Journal of Mathematics and Statistics 9 (3): 161-168, 2013
ISSN: 1549-3644

doi:10.3844/jmssp.2013.161.168 Published Online 9 (3) 2013 (http://www.thescipub.com/jmss.toc)

SMOOTHING SPLINE IN SEMIPARAMETRIC ADDITIVE REGRESSION MODEL WITH BAYESIAN APPROACH

[1]Rita Diana, [2]I. Nyoman Budiantara, [2]Purhadi and [3]Satwiko Darmesto

[1,2]Department of Statistics, Faculty of Mathematics and Natural Sciences,
Sepuluh Nopember Institute of Technology, Jl. Arif Rahman Hakim, Surabaya 60111, Indonesia
[1,3]The Indonesian Central Bureau of Statistics, Jakarta, Indonesia

Received 2013-02-13, Revised 2013-04-05; Accepted 2013-06-12

ABSTRACT

Semiparametric additive regression model is a combination of parametric and nonparametric regression models. The parametric components are not linear but following a polynomial pattern, while the nonparametric components are unknown pattern and assumed to be contained in the Sobolev space. The nonparametric components can be approximated by smoothing spline functions. In the development of smoothing spline, the classical statistical approach cannot be applied for solving the inference problem such as constructing confidence intervals for the regression curve. To construct confidence interval of smoothing spline curve in the semiparametric additive regression model, we propose to use Bayesian approach, by assuming improper Gaussian distribution for prior distribution in nonparametric components and multivariate normal distribution for parametric components. In this study, we obtain parameter estimators for parametric component and smoothing spline estimators for the nonparametric component in semiparametric additive regression model. Moreover, we also develop a smoothing parameters selection method simultaneously using Generalized Maximum Likelihood (GML) and confidence intervals for the parameters of the parametric component and the smoothing spline functions of the nonparametric component using Bayesian approach. By computing each posterior mean and posterior variance of parametric component parameters and smoothing spline functions, confidence intervals can be constructed for the parametric component parameters and confidence interval smoothing spline functions for nonparametric components in semiparametric additive regression models. We create R-code to implement estimation model and inference procedure. Our simulation studies reveal estimation and inference method perform reasonably well.

Keywords: Bayesian, GML, Confidence Interval, Semiparametric Additive Regression Model, Smoothing Spline

1. INTRODUCTION

In regression model, there are some components which have sufficient information to describe the relationship pattern between the predictors and the response variables. However, there are also vague or nuisance components. Hence, in this study, the semiparametric additive regression model is used to overcome these difficulties. This model is a combination of parametric and nonparametric regression models. The parametric components are not linear but follow a polynomial pattern while the nonparametric components are unknown pattern and assumed to be contained in the Sobolev space. The nonparametric component can be approximated by using functions such as spline, polynomial local or kernel. Among these approximations, spline function has high flexibility and capability to handle the data with changing behavior in certain sub-intervals (Eubank, 1999). It has also been pointed out by (Liang, 2006; Aydin, 2008) that compares

Corresponding Author: Rita Diana, Department of Statistics, Faculty of Mathematics and Natural Sciences,
Sepuluh Nopember Institute of Technology, Jl. Arif Rahman Hakim, Surabaya 60111, Indonesia
Tel: +6285232078886 Fax: +62315922940

the smoothing spline technique with kernel in semiparametric and nonparametric regression.

There are three common approaches for estimations the regression function by using spline in semiparametric regression model i.e., are regression spline (truncated spline, cubic spline, B-spline), penalized spline (p-spline) and smoothing spline. For penalized spline and regression spline, we need to be careful in determining the number and location of the knots, whereas the smoothing spline does not require the selection of knots. Furthermore, the performance of smoothing spline in semiparametric regression model is better and more flexible than the penalized and regression spline (Aydin and Tuzemen, 2010). Smoothing spline estimator can be obtained by classical approaches such as Penalized Least Square (PLS), Penalized Maximum Likelihood or Penalized Likelihood. Beside the classical approach, these estimators in semiparametric regression can also be estimated by Bayesian approach (Wang, 2011).

Penalized and regression spline research in regression model of semiparametric additive with Bayesian approach has been developed by many experts. Among them, for instance: (Wong and Kohn, 1996; Li, 2000; Smith *et al*., 2000; Kandala *et al*., 2001; Panagiotelis and Smith, 2008; Ryu *et al*., 2011) were those who use regression spline with Bayesian approach, while (Lang and Brezger, 2004; Jerak and Wagner, 2006; Nott, 2006; Costa, 2008; Marley and Wand, 2010; Shen, 2011) were those who use p-spline with Bayesian approach. Wang (2011) used a smoothing spline in semiparametric regression model whose parametric components are linear patterned with Bayesian approach. However estimation of regression function by smoothing spline in semiparametric additive regression model which parametric components are not linear (polynomial) using Bayesian approach have not existed yet.

In the development of smoothing spline, the classical statistical approach cannot be applied for solving the inference problem such as constructing a confidence intervals for the regression curve. Therefore, some researchers use Bayesian approach for building such confidence intervals for smoothing spline function. (Wahba, 1983; Nychka, 1988) have used a Bayesian approach to construct confidence intervals of smoothing spline on nonparametric model. In this study we developed approaches about the smoothing spline in semiparametric additive regression models with parametric components which are not linear (polynomial) by using Bayesian approach. We also developed some methods for selecting optimal smoothing parameters simultaneously in semiparametric additive regression models as well as building confidence intervals for the parametric component parameters and smoothing spline functions. We proposed this method with simulation data.

2. MATERIALS AND METHODS

This chapter discusses some theories used for building smoothing spline estimators in semiparametric additive regression model with Bayesian approach.

2.1. Semiparametric Additive Regression Model

Suppose sampling observations $(x_{1j},...,x_{pj}, z_{1j},...,z_{qj}, y_j)$ with y_j as response variable and $j = 1,2,\ldots,n$ shows the amount of observations. X_{ij} are predictor variables for $i = 1,2,\ldots,p$ which have not linier patterns to response variable but follow the polynomial pattern. z_{kj} are predictor variables for $k = 1,2,\ldots,q$ which have unknown relationship pattern with response variable. The relationship between $(x_{1j},...,x_{pj})$, $(z_{1j},...,z_{qj})$ and y_j are modeled by semiparametric additive regression:

$$y_j = \sum_{i=1}^{p}\left(\sum_{h=0}^{r}\gamma_{hi}x_{ij}^{h}\right) + \sum_{k=1}^{q} f_k(z_{kj}) + \varepsilon_j \qquad (1)$$

The parameter γ_{hi} is parameter vector of the unknown parametric components for $h = 0,1,2,\ldots,r$; $i = 1,2,\ldots,p$. Random error ε_j are assumed mutually independent and normally distributed with zero mean and variance σ^2. According to Wang (2011) the shape of regression curve f_k is unknown and assumed to be contained in the Sobolev space:

$$W_2^m[a_k, b_k] = \left\{f_k ; \int_{a_k}^{b_k} (f_k^{(m)}(z_k))^2 \, dz_k < \infty\right\}$$

The Equation 1 can be written as follow:

$$y_j = \sum_{i=1}^{p} x_{ij}^{T}\gamma_i + \sum_{k=1}^{q} f_k(z_{kj}) + \varepsilon_j$$

where, $\gamma_i = (\gamma_{0i}, \gamma_{1i}, \gamma_{2i},\ldots, \gamma_{ri})^T$ and $x_{ij}^T = (1, x_{ij}, x_{ij}^2,\ldots, x_{ij}^r)$ for $i=1,2,\ldots,p$. Nonparametric regression curve f_k is estimated by PLS method by minimizing:

$$n^{-1}\sum_{j=1}^{n}(y_i - \sum_{i=1}^{p} x_{ij}^{T}\gamma_i - \sum_{k=1}^{q} f_k(z_{kj}))^2 + \lambda\sum_{k=1}^{q}\theta_k^{-1}\int_{a_k}^{b_k}(f_k^{(m)}(z_k))^2 dz_k \qquad (2)$$

where, the parameter $\lambda_k = \lambda/\theta_k$, k = 1,2,…,q controls the balance between goodness of fit and smoothness measure function.

To get the smoothing spline estimators in semiparametric additive regression model, Wang (2011) used an extension from Wahba (1990). The form of general spline function used is:

$$\sum_{k=1}^{q} f_k(z_k) = \sum_{k=1}^{q}\sum_{v=1}^{m_k} \alpha_{kv}\phi_{kv}(z_k) + \sum_{k=1}^{q}\sum_{j=1,l=1}^{n} \beta_{kj}\theta_k\psi_k(z_{kj}, z_{kl}) \quad (3)$$

Let $\alpha_k = (\alpha_{k1},...,\alpha_{km_k})^T$, $\beta_k = (\beta_{k1},\ldots,\beta_{kn})^T$, $X_i = \left\{x_{ij}^h\right\}_{h=0,j=1}^{r_i,n}$, $T_k = \{\phi_{kv}(z_{kj})\}_{j=1,v=1}^{n,m_k}$ where $\phi_{kv}(z_{kj}) = z_{kj}^{v-1}/(v-1)!$, $v = 1,2,...,m_k$ and $\theta_k V_k = \theta_k\{\psi_k(z_{kj}, z_{kl})\}_{j=1,l=1}^{n,n}$ with $\psi_k(z_{kj}, z_{kl}) = \int_{a_k}^{b_k} \frac{(z_{kj}-u)_+^{m-1}(z_{kl}-u)_+^{m-1}}{[(m-1)!]^2}du$ then Equation 3 can be written in matrix form $f_k = T_k\alpha_k + \theta_k V_k\beta_k$ for k=1,2,...,q and Equation 2 can be solved by minimizing:

$$\begin{aligned} &n^{-1}(y - \sum_{i=1}^{p} X_i\gamma_i - \sum_{k=1}^{q} T_k\alpha_k - \sum_{k=1}^{q}\theta_k V_k\beta_k)^T \\ &\times(y - \sum_{i=1}^{p} X_i\gamma_i - \sum_{k=1}^{q} T_k\alpha_k - \sum_{k=1}^{q}\theta_k V_k\beta_k) + \lambda\sum_{k=1}^{q}\beta_k^T\theta_k V_k\beta_k \end{aligned} \quad (4)$$

If $\mu = (\gamma^T,\alpha^T)^T$, $\gamma = (\gamma_1^T,...,\gamma_p^T)^T$, $\alpha = (\alpha_1^T,...,\alpha_q^T)^T$, $\beta = (\beta_1,\ldots,\beta_n)^T$, S = (X T), X = ($X_1,\ldots,X_p$), T = ($T_1,\ldots T_q$) and $V_\theta = \theta_1V_1+\ldots+\theta_qV_q$ then Equation 4 can be written as:

$$n^{-1}(y - S\mu - V_\theta\beta)^T(y - S\mu - V_\theta\beta) + \lambda\beta^T V_\theta\beta \quad (5)$$

By taking partially derivative of Equation 5 with respect to μ and β then the results are to be equal to zero, we obtain:

$$\hat{\beta} = M^{-1}(y - S\hat{\mu}) = (M^{-1} - M^{-1}S(S^TM^{-1}S)^{-1}S^TM^{-1})y \quad (6)$$

$$\hat{\mu} = (\hat{\gamma}^T,\hat{\alpha}^T)^T = (S^TM^{-1}S)^{-1}S^TM^{-1}y \quad (7)$$

with I is the identity matrix and $M = V_\theta + n\lambda I$. Based on Equation 6 and 7, we obtain estimators for smoothing spline in semiparametric multivariabel regression model that can be expressed as follows $\sum_{k=1}^{q}\hat{f}_k = T\hat{\alpha} + V_\theta\hat{\beta}$ and $\hat{h} = X\hat{\gamma} + \sum_{k=1}^{q}\hat{f}_k$ or $\hat{h} = X\hat{\gamma} + T\hat{\alpha} + V_\theta\hat{\beta}$.

3. RESULTS

3.1. Parameter Estimation

In Bayesian approach, selection of the prior distribution is very important. The prior distribution used in this study is restricted to improper Gaussian prior distribution for the nonparametric component and multivariate normal distribution for parametric components. In Bayesian approach case, the point estimation is obtained from posterior mean and the interval estimation is obtained from its posterior variance.

Given sampling observation (x_{ij}, z_{kj}, y_j), j = 1,…n; i = 1,…,p; k = 1,…,q can be obtained from stochastic process {y(x, z), x, z ∈ [a,b]} and follows the model (1). Prior distribution f_k is defined as Equation 8:

$$\sum_{v=1}^{m_k}\alpha_{kv}\phi_{kv}(z_k) + \eta^{1/2}\theta_k^{1/2}g_k(z_k) \quad (8)$$

where, $\alpha \sim N(0, \tau I)$, $\tau \to \infty$, η is positive constant and $\{g_k(z_k); z_k \in [a_k,b_k]\}$ is integrated Weiner process with:

$$g_k(z_k) = \int_{a_k}^{z_k} \frac{(z_k - u)^{m-1}}{(m-1)!}dW(u)$$

where, W(u) is Weiner process with zero mean and reproducing kernel covariance θ_kV_k. Moreover, α and $g_k(z_k)$ are mutually independent. Hence, {h(x,z),x,z ∈ [a,b]} have prior distribution of improper:

$$\begin{aligned} h(x,z) &= \sum_{i=1}^{p} x_i^T\gamma_i + \sum_{k=1}^{q} f_k(z_k) \\ &= \sum_{i=1}^{p} x_i^T\gamma_i + \sum_{k=1}^{q}\sum_{v=1}^{m_k}\alpha_{kv}\phi_{kv}(z_k) + \eta^{1/2}\sum_{k=1}^{q}\theta_k^{1/2}g_k(z_k) \end{aligned} \quad (9)$$

where, $\gamma \sim N(0, \delta I)$, $\delta \to \infty$, $\eta\theta_k = \sigma^2/\lambda_k$ and {ε(z)} is Gaussian process with zero mean and $Cov(\varepsilon(z_j),\varepsilon(z_1)) = \sigma^2$ for j=l and zero for others.

Let y, h and ε given as Gaussian random vector with zero mean and follow model (1) in which $E(\varepsilon\varepsilon^T)=\sigma^2I$, $E(h\varepsilon^T) = 0$ and $E(hh^T) = \eta\sum_{k=1}^{q}\theta_k V_{h_k} = \eta V_{\theta_h}$. If m has normal multivariate distribution with E(m) = 0, $E(m\varepsilon^T) = 0$, $E(mm^T) = \eta\sum_{k=1}^{q}\theta_k V_{m_k} = \eta V_{\theta_m}$ and

$E(mh^T) = \eta\sum_{k=1}^{q}\theta_k V_{m_k h_k} = \eta V_{\theta_{mh}}$ then joint distribution between m and y i.e.:

$$\begin{pmatrix} m \\ y \end{pmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \eta\sum_{k=1}^{q}\theta_k V_{m_k} & \eta\sum_{k=1}^{q}\theta_k V_{m_k h_k} \\ \eta\sum_{k=1}^{q}\theta_k V_{h_k m_k} & \eta\sum_{k=1}^{q}\theta_k V_{h_k} + \sigma^2 I \end{bmatrix}\right)$$

Hence, by using a standard result on multivariate normal distribution e.g., Johnson and Wichern (2001) (Result 4.6), the conditional distribution of m given y and X is normal with each of mean and covariance Equation 10 and 11:

$$\begin{aligned} E(m \mid y) &= \sum_{k=1}^{q}\theta_k V_{m_k h_k}\left(\sum_{k=1}^{q}\theta_k V_{h_k} + n\lambda I\right)^{-1} y \\ &= V_{\theta_{mh}}(V_{\theta_h} + n\lambda I)^{-1} y, \end{aligned} \quad (10)$$

and:

$$\begin{aligned} Var(m \mid y) &= \eta\Big(\sum_{k=1}^{q}\theta_k V_{m_k} - \sum_{k=1}^{q}\theta_k V_{m_k h_k} \\ &\quad \times \Big(\sum_{k=1}^{q}\theta_k V_{h_k} + n\lambda I\Big)^{-1}\sum_{k=1}^{q}\theta_k V_{h_k m_k}\Big) \end{aligned}$$

$$Var(m \mid y) = \eta(V_{\theta_m} - V_{\theta_{mh}}(V_{\theta_h} + n\lambda I)^{-1} V_{\theta_{hm}}) \quad (11)$$

with $\lambda = \sigma^2/n\eta$.

Equation 9 can be written in matrix form $h = X\gamma + T\alpha + \eta^{1/2}\sum_{k=1}^{q}\theta_k^{1/2} g_k(z_k)$ or $h = S\mu + \eta^{1/2}\sum_{k=1}^{q}\theta_k^{1/2} g_k(z_k)$ where $\mu \sim N(0, \pi I)$, $\pi \to \infty$. Assumed that μ with $g_k(z_k)$ are mutually independent, then we obtained $V_{\theta_h} = \zeta SS^T + \eta\sum_{k=1}^{q}\theta_k V_k$ with $\zeta = \pi/\eta$. If we take m=h then we obtain $V_{\theta_{mh}} = \zeta s^T S^T + \sum_{k=1}^{q}\theta_k \psi_k^T$ and $V_{\theta_m} = \zeta s^T s + \sum_{k=1}^{q}\theta_k \psi_k$.

Using quadratic loss function, we find the estimator bayes is the posterior mean h, hence we get:

$$E(h \mid y) = V_{\theta_{mh}}(V_{\theta_h} + n\lambda I)^{-1} y$$

$$E(h \mid y) = (\zeta s^T S^T + \sum_{k=1}^{q}\theta_k \psi_k^T)(\zeta SS^T + \sum_{k=1}^{q}\theta_k V_k + n\lambda I)^{-1} y$$

$$E(h \mid y) = \zeta s^T S^T(\zeta SS^T + M)^{-1} y + \sum_{k=1}^{q}\theta_k \psi_k^T(\zeta SS^T + M)^{-1} y \quad (12)$$

with $M = \sum_{k=1}^{q}\theta_k V_k + n\lambda I$.

If the limit of posterior mean value h is taken for $\zeta \to \infty$ we find that:

$$\lim_{\zeta\to\infty} \zeta S^T(\zeta SS^T + M)^{-1} y = (S^T M^{-1} S)^{-1} S^T M^{-1} y$$

$$\lim_{\zeta\to\infty} (\zeta SS^T + M)^{-1} y = (M^{-1} - M^{-1}S(S^T M^{-1} S)^{-1} S^T M^{-1}) y$$

by combining this results with Equation 6 and 7, we have $\lim_{\zeta\to\infty} E(h \mid y) = \hat{h}$. This result is identical with smoothing spline estimator that obtained from PLS approach. Based on prior (9) and Equation 12 obtained $\lim_{\zeta\to\infty} E(\gamma \mid y) = \hat{\gamma}$ and $\lim_{\zeta\to\infty} E(f_k \mid y) = \hat{f}_k$.

3.2. Smoothing Parameters Selection Method

Smoothing spline estimators depend on the smoothing parameter. Hence, the smoothing parameters selection are crucial for the performance of smoothing spline function estimates. A selection method of the smoothing parameters ($\lambda_1, \ldots, \lambda_q$) for smoothing spline estimators in semiparametric additive regression model using Bayesian approach, that is Generalized Maximum Likelihood (GML) is given as follows. The basic idea of using GML method was firstly done by Wahba (1985) in a nonparametric regression model.

If given w_1 and w_2 with decompotition as follows:

$$\begin{pmatrix} w_1 \\ \cdots \\ w_2 \end{pmatrix} = \begin{pmatrix} F^T \\ \cdots \\ \varphi^{-1/2} S^T \end{pmatrix} y,$$

where S is a $(n \times (\sum_{i=1}^{p} r_i + \sum_{k=1}^{q} m_k))$ matrix and F is a $(n \times (n - \sum_{i=1}^{p} r_i - \sum_{k=1}^{q} m_k))$ matrix that satisfies $F^T F = I$, $F^T S = 0$, $\varphi = \tau/\eta$ and h has improper prior distribution (9) then:

$$w_1 \sim N(0, \eta(F^T M F)$$

and:

$$w_2 \sim N(0, \eta(S^T S)(S^T S))$$

Determining both of these distributions, we found that only w_1 contain the smoothing parameters $\lambda_1,\ldots, \lambda_q$. Based on the distribution of w_1, the log likelihood function can be obtained as follows:

$$\log L(\lambda_1,\ldots,\lambda_q,\eta|w_1) = -\frac{n-\sum_{i=1}^{p} r_i-\sum_{k=1}^{q} m_k}{2}\log\eta + -\frac{1}{2}\log|F^TMF| - \frac{1}{2\eta}w_1^T(F^TMF)^{-1}w_1 + K_1$$

The log likelihood function gives the maximum likelihood estimator, which is:

$$\hat{\eta} = \frac{w_1^T(F^TMF)^{-1}w_1}{n-\sum_{i=1}^{p} r_i-\sum_{k=1}^{q} m_k}.$$

By subsitution of $\hat{\eta}$ into log likelihood function, we get:

$$\log L(\lambda_1,\ldots,\lambda_q|w_1) = -\frac{n-\sum_{i=1}^{p} r_i-\sum_{k=1}^{q} m_k}{2}\log\left(\frac{w_1^T(F^TMF)^{-1}w_1}{\left|(F^TMF)^{-1}\right|^{1/\left(n-\sum_{i=1}^{p} r_i-\sum_{k=1}^{q} m_k\right)}}\right) + K_2$$

with K_1 and K_2 are constants which not depend on $\lambda_1,\ldots,\lambda_q$ and η. Maximizing $\log L(\lambda_1,\ldots,\lambda_q|w_1)$ is equivalent to minimizing:

$$GML(\lambda_1,\ldots,\lambda_q) = \frac{w_1^T(F^TMF)^{-1}w_1}{\left|(F^TMF)^{-1}\right|^{1/\left(n-\sum_{i=1}^{p} r_i-\sum_{k=1}^{q} m_k\right)}}.$$

where $w_1 = F^Ty$. The values of $\lambda_1,\ldots,\lambda_q$ are optimal by minimizing of $GML(\lambda_1,\ldots,\lambda_q)$.

3.3. Confidence Interval

One way for constructing confidence intervals for semiparametric estimates is bootstrap and Bayesian approach. The disadvantage of the bootstrap confidence interval is that they are more computationally intensive. Hence, to compute the confidence interval of the smoothing spline functions f_k and the parameter γ in semiparametric additive regression model, we can use Bayesian approach. Based on the prior (9), Equation 11 and 12, we obtain:

$$\lim_{\zeta\to\infty} E(\gamma|y) = \hat{\gamma},\ \lim_{\zeta\to\infty} E(f_k|y) = \hat{f}_k$$

and

$$Var(m|y) = \eta\left\{\zeta s^Ts + \sum_{k=1}^{q}\theta_k\psi_k - (\zeta s^TS^T + \sum_{k=1}^{q}\theta_k\psi_k^T)\times (\zeta SS^T + M)^{-1}(\zeta Ss + \sum_{k=1}^{q}\theta_k\psi_k)\right\}.$$

Let m = h then we have:

$$Var(h|y) = \eta\left\{\sum_{k=1}^{q}\theta_k\psi_k + s^T(\zeta I - \zeta^2S^T(\zeta SS^T + M)^{-1}S)s + -2s^T(\zeta S(\zeta SS^T + M)^{-1})\sum_{k=1}^{q}\theta_k\psi_k^T + -\sum_{k=1}^{q}\theta_k\psi_k^T(\zeta SS^T + M)^{-1}\sum_{k=1}^{q}\theta_k\psi_k^T\right\}.$$

If the limit value of posterior variance h is taken for $\zeta\to\infty$ we find that:

$$\lim_{\zeta\to\infty} Var(h|y) = \eta\left\{\sum_{k=1}^{q}\theta_k\psi_k + s^T\left(S^TM^{-1}S\right)^{-1}s + -2s^T\left(S^TM^{-1}S\right)^{-1}S^TM^{-1}\sum_{k=1}^{q}\theta_k\psi_k + -\sum_{k=1}^{q}\theta_k\psi_k^T\left(M^{-1} - M^{-1}S\left(S^TM^{-1}S\right)^{-1}S^TM^{-1}\right)\sum_{k=1}^{q}\theta_k\psi_k\right\}.$$

Hence:

$$\lim_{\zeta\to\infty} Var(\gamma|y) = \eta A_{ii}.$$

where, A_{ii} is a matrix corresponds to rows 1 to $\sum_{w=1}^{p} r_w + 1$ and coloms 1 to $\sum_{w=1}^{p} r_w + 1$ of matrix $\left(S^TM^{-1}S\right)^{-1}$, and:

$$\lim_{\zeta\to\infty} Var(f_k|y) = \eta\left\{\theta_k\psi_k + \phi_k^TB_{kk}\phi_k - 2\phi_k^TC_k + -\sum_{k=1}^{q}\theta_k\psi_k^T\left(M^{-1} - M^{-1}S\left(S^TM^{-1}S\right)^{-1}S^TM^{-1}\right)\sum_{k=1}^{q}\theta_k\psi_k\right\}$$

where, B_{kk} is a matrix corresponds to rows $\sum_{w=1}^{k-1} m_w + 1$ to $\sum_{w=1}^{k} m_w$ and columns $\sum_{w=1}^{k-1} m_w + 1$ to $\sum_{w=1}^{k} m_w$ of matrix

$\left(S^TM^{-1}S\right)^{-1}$ and C_k is a subvector of $\left(S^TM^{-1}S\right)^{-1}S^TM^{-1}\sum_{k=1}^{q}\theta_k\psi_k$ with elements from $\sum_{w=1}^{k-1}m_w+1$ to $\sum_{w=1}^{k}m_w$.

Therefore, by calculating each posterior mean and posterior variance of parameters γ and f_k, we can construct confidence intervals for the parameters γ and confidence intervals for smoothing spline functions f_k in semiparametric additive regression model.

3.4. Simulation Study

In the simulations, we generated our data from the semiparametric additive regression model in (1) with n = 100, p = 4, r = 2, q = 2 and m = 2. For the parametric part, we set parameter of γ = (0.0, 1.0, 0.8, 1.4, 0.6, 1.2, 0.9, 1.1, 1.2)^T. The x_i's were generated from the multivariate normal distribution with zero mean and $Cov(x_{ij}, x_{ik}) = 0.5^{|j-k|}$. For the nonparametric part, the true functions were set to be:

$$f_1(z_1) = -2\sin(2\pi z_1)$$

and

$$f_2(z_2) = -8.03 + 9\exp(-(z_2 - 0.3)^2)$$

The z_{ik}'s were generated independently from the uniform distribution on [0, 1]. The random errors ε_j were generated from the normal distribution with zero mean and standard deviation σ = 0.9.

Based on GML method, we create R-code for choosing smoothing parameters simultaneously. The optimum smoothing parameter is used to obtain parameter estimator for parametric components and smoothing spline estimator for the nonparametric component in semiparametric additive regression model. The GML values with a number of different smoothing parameters for the simulation data is depicted in **Table 1**. We can see the changing of GML values from small (line 1) to large smoothing parameter (line 8). According to the eight smoothing parameters combination that are applied in the model, the optimal smoothing parameters are λ_1 = 2.083E-05 and λ_2 = 5.000E-05 having the smallest GML value of 48.01812 (line 4). Next, using the optimal smoothing parameter, the estimation of the parametric component γ and the 95% confidence interval of the parametric component appropriate can be seen from **Table 2**.

Figure 1 shows the estimation of smoothing spline function as well as the 95% confidence interval for σ = 0.9. The rounded lines are the true functions, solid lines represent the smoothing spline curves with optimal smoothing parameter and the dashed lines denote the confidence interval for lower and upper smoothing spline. If we compare it with the smoothing spline function estimates as well as the 95% confidence interval for σ = 1.0 and σ = 0.6 in **Fig. 2**, it is seen that the smoothing spline function estimation for σ = 0.6 is closer to the true function and also the perform reasonably well. It can also be seen from the Mean Square Error (MSE) for for three different error standar deviation is used in **Table 3**.

Table 1. The smoothing parameter and GML value

λ_1	λ_2	GML(λ_1, λ_2)
2.083E-02	5.000E-02	126.66030
2.083E-03	5.000E-03	116.24130
2.083E-04	5.000E-04	78.30659
2.083E-05	5.000E-05	48.01812
2.083E-06	5.000E-06	48.21623
2.083E-07	5.000E-07	48.37583
2.083E-08	5.000E-08	68.51872
2.083E-09	5.000E-09	123.57630

Table 2. Estimation parameters and 95% confidence intervals for parametric component

Parameter	Estimated value	95% Confidence Interval
γ_0	0.250197	(-0.134254, 0.634648)
γ_{11}	1.030157	(0.822267, 1.238047)
γ_{21}	0.772059	(0.654121, 0.889997)
γ_{12}	1.324766	(1.062530, 1.587002)
γ_{22}	0.533844	(0.407727, 0.659960)
γ_{13}	1.344948	(1.103840, 1.586056)
γ_{23}	0.861096	(0.716370, 1.005822)
γ_{14}	1.165267	(0.925958, 1.404576)
γ_{24}	1.180716	(1.016211, 1.345222)

Table 3. Mean coverage of 95% confidence interval and MSE

	Mean Coverage		
n = 100	$f_1(z_1)$	$f_2(z_2)$	MSE
σ = 0.6	1.00	1.00	0,21858
σ = 0.9	1.00	1.00	0,48275
σ = 1.0	0,98	1.00	0,59420

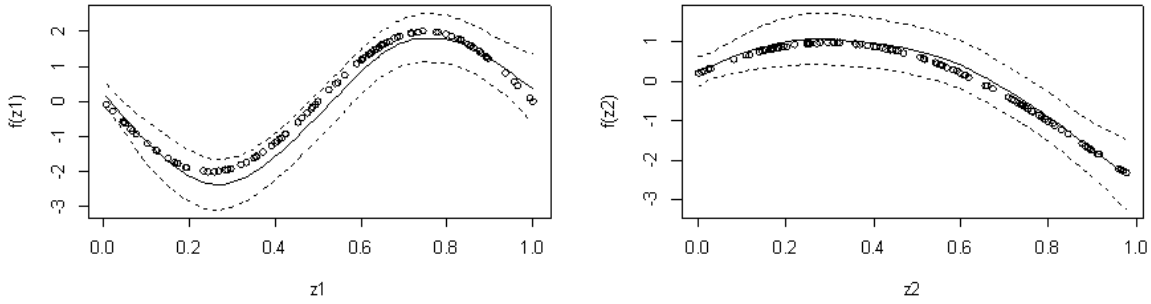

Fig. 1. Estimation for nonparametric components and its 95% confidence interval for $\sigma = 0.9$

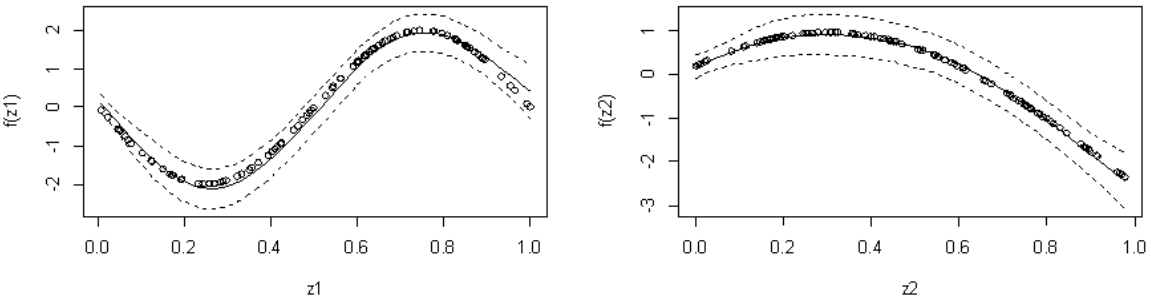

Fig. 2. Estimation for nonparametric components and its 95% confidence interval for $\sigma = 0.6$

4. DISCUSSION

The prior distributions used in the model in this study are improper Gaussian distribution for nonparametric components and multivariate normal distribution for parametric components. The smoothing spline estimators were obtained by using Bayesian approach. Our result is identical to smoothing spline estimator obtained by the PLS approach. The confidence intervals for the parameters of the parametric component and confidence intervals for smoothing spline functions of nonparametric component in semiparametric additive regression model can be constructed through Bayesian approach. By using the Bayesian approach, we obtain the GML method for selecting optimal smoothing parameters simultaneously where the shape of the smoothing parameters is fixed.

5. CONCLUSION

The smoothing spline in semiparametric additive regression model with Bayesian approach is a development of Bayesian smoothing spline for nonparametric component. Using Bayesian approach we obtain parametric component parameter estimators and smoothing spline estimators for nonparametric components and smoothing parameters selection method simultaneously with GML in semiparametric additive regression model. In addition, by computing each posterior mean and posterior variance of parameters $\hat{\gamma}$ and $\hat{f}_k$ the confidence intervals can be constructed for the parametric component parameters and confidence interval smoothing spline functions for nonparametric components in semiparametric additive regression models. Numerical example shows that estimation and inference method can be applied well using simulation data. The problem remaining is to apply this model in real life problem. The further study may consider to estimate the smoothing parameters simultaneously through Markov Chain Monte Carlo (MCMC) and to use other prior distribution in semiparametric additive regression models.

6. ACKNOWLEDGEMENT

The first research would like to thanks The Indonesian Central Bureau of Statistics (BPS) Indonesia

for the Ph.D. grant through Human Resources Development Project in cooperation with Institut Teknologi Sepuluh Nopember (ITS), Indonesia with contract number: 19/KS/01-VII/2010.

7. REFERENCES

Aydin, D. and M.S. Tuzemen, 2010. Estimation in semi-parametric and additive regression using smoothing and regression spline. Proceedings of the 2nd International Conference on Computer Research and Development, May 7-10, IEEE Xplore Press, Kuala Lumpur, pp: 465-469. DOI: 10.1109/ICCRD.2010.101

Aydin, D., 2008. A comparison of the nonparametric regression models using smoothing spline and kernel regression. Int. J. Math., Phys. Eng. Sci., 2: 75-79.

Costa, M.J., 2008. Penalized spline models and applications. Ph.D. Thesis, University of Warwick.

Eubank, R.L., 1999. Nonparametric regression and spline smoothing.

Jerak, A. and S. Wagner, 2006. Modeling probabilities of patent oppositions in a Bayesian semiparametric regression framework. Empirical Econ., 31: 513-533. DOI: 10.1007/s00181 -005-0047-0

Johnson, R.A. and D.W. Wichern, 2001. Applied Multivariate Statistical Analysis. 5th Edn., Prentice Hall, Upper Saddle, New Jersey, ISBN-10: 0130925535, pp: 767.

Kandala, N.B., S. Lang and S. Klasen, 2001. Semiparametric Analysis of Childhood Undernutrition in Developing Countries. University of Munich.

Lang, S. and A. Brezger, 2004. Bayesian P-splines. J. Comput. Grap. Statist., 13: 183-212. DOI: 10.1198/1061860043010

Li, Q., 2000. Efficient estimation of additive partially linear models. Int. Econ. Rev., 41: 1073-1092.

Liang, H., 2006. Estimation in partially linear models and numerical comparisons. Comput. Stat. Data Anal., 50: 675-687. DOI: 10.1016/J.CSDA.2004.10.007.

Marley, J.K. and M.P. Wand, 2010. Non-standard semiparametric regression via brugs. J. Statist. Software, 37: 1-28.

Nott, D., 2006. Semiparametric estimation of mean and variance functions for non-Gaussian data. Comput. Statist., 21: 603-620. DOI: 10.1007/s00180-006-0017-9

Nychka, D., 1988. Bayesian confidence intervals for smoothing splines. J. Am. Statist. Assoc., 83: 1134-1143.

Panagiotelis, A. and M. Smith, 2008. Bayesian identification, selection and estimation of semiparametric functions in high-dimensional additive models. J. Econ., 143: 291-316. DOI: 10.1016/J.JECONOM.2007.10.003

Ryu, D., B.K. Mallick and E. Li, 2011. Bayesian nonparametric regression analysis of data with random effects covariates from longitudinal measurements. Biometrics, 67: 454-466. DOI: 10.1111/j.1541-0420.2010.01489.x

Shen, J., 2011. Additive Mixed Modeling of HIV Patient Outcomes Across Multiple Studies. University of California.

Smith, M., R. Kohn and S.K. Mathur, 2000. Bayesian semiparametric regression: An exposition and application to print advertising data. J. Bus. Res., 49: 229-244. DOI: 10.1016/S0148-2963(99)00055-7

Wahba, G., 1983. Bayesian "confidence intervals" for the cross-validated smoothing spline. J. Royal Statist. Soc., Seri. B, 45: 133-150.

Wahba, G., 1985. A comparison of GCV and GML for choosing the smoothing parameter in generalized spline smoothing problem. Ann. Statist., 13: 1378-1402. DOI: 10.1214/aos/117649743

Wahba, G., 1990. Spline Models for Observational Data. 1st Edn., SIAM, Philadelphia, ISBN-10: 0898712440, pp: 180.

Wang, Y., 2011. Smoothing Splines: Methods and Applications. 1st Edn., CRC Press Inc, ISBN-10: 1420077554, pp: 384.

Wong, C. and R. Kohn, 1996. A Bayesian approach to additive semiparametric regression. J. Econ., 74: 209-235. DOI: 10.1016/0304-4076(95)01743-7.