Communications for Statistical Applications and Methods
2021, Vol. 28, No. 5, 537–551

https://doi.org/10.29220/CSAM.2021.28.5.537
Print ISSN 2287-7843 / Online ISSN 2383-4757

Pliable regression spline estimator using auxiliary variables

Jae-Kwon Oh[a], Jae-Hwan Jhong[1,a]

[a]Department of Information Statistics, Chungbuk National University, Korea

Abstract

We conducted a study on a regression spline estimator with a few pre-specified auxiliary variables. For the implementation of the proposed estimators, we adapted a coordinate descent algorithm. This was implemented by considering a structure of the sum of the residuals squared objective function determined by the B-spline and the auxiliary coefficients. We also considered an efficient stepwise knot selection algorithm based on the Bayesian information criterion. This was to adaptively select smoothly functioning estimator data. Numerical studies using both simulated and real data sets were conducted to illustrate the proposed method's performance. An R software package `psav` is available.

Keywords: auxiliary variable, B-spline, coordinate descent algorithm, knot selection, nonparametric regression

1. Introduction

A nonparametric function estimation is a statistical method that aims to estimate a function based on observed data, assuming that the function belongs to an infinite dimensional parameter space. The function estimation methods that are suitable for various types of data are being studied. Renowned methods are the kernel density estimation, local polynomial and spline; see Fan and Gijbels (1996), Efromovich (2008), Green and Silverman (1993) and Tsybakov (2008).

A representative method used for the function estimation is the basis function method. Since it is impossible to estimate the infinite parameters using finite data in a function estimation, a function space and base functions are introduced. The estimated target function is represented by a linear combination of the basis functions that spans the appropriate function space. Once the function space and the basis functions are defined, the target function can be expressed by using its predictor variables. Thus, the basis function method allows us to consider the problem of estimating the target function as a result of estimating the regression coefficients. The basis function has the advantage of being able to apply statistical methodologies. These include the least squares, the absolute deviations, and the likelihood problems.

There is a spline basis out of many basis functions that are used for interpolating the data or fitting smooth curves. The spline basis function is defined as a differentiable piecewise polynomial for each given knot interval. The main basis techniques are the B-splines and the truncated power basis splines (De Boor, 1978). The truncated power basis spline has the advantages of possessing

This research was supported by Chungbuk National University Korea National University Development Project (2020).
[1] Corresponding author: Department of Information Statistics, Chungbuk National University, Chungdae-ro 1, Seowon-Gu, Cheongju, Chungbuk 28644, Korea. E-mail: jjh25@cbnu.ac.kr

Published 30 September 2021 / journal homepage: http://csam.or.kr

a simple construction and easily interpreting the parameters in the model. However, the coefficients are correlated in the model, where there are many overlapping intervals. When the predictor is large, the curves of that basis become almost vertical and parallel. Therefore, an incorrect fitting occurs. In contrast, the B-spline basis is more complex than the truncated power basis. The reason for using the B-spline is mainly for computational problems. The B-spline has minimal support (or compact support) that minimizes the amount of the overlap between the spline basis. Therefore, it enables stable calculations (Yee, 2015).

In the basis function methodology, the objective function to be optimized is primarily a convex or concave function. Its domain is a coefficient vector of the bases constituting the estimated function. The coordinate descent algorithm (Wright, 2015) is simple, efficient and useful for optimizing these objective functions. The concept of the algorithm is to optimize the solution of the convex (concave) function minimization (maximization) problem for multi-dimensional vectors. A coefficient update holds the remaining coefficients as constants. It considers the objective function as a one-dimensional function.

Knot selection in the regression spline (PSE) is a significant challenge. The model's performance is highly influential depending on the location and the number of the knots. The knot selection is the same as the variable selection in a multiple regression. Many research have been conducted on knot selection. Osborn *et al.* (1998) proposed an algorithm that allows for the efficient calculation of the lasso (Tibshirani, 1996) estimator for a knot selection. Leitenstorfer and Tuz (2007) considered the boosting techniques used to select variables in the knot selection. Garton *et al.* (2020) proposed a method for selecting the number and location of the knots when the data are Gaussian and non-Gaussian.

The qualitative or categorical factors, such as gender and race, sometimes perform as predictors that are very useful in explaining a regression's response variables. These qualitative factors are used as predictors in the form of indicators or dummy variables. The dummy variables have a variety of uses, and can always be used whenever the qualitative factors impact a regression relationship (Chatterjee and Hadi, 2015). Tibshirani and Friedman (2020) proposed an extended lasso by adding modifying variables. These include gender, age, and time to response variables and predictors. They allow for the possibility that some or all of the coefficients vary following each category.

In this paper, we present a new statistical learning theory for the modeling and analysis of data, This theory consists of the auxiliary variables besides the response and predictor variables. The proposed model was performed with only one fit, regardless of the number of categories. We allow the coefficients according to the categories of the auxiliary variables to have different values. We express the estimator with a linear combination of the main predictor and the auxiliary term of the B-spline. The coefficient is estimated by applying a coordinate descent algorithm to minimize the residual sum of squares. In regards of the knot selection to reduce the computational cost, an improved stepwise selection is introduced. We consider five types of simulation data to measure the performance of the proposed method and conduct a real-data analysis involving the auxiliary variables.

This paper is organized as follows. In Section 2, we define the B-spline regression estimators including its auxiliary variables. In Section 3, the process of updating the coefficients based on the coordinate descent algorithm and the specific implementation of the knot selection are explained. We validate the performance of our proposed model using the simulations and real data in Section 4. Section 5 summarizes the paper's conclusions. An `R` software package `psav` is available at `https://github.com/OJKda/psav-package`.

2. Model and estimator

Consider a nonparametric regression model

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

where $\{x_i\}_{i=1}^n$ belongs to a subinterval of $\mathbb{R}$, $\{y_i\}_{i=1}^n \in \mathbb{R}$ are the responses, and $\{\varepsilon_i\}_{i=1}^n$ are the random errors with a mean of zero and a variance of $\sigma^2 > 0$. For the notational convenience, we fix the subinterval as $[0, 1]$ in the remainder of this paper.

In this model, we have the measurements of one or more auxiliary binary variables $\{z_i\}_{i=1}^n$ for $z_i = (z_{i1}, \ldots, z_{iK}) \in \mathbb{R}^K$. Here, K is the auxiliary variable's number of levels. For example, it may be sex for $K = 2$, that are composed of male and female. Thereafter, we allow the possibility that the regression curves are different among males and females. We discuss the spinal bone's mineral's densities for the North American adolescents with their sex and ethnicity/race for the auxiliary variables in Section 4.2. The goal is to estimate f based on the given observations $(x_1, z_1, y_1), \ldots, (x_n, z_n, y_n)$.

Let $B_1, \ldots, B_J$ be the B-spline on the basis functions for a univariate variable of the order r with the interior knots $\{\xi_1, \ldots, \xi_L\}$ over $[0, 1]$. This is such that

$$t_1 \leq \cdots \leq t_r \leq 0 < \xi_1 < \cdots < \xi_L < 1 \leq t_{r+1} \leq \cdots \leq t_{2r},$$

where $\{t_i\}_{i=1}^{2r}$ are the boundary knots. A B-spline basis of the order r spans the linear space of the piecewise polynomials of the degree $r - 1$ with continuous derivatives in the order of $r - 2$. Since the B-spline basis functions have small supports, it can provide a numerically efficient to others such as the truncated power splines. B-splines are also known to be non-negative and have a sum of one (De Boor, 2001).

For $u \in [0, 1]$ and the binary vector $v = (v_1, \ldots, v_K)$ with length K, define

$$\mathsf{f}(u, v; \theta) = \sum_{j=1}^{J} \beta_j B_j(u) + \sum_{j=1}^{J} \sum_{k=1}^{K} \gamma_{jk} v_k B_j(u),$$

where $\theta = (\beta, \gamma_1, \ldots, \gamma_J)$ is a coefficient vector with $\beta = (\beta_1, \ldots, \beta_J) \in \mathbb{R}^J$ and $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jK}) \in \mathbb{R}^K$ for $j = 1, \ldots, J$. We note that β_j is a scalar, where as γ_j is a vector of the length of K. Here, β is a coefficient vector that controls the effect of the main predictor value u. $\{\gamma_{jk}\}$, $j = 1, \ldots, J$, and $k = 1, \ldots, K$ are coefficients that affect an interaction of u and the auxiliary information v.

We consider the following residual sum of squares objective function

$$R(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \{y_i - \mathsf{f}(x_i, z_i; \theta)\}^2, \tag{2.2}$$

and define

$$\hat{\theta} = \left(\hat{\beta}, \hat{\gamma}\right) = \operatorname*{argmin}_{\theta} R(\theta).$$

Thereafter, the proposed estimator, that we label *pliable spline estimator (PSE)* is given by

$$\hat{f} = \mathsf{f}\left(\cdot; \hat{\theta}\right).$$

3. Implementation

3.1. Coordinate wise update

Since the objective function (2.2) is convex regarding θ, one can adapt a coordinate descent algorithm to compute $\hat{\theta}$. For $j = 1, \ldots, J,\ k = 1, \ldots, K$, we denote a univariate objective function of β_j and γ_{jk}

$$r_j(\beta_j) = R\left(\tilde{\beta}^{(-j)}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_J\right) \quad \text{and} \quad r_{jk}(\gamma_{jk}) = R\left(\tilde{\beta}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_{j-1}, \tilde{\gamma}_j^{(-k)}, \tilde{\gamma}_{j+1}, \ldots, \tilde{\gamma}_J\right), \tag{3.1}$$

where $\tilde{\beta} = (\tilde{\beta}_1, \ldots, \tilde{\beta}_J) \in \mathbb{R}^J$ and $\tilde{\gamma}_j = (\tilde{\gamma}_{j1}, \ldots, \tilde{\gamma}_{jK}) \in \mathbb{R}^K$ are initial vectors for β and γ_j, respectively. Additionally, let

$$\tilde{\beta}^{(-j)} = \left(\tilde{\beta}_1, \ldots, \tilde{\beta}_{j-1}, \beta_j, \tilde{\beta}_{j+1}, \ldots, \tilde{\beta}_J\right) \quad \text{and} \quad \tilde{\gamma}_j^{(-k)} = \left(\tilde{\gamma}_{j1}, \ldots, \tilde{\gamma}_{j(k-1)}, \gamma_{jk}, \tilde{\gamma}_{j(k+1)}, \ldots, \tilde{\gamma}_{jK}\right),$$

be the vectors with their initial values substituted for the j^{th} and jk^{th} coefficients, respectively. The coordinate-wise update has the form

$$\tilde{\beta}_j \leftarrow \underset{\beta_j \in \mathbb{R}}{\operatorname{argmin}}\, r_j(\beta_j) \quad \text{and} \quad \tilde{\gamma}_{jk} \leftarrow \underset{\gamma_{jk} \in \mathbb{R}}{\operatorname{argmin}}\, r_{jk}(\gamma_{jk}). \tag{3.2}$$

Thus, the algorithm iteratively and coordinately updates the coefficients by the minimum of a univariate objective function (3.1) as in (3.2) until its convergence. The iteration halts when the difference in the current and the updated value of the objective function is less than $\epsilon = 10^{-5}$.

3.2. Minimizing of the univariate objective functions

The solution to the optimization problem is to obtain a quadratic solution to the β_j. We select a coordinate index $j \in \{1, \ldots, J\}$ and observe

$$\begin{aligned} r_j(\beta_j) =& \frac{1}{2n} \sum_{i=1}^{n} \left\{ y_{ij} - \beta_j B_j(x_i) \right\}^2 \\ =& \frac{\sum_{i=1}^n B_j^2(x_i)}{2n} \left(\beta_j - \frac{\sum_{i=1}^n y_{ij} B_j(x_i)}{\sum_{i=1}^n B_j^2(x_i)} \right)^2 + (\text{terms independent for } \beta_j), \end{aligned}$$

where

$$y_{ij} = y_i - \sum_{l \neq j} \tilde{\beta}_l B_l(x_i) - \sum_{j=1}^{J} \sum_{k=1}^{K} \tilde{\gamma}_{jk} z_{ik} B_j(x_i),$$

is a partial residual. Thus, we update

$$\tilde{\beta}_j \leftarrow \frac{\sum_{i=1}^n y_{ij} B_j(x_i)}{\sum_{i=1}^n B_j^2(x_i)} \quad \text{for} \quad j = 1, \ldots, J. \tag{3.3}$$

Similarly, r_{jk} can be expressed as

$$r_{jk}(\gamma_{jk}) = \frac{\sum_{i=1}^n z_{ik}^2 B_j^2(x_i)}{2n} \left\{ \gamma_{jk} - \frac{\sum_{i=1}^n y_{ijk} z_{ik} B_j(x_i)}{\sum_{i=1}^n z_{ik}^2 B_j^2(x_i)} \right\}^2 + (\text{ terms independent for } \gamma_{jk}),$$

where

$$y_{ijk} = y_i - \sum_{j=1}^{J} \tilde{\beta}_j B_j(x_i) - \sum_{l \neq j} \sum_{k=1}^{K} \tilde{\gamma}_{lk} z_{ik} B_l(x_i) - \sum_{m \neq k} \tilde{\gamma}_{jm} z_{im} B_j(x_i).$$

Ignoring the terms that are independent of γ_{jk}, one can update $\tilde{\gamma}_{jk}$ as

$$\tilde{\gamma}_{jk} \leftarrow \frac{\sum_{i=1}^{n} y_{ijk} z_{ik} B_j(x_i)}{\sum_{i=1}^{n} z_{ik}^2 B_j^2(x_i)} \quad \text{for} \quad j = 1, \ldots, J, \ k = 1, \ldots, K. \tag{3.4}$$

In (3.4), We note that the updated formula cannot be defined when the denominator terms become zero. This case may occur when there are no observations with a specific k^{th} auxiliary variable index in support of the j^{th} B-spline basis function. However, in this case, since r_{jk} is mathematically expressed as a constant function regarding γ_{jk}, $\tilde{\gamma}_{jk}$ does not perform a role in the model. Consequently, there is no problem in updating $\tilde{\gamma}_{jk}$ to any real value. In general, as the number of the knots and the order of the r of the spline increases, the support decreases. Thus, this phenomenon may occur when processing observations with a high complexity.

3.3. Stepwise knot selection

We apply the stepwise selection (Efroymson, 1960) that is a variable selection method used in a multiple regression analysis, for the knot selection. In the multiple regression analyses, when the number of variables is large, a fitting regression model can lead to overfitting. This results in a problem where the variance of the coefficient estimates increase. To solve this problem, a stepwise selection is applied to select an optimal subset from a set of predictors. In an RS, the knot selection results in the same challenge. If the number of the knots is insufficient, the regression model is underfitting, and if there are too many knots, overfitting occurs. Therefore, the knot selection is important as the number of the knots significantly affects the fitted model.

A stepwise selection begins with a model without knots, and then iteratively adds and deletes the knots to the model individually until the best model is found. The Bayesian information criterion (BIC) (Schwarz *et al.*, 1978) is considered as an evaluation criterion when comparing models. The equation is as follows,

$$BIC = n \log\left(\frac{1}{n} \sum_{i=1}^{n} (y_i - \mathrm{f}(x_i, z_i))^2\right) + p \log(n),$$

where p is the number of nonzero $\hat{\beta}_j$'s.

The stepwise selection procedure is provided in Algorithm 1. First, we begin a null model $\mathcal{M}_0$ without the knots. Thereafter, the algorithm fits all possible models with a single knot based on the K candidates of the knots sequence and we select $\mathcal{M}_1$ with the smallest BIC. If $BIC(\mathcal{M}_0) < BIC(\mathcal{M}_1)$, the algorithm selects $\mathcal{M}_b = \mathcal{M}_0$ as the optimal result. Alternatively, we can determine $\mathcal{A}_2$ that is the optimal fit over all the possible models with two knots, including the knot we selected in previous step. For the deletion process, we fit all possible models by replacing the oldest knot in $\mathcal{A}_2$ with each remaining candidate knot. The reason for excluding the oldest knots in the deletion step is to reduce the computational cost of the algorithm. Thereafter, denote $\mathcal{D}_2$ as the smaller BIC among these models. Thus, the proposed stepwise algorithm works iteratively to increase $k = 2, \ldots, K$ until $\mathcal{M}_{k-1}$ is better than $\mathcal{A}_k$ or $\mathcal{D}_k$ in the BIC sense.

Algorithm 1: Stepwise knot selection Algorithm

$\mathcal{M}_k$ is a model with k knots for $k = 0, \ldots, K$.
$\mathcal{A}_k$ is a model with one knot added to the $\mathcal{M}_{k-1}$.
$\mathcal{D}_k$ is a model with the oldest knot removed and the new knot adds to $\mathcal{A}_k$.
$\mathcal{M}_b$ is the best model.

1. Let $\mathcal{M}_0$ denote the model without the knots.
2. Compute $\mathcal{M}_1$.
3. If $BIC(\mathcal{M}_0) > BIC(\mathcal{M}_1)$ then we go to step 4.
4. For $k = 2, \ldots, K$:
   (a) Construct $\mathcal{A}_k$, $\mathcal{D}_k$.
   (b) $\mathcal{M}_k = \underset{\mathcal{M}=(\mathcal{A}_k, \mathcal{D}_k)}{\operatorname{argmin}} BIC(\mathcal{M})$
   (c) $BIC(\mathcal{M}_{k-1}) > BIC(\mathcal{M}_k)$ then go to next loop.
5. $\mathcal{M}_b = \mathcal{M}_{k-1}$

3.4. Algorithm details

Algorithm 2 represents the implementation of the proposed algorithm. Given the observed data and some input parameters, first we initialize the coefficient to zero. Thereafter, we compute the initial residual sum of squares, that is $(1/2n)\sum_{i=1}^{n} y_i^2$. In the implementation, we check the j, k index at which the denominator values become zero before their iterations. These values can be obtained from the observations without the coefficients. Thereafter, the algorithm only updates the active coefficients whose denominator is not zero. It keeps other coefficients at the initial value of zero. This algorithm is an efficient way to reduce the running time by reducing the number of the iterations during the update. For each iteration, we update a single coefficient sequentially in a coordinated manner. If the difference between the previous and updated value of the objective function is less than a small positive quantity, we state that $\epsilon = 10^{-5}$. We then complete the algorithm.

4. Numerical studies

4.1. Simulation

In this section, we illustrate the performance of the proposed method on the simulated examples. We generate predictors as sample size n sequences between $[0, 1]$. ϵ_i is generated from $N(0, \sigma^2)$ for $i = 1, \ldots, n$. σ is determined by signal-to-noise ratio (SNR) (Meier *et al.*, 2009), representing the variance's ratio of the true function over the σ^2. The auxiliary variable is implemented as a one-hot encoding. For example, the observed value of the auxiliary variable with two categories is a vector consisting of 1 and 0, such as $(1, 0)$ and $(0, 1)$. If there are four categories, they are expressed as $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, $(0, 0, 0, 1)$.

The four examples are as follows: Examples 1–2 consider a piecewise constant function and a piecewise linear function to set up the exact location of the true knots. The true knots are $0.4, 0.7$ in Example 1 and $0.2, 0.45, 0.75$ in Example 2. Example 3 is designed as a nonlinear function that

Algorithm 2: Coordinate descent Algorithm (CDA)

Input: X : B-spline basis matrix $\in \mathbb{R}^{n\times J}$,
Z : auxiliary variable matrix $\in \mathbb{R}^{n\times K}$,
y : response vector $\in \mathbb{R}^n$,
r : order of B-spline,
m : maximum number of iterations,
ϵ : stopping criterion
Initialization: $\tilde{\beta} \in \mathbb{R}^J, \tilde{\gamma}_j \in \mathbb{R}^K$ for $j = 1, \ldots, J$
$RSS_{\text{old}} = 1/2n \sum_{i=1}^{n} y_i^2$
for *iteration* = 1 *to* m **do**
  for $j = 1$ *to* J **do**
    Update $\tilde{\beta}_j$ by (3.3)
  end
  for $(j, k) = (1, 1)$ *to* (J, K) **do**
    Update $\tilde{\gamma}_{jk}$ by (3.4)
  end
  Compute $RSS_{\text{new}} = R(\tilde{\beta}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_J)$
  if $|RSS_{old} - RSS_{new}| < \epsilon$ **then**
    break
  else
    $RSS_{\text{old}} = RSS_{\text{new}}$
  end
end
Output: $\tilde{\beta}, \tilde{\gamma}_1, \ldots, \tilde{\gamma}_J$

may appear in the real data. We evaluated the performance of the proposed model when the number of the auxiliary variable categories increased to four in Example 4. Figure 1 displays the underlying functions of Examples 1-4.

Example 1.

$$f(x,z) = \begin{cases} 0.3, & \text{if} \quad 0 \le x < 0.4 \\ -0.5, & \text{if} \quad 0.4 \le x < 0.7 \\ 0.7, & \text{if} \quad 0.7 \le x \le 1, \end{cases} \quad \text{for} \quad z = (1,0)$$

$$f(x,z) = \begin{cases} 0.6, & \text{if} \quad 0 \le x < 0.4 \\ -1, & \text{if} \quad 0.4 \le x < 0.7 \\ 1.4, & \text{if} \quad 0.7 \le x \le 1 \end{cases} \quad \text{for} \quad z = (0,1)$$

Example 2.

$$f(x,z)=\begin{cases}-0.5x+0.2, & \text{if} \quad 0\le x<0.2\\ 1.6x-0.22, & \text{if} \quad 0.2\le x<0.45\\ 0.125x+0.44375, & \text{if} \quad 0.45\le x<0.65\\ -0.2x+0.655, & \text{if} \quad 0.65\le x\le 1\end{cases} \quad \text{for} \quad z=(1,0)$$

$$f(x,z)=\begin{cases}-0.5x+0.4, & \text{if} \quad 0\le x<0.2\\ 0.2x+0.24, & \text{if} \quad 0.2\le x<0.45\\ -0.75x+0.6875, & \text{if} \quad 0.45\le x<0.65\\ 2x-1.1, & \text{if} \quad 0.65\le x\le 1\end{cases} \quad \text{for} \quad z=(0,1)$$

Example 3.

$$f(x,z)=\begin{cases}\sqrt{x}+\cos(7(x-1))+\sin(3(x-2))+\cos(14x)+\sin(6(x-2)), & \text{for} \quad z=(1,0)\\ \sqrt{x}+2\cos(7(x-1))+\sin(3(x-2))+\cos(14x), & \text{for} \quad z=(0,1)\end{cases}$$

Example 4.

$$f(x,z)=\begin{cases}\sqrt{x}+\cos(7(x-1))+\sin(3(x-2))+\cos(14x)+\sin(6(x-2)), & \text{for} \quad z=(1,0,0,0)\\ \sqrt{x}+\sin(3(x-2))+\cos(14x), & \text{for} \quad z=(0,1,0,0)\\ \sqrt{x}+\cos(7(x-1))+2\sin(11(x-2)), & \text{for} \quad z=(0,0,1,0)\\ \sqrt{x}+\cos(7(x-1))+\sin(3(x-2))+\cos(14x), & \text{for} \quad z=(0,0,0,1)\end{cases}$$

We compare the proposed method with the regression spline (RS) of Marsh and Cormier (2001), smoothing spline (SS) of Wahba (1990), local regression (LR) of Loader (2006), local polynomial (LP) of Fan and Gijbels (1996), and categorical regression splines (CRS) of Nie and Racine (2012). The RS is obtained by the linear fitting with the B-spline basis function without an auxiliary variable. The SS is a method of applying the smoothing and shrinkage techniques. The LR and LP are generalizations of the moving average and polynomial regression. CRS is computing nonparametric regression splines in the presence of both continuous and categorical predictors. RS, SS and LR are provided in the `stats` package in `R`. LP is provided in the `KernSmooth` package in `R` (Wand, 2020). CRS is provided in the `crs` package in `R` (Racine and Nie, 2021). To reduce the computational burden, we computed all the estimators by using the default settings for each package.

We consider the mean squared error (MSE), mean absolute error (MAE) criterion for the discrepancy measure between the underlying function f, and each function estimator, which are given as follows,

$$\text{MSE}\left(\hat{f}\right)=\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(x_i)-f(x_i)\right)^2 \quad \text{and} \quad \text{MAE}\left(\hat{f}\right)=\frac{1}{n}\sum_{i=1}^{n}\left|\hat{f}(x_i)-f(x_i)\right|.$$

The PSE was simulated by fixing the number of the initial knots at 50. In addition, the suitable order r of B-spline basis for each example was assigned. Examples 1–2 sets the order $r=1,2$, respectively, and Examples 3–4 set the order $r=4$. We assign each sigma corresponding to our fixed SNR and fit the model with an algorithm designed after converting the auxiliary variable into a binary variable. The data were divided according to the number of the categories of the auxiliary variables

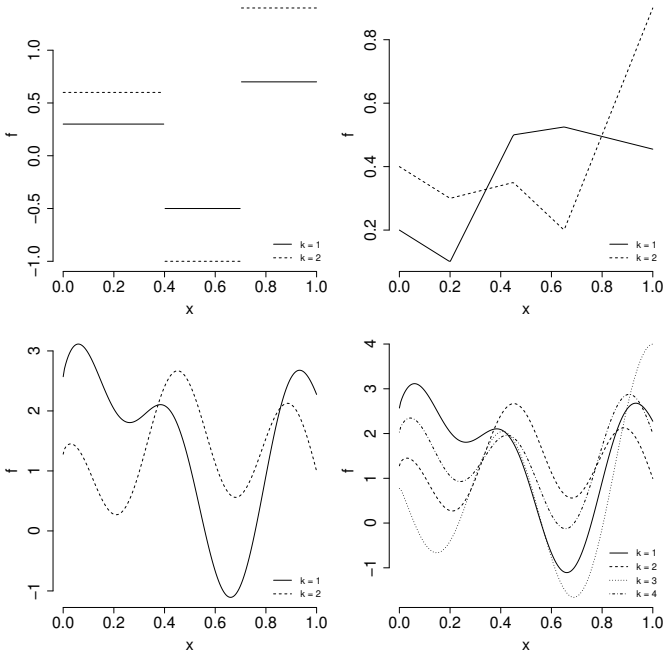

Figure 1: *The top left panel indicates the underlying function for Example 1. The true location of the knot was set to* $0.4, 0.7$. *The solid line is a function when* $k = 1$ ($z = (1, 0)$) *and the dashed line is a function when* $k = 2$ ($z = (0, 1)$). *The top right panel demonstrates the true underlying function for Example 2. The true locations of the knots are* $0.2, 0.45$, *and* 0.75. *The bottom left panel indicates the underlying function for Example 3. In the top right and bottom left panels, the lines represent the same category as the top left panel. The bottom right panel presents the underlying function for Example 4. The solid line is a function of* $k = 1$ ($z = (1, 0, 0, 0)$), *the dashed line is* $k = 2$ ($z = (0, 1, 0, 0)$), *the dotted line is* $k = 3$ ($z = (0, 0, 1, 0)$), *and the dotdash line is* $k = 4$ ($z = (0, 0, 0, 1)$).

to apply the different methods. In the case of the RS, we consider a model that is a linear fit of the B-spline basis function obtained by the selected knot in the PSE. It must not contain any auxiliary variables.

Tables 1–4 are experimental results of the various scenarios in each of the examples. Sample size n and SNR were fixed to apply four scenarios per table. Considering the four categories, Example 4 is tested by increasing the size of the sample to 500 and 1,000 to ensure that there are sufficient samples for each category.

As described in Tables 1–2, the PSE indicates a better performance than the other methods. Since the PSE can be fitted with the degree 0 or 1, estimators similar to the underlying function are obtained. The RS, that uses the same basis function as the PSE, can be fitted with the degrees 0 and 1. However, the performance is poor because the auxiliary variables are not included in the model. The SS, LR, LP and CRS are fitted with curves, therefore, it is not a good model in Examples 1–2. Since the PSE provided the initial knots as the quantile of the predictor variable, it chooses knots that is adjacent to the true knots. Accordingly, the PSE demonstrates a good performance in Examples 1–2.

In Tables 3–4, the PSE dose not delay the other nonlinear function estimation methods. The PSE has a much better performance than the LR and LP and a similar performance to the SS and CRS. Unlike the SS, that forms knots in all inputs and uses the shrinkage method, the PSE can estimate the nonlinear functions with fewer knots through a knot selection. Furthermore, the SS, LR, and LP

Table 1: The average of each criterion (×100) over 100 runs of OUR, RS, SS, LR, LP and CRS for Example 1 with a sample size $n = 200, 500$ and a signal-to-noise ratio $SNR = 5, 15$, with the standard error in parentheses. The bold text is the smallest criterion in each scenario

n	SNR	Method	MSE	MAE	n	SNR	Method	MSE	MAE
200	5	PSE	**2.76(1e-03)**	**7.30(2e-03)**	500	5	PSE	**1.98(5e-04)**	**4.82(1e-03)**
		RS	9.14(6e-04)	25.48(6e-04)			RS	8.29(4e-04)	25.19(3e-04)
		SS	4.09(8e-04)	14.10(2e-03)			SS	2.44(3e-04)	10.60(9e-04)
		LR	14.16(1e-03)	26.32(2e-03)			LR	13.91(8e-04)	25.77(9e-04)
		LP	10.78(5e-03)	17.61(3e-03)			LP	5.99(3e-03)	12.17(2e-03)
		CRS	7.11(2e-03)	18.47(3e-03)			CRS	4.55(7e-04)	13.66(1e-03)
	15	PSE	**1.84(8e-04)**	**4.77(1e-03)**		15	PSE	1.73(5e-04)	**3.73(7e-04)**
		RS	8.46(5e-04)	25.18(6e-04)			RS	8.10(3e-04)	25.20(3e-04)
		SS	2.12(5e-04)	9.63(1e-03)			SS	**1.40(2e-04)**	7.57(5e-04)
		LR	13.63(1e-03)	25.45(1e-03)			LR	13.84(8e-04)	25.59(7e-04)
		LP	11.13(6e-03)	14.77(3e-03)			LP	7.37(5e-03)	10.56(3e-03)
		CRS	5.24(2e-03)	14.59(3e-03)			CRS	4.10(5e-04)	12.00(9e-04)

Table 2: The average of each criterion (×100) over 100 runs of OUR, RS, SS, LR, LP and CRS for Example 2 with a sample size $n = 200, 500$ and a signal-to-noise ratio $SNR = 5, 15$, with the standard error in parentheses. The bold text is the smallest criterion in each scenario

n	SNR	Method	MSE	MAE	n	SNR	Method	MSE	MAE
200	5	PSE	**0.05(2e-05)**	**1.66(4e-04)**	500	5	PSE	**0.02(8e-06)**	**1.01(2e-04)**
		RS	1.03(3e-05)	8.77(2e-04)			RS	1.03(1e-05)	8.87(8e-05)
		SS	**0.05(2e-05)**	1.72(3e-04)			SS	**0.02(7e-06)**	1.14(2e-04)
		LR	0.10(2e-05)	2.38(2e-04)			LR	0.09(9e-06)	2.22(1e-04)
		LP	0.13(8e-05)	2.58(6e-04)			LP	0.05(3e-05)	1.67(3e-04)
		CRS	0.06(2e-05)	1.90(3e-04)			CRS	0.03(7e-06)	1.28(2e-04)
	15	PSE	**0.02(7e-06)**	**0.97(2e-04)**		15	PSE	**0.01(3e-06)**	**0.62(1e-04)**
		RS	1.00(3e-05)	8.70(2e-04)			RS	1.02(1e-05)	8.86(7e-05)
		SS	**0.02(5e-06)**	1.07(2e-04)			SS	**0.01(3e-06)**	0.75(1e-04)
		LR	0.08(1e-05)	2.15(2e-04)			LR	0.08(6e-06)	2.13(1e-04)
		LP	0.09(6e-05)	2.09(6e-04)			LP	0.04(2e-05)	1.39(3e-04)
		CRS	**0.02(7e-06)**	1.23(2e-04)			CRS	**0.01(3e-06)**	0.87(1e-04)

Table 3: The average of each criterion (×100) over 100 runs of OUR, RS, SS, LR, LP and CRS for Example 3 with a sample size $n = 200, 500$ and a signal-to-noise ratio $SNR = 5, 15$, with the standard error in parentheses. The bold text is the smallest criterion in each scenario

n	SNR	Method	MSE	MAE	n	SNR	Method	MSE	MAE
200	5	PSE	**2.28(7e-04)**	**11.67(2e-03)**	500	5	PSE	1.13(3e-04)	8.23(1e-03)
		RS	42.96(1e-03)	57.18(1e-03)			RS	43.38(7e-04)	58.38(7e-04)
		SS	2.38(8e-04)	12.08(2e-03)			SS	1.07(3e-04)	8.06(1e-03)
		LR	21.74(1e-03)	39.07(1e-03)			LR	21.70(8e-04)	39.36(8e-04)
		LP	12.26(9e-03)	25.41(8e-03)			LP	4.86(4e-03)	16.32(5e-03)
		CRS	2.42(8e-04)	12.00(2e-03)			CRS	**1.05(3e-04)**	**7.88(1e-03)**
	15	PSE	**0.85(3e-04)**	**7.10(1e-03)**		15	PSE	0.41(1e-04)	4.89(8e-04)
		RS	41.88(1e-03)	56.89(1e-03)			RS	42.73(5e-04)	58.12(6e-04)
		SS	0.89(3e-04)	7.30(1e-03)			SS	0.42(1e-04)	5.01(8e-04)
		LR	21.21(1e-03)	38.66(1e-03)			LR	21.41(7e-04)	39.19(7e-04)
		LP	9.66(7e-03)	21.49(7e-03)			LP	4.05(3e-03)	14.39(5e-03)
		CRS	0.87(3e-04)	7.20(1e-03)			CRS	**0.38(1e-04)**	**4.67(8e-04)**

require as many models as the number of levels of the auxiliary variables. However, the PSE is an efficient method since it can fit as one model regardless of the number of categories.

Table 4: The average of each criterion (×100) over 100 runs of OUR, RS, SS, LR, LP and CRS for Example 4 with a sample size $n = 500, 1{,}000$ and a signal-to-noise ratio $SNR = 5, 15$, with the standard error in parentheses. The bold text is the smallest criterion in each scenario

n	SNR	Method	MSE	MAE	n	SNR	Method	MSE	MAE
500	5	PSE	**2.40(5e-04)**	**11.87(1e-03)**	1000	5	PSE	**1.24(3e-04)**	8.54(9e-04)
		RS	55.76(3e-03)	58.79(2e-03)			RS	56.32(2e-03)	59.17(1e-03)
		SS	2.42(5e-04)	12.06(1e-03)			SS	1.26(3e-04)	8.74(1e-03)
		LR	23.51(1e-03)	40.43(9e-04)			LR	23.71(8e-04)	40.90(7e-04)
		LP	14.54(7e-03)	27.81(6e-03)			LP	7.91(3e-03)	20.54(4e-03)
		CRS	2.45(6e-04)	11.92(1e-03)			CRS	**1.24(3e-04)**	**8.52(1e-03)**
	15	PSE	0.86(2e-04)	7.09(9e-04)		15	PSE	**0.43(1e-04)**	**5.04(7e-04)**
		RS	55.66(3e-03)	58.74(2e-03)			RS	55.67(2e-03)	58.80(1e-03)
		SS	0.93(2e-04)	7.49(9e-04)			SS	0.50(1e-04)	5.50(6e-04)
		LR	22.98(9e-04)	40.20(9e-04)			LR	23.35(7e-04)	40.68(7e-04)
		LP	13.80(7e-03)	25.92(6e-03)			LP	7.40(4e-03)	18.94(5e-03)
		CRS	**0.85(2e-04)**	**7.05(9e-04)**			CRS	0.44(1e-04)	5.13(7e-04)

Table 5: The average of each criterion (×100) over 100 runs of OUR, RS, SS, LR, LP and CRS for Example 5 with a sample size $n = 500, 1{,}000$ and a signal-to-noise ratio $SNR = 5, 15$, with the standard error in parentheses. The bold text is the smallest criterion in each scenario

n	SNR	Method	MSE	MAE	n	SNR	Method	MSE	MAE
500	5	PSE	3.49(8e-04)	14.17(2e-03)	1000	5	PSE	1.77(4e-04)	10.14(1e-03)
		RS	144.56(8e-03)	91.51(3e-03)			RS	145.23(5e-03)	91.76(2e-03)
		SS	**3.12(8e-04)**	**13.70(2e-03)**			SS	**1.66(3e-04)**	**9.99(9e-04)**
		LR	12.57(1e-03)	26.85(1e-03)			LR	12.12(7e-04)	26.32(9e-04)
		LP	24.62(1e-02)	33.32(6e-03)			LP	12.37(5e-03)	23.91(4e-03)
		CRS	3.36(8e-04)	14.08(2e-03)			CRS	1.75(4e-04)	10.18(1e-03)
	15	PSE	1.19(3e-04)	8.33(1e-03)		15	PSE	**0.63(1e-04)**	**6.12(6e-04)**
		RS	142.86(8e-03)	90.88(2e-03)			RS	145.64(5e-03)	92.02(2e-03)
		SS	1.18(3e-04)	8.41(1e-03)			SS	0.66(1e-04)	6.33(6e-04)
		LR	11.60(1e-03)	25.39(1e-03)			LR	11.75(6e-04)	25.82(7e-04)
		LP	21.71(1e-02)	30.19(6e-03)			LP	11.80(5e-03)	22.16(4e-03)
		CRS	**1.13(2e-04)**	**8.20(1e-03)**			CRS	**0.63(1e-04)**	6.16(6e-04)

We consider Example 5 with two auxiliary variables, unlike the examples above. These are set as variables that determine shape and amplitude of the underlying function. The shape variable has two levels, cosine and x-axis symmetric cosine. The other one determines the amplitude and consists of three levels. The underlying function is demonstrated in Figure 2 in the form of six curves as a combination of the levels of the two variables.

Table 5 is the experimental results of Example 5 in the same scenario as Table 4. When the sample size $n = 1{,}000$ and $SNR = 15$, the PSE performs well. In other scenarios, PSE, SS and CRS show similar performance. PSE has advantages that SS does not have, as described above. In CRS, it is cumbersome to set the appropriate option to find the optimal fit. Judging from the results, it can be said that PSE is comparable to other competitive methods for function estimation models.

4.2. Bone mineral density (BMD) data analysis

The BMD's data was obtained in Bachrach *et al.* (1999). The data were obtained from `loon.data` library of R, where they are saved under the name "bone". The data is relative to the spinal bone mineral density measurements on 261 North American adolescents. The data consist of 485 observations. Five variables identify the subject (idnum), age, sex, relative spinal bone mineral density (rspnbmd),

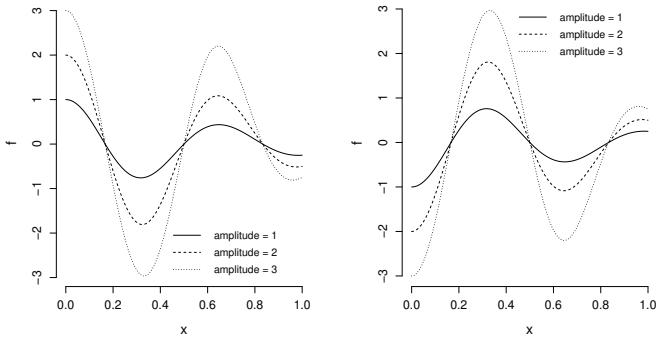

Figure 2: *The underlying function of Example 5. The left panel has a cosine shape, and the right panel has an x-axis symmetric cosine shape. Identical lines on both panels have the same amplitude.*

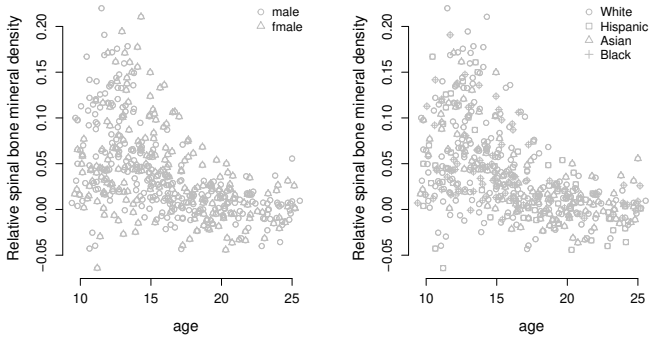

Figure 3: *Plot of the scatterplot by the auxiliary variable in the bone mineral data. The left panel is scatterplot when auxiliary variable is sex. The circles and triangles represent males and females, respectively. The right panel is a scatterplot when the auxiliary variable is ethnicity. Circles, quadrangles, triangles and crosses mean White, Hispanic, Asian and Black, respectively.*

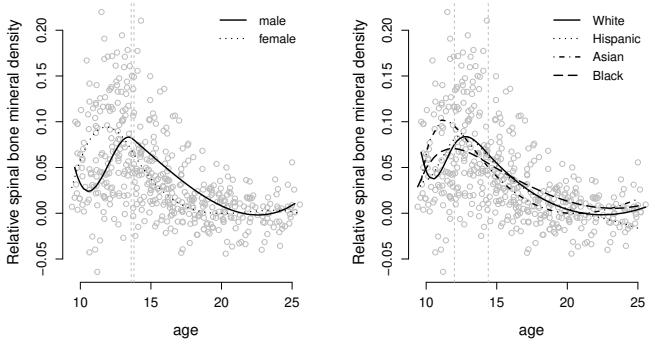

Figure 4: *The plot represents the estimators using the auxiliary variables applied by the PSE. The gray points are the scatterplots of BMD data. The left panel is PSE when sex is used as an auxiliary variables, and the right panel is used as ethnicity. The gray vertical dashed line represents the position of the selected knots.*

ethnicity/race (ethnic).

In the proposed method, we apply the predictors as age, and the responses as rspnbmd. The

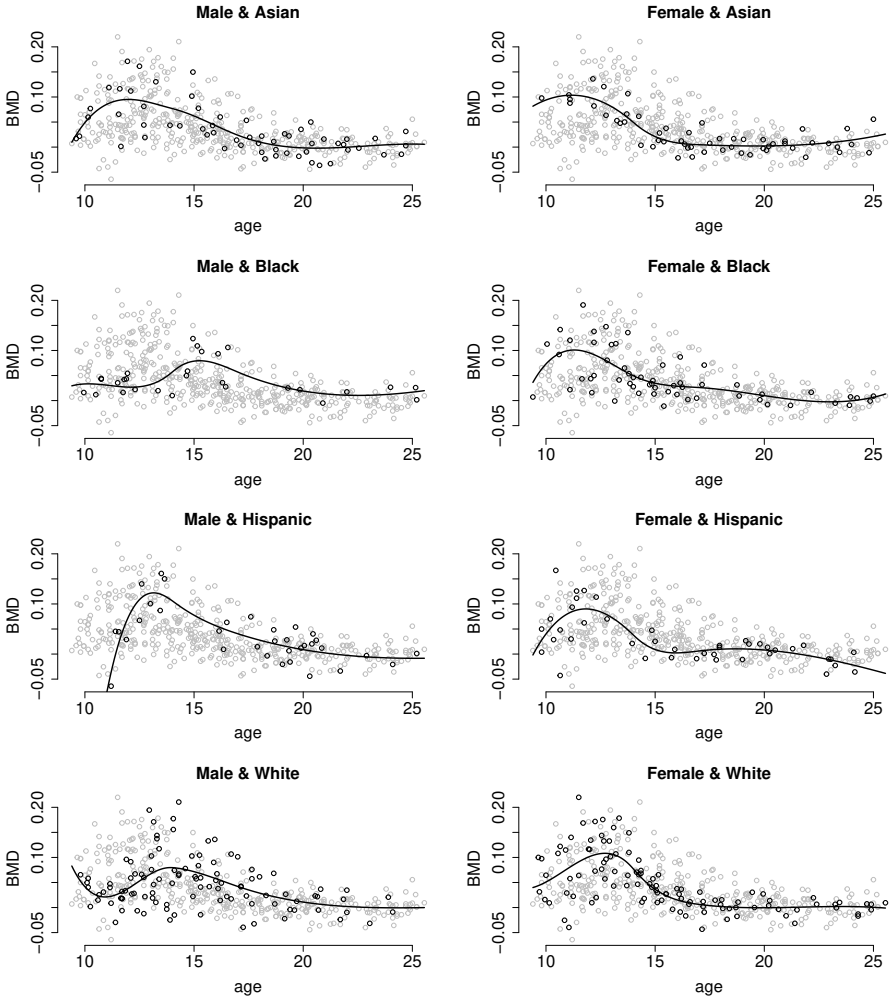

Figure 5: *Plot of the PSE with both sex and ethnic applied as auxiliary variables. In each plot, the black points are the data corresponding to that category. The solid line represents the estimator for each category.*

two variables are used as the auxiliary variables. In one case, a sex variable with two categories (male and female) and in the second case an ethnicity variable with four categories (Asian, Black, Hispanic, White) are applied. Finally, we contemplate a model that contains both variables being eight categories. We consider cubic splines in all cases with order $r = 4$.

Figure 3 displays the scatterplots when the auxiliary variable is "sex" and "ethnic", respectively. On the left panel, the data are divided into males and females, it is indicated that females aged 10 to 15 years are growing at an age when bone density levels are relatively faster than that of males. The right panel is divided by race, which is difficult to distinguish.

Figure 4 indicates the shape of the estimator when the auxiliary variables are used. When the auxiliary variable is "sex" (left), it can be stated that in the adolescent phase (between the ages of 10 and 13 years), females tend to increase their BMD relatively earlier than that of their male counterparts. When the auxiliary variable is "race" (right), it displays that Asians have a relatively higher BMD in

their adolescent phase (between the ages of 10 and 13 years) as compared to that of other races. White and Hispanic adolescents demonstrate similar trends. Figure 5 is an estimator by the combined sex and ethnicity. The Male & Black curve is flattened, with no soaring sections as compared to the other categories. Additionally, the Female & Asian curve demonstrates that Asian women have a higher BMD at a young age than other races.

5. Conclusion

In this article, we have developed a nonparametric regression function estimation method using the B-splines if there are auxiliary variables, in addition to the predictor variable. We devised a coordinate-wise update scheme to efficiently optimize the objective function. It was confirmed that the optimal knots of the spline function were adaptively selected by the proposed stepwise algorithm. The performance of the proposed estimator has been depicted with the simulated and real data analysis.

The results of this paper are expected to provide a foundation for further studies. They can be generalized and extended in several mannerisms.

First, one may consider the nonparametric quantile regression function estimator. Switching from the sum of the residuals squared objective function to the absolute deviation loss function allows the coordinate-descent-based algorithms to obtain a specified τ's quantile. This will occur instead of the mean for $\tau \in [0, 1]$ (Jhong and Koo, 2019).

Second, it is expected that the regularization method can be applied with the addition of an appropriate penalty term for the knot selection. As a suitable penalty term, there is the total variation of the $(r-1)$-th derivative of the spline function with the order r. Furthermore, there are the penalization methodologies, using the ℓ_1 or ℓ_2 norm of this total variation (Jhong *et al.*, 2017; Meyer, 2012; Mammen *et al.*, 1997). It appears to be an interesting topic that adds a penalty term for the selection of the auxiliary variables, and the knots.

Acknowledgments

This research was supported by Chungbuk National University Korea National University Development Project (2020).

References

Bachrach LK, Hastie T, Wang MC, Narasimhan B, and Marcus R (1999). Bone mineral acquisition in healthy Asian, Hispanic, black, and Caucasian youth: a longitudinal study, *The Journal of Clinical Endocrinology & Metabolism*, **84**, Oxford University Press, 4702–4712.

Chatterjee S and Hadi AS (2015). *Regression Analysis By Example*, John Wiley & Sons.

De Boor CR (1978). *A Practical Guide to Splines*, **27**, Springer-Verlag, New York.

Efroymson MA (1960). Multiple regression analysis, *Mathematical Methods for Digital Computers*, John Wiley & Sons, 191–203.

Efromovich S (2008). *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer Science & Business Media.

Fan J, Gijbels I (1996). *Local Polynomial Modelling and its Applications: Monographs on Statistics and Applied Probability 66*, **66**, CRC Press.

Garton N, Niemi J, and Carriquiry A (2020). *Knot Selection in Sparse Gaussian Processes*, arXiv preprint arXiv:2002.09538.

Green PJ and Silverman W (1993). *Nonparametric Regression and Generalized Linear Models: A*

Roughness Penalty Approach(1st ed), CRC Press.

Jhong JH and Koo JY (2019). Simultaneous estimation of quantile regression functions using B-splines and total variation penalty, *Computational Statistics & Data Analysis*, **133**, Elsevier, 228–244.

Jhong JH, Koo JY, and Lee SW (2017). Penalized B-spline estimator for regression functions using total variation penalty, *Journal of Statistical Planning and Inference*, **184**, Elsevier, 77–93.

Loader C (2006). *Local regression and likelihood*, Springer Science & Business Media.

Leitenstorfer F and Tutz G (2007). Knot selection by boosting techniques, *Computational Statistics & Data Analysis*, **51**, Elsevier, 4605–4621.

Marsh L and Cormier DR (2001). *Spline regression models*, **137**, Sage.

Wand M (2020). *KernSmooth: Functions for Kernel Smoothing Supporting Wand & Jones 1995*.

Meier L, Van de Geer S, and Bühlmann P (2009). High-dimensional additive modeling, *The Annals of Statistics*, **37**, Institute of Mathematical Statistics, 3779–3821.

Meyer MC (2012). Constrained penalized splines, *Canadian Journal of Statistics*, **40**, Wiley Online Library, 190–206.

Mammen E, and Van de Geer S (1997). Locally adaptive regression splines, *Annals of Statistics*, **25**, Institute of Mathematical Statistics, 387–413.

Nie Z and Racine JS (2012). The crs Package: Nonparametric Regression Splines for Continuous and Categorical Predictors, *R Journal*, **4**.

Osborne MR, Presnell B, and Turlach BA (1998). Knot selection for regression splines via the lasso, *Computing Science and Statistics*, 44–49.

Racine JS and Nie Z (2021). *CRS: Categorical Regression Splines, 2021*. R package version 0, 15–33.

Schwarz G (1978). Estimating the dimension of a model, *Annals of Statistics*, **6**, Institute of Mathematical Statistics, 461–464.

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

Tibshirani R and Friedman J (2020). A pliable lasso, *Journal of Computational and Graphical Statistics*, **29**, 215–225.

Tsybakov AB (2008). *Introduction to Nonparametric Estimation*, Springer, London.

Wahba G (1990). *Spline Models for Observational Data*, SIAM, Philadelphia.

Wright SJ (2015). Coordinate descent algorithms, *Mathematical Programming*, **151**, 3–34.

Yee TW (2015). *Vector Generalized Linear and Additive Models: With An Implementation in R*, Springer, New York.

Received April 15, 2021; Revised June 29, 2021; Accepted August 05, 2021