Model Selection in Additive Partial Linear Models

Using Local False Discovery Rate

Changhwan Lee

The Graduate School

Yonsei University

Department of Applied Statistics

Model Selection in Additive Partial Linear Models

Using Local False Discovery Rate

A Master's Thesis

Submitted to the Department of Applied Statistics

and the Graduate School of Yonsei University

in partial fulfillment of the

requirements for the degree of

Master of Arts

Changhwan Lee

Feburary 2019

This certifies that Master's thesis of Changhwan Lee is approved.

Thesis Supervisor: Prof. Taeyoung Park

Committee Member: Prof. Sangwook Kang

Committee Member: Prof. Chul-Eung Kim

The Graduate School

Yonsei University

Feburary 2019

Contents

List of Figures

List of Tables

ABSTRACT

Model Selection in Additive Partial Linear Models

Using Local False Discovery Rate

An additive partial linear model has been used to explain any given data with advantages of simplicity stemming from the parametric components and flexibility stemming from the nonparametric components. Especially, this paper accounts for building an additive partial linear model by variable selection method based on local false discovery rate. The Bayesian LASSO method with zero-inflated mixture prior replaces the traditional knot selecting method when fitting nonparametric components. In addition, the ability to detect a linear effect when an explanatory variable was put in nonparametric components helps the model be used more flexible in the way that a researcher doesn't need to predetermine which variable belong to parametric or nonparametric components. Plus, the proposed model can alleviate the overfitting problem. The proposed methodology is demonstrated through a Monte Carlo Markov Chain(MCMC) simulation study and an epidemiological data analysis.

Key words : Additive partial linear model, Bayesian LASSO, Local false discovery rate, Zero inflated mixture prior, Variable selection

Chapter 1

Introduction

As an intermediate model between multiple linear regression models and additive nonparametric regression models, an additive partial linear model(APLMs) consists of two models to represent parametric components and nonparametric components. Therefore, APLMs has advantages of multiple linear regression models such as easy interpretation and those of additive nonparametric regression models such as flexibility needed when fitting the model.

Especially, when estimating appropriate forms for complicated functions, the nonparametric regression method has been appealed to many researchers because it helps the researchers to pursuit more flexibility and accuracy. Also, it has enabled them to overcome the barriers caused by the parametric assumptions, which means the nonparametric regression method relaxes the strong model assumption included in the parametric regression methods. One of the famous methods having been studied by many researchers is kernel smoothing or local polynomials. These methods estimate true functions by imposing weights locally. Also, regression splines and smoothing splines are used frequently in nonparametric regression. When using the regression spline method, how to select knots for accurate curve fitting is an important issue. There have been

suggested various methods for it. One group of methods is conventional knot selection such as stepwise selection or branch and bound. Others are stochastic knot selection such as free-knot selection method, which means the number and location of knots are determined by data. It shows much better accurate results than those of conventional methods. However, it also has drawbacks that it is time-consuming for convergence (Miyata and Shen, 2003).

Therefore, when fitting curve using spline, another method is suggested to alleviate the shortcomings of free-knot selection mentioned above like predetermining the number and locations of knots. When considering that each basis function in a basis has a corresponding coefficient, it is clear that selecting appropriate knots to fit curve means deciding which basis function's coefficient significant is. Therefore, knot selecting problem can be replaced to variable selection and multiple hypotheses testing under the situation. Traditionally, there have been many methods to select meaningful variables using various penalties. Ridge regression suggested by Hoerl and Kennard (1970) and LASSO regression suggested by Tibshirani (1996) select variables imposing penalty on each coefficient's size. Especially, LASSO regression is perceived as a useful tool for variable selection. It uses absolute value as a penalty when choosing the coefficients in a model. That model was further extended by other researchers to develop some methods such as Group LASSO (Yuan and Lin, 2006), Fused LASSO (Tibshirani et al., 2005) and Elastic net (Zou and Hastie, 2005) to solve more complicated problems.

By assigning appropriate prior density, LASSO can work in a Bayesian method as Park and Casella (2008) suggested. They constructed the Bayesian model for the parameters in the LASSO model and suggested that the square of tuning parameter follows a gamma distribution. Based on the Bayesian LASSO, many methods have been tried. Researchers started to use spike and slab prior distribution instead of using multivariate laplace distribution as prior. Zhao and Sarkar (2015) suggested a model constructing reliable confidence interval in the sparsity situation by changing the prior density to

zero-inflated mixture prior as an example of spike and slab prior to fit the coefficients. In addition, as a criterion to decide which variables significant are, Zhao and Sarkar (2015) used local false discovery rate representing how significant a variable is. This concept was originated from false discovery rate suggested by Benjamini and Hochberg (1995). False discovery rate is associated with the Bonferroni correction in multiple testing in the way of controlling the overall error rate. Efron et al. (2001) focused on whether each coefficient is significant and extended false discovery rate to locally one and named it local false discovery rate.

We applied the Bayesian LASSO with zero-inflated mixture prior to an additive partial linear model and using the concept of the local false discovery rate, we selected the significant variables under the overall model level. When using the nutritional epidemiology study data, we realized that it needs to improve the flexibility of the proposed model because it is often found that there are many cases that true relationships among variables are not known and they change as the further researches progress. Therefore, our proposed model has added a basis function representing a linear relationship between the dependent variable and explanatory variable in the basis. This change allows the model to detect the linear relationship between the dependent variable and covariates in the long run although covariates were fitted in the nonparametric components the beginning of the study because of the uncertainty of relationship and remaining parametric components are used for categorical or ordinal variables which are hard to assume non-linear relationship as suggested by Hastie (2017). In sum, we proposed the model selecting coefficients in the additive partial linear model with a basis including linear relationship, the prior suggested by Zhao and Sarkar (2015) and the concept of local false discovery rate in the overall model level. In this process, our proposed model could reduce overfitting problem.

This paper is organized as follows. Chapter 2 describes the additive partial linear model concretely and zero-inflated mixture prior. Also, the variable selection method

based on the local false discovery rate is illustrated in the chapter. Chapter 3 illustrates prior distributions of the proposed model and posterior sampling. In Chapter 4, a simulation study for validation of the proposed model is implemented. The proposed model is applied in epidemiology data which the result about significant variables is controversial in Chapter 5 and Chapter 6 concludes with a discussion.

Chapter 2

Model Specification

2.1 Addtitive Partial Linear Model

Assume that there is a dependent variable $\mathbf{Y} = \{y_1, \cdots, y_n\}^\top$ and there are covariates $\mathbf{X} = \{X_1, \cdots, X_L\}^\top$ and $\mathbf{Z} = \{Z_1, \cdots, Z_K\}^\top$. An additive partial linear model from the data is given by

$$y_i = \beta_0 + \sum_{l=1}^{L} f_l(X_l) + \sum_{k=1}^{K} \beta_k Z_k + \epsilon_i, \epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), i = 1, \cdots, n \qquad (2.1)$$

where $\mathbf{X}$ and $\mathbf{Z}$ represent nonparametric and linear components respectively, $f_1, \cdots, f_L$ are unknown smooth functions corresponding to $X_1, \cdots, X_L$, β_0 is a intercept, $\beta_1, \cdots, \beta_K$ are coefficients of $Z_1, \cdots, Z_K$ and ϵ is a error following normal distribution with mean 0 and unknown variance σ^2.

Among various regression spline, we use radial basis functions to appproximate some unknown smooth functions and add first order term in the basis fucntion to detect the linear effect in the nonparametric components.

$$\mathbf{b}(x) = \{x, |x - \tau_1|^3, \cdots, |x - \tau_\kappa|^3\}^\top$$

where τ_κs are fixed knots and located with equal intervals, $\min_{1\le i\le n} x_i < \tau_1 < \cdots < \tau_\kappa < \max_{1\le i\le n} x_i$.

To apply the Bayesian LASSO, we change the the model (2.1),

$$\mathbf{Y} - \bar{y}\mathbf{1}_n = \mathbf{b}^*(\mathbf{X})\boldsymbol{\gamma} + \mathbf{Z}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \overset{iid}{\sim} \mathrm{N}(0, \sigma^2)$$

where $\boldsymbol{\gamma} = (\gamma_1, \cdots, \gamma_L)^\top$ is regression parameters for estimating unknown smooth functions, $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_K)^\top$ is regression parameters, $\mathbf{Y}$ is n $\times$ 1 vector, $\mathbf{X}$ is the $n \times L$ matrix, $\mathbf{Z}$ is $n \times K$ matrix, σ^2 is unknown variance of ϵ and $\mathbf{b}^*(x)$ and $\mathbf{Z}^*$ are standardized $\mathbf{b}(x)$ and $\mathbf{Z}$. These changes allow the model to ensure identifiability of smooth functions by satisfying the assumption, $E\{f_l(X_l)\} = 0$ for $l = 1, \cdots, L$.

Under the situation, the way to select knots in the polynomial functions for fitting the unknown smooth functions was used traditionally. However, we approximate the unknown smooth functions by determining which coefficients of the polynomial function 0 is instead of selecting the location of knots when considering that all coefficients are not used to fit the smooth functions.

2.2 Zero Inflated Mixture Prior

Zhao and Sarkar (2015) developed a model based on the Bayesian LASSO model suggested by Park and Casella (2008). Park and Casella used Gaussian distribution for coefficients as a prior. However, it is hard for the prior to produce exact zero. The model suggested by Zhao and Sarkar (2015) has a particular prior called zero-inflated mixture prior(ZIMP), one of the slab and spike prior, to deal with the problem of sparsity that some modern data have. Suppose $\mathbf{Y}$ follows a probability distribution,

$$\mathbf{Y}|\boldsymbol{\beta} \sim p(\mathbf{y}|\boldsymbol{\beta})$$

where coefficient vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^\top$ and $\boldsymbol{\beta}$ has the prior as follows.

$$\beta_i \sim \begin{cases} 0 & \text{with probability } \pi_0 \\ \psi(\beta_i) & \text{with probability } 1-\pi_0 \end{cases}$$

where $\psi(\beta_i)$ is a known distribution when $\beta_i \neq 0$. The prior was applied to $\boldsymbol{\beta}$ commonly in the Bayesian LASSO model to solve the problem that many variables are not significant because of the lack of data. We consider that situation is related to the fitting smooth functions because all of coefficients are not used and a few coefficients shows significance when fitting the smooth functions. We use the prior not only for nonparametric components but also for linear compoenents to determine a model.

2.3 Local False Discovery Rate and Variable Selection Based Local FDR

Benjamini and Hochberg (1995) introduced a false discovery rate(FDR) to overcome the downside of the familywise error rate caused by repetitive single inference procedures. It was defined as a proportion of the null hypotheses that is true in real among the hypotheses declared significant. In other words, FDR is the expected proportion of wrongly rejected hypotheses.

In many cases, researchers are interested in determining the probability that each coefficient is zero which implies the corresponding null hypothesis is true instead of controlling the overall error rate. Therefore, Efron et al. (2001) introduced local FDR which implies a probability that a hypothesis is true given the data. In other words, a local FDR is defined as a posterior probability that the corresponding the null hypothesis can not be rejected. Suppose that a hypothesis testing as follows.

$$X \sim \begin{cases} f_0(X) & \text{with probability } p_0 \\ f_1(X) & \text{with probability } 1-p_0 (= p_1) \end{cases}$$

where X has a density $f_0(x)$ if the the null hypothesis is true, otherwise X has a density $f_1(x)$. Therefore, we can calculate the local false discovery rate as following.

$$P(H_0 \text{ is true}|x) = \frac{p_0 f_0(x)}{p_0 f_0(x) + p_1 f_1(x)}$$

Combined with the ZIMP, we can consider the process of determining the values of coefficients for fitting given smooth functions as multiple hypotheses testing of $\boldsymbol{\beta}$ because the posterior distribution follows a mixture distribution and the simulation is equivalent to select a distribution of the two distributions. The null hypothesis is that $\beta_i = 0$ and the alternative hypothesis is that β_i follows a Gaussian distribution with mean 0 and unknown variance. As a result, we use the local FDR as a criterion to determine whether each coefficient is 0 or not to approximate the given smooth functions. To be specific, the significance of each parameter is determined by multiple testing procedure controlling Bayes FDR(Sarkar et al., 2008) at overall significant level 0.15. To select appropriate variables, we conducted variable selection based on the local false discovery rates used by Bonato et al. (2010). This method uses all the samples from the MCMC simulation instead of using only a few portions of samples. First, conduct MCMC simulation. Then, we can notice which variables are used to build the model in each iteration and transform the sample to 1 if the coefficients have a non-zero value. As a result, we can compute the following p_j for each jth parameter.

$$p_j \equiv \frac{1}{K} \sum_{k=1}^{K} \mathrm{I}(\zeta_j \in \mathbf{X}^{(k)})$$

ζ_j means jth covariate and $\mathbf{X}^{(k)}$ represents the set of covariates used to the build the model in the kth iteration. Therefore, $1-p_j$ is the estimate of the jth local false discovery rate. Next, set the global FDR bound α in (0,1). Then, sort the local FDRs calculated from the previous step in ascending order and choose m local FDRs such that

$$\underset{1\leq m\leq p}{\arg\max}\left\{m\Big|\frac{1}{m}\sum_{i=1}^{m}\big(1-p_{(i)}\big)\leq\alpha\right\}$$

where $p_{(i)}$ represents the estimate of ith smallest local FDR. Let $p_{(m)}=\phi_\alpha$. Next, Claim set $\mathcal{X}_{\phi_\alpha}=\{j:p_j>\phi_\alpha\}$ is significant which is suggested by Morris et al. (2008). Finally, the parameters having the index j in $\mathcal{X}_{\phi_\alpha}$ are selected. In other words, we select the $\beta_{(1)},\cdots,\beta_{(m)}$ corresponding to $\mathcal{X}_{\phi_\alpha}$ not being likely to 0 when considering the local false discovery rate.

Chapter 3

Bayesian Inference

3.1 Prior Specification

Park and Casella (2008) established the Bayesian model to determine which coefficient is significant. The prior for coefficients was changed to mixture distribution called ZIMP prior used by Zhao and Sarkar (2015). Therefore, considering the similarity of our model to estimate appropriate coefficients, we choose priors for some parameters as follows based on the model with ZIMP. The prior of μ is a flat prior. In case of σ^2, although an improper prior is used for σ^2, the posterior density for σ^2 is proper.

$$
\begin{aligned}
p(\mu) &\propto 1 \\
\sigma^2 &\sim \text{Inv-Gamma}(a, b), a = 0, b = 1
\end{aligned}
$$

The prior distribution for the probability that a coefficient is equal to 0 follows a beta distribution. a_{π_g} and b_{π_g} are constants. g indicates the covariate or coefficients sets for fitting gth smooth function. Therefore, π_g means prior probabilty for gth coefficients set.

$$\pi_g \sim \text{Beta}\left(a_{\pi_g}, b_{\pi_g}\right), g = 1, \cdots, G$$

$\mathbf{\Theta} = (\theta_{11}, \cdots, \theta_{gj}, \cdots, \theta_{Gm_G})^\top$ follows a prior mixture distribution as follows. gj indicates jth individual covariate or coefficient in the gth set.

$$\theta_{gj} \sim \pi_g \text{I}(\theta_{gj} = 0) + (1 - \pi_g)\text{N}(0, \sigma^2\nu_{gj}^2), g = 1, \cdots, G, j = 1, \cdots, m_g$$

The coefficients variance vector $\boldsymbol{\nu} = (\nu_1^2, \cdots, \nu_{m_g}^2)$ follows the

$$\nu_1^2, \cdots, \nu_{m_g}^2 \sim \prod_{j=1}^{m_g} \frac{\lambda^2}{2} e^{-\lambda^2 \frac{\nu_j^2}{2}} d\nu_j^2, g = 1, \cdots, G$$

The square of LASSO tuning parameter follows a gamma distribution. a_{λ^2} and b_{λ^2} are constants.

$$\lambda^2 \sim \text{Gamma}\left(a_{\lambda^2}, \frac{1}{b_{\lambda^2}}\right)$$

3.2 Posterior Computation

Prior to computating posterior distribution, for convenience, the incoporated parameter set is redefined as follows.

$$\mathbf{\Theta} = \{\boldsymbol{\gamma}^\top, \boldsymbol{\beta}^\top\}^\top = \{\theta_{11}, \theta_{12}, \cdots, \theta_{1m_1}, \theta_{21}, \cdots, \theta_{Gm_G}\}^\top$$

and the incoporated covariate set is defined as follows.

$$\mathbf{W} = \{\mathbf{b}^*(\mathbf{X}), \mathbf{Z}^*\} = \{W_{11}, W_{12}, \cdots, W_{1m_1}, W_{21}, \cdots, W_{Gm_G}\}.$$

Finally, $\tilde{\mathbf{Y}}$ is defined as $\mathbf{Y} - \bar{y}\mathbf{1}_n$. We usually marginalize out μ because it is out of our

interest. We draw samples from the below target posterior density.

$$
\begin{aligned}
P(\boldsymbol{\Theta}, \boldsymbol{\nu}, \boldsymbol{\pi}, \lambda^2, \sigma^2 | \tilde{\mathbf{Y}}) &= \int P(\boldsymbol{\Theta}, \boldsymbol{\nu}, \boldsymbol{\pi}, \lambda^2, \sigma^2, \mu | \mathbf{Y}) d\mu \\
&\propto p(\tilde{\mathbf{Y}} | \boldsymbol{\Theta}, \boldsymbol{\nu}, \sigma^2) \prod_{g=1}^{G} \prod_{j=1}^{m_g} p(\theta_{gj} | \nu_{gj}, \sigma^2, \pi_g) \prod_{g=1}^{G} \prod_{j=1}^{m_g} p(\nu_{gj}^2 | \lambda^2) \\
&\quad \prod_{g=1}^{G} p(\pi_g | a_{\pi_g}, b_{\pi_g}) p(\lambda^2 | a_{\lambda^2}, b_{\lambda^2}) p(\sigma^2 | a, b)
\end{aligned}
$$

Step 1. Draw θ_{gj} from the following a mixture distribution,

$$
\boldsymbol{\Theta}_{gj} | \tilde{\mathbf{Y}}, \boldsymbol{\Theta}_{-gj}, \boldsymbol{\nu}, \lambda^2, \boldsymbol{\pi}, \sigma^2 \sim \begin{cases} 0 & \text{with } l_{gj} \\ \mathrm{N}\left(\dfrac{(\tilde{\mathbf{Y}} - \mathbf{W}_{-gj} \boldsymbol{\Theta}_{-gj})^\top \mathbf{W}_{gj}}{\mathbf{W}_{gj}^\top \mathbf{W}_{gj} + \dfrac{1}{\nu_{gj}^2}}, \dfrac{\sigma^2}{\mathbf{W}_{gj}^\top \mathbf{W}_{gj} + \dfrac{1}{\nu_{gj}^2}} \right) & \text{with } 1 - l_{gj} \end{cases}
$$

where

$$
l_{gj} = \frac{\pi_g}{\pi_g + (1 - \pi_g)(1 + \nu_{gj}^2 \mathbf{W}_{gj}^\top \mathbf{W}_{gj})^{-\frac{1}{2}} \exp\left(\dfrac{((\tilde{\mathbf{Y}} - \mathbf{W}_{-gj} \boldsymbol{\Theta}_{-gj})^\top \mathbf{W}_{gj})^2}{2\sigma^2 \left(\mathbf{W}_{gj}^\top \mathbf{W}_{gj} + \dfrac{1}{\nu_{gj}^2} \right)} \right)}
$$

for $g = 1, \cdots, G$ and $j = 1 \cdots, m_g$

Step 2. Draw π_g from $p(\pi_g | \tilde{\mathbf{Y}}, \boldsymbol{\Theta}, \boldsymbol{\nu}, \lambda^2, \boldsymbol{\pi}_{-g}, \sigma^2)$ which is a beta distribution,

$$
\pi_g | \tilde{\mathbf{Y}}, \boldsymbol{\Theta}, \boldsymbol{\nu}, \lambda^2, \boldsymbol{\pi}_{-g}, \sigma^2 \sim \text{Beta}\left(a_{\pi_g} + \sum_{j=1}^{m_g} \eta_{gj}, b_{\pi_g} + m_g - \sum_{j=1}^{m_g} \eta_{gj} \right)
$$

where

$$\eta_{gj} \equiv \mathrm{I}(\theta_{gj} = 0)$$

for $g = 1, \cdots, G$ and $j = 1 \cdots, m_g$

Step 3. Draw σ^2 from $p(\sigma^2|\tilde{\mathbf{Y}}, \boldsymbol{\Theta}, \boldsymbol{\nu}, \lambda^2, \boldsymbol{\pi})$ which is an inverse gamma distribution,

$$\sigma^2|\tilde{\mathbf{Y}}, \boldsymbol{\Theta}, \boldsymbol{\nu}, \lambda^2, \boldsymbol{\pi} \sim \mathrm{IG}\left(\frac{N}{2}, \frac{\|\tilde{\mathbf{Y}} - \mathbf{W}\boldsymbol{\Theta}\|_2^2 + \boldsymbol{\Theta}^\top \mathbf{D}_{\boldsymbol{\nu}}{}^{-1}\boldsymbol{\Theta}}{2}\right)$$

where

$$N \equiv n + \sum_{g=1}^{G} m_g - \sum_{g=1}^{G}\sum_{j=1}^{m_g} \eta_{gj} - 1$$

$$\mathbf{D}_{\boldsymbol{\nu}} \equiv \mathrm{diag}(\nu_{11}^2, \nu_{12}^2, \cdots, \nu_{Gm_G}^2)$$

Step 4. Draw ν_{gj}^2 from $p((\nu_{gj}^2)^{-1}|\tilde{\mathbf{Y}}, \boldsymbol{\Theta}, \boldsymbol{\nu}_{-gj}, \lambda^2, \boldsymbol{\pi}, \sigma^2)$

$$\sim \begin{cases} \text{Inv-Gamma}\left(1, \dfrac{\lambda^2}{2}\right), & \text{if } \theta_{gj} = 0 \\ \text{Inv-Gaussian}\left(\sqrt{\dfrac{\lambda^2\sigma^2}{\theta_{gj}^2}}, \lambda^2\right), & \text{if } \theta_{gj} \neq 0 \end{cases}$$

for $g = 1, \cdots, G$ and $j = 1 \cdots, m_g$

Step 5. Draw λ^2 from $p(\lambda^2|\tilde{\mathbf{Y}}, \boldsymbol{\Theta}, \boldsymbol{\nu}, \boldsymbol{\pi}, \sigma^2)$ which is a gamma distribution,

$$\lambda^2|\tilde{\mathbf{Y}}, \boldsymbol{\Theta}, \boldsymbol{\nu}, \boldsymbol{\pi}, \sigma^2 \sim \text{Gamma}\left(\sum_{g=1}^{G} m_g + a_{\lambda^2}, \left(\frac{\sum_{g=1}^{G}\sum_{j=1}^{m_g}\nu_{gj}^2}{2} + b_{\lambda^2}\right)^{-1}\right)$$

Chapter 4

Simulation Study

We conducted a simulation study to assess the performance of the proposed model using various functions from the functions presenting an insignificant effect on the functions including a complicated nonlinear effect. We set n = 800 and 14 knots for nonparametric components are located with the same interval in [0,1]. The locations of knots are fixed and σ for ϵ was set to 1. We ran the simulation with 100,000 iterations and used 10,000 periods as burn-in. The model is composed of 8 nonparametric components and 6 parametric components. Continuous variables belong to nonparametric components and categorical variables belong to parametric components. Continuous variables are defined in [0,1], odd-numbered categorical variables are generated to 1 or 2 with probability 0.5 and even-numbered categorical variables are generated to 1,2 or 3 with probability $\frac{1}{3}$. The functions in nonparametric components represent 4 insignificant, 2 linearly significant and 2 nonlinearly significant relationships between explanatory variables and dependent variable. Also, parametric components include 6 covariates composed of 2 linearly significant and 4 insignificant effects respectively.

$$y_i = \beta_0 + \sum_{l=1}^{8} f_l(X_l) + \sum_{k=1}^{6} \beta_k Z_k + \epsilon_i, \epsilon_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2), i = 1, \cdots, 800$$

$$
\begin{aligned}
f_1(x) &= 3\exp(-30(x-0.2)^2) + \exp(-50(x-0.7)^2) \\
f_2(x) &= \sin(2\pi x) \\
f_3(x) &= x \\
f_4(x) &= -0.75x \\
f_5(x) = \cdots = f_8(x) &= 0 \\
(\beta_0, \beta_1, \beta_2, \cdots, \beta_6)^\top &= (0, 0.6, -1, 0, 0, 0, 0)^\top
\end{aligned}
$$

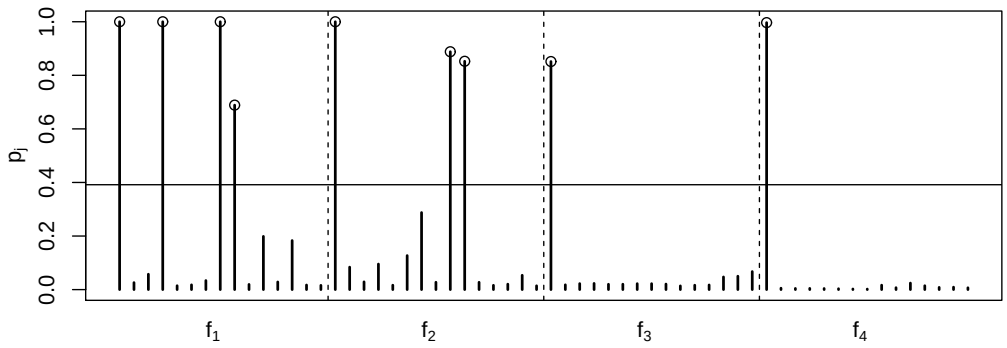

Figure 4.1: Plot of estimates of local FDRs related to covariates having nonlinear effect and linear effect. The local FDRs circled on the top were selected.

Figure 4.1 and 4.2 represent the plot of the whole local false discovery rates. Vertical dotted lines are used to identify local false discovery rate group belonging to each function defined above. Therefore, there are 15 local false discovery rates in each group and the last 6 local false discovery rates correspond to parametric components. The horizontal lines mean ϕ_α determined based on the whole false discovery rates. α was set to 0.15. As a result, ϕ_α was calculated to 0.392. Although some samples corresponding to a coefficient are not zero from the MCMC simulations, we don't use the samples if the

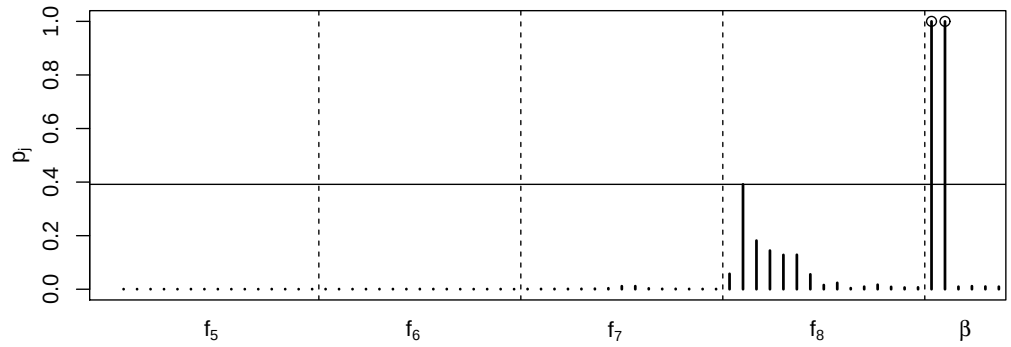

Figure 4.2: Plot of estimates of local FDRs related to covariates having insignificant effect and categorical covariates. The local FDRs circled on the top were selected.

corresponding local false discovery rate doesn't exceed the criterion based on variable selection method using the local false discovery rate and determine the coefficient to 0.

As shown in Figure 4.1, four,three,one and one significant coefficients were selected respectively among the coefficients groups corresponding to f_1, f_2, f_3 and f_4. Also, coefficients corresponding to f_5, f_6, f_7, f_8 are not selected accroding to Figure 4.2, which means they have insignificant effect. In case of categorical variables, β_1, β_2 were selected and $\beta_3, \beta_4, \beta_5, \beta_6$ were not selected. In sum, we started the whole model with 126 coefficients but have built the model with 11 coefficients.

Figure 4.3 shows the plot of nonparametric components having significant effect. The plots shown in the first row of Figure 4.3 are parts of Figure 4.1. The plots were divided according to the functions and the horizontal dashed line is equal to ϕ_α. The second row of Figure 4.3 shows fitted smoothing curves before choosing significant coefficients using local false discovery rate. Also, the third row of Figure 4.3 shows fitted smoothing curves after choosing significant coefficients by our proposed model. Meanwhile, the first and second columns correspond to functions where the nonlinear effect is included and the third and fourth columns correspond to functions where the

linear effect is included. In case of f_1 and f_2, the plots in the second row in Figure 4.3 shows the fitted smoothing curve doesn't cover the true function in part although all variables were used to fit smoothing curve, which implies overfitting. After the variable selection method that we proposed, overfitting problem became moderate the overfitting problem.

Figure 4.4 shows the plot of nonparametric components having insignificant effect. The meaning of each row is equal to those of Figure 4.3. Among the functions, f_8 shows overfitting problem like the cases of f_1 and f_2. The result of the simulation of f_8 shows a nonlinear effect when the variable selection method was not used although f_8 has an insignificant effect. However, our proposed model fitted the smoothing curve to the true function well by excluding all the variables corresponding to f_8 based on the variable selection method.

Meanwhile, Figure 4.5 shows the posterior distributions of parametric components and Figure 4.6 shows that of error variance. The dashed line represents the true value of each parameter. In Figure 4.5 and Figure 4.6, the marginal posterior distribution are well covered the true vertical line, which reinforces the validity of the model that we proposed. In conclusion, we can select the final model without determining in advance which variables have to be belonged to nonparametric or parametric components by our proposed model. In addition, we can not only build the additive partial linear model parsimoniously but also alleviate the overfitting problem.

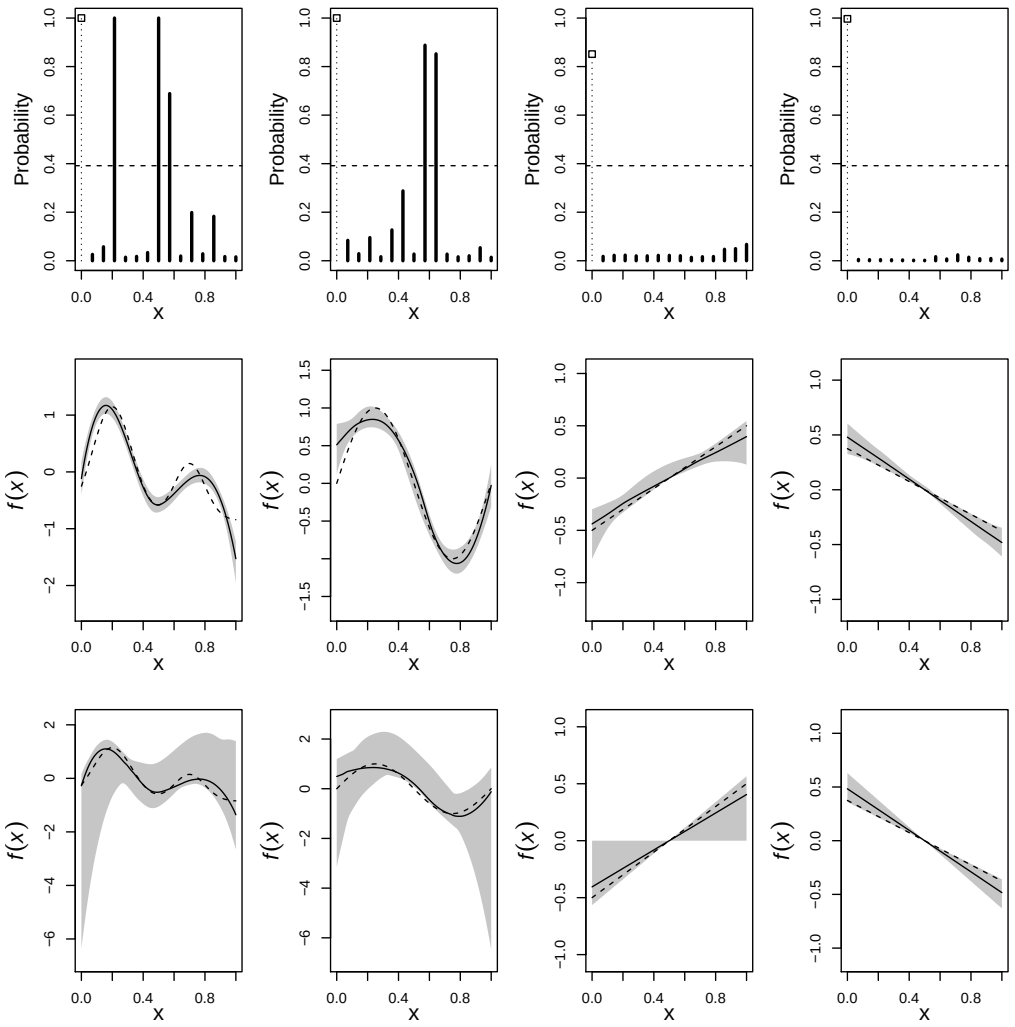

Figure 4.3: Simulation result plot for significant nonparametric components. The above panels show the probabilities that each coefficient is chosen and the below panels show the true functions(dashed lines), point-wise medians of the functions(solid lines) and point-wise 95% intervals(gray regions).

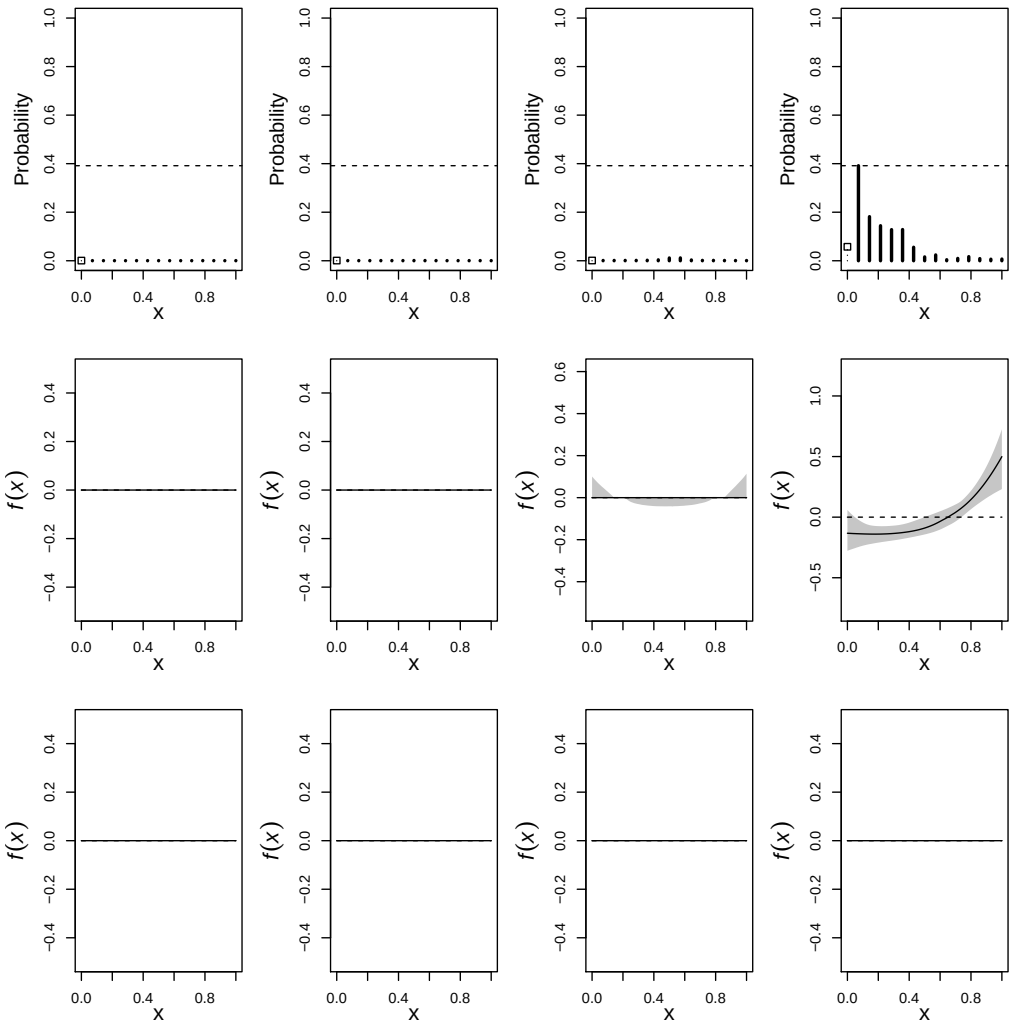

Figure 4.4: Simulation result plot for insignificant nonparametric components. The above panels show the probabilities that each coefficient is chosen and the below panels show the true functions(dashed lines), point-wise medians of the functions(solid lines) and point-wise 95% intervals(gray regions).

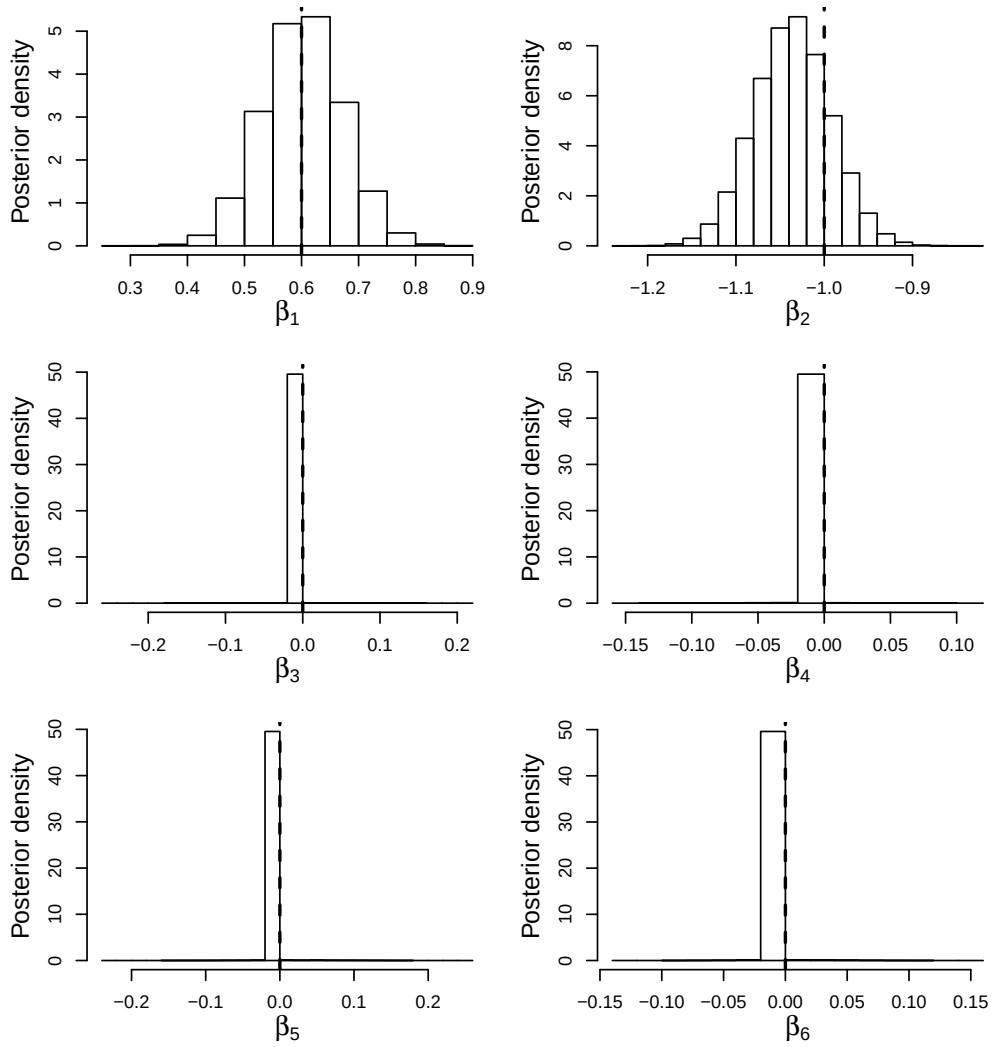

Figure 4.5: Historgram of marginal posterior distributions for parametric components. The vertical dotted line represents true value.

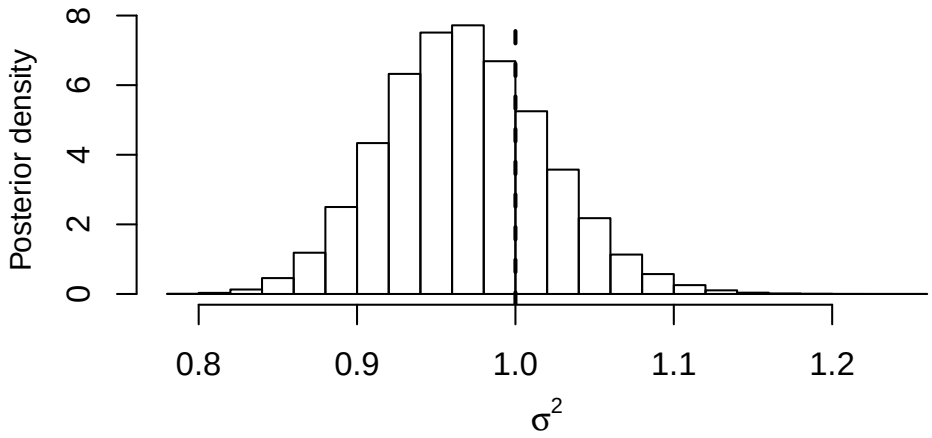

Figure 4.6: Historgram of marginal posterior distributions for error variance. The vertical dotted line represents true value.

Parameter	True	LASSO	SCAD	Proposed
β_1	0.6	0.572	0.589	0.603
β_2	-1	-0.980	-0.975	-1.046
β_3	0	0.008	0	0
β_4	0	1.5e-4	-0.098	0
β_5	0	-0.052	-0.013	0
β_6	0	-0.002	0	0

Table 4.1: Comparison of methods in categorical variables. The values of the proposed model is the median of simulation results.

To reinforce the validity of the proposed model, we compared the model with other models using the simulation data. With the models using LASSO method and SCAD method, we reran the simulation. As did in the proposed model, we did not predetermine the expected effects of covariates before doing the simulation.

The Figure 4.7 and Figure 4.8 were constructed by adding the curves made by other methods such as LASSO and SCAD on the third row plots of Figure 4.3 and Figure 4.4. In the case of functions having significant effects, they show a similar result. However, in the case of the functions having insignificant effects, other models have difficulty in detecting insignificant effects. Except for the result of SCAD in f_6, other compared methods failed to produce exact zero. Also, in the case of parametric components composed of categorical variables as shown in the Table 4.1, other methods could not make exactly zero. Among the 4 insignificant effects, SCAD could detect 2 variables as insignificant and Lasso detected 0 variables as insignificant by not producing zero but producing tiny values. As a result, we started the whole model with 126 coefficients but have built the model with 17 coefficients when using the SCAD and 35 coefficients when using the Lasso. Therefore, we can approximate the curves and select the variables more parsimoniously with the proposed model.

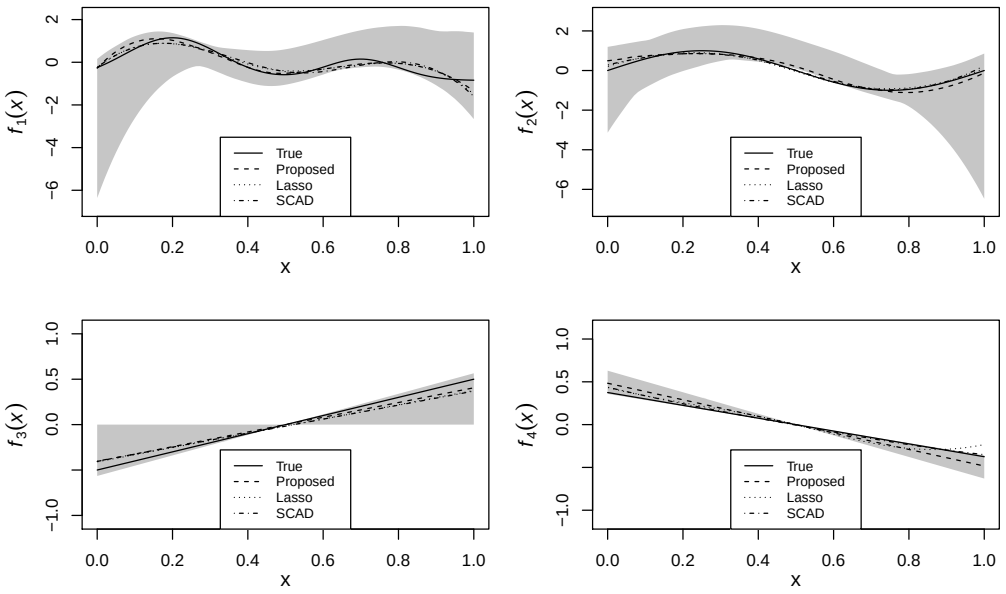

Figure 4.7: Simulation data plot comparing with other models for significant nonparametric components. The gray regions represent point-wise 95% intervals.

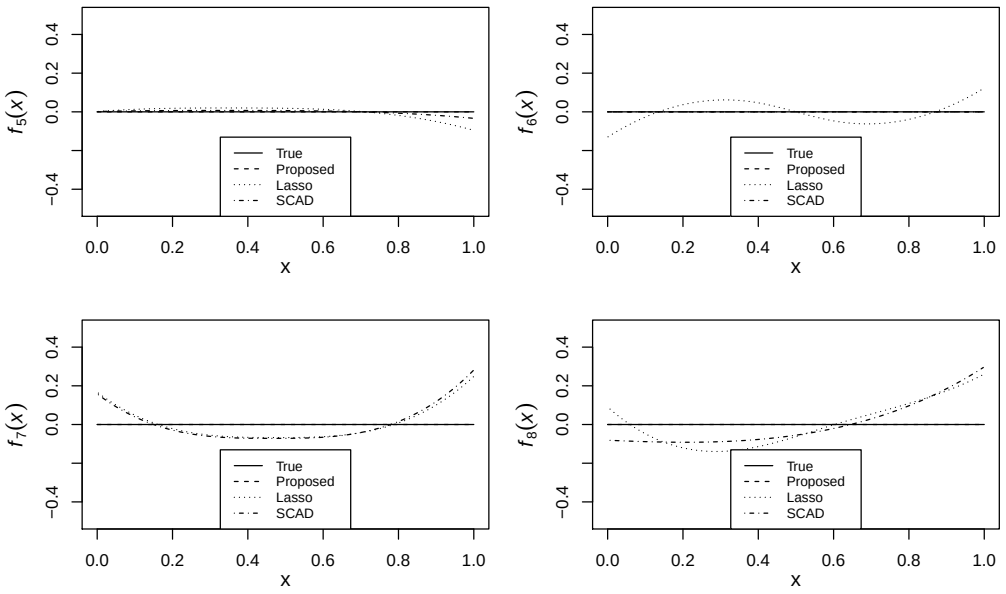

Figure 4.8: Simulation data plot comparing with other models for insignificant nonparametric components.

Chapter 5

Real Data Analysis

There have been many kinds of study that discuss the relationship between cancers and beta-carotene. Such epidemiological studies have shown that the influential antioxidant properties beta-carotene has can help prevent cancer. In addition, beta-carotene has the ability to purge the body of free radicals which can lead to cancer. Therefore, many researchers such as clinicians and nutritionists have tried to investigate the relationship between concentrations of beta-carotene and other factors like age, alcohol consumption, dietary intake, a smoking status that can help them to make a clinical decision and to customize the therapy. For example, Nierenberg et al. (1989) found that there was a positive relationship between beta-carotene and not only sex but also dietary carotene, whereas smoking status and body mass index(BMI) has a negative relationship. However, Faure et al. (2006) found that beta-carotene was related to age, gender, smoking status. On the other hand, Banerjee and Ghosal (2014) chose the smokestat, bmi, vituse, fat and fiber as linear effects and cholesterol as a nonlinear effect. In sum, there have been various results which make themselves less convincing.

The nutritional epidemiology data was made from a cross-sectional study. The dependent variable is concentrations of beta-carotene from 315 subjects and it was trans-

formed to $\log$(beta-carotene). In this procedure, a subject was deleted because of a numerical error caused by logarithm transformation. Liu et al. (2011) predetermined and the relationships between the beta-carotene and other variables. Especially, AGE and CHOL variables were assumed that they had a nonlinear effect based on the previous investigation. We added some variables such as RETDIET and VITUSE as Banerjee and Ghosal (2014) did but did not preassume their effects. The factors for investigating relationship among them are as follows: AGE(years), BMI(weight/height^2), CALORIES(number of calories consumed per day), FAT(grams of fat consumed per day), FIBER(grams of fiber consumed per day), ALCOHOL(number of alcohol drinks per day), CHOL(mg/day), BETADIET(mcg/day), RETDIET(mcg/day), SEX(1=Male, 2=Female), SMOKESTAT(1=Never, 2=Former, 3=Current Smoker), VITUSE(vitamin use, 1=fairly often, 2=not often, 3=no). To this end, the used model is as follows.

$$\begin{aligned}\log(\text{beta-carotene}) = \beta_0 &+ f_1(\text{AGE}) + f_2(\text{BMI}) + f_3(\text{CALORIES}) + f_4(\text{FAT}) \\ &+ f_5(\text{FIBER}) + f_6(\text{ALCOHOL}) + f_7(\text{CHOL}) + f_8(\text{BETADIET}) \\ &+ f_9(\text{RETDIET}) + \beta_1\text{SEX} + \beta_2\text{SMOKESTAT} + \beta_3\text{VITUSE} + \epsilon\end{aligned}$$

The radial basis was used and the number of knots was determined to 4. Therefore, the number of basis terms in the basis functions for each nonparametric component is 5. Also, we ran the simulation with 100,000 iterations. Except for the categorical variables such as SEX, SMOKE and VITUSE, we consider all variables as nonparametric components expecting that the proposed model can detect the linear effects although the variables are fitted by the nonparametric components.

As shown in Figure 5.1, ϕ_α is 0.526 when the α was set to 0.15 and 3 basis terms in AGE variable were selected, which implies AGE has a nonlinear effect. Also, only the first basis term representing linear effect was chosen in BMI and FIBER variables which suggest they have a linear effect. Other variables fitted by the nonparametric component doesn't represent significance. In the case of categorical variables, SMOKESTAT and

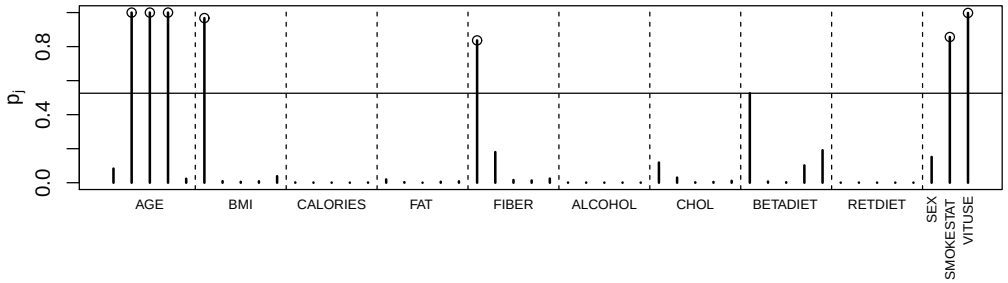

Figure 5.1: Plot of estimates of all p_js for real data. The p_js circled on the top were selected.

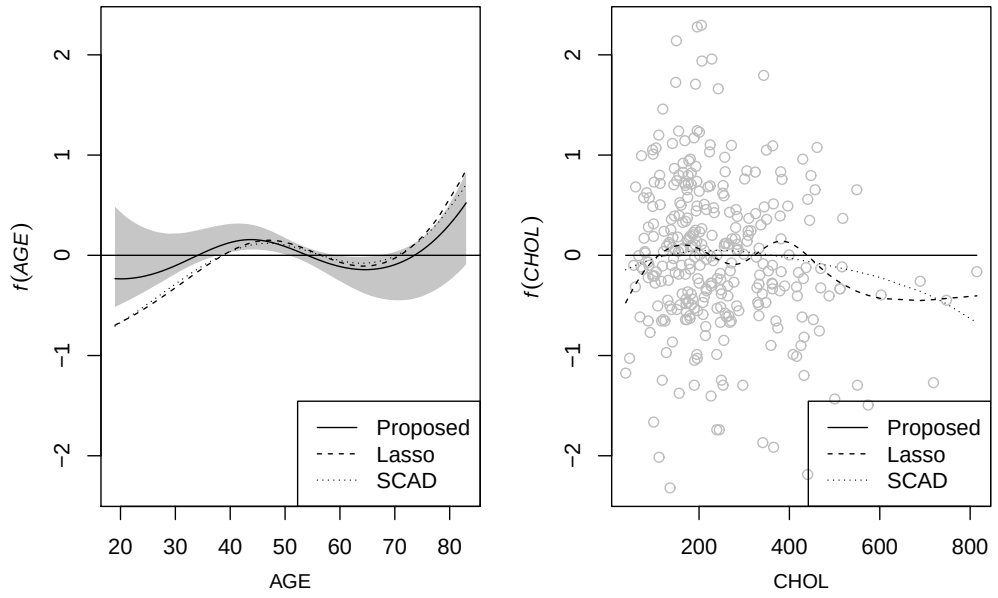

Figure 5.2: Real data plot comparing with other models. The left panel shows point-wise 95% intervals(gray regions). The right panel shows results on the scatter plot.

VITUSE variables show a linear effect. Among the 48 coefficients, 7 coefficients were selected.

Meanwhile, we compared the proposed model with other methods selecting variables such as LASSO and SCAD. Unlike the proposed model, they predetermined which vari-

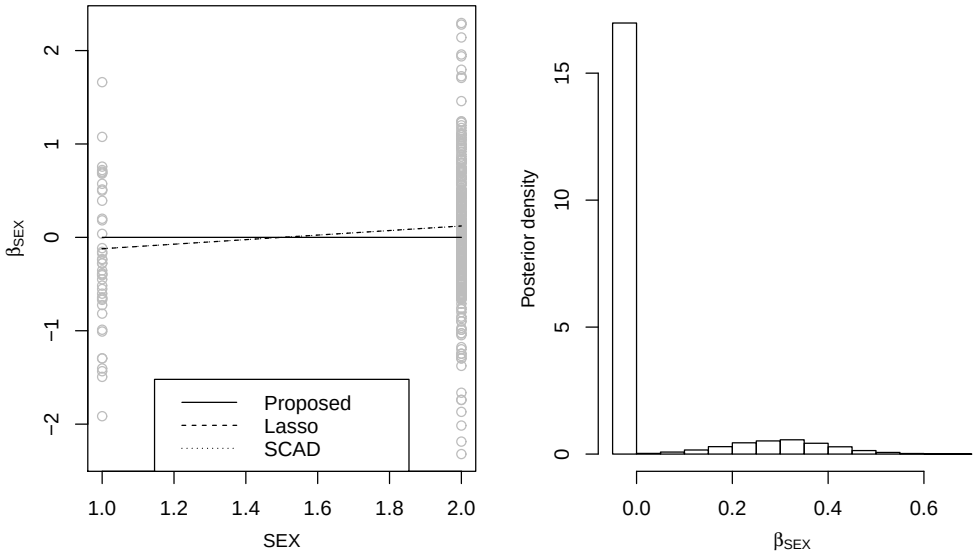

Figure 5.3: Nutritional epidemiology data plot and histogram comparing with other models for SEX variable

able a nonlinear effect has through a prior investigation. Although the proposed model doesn't assume any effect in the variables, it shows a similar result as other 2 methods do. The significant difference between the models is CHOL and SEX. SCAD and LASSO method has assumed that CHOL and AGE included nonlinear effect respectively, but the fitted smoothing curve of CHOL varies from method to method while fitted smoothing curves of AGE variable show similar results as shown in Figure 5.2. SEX variable was selected in SCAD and LASSO methods except for the proposed model. The histogram shown in Figure 5.3 means that the posterior distribution of the coefficient for SEX is a mixture distribution of two distributions. Our proposed model determined coefficients for SEX is closer to zero between the two distributions when the given data are considered. Another posterior distribution is located near the values inferred by other methods and Table 5.1 shows the values from other models. From the first to seventh row of Table

Covariates	$p*$	Method		
		SCAD	LASSO	Proposed
BMI	1	-0.032	-0.031	-0.035
CALORIES	0	-1.6e-04	-8.5e-05	0
FAT	0	0	-4.4e-04	0
FIBER	1	0.024	0.024	0.034
ALCOHOL	0	5.0e-03	3.2e-03	0
BETADIET	0	5.0e-05	4.1e-05	0
RETDIET	0	3.5e-05	3.9e-05	0
SEX	0	0.244	0.242	0
SMOKESTAT	1	-0.138	-0.132	-0.201
VITUSE	1	-0.137	-0.137	-0.215
AGE	3	nonlinear	nonlinear	nonlinear
CHOL	0	nonlinear	nonlinear	insignificant

Table 5.1: The result of proposed model and comparison with other models. $p*$ is the number of selected covariates from the proposed model. The median was used for the proposed model.

5.1 represent covariates determined to have a linear effect or no effect. Next 3 rows represent results of categorical variables and last 2 rows represent covariates predetermined as a nonlinear effect in other models. In Table 5.1, SCAD and LASSO methods show results close to zero but not exactly zero about some covariates, which implies overfitting

Parameter	mean	median	95% lower	95% upper
Intercept	4.960	4.960	4.885	5.032
BMI	-0.035	-0.035	-0.047	-0.022
FIBER	0.033	0.034	0.019	0.048
SMOKESTAT	-0.201	-0.201	-0.305	-0.098
VITUSE	-0.215	-0.215	-0.300	-0.129
σ^2	0.437	0.434	0.365	0.520

Table 5.2: Summary for parameters and covariates determined as linearly significant. The credible interval is constructed except for zero.

problems. Also, the confidence interval of the covariates judged as a linear effect was constructed except for the zero. The values inferred by other methods are located in the interval represented in Table 5.2. The intercept was introduced by normal distribution with mean of log(beta-carotene) and variance $\frac{\sigma^2}{n}$ (Park and Casella, 2008)

Chapter 6

Discussion

We proposed a methodology selecting variables in additive partial linear models using local false discovery rate in this paper. We focused on replacing free knot selection method used traditionally with selecting corresponding coefficients. In addition, we determine the effects later by a criterion based on the whole local false discovery rates calculated in the simulation without predetermining the attribute given variables such as the model suggested by Liu et al. (2011) or Banerjee and Ghosal (2014). In this procedure, we tried to let the model detect a linear effect with linear basis term, not nonlinear basis term by adding a linear basis term in the used spline. Through a simulation study, the proposed method was verified selecting significant variables and determining well not only the existences of effects but also the type of effect. Also, the overfitting problem was reduced by the proposed method. This method can be extended to other generalized linear models such as logit or log-linear model according to the type of dependent variable in the future study.

References

Banerjee, S. and Ghosal, S. (2014). Bayesian variable selection in generalized additive partial linear models. *Stat*, 3(1):363–378.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300.

Bonato, V., Baladandayuthapani, V., Broom, B. M., Sulman, E. P., Aldape, K. D., and Do, K.-A. (2010). Bayesian ensemble methods for survival prediction in gene expression data. *Bioinformatics*, 27(3):359–367.

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). Empirical bayes analysis of a microarray experiment. *Journal of the American Statistical Association*, 96(456):1151–1160.

Faure, H., Preziosi, P., Roussel, A., Bertrais, S., Galan, P., Hercberg, S., and Favier, A. (2006). Factors influencing blood concentration of retinol, α-tocopherol, vitamin c, and β-carotene in the french participants of the su. vi. max trial. *European Journal of Clinical Nutrition*, 60(6):706–717.

Hastie, T. J. (2017). Generalized additive models. In *Statistical models in S*, pages 249–307. Routledge.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Liu, X., Wang, L., and Liang, H. (2011). Estimation and variable selection for semiparametric additive partial linear models (ss-09-140). *Statistica Sinica*, 21(3):1225–1248.

Miyata, S. and Shen, X. (2003). Adaptive free-knot splines. *Journal of Computational and Graphical Statistics*, 12(1):197–213.

Morris, J. S., Brown, P. J., Herrick, R. C., Baggerly, K. A., and Coombes, K. R. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64(2):479–489.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R., and Group, S. C. P. S. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, 130(3):511–521.

Park, T. and Casella, G. (2008). The bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Sarkar, S. K., Zhou, T., and Ghosh, D. (2008). A general decision theoretic formulation of procedures controlling fdr and fnr from a bayesian perspective. *Statistica Sinica*, 18(3):925–945.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zhao, Z. and Sarkar, S. K. (2015). A bayesian approach to constructing multiple confidence intervals of selected parameters with sparse signals. *Statistica Sinica*, 25(2):725–741.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

국 문 요 약

지역 오발견율을 이용한

가법 부분 선형 모형에서의 변수선택

가법 부분 선형 모형은 모수적 구성요소와 비모수적 구성요소로 이루어져 있으며 주어진 데이터를 설명할 때 모수적 구성요소의 단순성과 비모수적 구성 요소의 유연성을 이용하기 위해 사용되어 왔다. 본 논문은 지역 오발견율에 기초한 변수 선택 방법에 의해 가법 부분 선형 모델을 구성하는 것을 설명한다. 영 과잉 혼합물 사전분포였를 이용하여 수정된 베이지안 라쏘 방법은 비모수적 구성요소를 추론할 때 필요한 전통적인 매듭 선택 방법을 대체한다. 또한, 설명변수가 모수적 또는 비모수적 구성요소에 의해 추론 될 때 선형효과를 감지하는 기능을 추가하였다. 제안된 방법론은 몬테카를로 마르코프 체인 시뮬레이션 연구 및 역학 데이터 분석을 통해 검증하였다. 본 연구를 통해 제안된 모형이 기존에 독립변수가 종속변수에 미치는 영향을 미리 정하지 않고도 정확한 효과를 판별해낼 수 있으며 과잉적합문제 또한 완화시킬 수 있음을 보였다.

핵심 용어: 비모수적 베이지안 추론, 베이지안 라쏘, 지역 오발견율, 영 과잉 혼합물 사전분포, 변수선택