Notice that we do not need the $\epsilon_i$'s to be normal. Lemma 3.1 is still true under the additional assumption:

**Assumption 3:** $\mathbb{E}(|\epsilon_i|^p) < \infty$. 

Theorem 3.2 Let \( \zeta = \sqrt{\hat{\sigma}^2 \text{tr}((\hat{\Lambda})^{-1})/n} \), \( \zeta_i = \zeta w_i^{-1/2} \). Let \( C(x, \alpha) = \hat{f}_i(x) \pm z_{\alpha/2} \zeta_i \). Under Assumptions 1-3,

\[
\frac{1}{n} \sum_{i=1}^{n} P(f(x_i) \in C(\alpha, x_i)) - P\left( \frac{(2m-1)(4m+1)}{8m^2} |U| \leq z_{\alpha/2} \right) \rightarrow 0
\]

uniformly in \( \alpha \) as \( n \rightarrow \infty \).

[Proof]

\[
\frac{1}{n} \sum_{i=1}^{n} P(f(x_i) \in C(\alpha, x_i)) = E\frac{1}{n} \sum_{i=1}^{n} I\left( \frac{|\hat{f}_i(x_i) - f(x_i)|}{\zeta_i} \leq z_{\alpha/2} \right) = E\frac{1}{n} \sum_{i=1}^{n} I\left( \frac{|\hat{f}_i(x_i) - \hat{f}_i(x_i)|}{\zeta} \leq z_{\alpha/2} \right).
\]

From Lemma 3.1 we have the conclusion.

Nychka (1988) argued that the diagonal elements of \( A(\hat{\Lambda}) \) are almost equal. Therefore

\[
\zeta_i^2 = \frac{\hat{\sigma}^2 \text{tr}(A(\hat{\Lambda}))}{w_i} \approx \frac{\hat{\sigma}^2 a_{ii}}{w_i} = \hat{\sigma}^2 [A(\hat{\Lambda}) W^{-1}]_{ii}.
\]

(3.4.2)

Combining (3.4.2) with Theorem 3.2, we conclude that these Bayesian confidence intervals approximately have the ACP property.

3.4.2 Bootstrap Confidence Intervals

The method for bootstrap confidence intervals for non-Gaussian data is similar to the development for the smoothing spline in section 3.2.3. The only difference is now the bootstrap samples are non-Gaussian. No approximation is involved after we get a spline fit. Therefore, we might expect the bootstrap confidence intervals to work better than Bayesian confidence intervals. We construct Nor, Per, Piv and BC bootstrap confidence intervals. Notice that in the case of Bernoulli data, there is no unknown scale parameter \( (\sigma = 1) \), and hence Piv intervals are the same as T-I intervals.

3.4.3 Simulation Study

We used the same experimental design as Gu (1992b). Bernoulli responses \( y_i \) were generated on \( x_i = (i - 0.5)/100, \ i = 1, \ldots, 100 \), according to a true logit function \( f(x) = 3[10^5 x^{11} (1 - x)^6 + 10^3 x^2 (1 - x)^{10}] - 2 \). \( B = 500 \). 100 trials were conducted. Option U of the UBR method was used to select \( \lambda \) (see (2.1.12)). We also repeated Gu's experiment for Bayesian confidence intervals, using UBR to select \( \lambda \), which allows direct comparison with the bootstrap intervals here.
Figure 3.4.1: Coverage percentages bootstrap intervals. Pluses: sample means; dotted lines: nominal coverage.

The coverage percentage of 95% and 90% intervals are plotted in Figure 3.4.1. The Nor and the Piv intervals work better than the Per and the BC intervals and are similar to Bayesian intervals. The Nor has smaller variance. The pointwise coverages are plotted in Figure 3.4.2. The bootstrap intervals are similar to Bayesian confidence intervals in the sense that the pointwise coverage is smaller than the nominal value at high curvature points. The Nor intervals are a little better than the Piv intervals in terms of dropping less than the Piv intervals at high curvature points. The Nor or the Bayesian intervals would be recommended on the basis of this particular experiment. A “typical” case is plotted in Figure 3.4.3.

3.5 Approximate Bayesian Confidence Intervals for PSA Estimates

3.5.1 Bayes Model for PSA estimates

In section 1.4, we introduced the PSA method for non-Gaussian data. As in Section 3.3, we want to construct confidence intervals for the estimates of the overall function and all components in the model.

We first extend Gaussian posterior calculations to the case where the sampling errors are non iid. Recall that $M = \mathcal{H}^0 \oplus \sum_{\beta=1}^{q} \mathcal{H}^\beta$ is our model space, $\mathcal{H}^0 = \text{span}\{\phi_1, \ldots, \phi_M\}$,
Figure 3.4.2: Stars are pointwise coverage of 90% intervals. Circles are pointwise coverage of 95% intervals. Dotted lines are nominal values 90% and 95%. Dashed curves are the magnitude of $|\hat{f}|$. 
Figure 3.4.3: Display of the 90% intervals in a “typical” case. Stars: data; solid lines: true function; dotted lines: spline fit; dashed lines: confidence intervals.
\( R_\beta(x, z) \) is the RK for \( \mathcal{H}^3 \) and \( R_\Theta(x, z) = \sum_{\beta=1}^{q} \theta_\beta R_\beta(x, z) \). Let

\[
F_\xi(x) = \sum_{\nu=1}^{M} \tau_\nu \phi_\nu(x) + b^3 \sum_{\beta=1}^{q} \sqrt{\theta_\beta} Z_\beta(x),
\]

where \( \tau = (\tau_1, \ldots, \tau_M)^T \sim N(0, \xi I) \), \( Z_\beta \) are independent, zero mean Gaussian stochastic processes, independent of \( \tau \), with \( \text{E}[Z_\beta(x)Z_\beta(z)] = R_\beta(x, z) \). Let \( Z(x) = \sum_{\beta=1}^{q} \sqrt{\theta_\beta} Z_\beta(x) \), then \( \text{E}[Z(\mathbf{x})Z(\mathbf{z})] = R_\Theta(x, z) \).

Suppose observations have the form \( y_i = F_\xi(x_i) + \epsilon_i, \ i = 1, \ldots, n \), where \( \epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \sigma^2 W^{-1}) \), with \( W \) positive definite and known. Similar to Gu (1992b), with \( n\lambda = \sigma^2/b \), we have \( f_{\lambda, \Theta}(x) = \lim_{\xi \to \infty} \text{E}(F_\xi(x)|\mathbf{y}) \), where \( f_{\lambda, \Theta}(x) \) is the solution to

\[
\min \left\{ \left( \mathbf{y} - \mathbf{f} \right)^T W \left( \mathbf{y} - \mathbf{f} \right) + n\lambda \sum_{\beta=1}^{q} \theta_\beta^{-1} \left\| P_\beta f \right\|^2 \right\},
\]

in \( \mathcal{M} \). The calculations of posterior means and covariances are the same as in Gu and Wahba (1993) with their \( M = Q_\Theta + n\lambda I \) replaced by \( M = Q_\Theta + n\lambda W^{-1} \). Since (1.5.11) and (1.5.12) of Wahba (1990) hold for all invertible matrix \( M \), Theorem 1 of Gu and Wahba (1993) can be easily extended to our case. We simply list the posterior means and covariances in the following theorem.

**Theorem 3.3** Let \( M = Q_\Theta + n\lambda W^{-1} \). Let \( g_{0, \nu}(x) = \tau_\nu \phi_\nu(x) \) and \( g_\beta(x) = b^3 \sqrt{\theta_\beta} Z_\beta(x), \ \nu = 1, \ldots, M, \ \beta = 1, \ldots, q \). Then

\[
\text{E}(g_{0, \nu}(x)|\mathbf{y}) = d_\nu \phi_\nu(x),
\]

\[
\text{E}(g_\beta(x)|\mathbf{y}) = \sum_{i=1}^{n} c_{i, \beta} R_\beta(x, x_i),
\]

\[
\frac{1}{b} \text{Cov}(g_{0, \nu}(z), g_{0, \nu}(x)|\mathbf{y}) = \phi_\nu(z)\phi_\nu(x) e_\nu^T (S^T M^{-1} S)^{-1} e_\nu,
\]

\[
\frac{1}{b} \text{Cov}(g_\beta(z), g_{0, \nu}(x)|\mathbf{y}) = -d_\nu \phi_\nu(x),
\]

\[
\frac{1}{b} \text{Cov}(g_\beta(z), g_\beta(x)|\mathbf{y}) = \theta_\beta R_\beta(z, x) - \sum_{i=1}^{n} c_{i, \beta}(z)\theta_\beta R_\beta(x, x_i),
\]

\[
\frac{1}{b} \text{Cov}(g_\gamma(z), g_\beta(x)|\mathbf{y}) = -\sum_{i=1}^{n} c_{i, \gamma}(z)\theta_\beta R_\beta(x, x_i),
\]

where \( e_\nu \) is the \( \nu \)th unit vector, and \( (d_1, \theta_1(z), \ldots, d_M, \theta_M(z)) = d_\beta(z)^T \) and \( (c_1, \theta_1(z), \ldots, c_n, \theta_n(z)) = c_\beta(z)^T \) are given by

\[
d_\beta(z) = (S^T M^{-1} S)^{-1} S^T M^{-1} \begin{pmatrix} \theta_\beta R_\beta(z, x_1) \\ \vdots \\ \theta_\beta R_\beta(z, x_n) \end{pmatrix},
\]
Next, we try to approximate the posterior distribution based on non-Gaussian data using Laplace method (see Gu (1992b) and the references therein). Suppose the sampling likelihood of \( y \) is proportional to \( \exp \{- \frac{1}{\sigma^2} \mathcal{L}_y(f) \} \), where \( f^T = (f(x_1), \ldots, f(x_n)) \). \( \mathcal{L}_y(\cdot) \) is convex and completely specified, and \( \sigma^2 \) is a possibly unknown dispersion parameter. Let \( \zeta, \eta \) be any one of \( \tau, \phi(x) \), \( \tau, \phi(z) \), \( \sqrt{\theta_\alpha Z_\alpha(x)} \) or \( \sqrt{\theta_\alpha Z_\alpha(z)} \) for arbitrary points \( x \) and \( z \). Under the prior (3.5.1) for \( f \), with \( \xi \to \infty \), the joint probability density function of \( (y, \zeta, \eta, f, \tau) \) given \( b \) and \( \sigma^2 \) is

\[
P(f|y)\hat{q}(f|\tau)\hat{r}(\zeta, \eta|f),
\]

where

\[
p(f|y) \propto \exp \left\{ -\frac{1}{\sigma^2} \mathcal{L}_y(f) \right\},
\]

\[
\hat{q}(f|\tau) \propto \exp \left\{ -\frac{1}{2b} (f - S\tau)^T Q_0^{-1} (f - S\tau) \right\},
\]

and \( \hat{r}(\zeta, \eta|f) \) is Gaussian with mean and covariance given in Theorem 3.3 with \( \sigma^2 = 0 \) and \( y = f \). Integrating out \( \tau \) from \( \hat{q}(f|\tau) \) yields

\[
q(f) \propto \exp \left\{ -\frac{1}{2b} f^T (Q_0^{-1} - Q_0^{-1} S (S^T Q_0^{-1} S)^{-1} S^T Q_0^{-1}) f \right\}.
\]

The posterior density function of interest is

\[
\pi(\zeta, \eta|y) \propto \int p(f|y)q(f)r(\zeta, \eta|f)df.
\]

This integral is not easy to calculate since \( p(f|y) \) is not Gaussian. Both \( q(f) \) and \( r(\zeta, \eta|f) \) are Gaussian. Thus, we only need to approximate \( p(f|y) \) by a Gaussian density using the Laplace method. Expanding \( \log p(f|y) \) via a Taylor series centered at the mode \( f_* \) of \( p(f|y)q(f) \) and neglecting the cubic and higher order terms, we have an approximation \( \hat{p}(f|y) \) of \( p(f|y) \):

\[
\log \hat{p}(f|y) = -\frac{1}{2\sigma^2} (f - (f_* - W_*^{-1} u_*))^T W_* (f - (f_* - W_*^{-1} u_*)) + C,
\]

where \( u_* = (\partial \mathcal{L}/\partial f)|_{f_*} \), \( W_* = (\partial^2 \mathcal{L}/\partial f \partial f^T)|_{f_*} \), and \( C \) is a constant independent of \( f \). Notice that \( f_* \) does not depend on \( \zeta \) and \( \eta \). We approximate the posterior density function \( \pi(\zeta, \eta|y) \) by

\[
\hat{\pi}(\zeta, \eta|y) \propto \int \hat{p}(f|y)q(f)r(\zeta, \eta|f)df.
\]
Theorem 3.4 The approximate posterior density \( \hat{\pi}(\zeta, \eta | y) \) is Gaussian with mean and covariance given in Theorem 3.3.

[Proof] The approximate likelihood \( \hat{p}(f | y) \) is identical to a Gaussian sampling likelihood with covariance \( \sigma^2 W^{-1} \) and observations \( Y = f - W^{-1}u \). Thus, everything is the same as Theorem 3.3 except for replacing \( W \) by \( W^* \). Gu (1992b) proved that \( f^* \) is the fixed point of equation 1.45. Therefore, \( u^* \) and \( W^* \) are the same as \( u \) and \( W \) calculated at fixed point of equation 1.45.

To get the Bayesian confidence intervals, we need to calculate posterior covariances in Theorem 3.3. Gu and Wahba (1993) discussed how to calculate these quantities when \( W = I \). Let \( Q = W^{1/2}Q_0 W^{1/2} \), \( S = W^{1/2}S, R_\beta(x, x_i) = \sqrt{w_i} R_\beta(x, x_i) \), then \( \hat{M} = \hat{Q} + n\lambda I \). We can calculate \((\hat{S}^T \hat{M}^{-1} \hat{S})^{-1}, \hat{d}_\beta(x) \) and \( \hat{c}_\beta(x) \) the same way as Gu and Wahba (1993). We then have \((S^T M^{-1} S)^{-1} = (\hat{S}^T \hat{M}^{-1} \hat{S})^{-1}, d_\beta(x) = \hat{d}_\beta(x) \) and \( c_\beta(x) = W^{1/2} \hat{c}_\beta(x) \).

3.5.2 Simulations

In this section, we use the same 6 simulations in section 2.3 to evaluate the performance of the approximate Bayesian confidence intervals for PSA estimates. Bootstrap confidence intervals can also be constructed. We do not include them in our simulation since they are too computer intensive. According to the previous sections, we postulate that the bootstrap confidence intervals work similarly to these approximate Bayesian confidence intervals.

Smoothing parameters are chosen by the UBR method (see (2.1.12)). Box-plots of the coverage of the overall and component-wise functions are shown in Figure 3.5.1 for Case I and Case II and in Figure 3.5.2 for Case III. Notice that for Case I, two components \( f_1 \) and \( f_2 \) are linear and go through the point zero, so the confidence intervals cover nearly all or none of the true function (see Gu and Wahba (1993)). These box-plots indicate that usually the coverages of the overall functions are better than the coverage component-wise. The average coverages are not far from the nominal values but some components have big variances. “Typical” 90% Bayesian intervals for each case are shown in Figure 3.5.3 for Case I and Case II and Figure 3.5.4 for Case III. For Case III, “Typical” 90% Bayesian intervals for slices of the probability function as a function of \( x_2 \) are plotted in Figure 3.5.5 for \( n = 200 \) and Figure 3.5.6 for \( n = 400 \).

3.6 Conclusions

We have compared the performance of several versions of bootstrap confidence intervals with each other and with Bayesian confidence intervals. Bootstrap confidence intervals work as well as Bayesian intervals from an ACP point of view and appear to be better for small sample sizes. We find it reassuring that the best variations of
Figure 3.5.1: Coverage percentages of 95% Bayesian intervals. Pluses: sample means; dotted lines: nominal coverage.
Nominal Coverage: 95 %   Nominal Coverage: 90 %

Figure 3.5.2: Coverage percentages of 95% and 90% Bayesian intervals. Pluses: sample means; dotted lines: nominal coverage.
Figure 3.5.3: Display of the 90% intervals. Solid lines: true functions; dotted lines: n=200 intervals; dashed lines: n=400 intervals.
Figure 3.5.4: Display of the 90% intervals. Solid lines: true functions; dotted lines: n=200 intervals; dashed lines: n=400 intervals.
Figure 3.5.5: Display of the 90% intervals for $n = 200$. Solid lines: true functions; dashed lines: estimates; dotted lines: Bayesian intervals.
Figure 3.5.6: Display of the 90% intervals for \( n = 400 \). Solid lines: true functions; dashed lines: estimates; dotted lines: Bayesian intervals.
bootstrap confidence intervals and the Bayesian confidence intervals give such similar results. This similarity lends credence to both methods. The advantages of bootstrap confidence intervals are:

1) They are easy to understand, even by an unsophisticated user. They can be used easily with any distribution;
2) They appear to have better coverage in small samples in the examples tried.

The disadvantage of bootstrap confidence intervals is that they are very computer intensive, especially for SS ANOVA and non-Gaussian data. But compared to typical data collection costs, the cost of several minutes or even several hours of CPU time is small.

Just like Bayesian intervals, these bootstrap confidence intervals should be interpreted as across the curve, instead of pointwise.

Even though the bootstrap confidence intervals are essentially an automatic method, they should be implemented carefully. If the bootstrap method is used, we recommend using either T-I or Nor intervals for Gaussian data, and Nor intervals for Non-Gaussian data. The commonly used Per intervals work well, but are inferior to T-I or Nor intervals in our simulations. When bootstrapping for small sample sizes and using GCV to select smoothing parameter(s), one should exclude interpolating cases, especially when using T intervals.
Chapter 4

Analyses of Data From the Wisconsin Epidemiology Study of Diabetic Retinopathy

4.1 Introduction

Due to its flexibility, the PSA method can be used in several stages of data analyses. In the previous chapters, we explained the PSA method as a tool for model building and estimation. In practice, we can also use the PSA method to explore the behavior of raw data. For example, we can get a marginal estimate for each covariate to investigate whether a covariate has a marginal nonlinear effect. Alternatively, we can get some marginal estimates on pairs of 2 or more covariates to investigate their interactions. Further, the PSA method can be used as a diagnostic tool (Fowlkes, 1987).

In this chapter, we use data from the Wisconsin Epidemiology Study of Diabetic Retinopathy (WESDR) to demonstrate the PSA method. See Klein, Klein, Moss, Davis and DeMets (1988), Klein, Klein, Moss, Davis and DeMets (1989) and references therein for a detailed description of the data and some analyses using GLIM.

In brief, a sample of 2990 diabetic patients was selected in an 11-county area in southern Wisconsin. This sample was composed of two groups. The first group consisted of 1210 patients diagnosed as having diabetes before 30 years of age and who took insulin ("younger onset group"). The second group consisted of 1780 patients who had the diagnosis of diabetes made at 30 years of age or older. Of these, 824 were taking insulin ("older onset taking insulin group") and 956 were not ("older onset not taking insulin group"). Of the 2990 eligible patients, 2366 participated in the baseline examination from 1980 to 1982.

Several covariates are recorded. We only list the variables pertinent to our analyses:

1. Age: age in years at the time of baseline examination;
2. **Duration**: duration of diabetes at the time of baseline examination;

3. **Glycosylated hemoglobin**: a measure of hyperglycemia;

4. **Body mass index (BMI)**: weight in kg / (height in m)$^2$;

5. **Systolic blood pressure** in mmHg;

6. **Pulse rate** counted for 30 seconds;

7. **Baseline retinopathy severity levels** (base-level). See explanation below.

1878 patients participated in the follow-up examination (1984 to 1986). At the baseline and follow-up examinations, each eye was graded as one of the 6 levels: 10, 21, 31, 41, 51 and 60+, in order of increasing retinopathy severity with 10 indicating no retinopathy and 60+ indicating proliferative retinopathy. The retinopathy level for a participant was derived by giving the eye with the higher level greater weight. For example, the level for a participant with level 31 retinopathy in each eye is specified by the notation “level 31/31”, whereas that for a participant with level 31 in one eye and less severe retinopathy in the other eye is noted as “level 31/<31”. This scheme provided an 11-step scale: 10/10, 21/<21, 21/21, 31/<31, 31/31, 41/<41, 41/41, 51/<51, 51/51, 60+/<60+ and 60+/60+. **Incidence** of any retinopathy is defined to be 1 for a participant with level 10/10 at the baseline examination and level 21/<21 or worse at the follow-up examination, and 0 otherwise. **Progression** is defined to be 1 for a participant with nonproliferative or no retinopathy if the participant had their baseline level increased two steps or more (10/10 to 21/21 for instance), and 0 otherwise. Progression to **proliferative** retinopathy is defined to be 1 for a participant if the participant had no or nonproliferative retinopathy (level 51/51 or better) at the baseline and proliferative retinopathy (level 60+/<60+ or worse) at the follow-up, and 0 otherwise. Our aims are to find risk factors and to build models for prediction of incidence, of progression, and of progression to proliferative retinopathy. Since there are too few events of progression to proliferative retinopathy in the older onset groups, we do not analyze these two cases.

### 4.2 Estimation of the Prediction Errors

Before analyzing the WESDR data, we first discuss some methods for estimating the prediction errors. These estimates of prediction errors can be used to select among candidates the model which gives the best prediction. For simplicity, we limit our discussion to binary data. All methods discussed here can be easily generalized to data from exponential families.

Assume that the design points $x$’s are independent and identically distributed from an unknown distribution $F$ on $X$. Binary data $y$’s are generated from the conditional
probability function:

\[ p(x) = P(y = 1|x). \]  

(4.2.1)

Let \( \hat{p}(x) \) be an estimate of the true probability function \( p(x) \). The true Kullback-Leibler discrepancy (KL) and mean squared error (MSE) between \( p(x) \) and \( \hat{p}(x) \) are defined as

\[ KL_{true} = \mathbb{E}_{x \sim F} \left[ p(x) \log \frac{p(x)}{\hat{p}(x)} + (1 - p(x)) \log \frac{1 - p(x)}{1 - \hat{p}(x)} \right] \]  

(4.2.2)

and

\[ MSE_{true} = \mathbb{E}_{x \sim F} (p(x) - \hat{p}(x))^2, \]  

(4.2.3)

where the expectation is taken over \( x \), which has distribution \( F \).

We only discuss methods for estimating \( KL_{true} \) in the remainder of this section. All these methods can be used to estimate \( MSE_{true} \).

Denote the whole sample by \( S = \{ (y_i, x_i), i = 1, \ldots, n \} \). If the sample size \( n \) is large, we can divide \( S \) into two parts: the training sample \( S_{\text{train}} \) and the test sample \( S_{\text{test}} \). We use the training sample to fit a candidate model and get an estimate function \( \hat{p}(x) \). Then the maximum likelihood estimate ("honest" estimate) of \( KL_{true} \) based on the test sample is

\[ \overline{KL}_{test} = \frac{1}{|S_{\text{test}}|} \sum_{j \in S_{\text{test}}} \left[ y_j \log \frac{1}{\hat{p}(x_j)} + (1 - y_j) \log \frac{1}{1 - \hat{p}(x_j)} \right], \]  

(4.2.4)

where \( |S_{\text{test}}| \) denotes the sample size of \( S_{\text{test}} \). Notice that \( \overline{KL}_{test} \) is biased by the amount

\[ \frac{1}{|S_{\text{test}}|} \sum_{j \in S_{\text{test}}} \{ p(x_j) \log(p(x_j)) + (1 - p(x_j)) \log(1 - p(x_j)) \}, \]  

(4.2.5)

which depends only on the underlying function, not on the model used for fitting. Thus when selecting a model, we should look at the difference of \( \overline{KL} \)'s between two models.

In practice, often we do not have a large enough sample size to split it into training sample and test sample, but we still want to select a model among some candidates based on reasonable estimates of the prediction errors. Splitting into training and test samples affects the PSA estimates and thus favors of simple models when the sample size is small. Instead if we use \( S \) as both training and test sample, the estimate of \( KL_{true} \) is

\[ \overline{KL}_{app} = \frac{1}{n} \sum_{i=1}^{n} \left[ y_i \log \frac{1}{\hat{p}_i} + (1 - y_i) \log \frac{1}{1 - \hat{p}_i} \right]. \]  

(4.2.6)

We call such an estimate as the apparent estimate because the same observations are used for fitting and assessing the prediction performance. \( \overline{KL}_{app} \) generally underestimates
$KL_{true}$. Model selection based on $\overline{KL}_{app}$ generally favors complex models. Efron (1982) and Gong (1986) suggested using the jackknife or the bootstrap to estimate and correct the bias of the apparent estimate. More precisely, write $KL_{true} = KL_{\hat{F}, F}$, where $\hat{F}$ is the empirical distribution of $F$ that puts mass $1/n$ at each point $x_1, \cdots, x_n$. The first distribution in the subscript of $KL_{true}$ denotes the training sample while the second denotes the test sample. It is easy to see that $\overline{KL}_{app} = \overline{KL}_{\hat{F}, \hat{F}}$.

Define the excess error to be

$$R(\hat{F}, F) = KL_{\hat{F}, \hat{F}} - KL_{F, F}$$

and the expected excess error to be

$$\rho = E_{\hat{F} \sim F} R(\hat{F}, F),$$

where the expectation is taken over $\hat{F}$, which is obtained from a sample $S$ generated by $F$ and (4.2.1). A bootstrap procedure to estimate $\rho$ is obtained by replacing $F$ by $\hat{F}$:

$$\hat{\rho}_{boot} = E_{\hat{F} \sim \hat{F}} R(\hat{F}^*, \hat{F}),$$

where $\hat{F}^*$ is the empirical distribution of a random sample $x_1^*, \cdots, x_n^*$ from $\hat{F}$. That is, we draw from $S$ $n$ times with replacement and write the bootstrap sample as $S^* = \{(y_i^*, x_i^*), i = 1, \cdots, n\}$. Denote the estimate of $p(x)$ based on $S^*$ by $\hat{p}^*(x)$. The bootstrap estimate of the excess error is

$$R^* = KL_{\hat{F}^*, \hat{F}^*} - KL_{\hat{F}^*, \hat{F}}.$$  

$$\begin{align*}
R^* &= \frac{1}{n} \sum_{i=1}^{n} \left[ y_i \log \frac{1}{\hat{p}^*(x_i)} + (1 - y_i) \log \frac{1}{1 - \hat{p}^*(x_i)} \right] - \\
&\quad \frac{1}{n} \sum_{i=1}^{n} \left[ y_i^* \log \frac{1}{\hat{p}^*(x_i^*)} + (1 - y_i^*) \log \frac{1}{1 - \hat{p}^*(x_i^*)} \right]. 
\end{align*}$$

(4.2.10)

Repeat this procedure $B$ times to get $KL^{b}_{\hat{F}^*, \hat{F}^*}, KL^{B}_{\hat{F}^*, \hat{F}}$ and $R^{b}_1, \cdots, R^{B}_B$. The bootstrap estimate of $KL_{true}$ is

$$\overline{KL}_{boot} = \frac{1}{B} \sum_{b=1}^{B} KL^{b}_{\hat{F}^*, \hat{F}}$$

(4.2.11)

with the standard deviation $\tilde{s}(\overline{KL}_{boot})$ of the $KL^{b}_{\hat{F}^*, \hat{F}}$'s as an estimate of the standard error $s(\overline{KL}_{boot})$. The bootstrap estimate of the expected excess error is

$$\hat{\rho}_{boot} = \frac{1}{B} \sum_{b=1}^{B} R^{b}_b.$$  

(4.2.12)

The bias corrected estimate of $KL_{true}$ is

$$\overline{KL}_{bc} = \overline{KL}_{app} + \hat{\rho}_{boot}.$$  

(4.2.13)
Another way to correct the bias of $\hat{KL}_{app}$ is to jackknife. The jackknife procedure randomly divides $S$ into $W$ parts: $S_1, \ldots, S_W$. Let $\hat{F}^{(w)}$ be the empirical distribution of the design points in $S - S_w$, $w = 1, \ldots, W$. Define pseudo-values

$$U_w = W \times KL_{\hat{F}, \hat{F}} - (W - 1) \times KL_{\hat{F}^{(w)}, \hat{F}^{(w)}}, \quad w = 1, \ldots, W. \tag{4.2.14}$$

The jackknife estimate of $KL_{true}$ is

$$\hat{KL}_{jack} = \frac{1}{W} \sum_{w=1}^{W} U_w \tag{4.2.15}$$

with an estimated standard error

$$\hat{s}(\hat{KL}_{jack}) = \sqrt{\frac{1}{W(W-1)} \sum_{w=1}^{W} (U_w - \hat{KL}_{jack})^2}. \tag{4.2.16}$$

We can also use V-fold cross-validation to get an estimate of $KL_{true}$. That is, randomly divide $S$ into $V$ parts: $S_1, \ldots, S_V$. Let $\hat{F}^{(v)}$ be the empirical distribution of the design points in $S - S_v$ and $\hat{F}^{(v)}$ be the empirical distribution of the design points in $S_v$, $v = 1, \ldots, V$. The cross-validation estimate of $KL_{true}$ is

$$\hat{KL}_{CV} = \frac{1}{V} \sum_{v=1}^{V} KL_{\hat{F}^{(v)}, \hat{F}^{(v)}}. \tag{4.2.17}$$

This is a two layer cross-validation if a cross-validation method is used to choose the smoothing parameters when fitting a PSA model based on $S - S_v$. It is very difficult to get an explicit estimate of the standard deviation $s(\hat{KL}_{CV})$ of $\hat{KL}_{CV}$, but one can use a Jackknife or bootstrap method. To jackknife $\hat{KL}_{CV}$, we randomly divide $S$ into $S_1, \ldots, S_W$. For a fixed $w$, we randomly divide $S - S'_w$ into $V$ parts: $S'^{(v)}_1, \ldots, S'^{(v)}_V$. Define the pseudo-values

$$U'_w = W \times KL_{CV} - (W - 1) \times \frac{1}{V} \sum_{v=1}^{V} KL_{\hat{F}^{(v)}, \hat{F}^{(v)}}, \tag{4.2.18}$$

where $\hat{F}^{(v)}_w$ and $\hat{F}^{(v)}_w$ are empirical distribution functions based on $S - S'_w$. The jackknifed estimate of $KL_{true}$ is

$$\hat{KL}_{CV-jack} = \frac{1}{W} \sum_{w=1}^{W} U'_w \tag{4.2.19}$$

with estimated standard error

$$\hat{s}(\hat{KL}_{CV-jack}) = \sqrt{\frac{1}{W(W-1)} \sum_{w=1}^{W} (U'_w - \hat{KL}_{CV-jack})^2}. \tag{4.2.20}$$
The bootstrap method simply replaces \( \hat{F} \) and \( \hat{F}^{(v)} \) in (4.2.17) by \( \hat{F}^{*} \) and \( \hat{F}^{*(v)} \), the bootstrap empirical distribution function based on \( \hat{F} \). That is, a bootstrap sample \( S^* \) with sample size \( n \) is drawn from \( S \) with replacement. Repeat the cross-validation procedure on \( S^* \) to get \( KL_{CV}^* \). Repeat this process \( B \) times to get \( KL_{CV}^*(1), \ldots, KL_{CV}^*(B) \). The mean

\[
KL_{CV-\text{boot}} = \frac{1}{B} \sum_{b=1}^{B} KL_{CV}^*(b) \quad (4.2.21)
\]
gives another estimate of \( KL_{true} \) and the standard deviation of \( KL_{CV}^*(b) \)'s gives an estimate of \( s(KL_{CV}) \).

We have introduced several methods for estimating the prediction errors. Gong (1986) found that for a parametric model, bootstrap bias corrected estimates are better than the jackknife estimates. It would be interesting to compare the performance of these methods under nonparametric models. We leave this for future research.

### 4.3 Incidence in the Younger Onset Group

There were 256 observations in the younger onset group after deleting missing values and the cases with pre-existing retinopathy (21/<21 or worse). Klein et al. (1989) reported that duration, glycosylated hemoglobin and pressure are significant using GLIM. Their final model from GLIM is (we call it Model I):

\[
\logit(P(\text{duration, glycosylated hemoglobin, pressure})) = \mu + a_1 \times \text{duration} + a_2 \times \text{glycosylated hemoglobin} + a_3 \times \text{pressure}. \quad (4.3.1)
\]

To investigate whether a linear model is appropriate, we obtained PSA estimates for age, duration, glycosylated hemoglobin and pressure separately with smoothing parameters selected by Option U (see (2.1.12)). These marginal estimates and proportions within each decile on the logit scale are plotted in Figure 4.3.1. We can see from these plots that the marginal effects of age and duration are important and nonlinear. We put the quadratic and some higher order polynomial terms in age and duration into the GLIM and found that the quadratic term in age is significant with a \( p \) value of 0.0071. So our final GLIM model is (we call it Model II):

\[
\logit(P(\text{age, duration, glycosylated hemoglobin, pressure})) = \mu + a_1 \times \text{age} + a_2 \times \text{age}^2 + a_3 \times \text{duration} + a_4 \times \text{glycosylated hemoglobin} + a_5 \times \text{pressure}. \quad (4.3.2)
\]
Figure 4.3.1: Estimations of the marginal effects. Solid lines: PSA estimates; segments: the logit of proportions within each decile.
With GLIM, we also found that the interaction term $\text{duration} \times \text{pressure}$ is borderline significant with a $p$ value of 0.0863. So we have another candidate model from GLIM (we call it Model III):

$$
\logit(P(\text{age, duration, glycosylated hemoglobin, pressure})) = \mu + a_1 \times \text{age} + a_2 \times \text{age}^2 + a_3 \times \text{duration} + a_4 \times \text{glycosylated hemoglobin} + a_5 \times \text{pressure} + a_6 \times \text{duration} \times \text{pressure}.
$$

(4.3.3)

The shape of the marginal effect of age suggests that a parametric model with lower order polynomials may not be good enough. We fitted an additive model with age, duration, glycosylated hemoglobin and pressure. We found that only the age main effect is nonlinear while the other main effects are linear. Therefore we have the following additive model (we call it Model IV):

$$
\logit(P(\text{age, duration, glycosylated hemoglobin, pressure})) = \mu + f_1(\text{age}) + a_1 \times \text{duration} + a_2 \times \text{glycosylated hemoglobin} + a_3 \times \text{pressure}.
$$

(4.3.4)

We could also ask whether the parametric multiplicative interaction term $\text{duration} \times \text{pressure}$ counts all interactions between duration and pressure. We fitted a PSA model with main effects of age, duration, glycosylated hemoglobin, pressure and all interaction terms between duration and pressure. That is, the interaction terms include $\text{linear}(\text{duration}) \odot \text{linear}(\text{pressure})$, $\text{smooth}(\text{duration}) \odot \text{linear}(\text{pressure})$, $\text{linear}(\text{duration}) \odot \text{smooth}(\text{pressure})$ and $\text{smooth}(\text{duration}) \odot \text{smooth}(\text{pressure})$. We found that except for the $\text{smooth}(\text{duration}) \odot \text{linear}(\text{pressure})$ term, the other three interactions terms are very small in scale (with range at least 3 orders of magnitude smaller than that of $\text{smooth}(\text{duration}) \odot \text{linear}(\text{pressure})$). We also found that except for age, all main effects are linear. Thus we have the following model with interactions (we call it Model V):

$$
\logit(P(\text{age, duration, glycosylated hemoglobin, pressure})) = \mu + f_1(\text{age}) + f_2(\text{duration}) + a_1 \times \text{glycosylated hemoglobin} + a_2 \times \text{pressure} + a_3 \times \text{age} \times \text{pressure} + \text{smooth}(\text{duration}) \odot \text{linear}(\text{pressure}).
$$

(4.3.5)

We want to select a model from the above 5 candidate models which is simple and gives the best possible prediction. We used the methods discussed in section 4.2 to estimate $K_{\text{true}}$ and $MSE_{\text{true}}$. Since we only have 256 observations, it is not appropriate to split the sample into a training sample and a test sample. In Table 4.3.1 and Table 4.3.2, we list the apparent estimates $\overline{K}_{\text{app}}$ and $MSE_{\text{app}}$, the bootstrap estimates $\overline{K}_{\text{boot}}$ and $MSE_{\text{boot}}$ with their standard deviations, the bias corrected
estimates $\hat{KL}_{bc}$ and $\hat{MSE}_{bc}$ via the bootstrap, the jackknifed estimates $\hat{KL}_{jack}$ and $\hat{MSE}_{jack}$ with their standard deviations, the 256-fold cross-validation (leave-out-one cross-validation) estimates $\hat{KL}_{CV-1}$ and $\hat{MSE}_{CV-1}$, the 16-fold cross-validation (leave-out-sixteen cross-validation) estimates $\hat{KL}_{CV-16}$ and $\hat{MSE}_{CV-16}$. We jackknifed the 16-fold cross-validation estimate with $W = 16$. These estimates are listed as $\hat{KL}_{CV-jack}$ with standard deviation $\hat{s}(\hat{KL}_{CV-jack})$. We also bootstrapped the 16-fold cross-validation estimate with $B = 100$ and listed the estimates as $\hat{KL}_{CV-boot}$ with standard deviation $\hat{s}(\hat{KL}_{CV-boot})$.

The leave-out-one cross-validation is the most efficient way to use the data. We believe that the results are less biased than other estimates. So we compare other estimates with them. For this particular example, we see that the $\hat{KL}_{bc}$'s are almost as good as the $\hat{KL}_{CV-1}$'s. The $\hat{KL}_{jack}$'s and the $\hat{KL}_{boot}$'s still have large bias. This agrees with Gong (1986) that bias correction via the bootstrap works better than bias correction via the jackknife. The $\hat{KL}_{CV-16}$'s are fine, though not as good as the $\hat{KL}_{bc}$'s. The Jackknifed cross-validation estimates $\hat{KL}_{CV-jack}$'s are poor with big standard deviations. The bootstrapped cross-validation estimates $\hat{KL}_{CV-boot}$ are also poor with large standard deviations.

From Table 4.3.1 and Table 4.3.2, we can see that Model IV gives the best prediction. Model II gives almost as good as prediction as Model IV and is simpler. The KL and MSE values give the same conclusions.

The main effects of Model II and Model IV along with the 90% Bayesian and bootstrap confidence intervals of Model IV are shown in Figure 4.3.2 and Figure 4.3.3. Again, all main effects of Model II are inside the confidence intervals of Model IV, indicating that these models are almost the same based on these data. The bootstrap confidence intervals are not smooth since we used $B = 100$, which is small due to the intensity of computation time. The behavior of the bootstrap confidence intervals is similar to that of Bayesian confidence intervals for nonlinear components. But it is different for linear components since in this case, the Bayesian confidence intervals are almost simultaneous confidence intervals.

The main new conclusions from our analyses is that patients of ages between 20 and 30 are at higher risk of incidence of diabetic retinopathy than younger or older patients.

### 4.4 Estimation of Odds Ratios

The PSA estimate of the probability function for binary data can also be used to calculate the odds ratios. Suppose we have two points $x$ and $z$. The odds ratio of $x$ and $z$ equals $OR(x/z) = \exp(f(x) - f(z))$, where the function $f$ is the logit of the probability function. A natural estimation of $OR(x/z)$ is $\overline{OR}(x/z) = \exp(\hat{f}(x) - \hat{f}(z))$. Based on the Bayes model (3.5.1) and calculations in Section 3.5, we can approximate the posterior distribution of $f(x) - f(z)|y$ by a Gaussian distribution with mean $\hat{f}(x) - \hat{f}(z)$ and variance $\delta^2 = \text{Var}(f(x) - f(z)|y)$. $\delta^2$ can be calculated from the formulas
### Table 4.3.1: Estimates of $KL_{true}$.

<table>
<thead>
<tr>
<th>Model</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>$KL_{app}$</td>
<td>.5668</td>
<td>.5504</td>
<td>.5446</td>
<td>.5506</td>
<td>.5433</td>
</tr>
<tr>
<td>$KL_{boot}$</td>
<td>.5757</td>
<td>.5639</td>
<td>.5602</td>
<td>.5616</td>
<td>.5568</td>
</tr>
<tr>
<td>$\hat{s}(KL_{boot})$</td>
<td>.0073</td>
<td>.0096</td>
<td>.0103</td>
<td>.0084</td>
<td>.0092</td>
</tr>
<tr>
<td>$KL_{bc}$</td>
<td>.5881</td>
<td>.5806</td>
<td>.5782</td>
<td>.5778</td>
<td>.5810</td>
</tr>
<tr>
<td>$KL_{jack}$</td>
<td>.5778</td>
<td>.5673</td>
<td>.5638</td>
<td>.5674</td>
<td>.5687</td>
</tr>
<tr>
<td>$\hat{s}(KL_{jack})$</td>
<td>.0299</td>
<td>.0315</td>
<td>.0321</td>
<td>.0323</td>
<td>.0309</td>
</tr>
<tr>
<td>$KL_{CV-1}$</td>
<td>.5849</td>
<td>.5807</td>
<td>.5780</td>
<td>.5770</td>
<td>.5840</td>
</tr>
<tr>
<td>$KL_{CV-16}$</td>
<td>.5900</td>
<td>.5879</td>
<td>.5875</td>
<td>.5871</td>
<td>.6001</td>
</tr>
<tr>
<td>$KL_{CV-jack}$</td>
<td>.6384</td>
<td>.6807</td>
<td>.6833</td>
<td>.6349</td>
<td>.7080</td>
</tr>
<tr>
<td>$\hat{s}(KL_{CV-jack})$</td>
<td>.0324</td>
<td>.0337</td>
<td>.0351</td>
<td>.0346</td>
<td>.1711</td>
</tr>
<tr>
<td>$KL_{CV-boot}$</td>
<td>.5718</td>
<td>.5596</td>
<td>.5567</td>
<td>.5292</td>
<td>.5285</td>
</tr>
<tr>
<td>$\hat{s}(KL_{CV-boot})$</td>
<td>.0295</td>
<td>.0312</td>
<td>.0304</td>
<td>.0396</td>
<td>.0280</td>
</tr>
</tbody>
</table>

### Table 4.3.2: Estimates of $MSE_{true}$.

<table>
<thead>
<tr>
<th>Model</th>
<th>I</th>
<th>II</th>
<th>III</th>
<th>IV</th>
<th>V</th>
</tr>
</thead>
<tbody>
<tr>
<td>$MSE_{app}$</td>
<td>.1932</td>
<td>.1863</td>
<td>.1841</td>
<td>.1863</td>
<td>.1842</td>
</tr>
<tr>
<td>$MSE_{boot}$</td>
<td>.1961</td>
<td>.1907</td>
<td>.1894</td>
<td>.1904</td>
<td>.1881</td>
</tr>
<tr>
<td>$\hat{s}(MSE_{boot})$</td>
<td>.0022</td>
<td>.0029</td>
<td>.0033</td>
<td>.0030</td>
<td>.0032</td>
</tr>
<tr>
<td>$MSE_{jack}$</td>
<td>.1974</td>
<td>.1930</td>
<td>.1917</td>
<td>.1929</td>
<td>.1948</td>
</tr>
<tr>
<td>$\hat{s}(MSE_{jack})$</td>
<td>.0124</td>
<td>.0124</td>
<td>.0122</td>
<td>.0128</td>
<td>.0121</td>
</tr>
<tr>
<td>$MSE_{CV-jack}$</td>
<td>.2186</td>
<td>.2377</td>
<td>.2481</td>
<td>.2208</td>
<td>.2463</td>
</tr>
<tr>
<td>$\hat{s}(MSE_{CV-jack})$</td>
<td>.0139</td>
<td>.0137</td>
<td>.0133</td>
<td>.0142</td>
<td>.0629</td>
</tr>
<tr>
<td>$MSE_{CV-boot}$</td>
<td>.1948</td>
<td>.1892</td>
<td>.1878</td>
<td>.1761</td>
<td>.1796</td>
</tr>
<tr>
<td>$\hat{s}(MSE_{CV-boot})$</td>
<td>.0123</td>
<td>.0128</td>
<td>.0128</td>
<td>.0145</td>
<td>.0115</td>
</tr>
</tbody>
</table>
Figure 4.3.2: Estimates of the main effects of Model II and Model IV. Dashed lines are 90% Bayesian confidence intervals of Model IV.
Figure 4.3.3: Estimates of the main effects of Model II and Model IV. Dashed lines are 90% bootstrap confidence intervals of Model IV.
in Theorem 3.3. Hence we can approximate the distribution of \( OR(x/z) \) by a log Normal distribution and construct the \((1 - \alpha) \times 100\%\) Bayesian confidence interval as \((\hat{OR}(x/z) \exp(-z_\alpha \delta), \hat{OR}(x/z) \exp(z_\alpha \delta))\).

For example, from the Model IV of Section 4.3, we see that patients with age from 20 to 30 are at higher risk. To compare the risk at some particular ages, one may want to calculate the odds ratios at these ages. Suppose we fix duration, glycosylated hemoglobin and pressure at their median values, and pick age = 25 as the base and compare its risk to age = 10, 15, 35 and 40. The estimated odds ratios and their Bayesian confidence intervals (inside parentheses) are listed in Table 4.4.1. We see that the odds at age = 25 is significantly higher than the odds at age = 10. The odds at age = 25 is not significantly higher than the odds at age = 15, 35 and 40.

Table 4.4.1: Odds ratios of different ages at the median values of duration, glycosylated hemoglobin and pressure and their 95% Bayesian confidence intervals.

<table>
<thead>
<tr>
<th>Points</th>
<th>25/10</th>
<th>25/15</th>
<th>25/35</th>
<th>25/40</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{OR} )</td>
<td>2.44 (1,09,5.46)</td>
<td>1.43 (0.82,2.48)</td>
<td>1.58 (0.74,3.41)</td>
<td>2.22 (0.58,8.53)</td>
</tr>
</tbody>
</table>

We conducted simulations to evaluate the performance of the estimates of the odds ratios and their Bayesian confidence intervals. In the first two simulations, we used two univariate logit functions:

Case A
\[
f(x) = \frac{1}{2} \beta_{10,5}(x) + \frac{1}{2} \beta_{7,7}(x) + \frac{1}{2} \beta_{5,10}(x) - 1,
\]

Case B
\[
f(x) = 3[10^5 x^{11}(1-x)^6 + 10^5 x^5(1-x)^{10}] - 2,
\]

where \( \beta_{p,q} \) is the Beta function and \( 0 \leq x \leq 1 \). The true probability functions of these two cases are plotted in Figure 4.4.1. Bernoulli responses \( y_i \) were generated on grid points \( x_i = (i - 0.5)/n, \ i = 1, \ldots, n, \) according to the true probability function, where \( n \) is the sample size. Two sample sizes are used: \( n = 100 \) and \( n = 200 \). In Case A, we use \( x = 0.2 \) as the base and calculate odds ratios at points \( x = 0.4, x = 0.6, x = 0.8 \) and \( x = 1 \). In Case B, we use \( x = 0.5 \) as the base and calculate odds ratios at points \( x = 0.1, x = 0.2, x = 0.3 \) and \( x = 0.4 \). In the third simulation, we use the estimated probability function of Model IV in Section 4.3 as the true model. The design is the same as the data. We call it Case C. As in the above odds ratio calculations, we use age = 25 as the base and calculate odds ratios at points age = 10, age = 15, age = 35 and age = 40.

We repeated all three simulations 100 times. In Table 4.4.2, the true odds ratios are listed in rows as \( OR \); medians (since some estimates of \( OR \)'s are very big, the median is more robust than the mean) of the 100 estimates of the odds ratios and the
Table 4.4.2: Odds ratios, estimates of odds ratios and coverages of Bayesian confidence intervals.

<table>
<thead>
<tr>
<th>Case</th>
<th>0.4/0.2</th>
<th>0.6/0.2</th>
<th>0.8/0.2</th>
<th>1/0.2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.45</td>
<td>6.77</td>
<td>0.47</td>
<td>0.08</td>
</tr>
<tr>
<td></td>
<td>0.64 (0.73)</td>
<td>6.66 (34.69)</td>
<td>0.65 (1.08)</td>
<td>0.06 (0.56)</td>
</tr>
<tr>
<td>Coverage</td>
<td>86 (90)</td>
<td>89 (94)</td>
<td>81 (89)</td>
<td>92 (95)</td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.54 (0.38)</td>
<td>6.65 (10.57)</td>
<td>0.66 (0.44)</td>
<td>0.06 (0.13)</td>
</tr>
<tr>
<td>Coverage</td>
<td>90 (95)</td>
<td>95 (97)</td>
<td>81 (92)</td>
<td>92 (97)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case</th>
<th>0.5/0.1</th>
<th>0.5/0.2</th>
<th>0.5/0.3</th>
<th>0.5/0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n=100</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>12.42</td>
<td>4.81</td>
<td>1.73</td>
<td>1.09</td>
</tr>
<tr>
<td></td>
<td>8.43 (4388.82)</td>
<td>4.16 (647.01)</td>
<td>1.99 (44.86)</td>
<td>1.19 (2.59)</td>
</tr>
<tr>
<td>Coverage</td>
<td>82 (90)</td>
<td>89 (93)</td>
<td>98 (98)</td>
<td>99 (99)</td>
</tr>
<tr>
<td>n=200</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>10.27 (9.97)</td>
<td>4.31 (3.87)</td>
<td>1.95 (0.86)</td>
<td>1.20 (0.28)</td>
</tr>
<tr>
<td>Coverage</td>
<td>88 (94)</td>
<td>91 (95)</td>
<td>92 (94)</td>
<td>98 (100)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case</th>
<th>25/10</th>
<th>25/15</th>
<th>25/35</th>
<th>25/40</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>OR</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2.11 (1.26)</td>
<td>1.41 (0.48)</td>
<td>1.21 (0.68)</td>
<td>1.28 (2.03)</td>
</tr>
<tr>
<td>Coverage</td>
<td>84 (92)</td>
<td>93 (99)</td>
<td>73 (80)</td>
<td>75 (79)</td>
</tr>
</tbody>
</table>
standard deviations are listed in the rows labelled $\hat{OR}$ with standard deviations inside parentheses; the number of times in the 100 replications that the 90% and 95% Bayesian confidence intervals covered the true odds ratios are listed in the rows labelled Coverage with the 95% coverage number inside parentheses. We conclude from this table that the estimates of $OR$'s are generally biased toward unity if the true odds ratio is far from unity. This bias is severe if the sample size is small and one of the two points of $OR$ is at the peak and the other is at the valley (for instance, .4/.2 from Case A). The bias may be severe at points where data are scarce (for instance, age = 40 of Case C). The coverages of Bayesian confidence intervals at high bias points are lower than the nominal value, while the coverages are higher at other points. So these Bayesian confidence intervals behave similarly to the Bayesian confidence intervals for the estimates of probabilities on the logit scale.

One should consider the possible bias when interpreting the estimated odds ratios and their Bayesian confidence intervals. For example, from Table 4.4.1, the odds at age = 25 is not significantly higher than the odds at age = 40. But since age = 25 is a peak point, it is likely that $\hat{OR}$ underestimates the true $OR$. This is supported by our simulation, where $\hat{OR} = 2.22$ and the median of $\hat{OR} = 1.28$ in Table 4.4.2. We can estimate the bias of log odds by $\hat{bias} = \ln OR - \ln \hat{OR} = \ln 2.22 - \ln 1.28 = 0.55$ (this is actually the bootstrap estimate of the bias, but again $\hat{bias}$ may underestimate the true $bias$). Then we can correct this bias in the original estimate: $\hat{OR}_{\text{corrected}} = 2.22 \times \exp(0.55) = 3.85$. The 95% Bayesian confidence interval for the corrected estimate of odds ratio becomes (1.00, 14.78), which is just significant.
4.5 Progression of the Younger Onset Group

After deleting the missing values, we have 669 observations for progression of the younger onset group. Klein et al. (1988) reported that only the glycosylated hemoglobin is significant in a GLIM model. The marginal PSA estimates of age, duration, glycosylated hemoglobin and bmi with proportions within each decile on the logit scale are plotted in Figure 4.5.1. They show that the effects in age, duration and bmi are strong and nonlinear. Using GLIM with linear glycosylated hemoglobin and polynomials in age, duration, bmi and their combinations, we found that the quadratic term in age, the interaction $age^2 \times duration$ and the interaction $age^2 \times bmi$ are significant.

Age, duration and bmi are highly correlated with each other. Also, since age must be greater than duration and age at diagnosis (age - duration) must be smaller than 30, the data are inside a band (see Figure 4.5.12 next). We decided to entertain three models: a model with age, glycosylated hemoglobin and bmi (Model I):

$$
\text{logit} \left( P(\text{age, glycosylated hemoglobin, bmi}) \right) = \mu + f_1(\text{age}) + f_2(\text{glycosylated hemoglobin}) + f_4(\text{bmi}) + f_{14}(\text{age, bmi}),
$$

(4.5.1)
a model with duration, glycosylated hemoglobin and bmi (Model II):

$$
\text{logit} \left( P(\text{duration, glycosylated hemoglobin, bmi}) \right) = \mu + f_3(\text{duration}) + f_3(\text{glycosylated hemoglobin}) + f_4(\text{bmi}) + f_{24}(\text{duration, bmi}),
$$

(4.5.2)
and a model with age, duration, glycosylated hemoglobin, bmi (model III):

$$
\text{logit} \left( P(\text{age, duration, glycosylated hemoglobin, bmi}) \right) = \mu + f_1(\text{age}) + f_2(\text{duration}) + f_3(\text{glycosylated hemoglobin}) + f_4(\text{bmi}) + f_{12}(\text{age, duration}) + f_{14}(\text{age, bmi}),
$$

(4.5.3)
where

$$
f_{12}(\text{age, duration}) = \text{linear}(\text{age}) \otimes \text{linear}(\text{duration}) + \\
\text{smooth}(\text{age}) \otimes \text{linear}(\text{duration}) + \\
\text{linear}(\text{age}) \otimes \text{smooth}(\text{duration}) + \\
\text{smooth}(\text{age}) \otimes \text{smooth}(\text{duration}),
$$

and similarly for $f_{14}(\text{age, bmi})$. We had 10 smoothing parameters to be estimated for model (4.5.3). We think this is probably the most smoothing parameters we can estimate with this sample size. In practice, we can reduce the number of smoothing
Figure 4.5.1: Estimations of the marginal effects. Solid lines: PSA estimates; segments: the logit of proportions within each decile.
parameters by eliminating some of the smooth terms. For example, from other analyses, we know that the effect of glycosylated hemoglobin is linear. We can replace $f_3$ (glycosylated hemoglobin) by a linear function in the model. We can eliminate one by one some of the smooth interaction terms which are almost zero in scale. The results after eliminating some smooth terms are similar to that of model (4.5.3).

For model I, we plot the age vs bmi on the left of Figure 4.5.2. Those patients who had progression of retinopathy are marked as stars (*) and those with no progression are marked as circles (o). We superimpose the contour lines of estimated posterior standard deviations at the median glycosylated hemoglobin. These contours agree well with the distribution of the observations. Thus they can be used to delineate a region in which the estimation of the probability function is deemed to be reliable. We decide to use the region with estimated posterior standard deviations less than or equal to 0.5. The probability function estimation at the median glycosylated hemoglobin is plotted on the right of Figure 4.5.2.

To see how the probability of progression depends on age and bmi, we plot the cross sections of the estimate as a function of age (Figure 4.5.3) and as a function of bmi.
Figure 4.5.3: Model I. Cross sections estimated probability of progression as a function of age, at four levels of glycósylated hemoglobin (gly) and four levels of bmi. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.5.4: Model I. Cross sections estimated probability of progression as a function of bmi, at four levels of glycosylated hemoglobin (gly) and four levels of age. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.5.5: Model I. Cross sections of estimated probability of progression as a function of age with their 90% Bayesian confidence intervals, at three levels of glycosylated hemoglobin (gly) and three levels of bmi. Low, median and high denote .25, .5 and .75 percentiles.
Figure 4.5.6: Model I. Cross sections of estimated probability of progression as a function of bmi with their 90% Bayesian confidence intervals, at three levels of glycosylated hemoglobin (gly) and three levels of age. Low, median and high denote .25, .5 and .75 percentiles.
Figure 4.5.7: Model II. Left: data and contours of constant posterior standard deviation at the median glycosylated hemoglobin as a function of duration and bmi. An * indicates an progression and a circle indicates a non-progression. Right: estimated probability in the defined region, as a function of duration and bmi at the median value of glycosylated hemoglobin.

(Figure 4.5.4). We also plot the cross sections along with their 90% Bayesian confidence intervals as a function of age (Figure 4.5.5) and as a function of bmi (Figure 4.5.6). Like incidence, we see that patients with age between 20 and 30 are at higher risk of progression. The risk of progression increases with the increase of body mass index and then tend to remain flat after bmi = 25. The estimated probabilities for bmi larger than 25 are not reliable since there are only a few observations there. This is demonstrated by the Bayesian confidence intervals becoming wider.

For model II, we plot the duration vs bmi on the left of Figure 4.5.7. Those patients who had progression of retinopathy are marked as stars (*) and those with no progression are marked as circles (o). We superimpose the contour lines of estimated posterior standard deviations at the median glycosylated hemoglobin. These contours agree well with the distribution of the observations. We decide to use the region with estimated posterior standard deviations less then or equal to 0.5. The probability function estimation at the median glycosylated hemoglobin is plotted on the right of Figure 4.5.7.

To see how the probability of progression depends on duration and bmi, we plot the
Figure 4.5.8: Model II. Cross sections estimated probability of progression as a function of duration, at four levels of glycosylated hemoglobin (gly) and four levels of bmi. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.5.9: Model II. Cross sections estimated probability of progression as a function of bmi, at four levels of glycosylated hemoglobin (gly) and four levels of duration. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.5.10: Model II. Cross sections of estimated probability of progression as a function of duration with their 90% Bayesian confidence intervals, at three levels of glycosylated hemoglobin (gly) and three levels of bmi. Low, median and high denote .25, .5 and .75 percentiles.
Figure 4.5.11: Model II. Cross sections of estimated probability of progression as a function of bmi with their 90% Bayesian confidence intervals, at three levels of glycosylated hemoglobin (gly) and three levels of duration. Low, median and high denote .25, .5 and .75 percentiles.
cross sections of the estimate as a function of duration (Figure 4.5.8) and as a function of bmi (Figure 4.5.9). We also plot the cross sections along with their 90% Bayesian confidence intervals as a function of duration (Figure 4.5.10) and as a function of bmi (Figure 4.5.11). Like incidence, the risk of progression increases with the increased duration up to about 15 years and then decreases. The estimated probabilities for duration bigger than 30 are not reliable since the Bayesian confidence intervals are too wide. The risk of progression increases with the increase of body mass index and then tend to remain flat after bmi = 25. The estimated probabilities for bmi larger than 25 are not reliable.

For model III, we plot the age vs duration in Figure 4.5.12. Those patients who had progression of retinopathy are marked as stars (*) and those with no progression are marked as circles (o).

To see how the probability of progression depends on age, duration and bmi, we plot the cross sections of the estimate at the median glycylated hemoglobin as a function of age (Figure 4.5.13), as a function of duration (Figure 4.5.14), and as a function of bmi (Figure 4.5.15). Notice that some parts of the cross sections are cut so that the joint age and duration is inside the band. We also plot cross sections along with their 90% Bayesian confidence intervals as a function of age (Figure 4.5.16), as a function of duration (Figure 4.5.17), and as a function of bmi (Figure 4.5.18). We see that patients with age between 20 and 30 are at higher risk of progression similar to incidence. The risk of progression increases with the increased duration up to about 15 years and then tend to remain flat. The estimated probabilities for duration bigger than 15 are not reliable since the Bayesian confidence intervals are too wide. The risk of progression increases with the increase of body mass index and then tend to remain flat after bmi = 25. Again, the estimated probabilities for bmi larger than 25 are not reliable.

All three models give the same conclusions about the risk factors. There are some differences between the shapes of estimates from three models, especially for duration. The decrease after duration $= 15$ from model II may be caused by the fact that we omitted the age variable, which is highly correlated with duration and the older patients are at lower risk.

### 4.6 Progression to Proliferative Retinopathy in the Younger Onset Group

After deleting the missing values, we have 677 observations. Six observations with duration bigger than 39.2 were deleted since they are far away from the remaining observations and thus are highly influential. We decided to delete these six observations by looking at a histogram of duration. A more objective method to detect influential points and outliers is needed and will be one of our topics for future research. See Eubank (1985) for a treatment of Gaussian data and Yandell and Green (1986). The
Figure 4.5.12: Model III. Data and contours of constant posterior standard deviation at the median glycosylated hemoglobin and median bmi, as a function of age and duration. An * indicates an progression and a circle indicates a non-progression. Two dotted lines are age - duration = 0 and age - duration = 30.
Figure 4.5.13: Model III. Cross sections estimated probability of progression at the median glycosylated hemoglobin as a function of age, at four levels of duration and four levels of bmi. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.5.14: Model III. Cross sections estimated probability of progression at the median glycosylated hemoglobin as a function of duration, at four levels of age and four levels of bmi. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.5.15: Model III. Cross sections of estimated probability of progression at the median glycosylated hemoglobin as a function of bmi, at four levels of age and four levels of duration. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.5.16: Model III. Cross sections of estimated probability of progression at the median glycosylated hemoglobin as a function of age with their 90% Bayesian confidence intervals, at two levels of duration and two levels of bmi. q2 and q3 are the quantiles at .375 and .625.
Figure 4.5.17: Model III. Cross sections of estimated probability of progression at the median glycosylated hemoglobin as a function of duration with their 90% Bayesian confidence intervals, at two levels of age and two levels of bmi. q2 and q3 are the quantiles at .375 and .625.
Figure 4.5.18: Model III. Cross sections of estimated probability of progression as a function of bmi with their Bayesian confidence intervals, at two levels of age and two levels of glycosylated hemoglobin. Low, median and high denote .25, .5 and .75 percentiles.
baseline level is divided into 4 categories: 10/10, 21/<21 and 21/21 as category 1, 31/<31 and 31/31 as category 2, 41/<41 and 41/41 as category 3, 51/<51 and 51/51 as category 4. Therefore, the variable base-level is a categorical variable. We use this example to illustrate how to deal with categorical variables.

Klein et al. (1989) found that glycosylated hemoglobin and base-level are significant without deleting the six observations using GLIM. We found that in addition to the above covariates, polynomials up to third order of duration are significant with or without deleting the six observations using GLIM. We decided to fit the model:

$$\text{logit}(P(\text{duration, glycosylated hemoglobin, base-level})) = \mu + f_1(\text{duration}) + a \times \text{glycosylated hemoglobin} + \sum_{k=2}^{4} \gamma_k I_k(\text{base-level}),$$

where $I_k(\text{base-level})$ is an indicator function which is 1 if base-level equals to category k and 0 otherwise. Notice that we set $\gamma_1 = 0$. We create three dummy variable for categories 2 to 4 and combine these dummy variable into the regression matrix $S$.

The estimates of the main effects and their 90% Bayesian confidence intervals are plotted in Figure 4.6.1. The estimates with duration greater than 30 are not reliable from the confidence intervals. The estimates of $\gamma_2, \gamma_3, \gamma_4$ and their posterior standard deviations (inside parentheses) are $2.18(0.55), 3.14(0.56)$ and $3.91(0.74)$. The main new conclusion from this analysis is that the risk of progression to proliferative retinopathy increases with increased duration up to about 13 years and decreases a little bit thereafter. It indicates that if a patient does not progress to proliferative retinopathy with 13 years of diagnosis, his/her risk is not likely to increase any more, although remaining high.

### 4.7 Incidence of the Older Onset Taking Insulin Group

After deleting the missing values, we have 143 observations. Klein et al. (1989) found that age and glycosylated hemoglobin are significant using a GLIM model. We found that quadratic terms of duration and pulse are significant using a GLIM model:

$$\text{logit}(P(\text{age, duration, glycosylated hemoglobin, pulse})) = \mu + a_1 \times \text{age} + a_2 \times \text{duration} + a_3 \times \text{duration}^2 + a_4 \times \text{glycosylated hemoglobin} + a_5 \times \text{pulse} + a_6 \times \text{pulse}^2.$$  

Fitting a PSA model with main effects of age, duration, glycosylated hemoglobin and pulse, we found that the main effects of age and glycosylated hemoglobin are
linear. Then we fitted the model:

$$\text{logit}(P(\text{age, duration, glycosylated hemoglobin, pulse}))$$
$$= \mu + a_1 \times \text{age} + f_1(\text{duration}) + a_2 \times \text{glycosylated hemoglobin} +$$
$$f_2(\text{pulse}).$$

(4.7.2)

The estimates of the main effects and their 90% Bayesian confidence intervals are plotted in Figure 4.7.1. We believe the ups and downs in the middle of the main effect of duration are caused by the poor choice of the smoothing parameter (too small) due to the small sample size. But the pattern of the main effect of duration is reliable and agrees with the previous conclusion. That is, the risk increases with the increasing duration up to about 6 years and does not increase any more thereafter. From the main effect of pulse, we conclude that the risk is higher for patients with pulse rate that is too fast. The fits from model (4.7.1) are well inside the Bayesian confidence intervals of the PSA estimates. So the model (4.7.1) is appropriate except the part at the low pulse rate. It is difficult to distinguish between these two models with such a small sample size.

4.8 Progression of the Older Onset Taking Insulin Group
Figure 4.7.1: Estimates of the main effects of incidence for older onset taking insulin. Dashed lines are 90% Bayesian confidence intervals of PSA estimates.
After deleting the missing values, we have 374 observations. Klein et al. (1988) found that age, duration, glycosylated hemoglobin and base-level are significant using a GLIM model. They concluded that the risk of progression is higher if duration is longer and if base-level is lower. We find that the effect of age is nonlinear (significant up to 4th order of polynomials in a GLIM model). The effect of bmi is borderline significant and nonlinear (a 4th order polynomial has a p value of 0.0609). We first divided base-level into 5 categories: 10/10 as category 1, 21/<21 and 21/21 as category 2, 31/<31 and 31/31 as category 3, 41/<41 and 41/41 as category 4, 51/<51 and 51/51 as category 5. We found that the difference of the effects of category 1 and of category 5 were not significant. Thus we combined them into category 1. So base-level has 4 categories.

We first fit a PSA model with main effects of age, duration, glycosylated hemoglobin, bmi and base-level, we found that the main effect of age and glycosylated hemoglobin are linear. Then we fitted the model:

$$
\text{logit} \left( P(\text{age, duration, glycosylated hemoglobin, bmi, base-level}) \right) = \mu + a_1 \times \text{age} + f_1(\text{duration}) + a_2 \times \text{glycosylated hemoglobin} + f_2(\text{bmi}) + \sum_{k=2}^{4} \gamma_k I_k(\text{base-level}).
$$

(4.8.1)

The estimates of the main effects and their 90% Bayesian confidence intervals are plotted in Figure 4.8.1. The estimates of $\gamma_2, \gamma_3, \gamma_4$ and their posterior standard deviations (inside parentheses) are 0.77(0.33), -0.64(0.34) and -1.00(0.37). Our conclusions basically agree with Klein et al. (1988) except some modifications:

1. The risk of progression increases with the increasing duration up to around 8 years and does not increase after that;

2. Body mass index has a moderate effect on progression. The risk increases with increasing bmi;

3. Baseline level has a significant effect on progression, but not monotonically. A patient with baseline level 10/10 has same risk of a patient with baseline level 51/<51 or 51/51. A patient with baseline level 10/10 has higher risk than a patient with baseline level 31/<31 to 41/41. A patient with baseline level 10/10 has lower risk than a patient with baseline level 21/<21 or 21/21. This may reflect the arbitrary division of the levels which may not be evenly spaced.

4.9 Incidence of the Older Onset Not Taking Insulin Group
Figure 4.8.1: Solid lines: main effects estimates of progression for the older onset taking insulin group. Dashed lines are 90% Bayesian confidence intervals.
After deleting the missing values, there were 297 observations. One observation of glycosylated hemoglobin is 23.6, which is much bigger than the others. We decided to delete this observation.

Klein et al. (1988) found that glycosylated hemoglobin is significant in a GLIM model. Using GLIM, we found that the effect of age is not linear and that there is a strong interaction between age and glycosylated hemoglobin.

We first fit a PSA model with the main effects of age and glycosylated hemoglobin and all their interaction terms. The main effect of glycosylated hemoglobin is linear. All interaction terms except smooth(age) × linear(glycosylated hemoglobin) are near zero. The final model is

\[
\begin{align*}
\text{logit } (P( \text{age, glycosylated hemoglobin } )) \\
= & \quad \mu + f_1(\text{age}) + a_1 \times \text{glycosylated hemoglobin} + \\
& \quad \text{smooth}(\text{age}) \times \text{linear}(\text{glycosylated hemoglobin}).
\end{align*}
\]  

(4.9.1)

The main effects and interaction of age and glycosylated hemoglobin are plotted in Figure 4.9.1.

We plot age vs glycosylated hemoglobin in Figure 4.9.2. Those patients who had an incidence of retinopathy are marked as *'s and those with no incidence are marked as circles. We superimpose the contour lines of estimated posterior standard deviations. These contours agree well with the distribution of the observations. Thus they can be used to delineate a region in which the estimate of the probability function is deemed to be reliable. We decided to use the region with estimated posterior standard deviations less then or equal to 1. The probability function estimate is plotted in Figure 4.9.3. We see that the risk is decreased if a patient is older and has higher glycosylated hemoglobin. This may be caused by mortality since such a patient is more likely to die during the observation period.

4.10 Progression of the Older Onset Not Taking Insulin Group

After deleting the missing values, there were 432 observations. Klein et al. (1988) found that age, duration and glycosylated hemoglobin are significant in a GLIM model. Using GLIM, we found that the duration main effect of polynomials up to the cubic is significant and the bmi main effect of polynomials up to the quartic is significant. We also found that the multiplication interaction between age and polynomials up to cubic of duration is significant. These indicate that a GLIM with lower order of polynomials may not be good enough. We decided to fit a PSA model:

\[
\begin{align*}
\text{logit } (P( \text{age, duration, glycosylated hemoglobin, bmi } ))
\end{align*}
\]
Figure 4.9.1: Main effects and interaction of age and glycosylated hemoglobin.
Figure 4.9.2: Data and contours of constant posterior standard deviation as a function of age and glycosylated hemoglobin. An * indicates an incidence and a circle indicates a non-incidence.
Figure 4.9.3: Estimated probability of incidence in the defined region, as a function of age and glycosylated hemoglobin.
\[
= \mu + f_1(\text{age}) + f_2(\text{duration}) + f_{1,2}(\text{age, duration}) + f_3 \times \text{glycosylated hemoglobin} + f_4(\text{bmi}). \tag{4.10.1}
\]

The main effects estimates are plotted in Figure 4.10.1. We see that the effects of glycosylated hemoglobin is the strongest and linear. The effect of bmi is moderate and linear. The effects of glycosylated hemoglobin and bmi are additive on logit scale. In the remaining plots, we fix glycosylated hemoglobin and bmi at their median values. The interaction estimate between age and duration is plotted in Figure 4.10.2.

We plot the age vs duration in Figure 4.10.3. Those patients who had progression of retinopathy are marked as *'s and those with no progression are marked as circles. We superimpose contour lines of the estimated posterior standard deviations as a function of age and duration with glycosylated hemoglobin and bmi fixed at their median values. These contours agree well with the distribution of the observations. We decided to use the region with estimated posterior standard deviations less than or equal to 0.5. The probability function estimate is plotted in Figure 4.10.4.

To see more clearly how the probability of progression depends on age and duration, we plot the cross sections of the estimate in Figure 4.10.5. The cross sections with their 90% Bayesian confidence intervals are plotted in Figure 4.10.6 and Figure 4.10.7. From these plots, we see that the risk decreases the increasing age and the shapes are different between the fourth quantile and other quantiles of duration due to interaction. The risk increases with increasing duration up to about 17 years and does not increase after that. The shape of the 4th quantile of age is different from the others.

### 4.11 Conclusions

We summarize the new findings from our analyses as follows:

1. **Age** effects are nonlinear and different for different groups. For incidence and progression in the younger onset group, the patients with age between 20 and 30 are at higher risk than younger or older patients. For progression of the older onset group, the risk decreases with increasing age. Since mortality may change the population under study, especially for the older onset group, further research is needed to verify that this decrease in age effect is real.

2. The effect of **duration** is consistent in all groups. The risk increases with increasing duration up to a certain point and does not increase after that. This indicates that if a patient has not had an event (incidence, progression or progression to proliferative) after some years of diabetes, his risk will not substantially increase after that, though it remains higher than new patients.

3. The effect of **glycosylated hemoglobin** is the strongest effect of those covariates studied and is linear, which agrees with the previous study. We find that there is a
Figure 4.10.1: Main effects estimates for progression of older onset not taking insulin.
Figure 4.10.2: Estimate of interaction between age and duration for progression of older onset not taking insulin.
Figure 4.10.3: Data and contours of constant posterior standard deviation as a function of age and duration at the median value of glycosylated hemoglobin and bmi.
Figure 4.10.4: Estimated probability of progression in the defined region, as a function of age and duration at the median value of glycosylated hemoglobin and bmi.
Figure 4.10.5: Cross sections of estimated probability of progression as a function of age and duration at the median value of glycosylated hemoglobin and bmi. q1, q2, q3 and q4 are the quantiles at .125, .375, .625 and .875.
Figure 4.10.6: Cross sections of estimated probability of progression as a function of age with their Bayesian confidence intervals, at four quantiles of duration and at the median value of glycosylated hemoglobin and bmi.
Figure 4.10.7: Cross sections of estimated probability of progression as a function of duration with their Bayesian confidence intervals, at four quantiles of duration and at the median value of glycosylated hemoglobin and bmi.
strong interaction between age and glycosylated hemoglobin for the incidence in the older onset not taking insulin group.

4. Body mass index has a moderate effect on progression of all groups. The risk increases with increasing BMI.

5. Pulse rate has a moderate effect on the incidence of retinopathy in older onset persons taking insulin. The patients with too high pulse rate are at higher risk.

6. Baseline retinopathy level has a significant and monotone effect on progression to proliferative retinopathy, which agrees with the previous study. We also find that the baseline retinopathy level has a significant effect on progression. Patients with baseline levels 21/<21 and 21/21 tend to have a higher risk and patients with baseline level 31/<31 to 41/41 tend to have a lower risk of progression. This may reflect the fact that the division into levels is not evenly spaced.
Chapter 5

PSA Models for Matched Case-Control Study

5.1 Introduction

A case-control study is described by Breslow and Day (1980) as the “backbone of epidemiology”. In a case-control study, some persons who have a specific disease (the cases) and some comparable persons who do not have the disease (the controls) are selected. Certain exposure information is obtained retrospectively. Epidemiologists are often interested in the predictive relationship between potential risk factors and prevalence of this disease.

Case-control studies are frequently conducted using a matched design. Often certain exposure variables are known a priori to affect prevalence but are of limited interest. To reduce confounding with other risk factors, these exposure variables are used as matching variables. That is, for a particular case, several controls are randomly selected from a pool of so that matching variables agree with the case’s. For simplicity of notation, we will assume that the same number of controls are selected for each case, but our methods apply to the general case where each matched set has different numbers of controls. Our methods also apply to situations where conditional likelihood should be used such as very fine stratification.

Suppose that we have $K$ matched sets and each matched set has one case and $R$ controls. So we have $n = K \times (R+1)$ observations in total. Define the response variable as

$$y_{rk} = \begin{cases} 1 & \text{if } r = 0, \\ 0 & \text{otherwise}, \end{cases} \quad r = 1, \cdots, R; \quad k = 1, \cdots, K.$$  

Denote the set of exposure variables by $x = (x_1, \cdots, x_d)$. We assume $x_i \in X^{(i)}$, where $X^{(i)}$ is a measurable space. The spaces $X^{(i)}$’s are very general in our setting. Each $X^{(i)}$ may be discrete, continuous or a region in an Euclidean space. Let $x_{0k}$ be the exposure
vector of the case in the $k$th matched set, and let $x_{rk}$ be the exposure vector of the $r$th control in the $k$th matched set.

Like Breslow and Day (1980) and Hastie and Pregibon (1988), we assume the model

$$\text{logit}(P(Y = 1|x)) = \alpha + f(x),$$

(5.1.1)

where $f \in \mathcal{M}$. $\mathcal{M}$ is a model space which we will discuss in the next section. The log odds ratio of an individual with exposure $x$ developing the disease relative to an individual with exposure $x'$ is simply $f(x) - f(x')$.

As discussed in Breslow and Day (1980), by ignoring matching, the unconditional logistic regression model has a large bias in estimation for matched case-control studies. Instead, the conditional probability that one observes the exposure vector $x_{ok}$ for the $k$th case and $x_{rk}$'s for the matched controls should be used to construct the likelihood. Like Breslow and Day (1980), we can represent this conditional probability as

$$\mu_{ok} = \frac{\exp(f(x_{0k}))}{\sum_{r=0}^{R} \exp(f(x_{rK}))}. \quad (5.1.2)$$

Let $f = (f(x_{01}), \ldots, f(x_{R1}), \ldots, f(x_{0K}), \ldots, f(x_{RK}))$. The full conditional log likelihood is

$$l(f) = \sum_{k=1}^{K} l_k, \quad (5.1.3)$$

where $l_k = f(x_{0k}) - \log(\sum_{r=0}^{R} \exp(f(x_{rK})))$.

Contingency tables and linear logistic regression models are the most commonly used methods for analyzing case-control studies. Contingency tables have the advantage that no restricted parametric models are assumed. But they become messy when dealing with several exposure factors and their interactions. A linear logistic regression model assumes that

$$f(x) = x_1 \beta_1 + \cdots + x_d \beta_d. \quad (5.1.4)$$

Maximum likelihood is used to estimate the coefficients $\beta_1, \cdots, \beta_d$. Linear logistic regression models can deal with several exposure factors and their interactions. They are easy to interpret. But in some applications, the rigid linear approximation is not appropriate or needs to be verified. To relax this linear assumption, Hastie and Pregibon (1988) used an additive model:

$$f(x) = f_1(x_1) + \cdots + f_d(x_d), \quad (5.1.5)$$

where $f_j(x_j)$ is a smooth function. They estimated each single function $f_j$ by solving a penalized conditional likelihood minimization problem. They used the backfitting algorithm to iteratively estimate all such functions.
In this chapter, we try to extend Hastie and Pregibon’s model by allowing interactions between exposure factors. Instead of using the backfitting algorithm, we solve a single penalized conditional likelihood minimization problem with multiple smoothing parameters. Generalized cross-validation or unbiased risk estimation is used to empirically assess the amount of smoothing at each Newton-Raphson iteration. In section 5.2, we will discuss the model space. We estimate $f$ by solving a minimization problem in the penalized conditional likelihood. We propose an algorithm based on Newton-Raphson iteration. Special structures of the weight matrix and observations are used to simplify the equation at each Newton-Raphson iteration. RKPACK is used to solve the simplified equation. The performance of this algorithm is compared to Hastie and Pregibon (1988)’s delta algorithm. In section 5.3, we construct Bayesian confidence intervals for the estimates of the overall function $f$ and its components. In section 5.4, we conduct simulations to evaluate the performance of the estimation and its Bayesian confidence intervals. In section 5.5, we apply our method to a case-control study of breast cancer.

5.2 Formulation and Algorithms

5.2.1 The Model Space

Let $X = X^{(1)} \otimes \cdots \otimes X^{(d)}$. As in section 1.3, consider the RKHS on $X$,

$$
\mathcal{H} = \sum_j \mathcal{H}^{(j)} \oplus \sum_{j < l} [\mathcal{H}^{(j)} \otimes \mathcal{H}^{(l)}] \oplus \cdots,
$$

(5.2.1)

where $\mathcal{H}^{(j)}$ is a RKHS on $X^{(j)}$ satisfying the side condition (1.3.2). Notice that we do not have the subspace which consists of constant functions. This restriction and the side conditions (1.3.2) are necessary for the identifiability of the components, which is obvious from (5.1.1). Each $\mathcal{H}^{(j)}$ can be further decomposed as $\mathcal{H}^{(j)} = \mathcal{H}^0_s \oplus \mathcal{H}^{(j)}_s$. After deciding on a model (that is, choosing which terms in (5.2.1) to retain), replacing $\mathcal{H}^{(j)}$, $\mathcal{H}^{(l)}$ and so forth in (5.2.1) by their decompositions, multiplying out and regrouping, we can write the model space as

$$
\mathcal{M} = \mathcal{H}^0 \oplus \sum_{j=1}^q \mathcal{H}^j,
$$

(5.2.2)

where $\mathcal{H}^0$ is a finite dimensional space containing functions which are not penalized, and the $\mathcal{H}^j$ are orthogonal. The additive model of Hastie and Pregibon (1988) is a special case when we only retain some of the main effects.

5.2.2 The Penalized Conditional Likelihood Estimation

After choosing a model space $\mathcal{M}$, we estimate $f$ and its components by minimizing the penalized log conditional likelihood:

$$
- \sum_{k=1}^K i_k + \frac{n}{2} \sum_{j=1}^q \theta_j^{-1} ||P^j f||^2
$$

(5.2.3)
in $\mathcal{M}$. $P^j$ is the orthogonal projection in $\mathcal{M}$ onto $\mathcal{H}^j$. $\|P^j f\|_2^2$ is a quadratic roughness penalty.

Let $\{\phi_j\}_{j=1}^M$ be a set of basis functions of $\mathcal{H}^0$, $\phi_j(x) = (\phi_1(x), \ldots, \phi_M(x))$. Let $R_j$ be the RK of $\mathcal{H}^j$. Define $R_\Theta = \sum_{j=1}^q \theta_j R_j$, where $\Theta = (\theta_1, \ldots, \theta_q)$. Let $\xi^T(x) = (R_\Theta(x_{01}, x), \ldots, R_\Theta(x_{R_1}, x), \ldots, R_\Theta(x_{0K}, x), \ldots, R_\Theta(x_{RK}, x))$. The minimizer of (5.2.3) has the form (Wahba 1990, O'Sullivan et al. 1986)

$$ f_{\lambda, \Theta}(x) = \phi(x)^T d + \xi(x)^T c, \quad (5.2.4) $$

where $c_{n \times 1}$ and $d_{M \times 1}$ are vectors of coefficients to be estimated. Substituting (5.2.4) into (5.2.3), we can solve for $c$ and $d$ by minimizing

$$ I(c, d) = -\sum_{k=1}^K l_k(c, d) + \frac{n}{2} \lambda c^T Q_\Theta c, \quad (5.2.5) $$

where $Q_\Theta$ is a $n \times n$ matrix with element $R_\Theta(x_{rk}, x_{r'k'})$ at the $(r+1)k$th row and $(r'+1)k'$th column, $r, r' = 0, 1, \ldots, R$; $k, k' = 1, \ldots, K$.

All $l_k$'s are concave functions of $c$ and $d$. We use the Newton-Raphson procedure to compute $c$ and $d$ iteratively. Let

\begin{align*}
\mu^T_1 &= (\mu_{0k}, \ldots, \mu_{Rk}), \\
\mu^T_2 &= (\mu_1^T, \ldots, \mu_K^T), \\
y^T_k &= (1, 0, \ldots, 0), \\
y^T &= (y_1^T, \ldots, y_K^T), \\
u &= \mu - y, \\
S &= (\phi(x_{01}), \ldots, \phi(x_{R1}), \ldots, \phi(x_{0K}), \ldots, \phi(x_{RK}))^T, \\
U_k &= \text{diag}(\mu_{0k}, \ldots, \mu_{Rk}), \\
W_k &= U_k - \mu_k \mu_k^T, \\
W &= \text{diag}(W_1, \ldots, W_K).
\end{align*}

It can be shown that

\begin{align*}
\partial I / \partial c &= Q_\Theta u + n\lambda Q_\Theta c, \quad (5.2.6) \\
\partial I / \partial d &= S^T u, \quad (5.2.7) \\
\partial^2 I / \partial c \partial c^T &= Q_\Theta W Q_\Theta + n\lambda Q_\Theta, \quad (5.2.8) \\
\partial^2 I / \partial c \partial d^T &= Q_\Theta W S, \quad (5.2.9) \\
\partial^2 I / \partial d \partial d^T &= S^T W S. \quad (5.2.10)
\end{align*}

The Newton-Raphson iteration satisfies the linear system

$$
\begin{pmatrix}
Q_\Theta W Q_\Theta + n\lambda Q_\Theta & Q_\Theta W S \\
S^T W Q_\Theta & S^T W S
\end{pmatrix}
\begin{pmatrix}
c - c_- \\
d - d_-
\end{pmatrix}
= 
\begin{pmatrix}
- Q_\Theta u_- - n\lambda Q_\Theta c_- \\
- S^T u_-
\end{pmatrix} \quad (5.2.11)
$$
where the subscript minus indicates quantities evaluated at the previous Newton-Raphson iteration. The solution to system (5.2.11) may not be unique since \( W_- \) is not of full rank. We can write (5.2.11) as

\[
\begin{pmatrix}
Q_\Theta W_- + n \lambda I & Q_\Theta W_- \\
S^T W_- & S^T W_-
\end{pmatrix}
\begin{pmatrix}
Q_\Theta c \\
S d
\end{pmatrix}
= 
\begin{pmatrix}
Q_\Theta (W_- f_- - u_) \\
S^T (W_- f_- - u_)
\end{pmatrix}. \tag{5.2.12}
\]

\( f = Q_\Theta c + S d \) is unique as long as the equation (5.2.12) is definite. It is obvious that the equation (5.2.12) is definite if \( S^T W_- \) is of full row rank. This condition is also necessary for the linear logistic model to have an unique solution with \( S \) as the regression matrix. The null space of \( W_- \) is spanned by \( \{ (a_1, \ldots, a_k) \} \), where \( 1 \) is a \((R+1) \times 1 \) vector with all elements equal to 1. \( S^T W_- \) is of full row rank if \( (a_1, \ldots, a_k) \notin \text{span}(S) \) and \( S \) is of full column rank, which is usually true. Notice that \( S \) does not contain the constant vector since we exclude the constant from our model space.

All we need is a solution of (5.2.11). If \( Q_\Theta \) is nonsingular, (5.2.11) is equivalent to the system

\[
\begin{align*}
(W_- Q_\Theta + n \lambda I) c + W_- S d &= (W_- f_- - u_), \\
S^T c &= 0. \tag{5.2.13}
\end{align*}
\]

If \( Q_\Theta \) is singular, any solution to (5.2.13) is also a solution to (5.2.11).

The equation (5.2.13) has the same form as equation (1.4.12), except that the weight matrix \( W_- \) here is not diagonal and is not of full rank. This imposes a challenge in calculation. We introduce an algorithm and compare it with the delta algorithm in the following sections.

### 5.2.3 Algorithms

We first develop an algorithm using the special structure of the weight matrix \( W \) and the observation \( y \). Write \( W_{k-} = (I - 1 \mu_{k-}^T) U_{k-} (I - 1 \mu_{k-}^T) = O_{k-}^T O_{k-} \), \( k = 1, \ldots, K \), where \( 1 \) is a column of \( R+1 \) ones, \( O_{k-} = U_{k-}^{1/2} (I - 1 \mu_{k-}^T) \). Denote \( O = \text{diag}(O_{1-}, \ldots, O_{K-}) \), then \( W_- = O^T O_- \). Let \( \tilde{Q}_\Theta = O Q_\Theta O_\Sigma^T, \tilde{S} = O S, \tilde{f} = O f \), and \( \tilde{y} = f_- - U_-^{-1/2} u_- \). Suppose \( \tilde{c} \) and \( \tilde{d} \) are solutions to

\[
\begin{align*}
(\tilde{Q}_\Theta + n \lambda I) \tilde{c} + \tilde{S} \tilde{d} &= \tilde{y}, \\
\tilde{S}^T \tilde{c} &= 0. \tag{5.2.14}
\end{align*}
\]

Let \( c = O^T \tilde{c} \) and \( d = \tilde{d} \). Notice that \( (I - 1 \mu_{k-}^T) u_{k-} = u_{k-} \), where \( u_{k-} = \mu_{k-} - y_{k-} \). It can easily be verified that \( c \) and \( d \) are solutions to (5.2.13). Equation (5.2.14) has exactly the same form as (1.4.13). We can use RKPACK to solve equation (5.2.14) and estimate the smoothing parameters. We describe the whole procedure in an algorithmic form.

**Algorithm 1** Assuming the inputs of matrices \( S \), the \( Q_j \)'s, the response vector \( y \) and the starting vector \( f_\circ \), perform:
1. Compute \( u_\cdot \) and \( O_\cdot \).

2. Compute transformations \( \hat{S}, \hat{Q}, \) and \( \hat{y} \).

3. Call RKPACK with inputs \( \hat{S}, \hat{Q}, \) and \( \hat{y} \) using GCV or UBR to choose smoothing parameters;

4. Compute the new \( f \) and deviance. Stop if \( f \) or the deviance does not change; otherwise go to step 1.

The GCV score we implicitly use in the above algorithm is

\[
V(\lambda, \Theta) = \frac{1}{n} \frac{\| (I - A(\lambda, \Theta)) U^{-1/2} \hat{y} \|^2}{\left[ (1/n) \text{tr}(I - A(\lambda, \Theta)) \right]^2},
\]

where \( A(\lambda, \Theta) \) satisfies

\[
U^{-1/2} f_{\lambda, \Theta} = A(\lambda, \Theta) U^{1/2} \hat{y},
\]

and \( f_{\lambda, \Theta} \) is the vector of estimates at the design points computed from the solution of (5.2.14). The UBR score is

\[
U(\lambda, \Theta) = \frac{1}{n} \| (I - A(\lambda, \Theta)) U^{-1/2} \hat{y} \|^2 + \frac{2}{n} \text{tr} A(\lambda, \Theta).
\]

We will compare the performance of the above algorithm with a special delta algorithm by simulation in section 5.4 and on a real data set in section 5.5. The idea of the delta algorithm is to replace the weight matrix \( W \) by an approximate matrix which is positive definite (Jorgenson, 1984). For an additive model such as (5.1.5), Hastie and Pregibon (1988) replaced the weight matrix \( W \) by its diagonal \( \hat{W} \). Extending their method to our model, \( \hat{W} = \text{diag}(\mu_{rk}(1 - \mu_{rk})) \). Replacing \( W \) in equation (5.2.13) by \( \hat{W} \) and letting \( \tilde{Q}_\Theta = \hat{W}^{-1/2} Q_\Theta \hat{W}^{1/2}, \tilde{S} = \hat{W}^{1/2} S, \tilde{c} = \hat{W}^{-1/2} c, \tilde{d} = d, \) and \( \tilde{y} = \hat{W}^{-1/2} (\hat{W} f - u_\cdot) \); (5.2.13) becomes

\[
(\tilde{Q}_\Theta + n\lambda I) \tilde{c} + \tilde{S} \tilde{d} = \tilde{y},
\]

\[
\tilde{S}^T \tilde{c} = 0.
\]

This equation has the same form as (5.2.14). We can use RKPACK to solve this equation and select smoothing parameters. We call this algorithm Algorithm 2.

### 5.3 Bayesian Confidence Intervals

Let the prior for \( f(x) \) be

\[
F_\xi(x) = \sum_{\nu=1}^{M} r_\nu \phi_\nu(x) + b^\cdot \sum_{\beta=1}^{q} \sqrt{\theta_\beta} Z_\beta(x),
\]

(5.3.1)
where \( \tau = (\tau_1, \cdots, \tau_M)^T \sim N(0, \xi I) \), and the \( Z_\beta \) are independent, zero mean Gaussian stochastic processes, independent of \( \tau \), with \( \mathrm{E}[Z_\beta(x)Z_\beta(z)] = R_\beta(x, z) \). Let \( Z(x) = \sum_{\beta=1}^2 \sqrt{\theta_\beta}Z_\beta(x) \), then \( \mathrm{E}[Z(x)Z(z)] = R_\Theta(x, z) \).

It is easy to check that Theorem 3.3 is true for a rank deficient matrix \( W \) with \( W^{-1} \) replaced by the Moore-Penrose inverse \( W^+ \) as long as \( Q_\Theta \) is invertible. We use the Moore-Penrose inverse to make it unique since \( (W^+)^+ = W \). Let \( \zeta, \eta \) be any one of \( \tau_\nu \phi_\nu(x), \tau_\mu \phi_\mu(z), \sqrt{\theta_\beta}Z_\beta(x) \) or \( \sqrt{\theta_\alpha}Z_\alpha(z) \) for arbitrary points \( x \) and \( z \). As in section 3.5.1, under the prior (5.3.1) for \( f \), with \( \xi \to \infty \), the joint probability density function of \( (y, \zeta, \eta, f) \) given \( b \) and \( \sigma^2 \) is

\[
p(f|y)q(f|\tau)r(\zeta, \eta|f),
\]

where

\[
p(f|y) \propto \exp\{l(f)\},
\]

\[
q(f) \propto \exp\left\{-\frac{1}{2\sigma^2} f^T (Q_\Theta^{-1} - Q_\Theta^{-1} S (S^T Q_\Theta^{-1} S)^{-1} S^T Q_\Theta^{-1}) f \right\},
\]

and \( r(\zeta, \eta|f) \) is Gaussian with mean and covariance given in Theorem 3.3 for \( \sigma^2 = 0 \), \( y = f \) and \( W^{-1} \) replaced by \( W^+ \).

This integral is not easy to calculate since \( p(f|y) \) is not Gaussian. Both \( q(f) \) and \( r(\zeta, \eta|f) \) are Gaussian, we only need to approximate \( p(f|y) \) by a Gaussian density using the Laplace method. Expanding \( \log p(f|y) \) via a Taylor series centered at the mode \( f_* \) of \( p(f|y)q(f) \) and neglecting the cubic and higher order terms, we have an approximation \( \hat{p}(f|y) \) of \( p(f|y) \):

\[
\log \hat{p}(f|y) = l(f_*) - u_*^T (f - f_*) - \frac{1}{2} (f - f_*)^T W_* (f - f_*)
\]

\[
= -\frac{1}{2} (f - \alpha)^T W_* (f - \alpha) + C,
\]

where \( u_* = (-\partial l/\partial f)|_{f_*}, W_* = (-\partial^2 l/\partial f \partial f^T)|_{f_*}, C \) is a constant independent of \( f \), and \( \alpha \) is a vector satisfying \( W_* \alpha = W_* f_* - u_* \). Notice that \( f_* \) does not depend on \( \zeta \) and \( \eta \). We approximate the posterior density function \( \pi(\zeta, \eta|y) \) by

\[
\hat{\pi}(\zeta, \eta|y) \propto \int \hat{p}(f|y)q(f)r(\zeta, \eta|f)df.
\]

**Theorem 5.1** The approximate posterior density \( \hat{\pi}(\zeta, \eta|y) \) is Gaussian with mean and covariance given in Theorem 3.3, with \( W \) replaced by \( W_* \) and the inverse replaced by the Moore-Penrose inverse.

[Proof] The approximate likelihood \( \hat{p}(f|y) \) is identical to a Gaussian sampling likelihood with covariance \( W_*^+ \) and observations \( \alpha \). The approximate likelihood \( \hat{p}(f|y) \) does not depend on the choice of \( \alpha \) since it only depends on \( W_* \alpha \). Everything is the same as
Theorem 3.3 except for replacing \( W \) by \( W_* \). Gu (1992b) proved that \( \mathbf{f}_* \) is the fixed point of equation (5.2.11). Thus \( \mathbf{u}_* \) and \( W_* \) are the same as \( \mathbf{u} \) and \( W \) calculated at a fixed point of equation (5.2.11).

To get the Bayesian confidence intervals, we need to calculate posterior covariances in Theorem 3.3. Gu and Wahba (1993) discuss how to calculate these quantities when \( W_* = I \). Let \( W_* = O_* O_*^T, \hat{Q}_\Theta = O_* Q_\Theta O_*^T, \hat{S} = O_* S, \rho_\beta = (\theta_1 R_\beta(x, x_{01}), \ldots, \theta_\delta R_\beta(x, x_{0\delta}), \ldots, \theta_1 R_\beta(x, x_{0\delta}), \ldots, \theta_\delta R_\beta(x, x_{0\delta})), \rho_\beta = O_* \rho_\beta, \) and \( \Omega = \hat{Q}_\Theta + n \lambda I \). It is easy to verify that \( W_*^+ = (O_*^+)^T O_*^+ \). Approximating the matrix \((S^T \Omega^{-1} S)^{-1}\) by \((\check{S}^T \check{\Omega}^{-1} \check{S})^{-1}\), we can calculate the posterior covariances the same way as Gu and Wahba (1993) did with \( Q_\Theta, S \) and \( \rho_\beta \)'s replaced by their tilde's. The Moore-Penrose inverses are not involved in this calculation.

### 5.4 Simulations

In this section, we conduct simulations to compare the performance of Algorithm 1 and Algorithm 2. In addition, we evaluate the coverage properties of Bayesian confidence intervals.

In our simulations, \( \mathbf{X} = X^{(1)} \otimes X^{(2)} \otimes X^{(3)} = [0, 1]^3 \) and \( f(x) = f_1(x_1) + f_2(x_2) + f_3(x_3) \), where \( f_1(x_1) = \log x_1 + 1, f_2(x_2) = 0, f_3(x_3) = x_3 - 0.5 \). \( x_3 \) is used as the matching variable. \( K \) matched sets are generated as follows. The density functions of \( x_1, x_2 \) and \( x_3 \) are independent. For the \( k \)th matched set, we first generate a case \( (Y = 1) \). Generate its covariate \( x_{i0k} \) randomly with a density function proportional to \( \exp\{f_i(x_i)\}, i = 1, 2, 3 \). We then generate \( R \) controls \( (Y = 0) \). For the \( r \)th control, generate covariate \( x_{irk} \) randomly with the uniform density on \([0, 1]\). The controls must have their third covariate satisfying \( |x_{3rk} - x_{30k}| < 0.1 \).

In the first simulation, we generated 200 matched pairs. In the second simulation, we generated 100 1:3 matched sets. 100 such data sets were generated and fitted with Algorithm 1 and Algorithm 2 respectively. For each repetition, the mean squared error and the deviance were calculated by \( \text{mse} = \sum_{k=1}^{K} \sum_{r=0}^{R} (f(x_{rk}) - f_{0,k}(x_{rk}))^2 \) and \( D = -2 \sum_{k=1}^{K} \log \mu_{0k} \) respectively. For the first simulation, the two algorithms almost have the same pattern of convergence for both mean squared error and deviance. The convergence pattern of a "typical" repetition is plotted on the left side of Figure 5.4.1. For the second simulation, the two algorithms almost have the same pattern of convergence for most repetitions. But in some repetitions Algorithm 1 performs better than Algorithm 2. In these repetitions, Algorithm 2 either has a slower convergence rate, or its estimates are inferior to those of Algorithm 1 at convergence, or it does not converge.

One particular repetition where Algorithm 1 outperforms Algorithm 2 is plotted on the right side of Figure 5.4.1.

Box-plots of the coverage percentages of 95% Bayesian intervals for the overall function and components are shown in Figure 5.4.2. Notice that \( f_2 \) is linear, the confidence intervals cover near all or none of the true function (Gu and Wahba, 1993). For the first simulation with pair matching, the average coverages are not far from the nominal
Figure 5.4.1: Convergence patterns. Both mse and deviance are on the log scale. On the left plot, we reduce the deviance by 0.3 to make the plot more compact.

values. For the second simulation with 1:3 matching, the average coverages are lower than the nominal values and have a large variance, suggesting that bigger sample size is needed. The 95% Bayesian intervals for a “typical” repetition from the pair matching are shown in Figure 5.4.3.

5.5 Analysis of Breast Cancer Data

In this section, we apply our method to a real data set. These data are taken from an ongoing collaborative study of risk factors for breast cancer, in which investigators from Wisconsin, Maine, Massachusetts, and New Hampshire are participating. Newcomb, Storer, Longnecker, Mittendorf, Greenberg, Clapp, P., Willett and MacMahon (1994) give a detail description of the study. For simplicity, we only use the subset of New Hampshire subjects interviewed during 9 months of 1990. We consider three covariates: age at interview, age at first full-term pregnancy (denoted as age at FFP) and age at menarche. The data is not matched a priori. We construct a 1:2 matched set based on age at interview. There are 318 observations. We do not intend to make any definite conclusions about breast cancer since some other risk factors are not considered (Newcomb et al., 1994). Rather, we demonstrate the ability of our method to capture the interaction and overall shape of the multivariate relative risk function.

The model we fit is

$$f(\text{age at FFP}, \text{age at menarche})$$
Figure 5.4.2: Coverage percentages of 95% Bayesian intervals. Plusses: sample means; dotted lines: nominal coverage.

Figure 5.4.3: Display of the 95% Bayesian intervals. Solid lines: true functions; dotted lines: estimate by Algorithm 1; dashed lines: 95% Bayesian intervals.
Figure 5.5.1: Left and middle: estimates of the main effects $f_1$ and $f_2$ with their 95% Bayesian intervals. Right: Estimate of the interaction $f_{12}$.

\[
= f_1(\text{age at FFP}) + f_2(\text{age at menarche}) + \\
f_{12}(\text{age at FFP, age at menarche}).
\] (5.5.1)

The estimates of the main effects $f_1$ and $f_2$ with their 95% Bayesian confidence intervals are plotted in Figure 5.5.1. The estimate of the interaction $f_{12}$ is plotted on the right side of Figure 5.5.1. We see that the interaction seems to be moderate.

We plot the age at FFP vs age at menarche on the left side of Figure 5.5.2. Cases are marked as *'s and controls are marked as circles. We superimpose contour lines of the estimated posterior standard deviations. These contours agree well with the distribution of the observations. We decide to use the region with estimated posterior standard deviations are less then or equal to 0.5. The estimate of the overall function $f$ in the defined region is plotted on the right side of Figure 5.5.2. Notice that we used sum-to-zero side conditions not set-to-zero conditions. Hence the overall functions is not the log odds ratio relative to a fixed exposure point. Instead, the trend of the overall function is important. The odds ratio of an New Hampshire woman with covariates $\mathbf{x} = (\text{age at FFP, age at menarche})$ developing breast cancer relative to a woman with covariates $\mathbf{x'} = (\text{age at FFP', age at menarche'})$ equals $\exp\{f(\mathbf{x}) - f(\mathbf{x'})\}$. The Bayesian confidence interval for this odds ratio can be calculated as section 4.4.

To see the shape of $f$ more clearly, we plot cross sections of the estimate in Figure 5.5.3. The risk of a New Hampshire woman increases dramatically if she has not had FFP at about 35 years old and her menarche was not late. The risk decreases with the
Figure 5.5.2: Left: data and contours of constant posterior standard deviation. Right: Estimation of the overall function $f$ in the defined region.

Age at menarche. The decrease is slightly faster if she has her FFP earlier.

Algorithm 1 is used to calculate the estimate given above. It uses 7 iterations and 251 seconds CPU time on a DEC 5000 to converge. We also tried to use Algorithm 2 to fit the same model. Its estimate is similar to that of Algorithm 1. It requires 49 iterations and 1333 seconds CPU time. Other practical examples we tried all indicate that Algorithm 1 is much faster than Algorithm 2. The convergence pattern of Algorithm 1 is quadratic. The convergence pattern of Algorithm 2 is linear after a few steps, which agrees with Hastie and Pregibon (1988).
Figure 5.5.3: Left: Cross sections of estimated function $f$ as a function of age at FFP. Right: Cross sections of estimated function $f$ as a function of age at menarche. $q_1$, $q_2$, $q_3$ and $q_4$ are the quantiles at .125, .375, .625 and .875.
Chapter 6

Concluding Remarks

6.1 Summary

We have proposed nonparametric models for data from exponential families (section 1.4) and matched case-control data (section 5.2) using the PSA method. Our model space is a subspace of a tensor product of RKHS's. We obtain estimates by solving a minimization problem involving the penalized likelihood. A Newton-Raphson method is used to solve this minimization problem. We use the GCV method and the unbiased risk method to choose smoothing parameters at each Newton-Raphson iteration. The algorithm is given in section 2.2. Our simulations indicate that both the GCV method and the unbiased risk method work well. When there is no over- or under-dispersion, option U of the unbiased risk method works better (section 2.3). We construct Bayesian and bootstrap confidence intervals for a PSA estimate. Our simulations indicate that both Bayesian and bootstrap confidence intervals work well. They behave similarly. That is, both of them have the ACP property, instead of pointwise property (Chapter 3).

The PSA method is an extension of the GLIM models. It has an analysis like standard ANOVA. That is, we can build a model as in ANOVA by looking at the estimates and confidence intervals of the components. Good estimates of the components need a moderate to large sample size. We illustrated how to use the PSA methods by analyzing binary data from the Wisconsin Epidemiological Study of Diabetic Retinopathy (Chapter 4). We found many nonlinear risk factors, which were not detected in previous analyses using GLIM models.

There are other recent researches on nonparametric methods for data from exponential families. Among them are the kernel method (Azzalini, Bowman and Härdle, 1989), generalized additive models (Hastie and Tibshirani, 1990), tree methods (Lo, 1993; Yang, 1993), MARS (Friedman, 1991) and projection pursuit (Roosen and Hastie, 1993). Each method has its own advantages. It will be interesting to compare the performance of all these methods.
6.2 Future Research

Hypothesis tests and model selection procedures based on some prediction criteria are important for the SS ANOVA method. The hypothesis or the underlying model is different from that in a parametric analysis. For a hypothesis test, the space generated by the null hypothesis may be infinite dimensional. For model selection, we do not assume the true model is inside our model spaces. We want to find a model which is nearest (in terms of prediction) to the true model. In section 4.2, we proposed several methods for estimating the prediction errors. We will evaluate the performance of these estimates and the performance of a model selection procedure based on these estimates in the future.

Theoretical results can provide insight into and justification for the PSA method. Large sample properties like consistency, convergence rate, strong or weak GCV theorem are desirable.

Due to human error or other reasons, we often have “bad” or high influential data points (sections 4.6 and 4.9). Usually the confidence intervals around these points are wide if we leave these points in the analysis. Thus we can delete the part of the estimates that exhibits wide confidence intervals. Alternatively, we can develop a diagnostic tool to detect influential points and outliers, and exclude them from the analysis.

We are currently doing research on using PSA method for a proportional-odds model of ordinal data:

$$\log\{\gamma_j/(1 - \gamma_j)\} = \theta_j - f(x), \quad j = 1, \ldots, k - 1,$$

where $\gamma_j = P(Y \leq j|x)$, $j = 1, \ldots, k$. $\theta_1 \leq \theta_2 \leq \cdots \theta_{k-1}$ are constants. $f(x)$ is in a model space $\mathcal{M}$. 
Bibliography

Abramovich, F. and Steinberg, D. (1993). Improved inference in nonparametric regression using $l_4$-smoothing splines, manuscript, Tel Aviv University.


