Hindawi Publishing Corporation
Journal of Probability and Statistics
Volume 2014, Article ID 203469, 11 pages
http://dx.doi.org/10.1155/2014/203469

Research Article

Direct Determination of Smoothing Parameter for Penalized Spline Regression

Takuma Yoshida

Graduate School of Science and Engineering, Kagoshima University, Kagoshima 890-8580, Japan

Correspondence should be addressed to Takuma Yoshida; yoshida@sci.kagoshima-u.ac.jp

Received 7 January 2014; Revised 31 March 2014; Accepted 31 March 2014; Published 22 April 2014

Academic Editor: Dejian Lai

Penalized spline estimator is one of the useful smoothing methods. To construct the estimator, having goodness of fit and smoothness, the smoothing parameter should be appropriately selected. The purpose of this paper is to select the smoothing parameter using the asymptotic property of the penalized splines. The new smoothing parameter selection method is established in the context of minimization asymptotic form of MISE of the penalized splines. The mathematical and the numerical properties of the proposed method are studied. First we organize the new method in univariate regression model. Next we extend to the additive models. A simulation study to confirm the efficiency of the proposed method is addressed.

1. Introduction

Penalized spline methods are a well-known efficient technique for nonparametric smoothing. Penalized splines were suggested by O'Sullivan [1] and Eilers and Marx [2]. In O'Sullivan [1], they used a cubic B-spline function and the penalty was the integrated squared second derivative of the B-spline function. On the other hand, Eilers and Marx [2] use a cubic B-spline function and a difference penalty on the spline coefficients. The Eilers and Marx's estimator is computationally efficient compared to smoothing splines and O'Sullivan's estimator since it removes the integration part of the penalty. Hence this paper focuses on the penalized spline estimator provided via Eilers and Marx [2]. The penalized spline method is efficient for both univariate regression and multiple regressions such as the additive model (see Marx and Eilers [3]). General properties usages and a description of the flexibility of penalized splines are described in Ruppert et al. [4].

When using penalized splines, the determination of the smoothing parameter is very important since it controls the trade-off between the goodness of fit and the smoothness of the fitted curve. As the classical method for achieving this, the grid search method is often used. The grid search method is selected by minimizing one criterion from candidate points of the smoothing parameter. Criteria for grid searches include cross-validation, generalized cross-validation, Mallow's C_p, and so forth. Although the grid search selection generally finds one optimal smoothing parameter, it is possible that the worth curve is obtained when not all the candidates are good. This tendency is especially striking in additive models since the number of the smoothing parameter is the same as that of the covariates. Several smoothing parameter selection methods using the grid search criteria have been developed by many authors such as Krivobokova [5], Reiss and Ogden [6], Wood [7], Wood [8], and Wood [9]. On the other hand, the mixed model representation of the spline smoothing has also been studied (see Lin and Zhang [10], Wand [11], and Ruppert et al. [4]). In mixed models, the grid search method is not necessary to obtain the final fit curve. The smoothing parameter in the mixed model can be written as the ratio of the variance of the random coefficient and the error. By estimating these unknown variances using a maximum likelihood method or a restricted maximum likelihood method (REML), the final fitted curves are obtained, yielding the estimated best linear unbiased predictor (EBLUP). Therefore the EBLUP does not require a grid search. However the fitted curve tends to theoretically oversmooth and the numerical stability is not guaranteed if a cubic spline is used (see Section 3). The Bayesian approach

to select the smoothing parameter has been studied by Fahrmeir et al. [12], Fahrmeir and Kneib [13], and Heinzl et al. [14]. Kauermann [15] compared some smoothing parameter selection methods.

In this paper, we propose a new method to determining the smoothing parameter using the asymptotic properties of the penalized splines. For the remainder of this paper, our new method will be known as the direct method. Before describing the outline of the direct method, we will briefly introduce the asymptotic studies of penalized splines. First, Hall and Opsomer [16] showed the consistency of the penalized spline estimator in white noise representation. Subsequently, Li and Ruppert [17], Claeskens et al. [18], Kauermann et al, [19], and Wang et al. [20] have developed the asymptotics for the penalized spline estimator in univariate regression. Yoshida and Naito [21] and Yoshida and Naito [22] have studied the asymptotics for penalized splines in additive regression models and generalized additive models, respectively. Xiao et al. [23] suggested a new penalized spline estimator, and developed its asymptotic properties in bivariate regression. Thus, the developments of the asymptotic theories of the penalized splines are relatively recent events. In addition, the smoothing parameter selection methods using asymptotic properties have not yet been studied. This motivates us to try to establish such methods.

The direct method is conducted by minimizing the mean integrated squared error (MISE) of the penalized spline estimator. In general, the MISE of the nonparametric estimator is divided into the integrated squared bias and the integrated variance of the estimator. Of course the penalized spline estimator is no exception and hence the direct method is stated by utilizing the expression of the asymptotic bias and variance of the penalized spline estimator, which have been derived by Claeskens et al. [18], Kauermann et al. [19], and Yoshida and Naito [22]. From their result, we see that the asymptotic order of the variance of the penalized spline estimator is only dependent on the sample size and the number of knots but not the smoothing parameter. However the second term of the asymptotic variance of the penalized spline estimator contains the smoothing parameter and we can see that the variance becomes small when the smoothing parameter increases. On the other hand, the squared bias of the penalized spline estimator increases if the smoothing parameter is reduced. Therefore the minimizer of the MISE of the penalized spline estimator can be seen as one of the optimal smoothing parameters. Since the MISE is asymptotically convex with respect to the smoothing parameter, the global minimum of the MISE can be found. This detection has been sufficiently developed for bandwidth selection in kernel regression (see Ruppert et al. [24], Wand and Jones [25], etc.). First the present paper focuses on univariate regression, and we next extend the direct method to the additive models. In both models, the mathematical and the numerical properties of the direct method are studied. In additive models, we need to select a smoothing parameter of the same number as the explanatory variable, such that the computational cost of the grid search becomes large. We expect that the computational cost of the direct method is dramatically reduced compared to the grid search.

The structure of this paper is as follows. In Section 2, we introduce a penalized spline estimator in a univariate regression model. Section 3 provides the direct method and related properties. Section 4 extends the direct method to the additive model. In Section 5, we confirm the performance of the direct method using a numerical study. We provide a discussion on the outlook and further studies in Section 6. The proofs of our theorems are provided in the appendix.

2. Penalized Spline Estimator

Consider the regression problem with n observations,

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

where Y_i is the response variable, x_i is the explanatory variable, f is the true regression function, and ε_i is the random error which is assumed to be independently distributed with mean 0 and variance σ^2. Throughout the paper we assume the explanatory variable $x_i \in [a, b]$ $(i = 1, \ldots, n)$ is not a random variable from which the expectation of Y_i can be expressed as $E[Y_i \mid x_i] = f(x_i)$. The support of the explanatory x_i can be relaxed as the real space $\mathbb{R}$. In order to simplify the way of locating the knots in the following, the support of the explanatory is assumed to be with compact space. We aim to estimate f via a nonparametric penalized spline method. We consider the knots $a = \kappa_0 < \kappa_1 < \cdots < \kappa_K < \kappa_{K+1} = b$ and, for $k = -p, \ldots, K-1$, let $B_k^{[p]}(x)$ be the pth degree B-spline basis function defined as

$$B_k^{[0]}(x) = \begin{cases} 1, & \kappa_k < x \le \kappa_{k+1}, \\ 0, & \text{otherwise}, \end{cases}$$

$$B_k^{[p]}(x) = \frac{x - \kappa_k}{\kappa_{k+p} - \kappa_k} B_k^{[p-1]}(x) + \frac{\kappa_{k+p+1} - x}{\kappa_{k+p+1} - \kappa_{k+1}} B_{k+1}^{[p-1]}(x) \tag{2}$$

associated with the above knots and the additional knots $\kappa_{-p} = \kappa_{-p+1} = \cdots = \kappa_0$ and $\kappa_{K+1} = \cdots = \kappa_{K+p+1}$. The B-spline function $B_k^{[p]}(x)$ is a piecewise pth degree polynomial on an interval $[\kappa_k, \kappa_{k+p+1}]$. The details of the B-spline basis functions are described in de Boor [26]. For simplicity, we write $B_k(x) = B_k^{[p]}(x)$ since we do not specify the p in the following sentence.

We use the linear combination of $\{B_k(x) \mid k = -p, \ldots, K-1\}$ and the unknown parameters b_k $(k = -p, \ldots, K-1)$ to approximate the regression function and consider the B-spline regression problem,

$$Y_i = s(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3}$$

where

$$s(x) = \sum_{k=-p}^{K-1} B_k(x) b_k. \tag{4}$$

The purpose is to estimate the parameter $\mathbf{b} = (b_{-p} \cdots b_{K-1})^T$ included in $s(x)$ instead of f directly.

The penalized spline estimator $\widehat{\mathbf{b}} = \left(\widehat{b}_{-p} \ \cdots \ \widehat{b}_{K-1}\right)^T$ of $\mathbf{b}$ is defined as the minimizer of

$$\sum_{i=1}^{n}\left\{Y_i - \sum_{k=-p}^{K-1} B_k(x)b_k\right\}^2 + \lambda \sum_{k=m-p}^{K-1}\left\{\Delta^m\left(b_k\right)\right\}^2, \tag{5}$$

where λ is the smoothing parameter and Δ is the backward difference operator defined as $\Delta b_k = b_k - b_{k-1}$ and

$$\Delta^m\left(b_k\right) = \Delta^{m-1}\left(\Delta b_k\right) = \sum_{j=0}^{m}(-1)^{m-j}{}_mC_j\, b_{k-m+j}. \tag{6}$$

Let $D_m = \left(d_{ij}^{(m)}\right)_{ij}$ be a $(K+p-m)\times(K+p)$ matrix, where $d_{ij}^{(m)} = (-1)^{|i-j|}{}_mC_{|i-j|}$ for $i \le j \le m+1$ and 0 otherwise. Using the notation $\mathbf{Y} = \left(Y_1 \ \cdots \ Y_n\right)^T$ and $Z = \left(B_{-p+j-1}(x_i)\right)_{ij}$, (5) can then be expressed as

$$\left(\mathbf{Y} - Z\mathbf{b}\right)^T\left(\mathbf{Y} - Z\mathbf{b}\right) + \lambda\mathbf{b}^T D_m^T D_m\mathbf{b}. \tag{7}$$

The minimum of (7) is obtained when

$$\widehat{\mathbf{b}} = \left(Z^TZ + \lambda D_m^TD_m\right)^{-1}Z^T\mathbf{Y}. \tag{8}$$

The penalized spline estimator $\widehat{f}(x)$ of $f(x)$ for $x \in [a,b]$ is defined as

$$\widehat{f}\left(x\right) = \sum_{k=-p}^{K-1} B_k\left(x\right)\widehat{b}_k = \mathbf{B}(x)^T\widehat{\mathbf{b}}, \tag{9}$$

where $\mathbf{B}(x) = \left(B_{-p}(x) \ \cdots \ B_{K-1}(x)\right)^T$.

If $\lambda \equiv 0$, $\widehat{f}(x)$ is reduced to the regression spline estimator, which is the spline estimator obtained via the least squares method. The regression spline estimator will lead to an oscillatory fit if the number of knots K is large. However the determination of K and the location of knots are very difficult problems. The advantage of the penalized spline smoothing is that the good smoothing parameter brings the estimator to the curve with the fitness and smoothness simultaneously without choosing the number and location of knots precisely. In the present paper, we use equidistant knots $\kappa_k = a + k/K$ and focus on the determination of the smoothing parameter. As the location of knots other than above, the quantiles of the data points $\{x_1, \ldots, x_n\}$ are often used (see Ruppert [27]). However it is known that the penalized spline estimator is hardly affected by the location of knots if K is not too small. Therefore we do not discuss the location of knots. We suggest the direct method for this in the next section.

3. Direct Determination of Smoothing Parameter

In this section, we provide the direct method for determining the smoothing parameter without a grid search. This direct method is given theoretical justification by asymptotic theory of the penalized spline estimator. To investigate the asymptotic property of the penalized spline estimator, we assume that $f \in C^{p+1}$, $K = o(n^{1/2})$, and $\lambda = O(n/K^{1-m})$.

For convenience we first give some notation. Let $G_n = n^{-1}Z^TZ$ and $\Lambda_n = G_n + (\lambda/n)D_m^TD_m$. Let $\mathbf{b}^*$ be a best L_∞ approximation to the true function f. This means that $\mathbf{b}^*$ satisfies

$$\begin{aligned}\sup_{x\in(a,b)}&\left|f\left(x\right) + K^{-(p+1)}b_a\left(x\right) - \mathbf{B}(x)^T\mathbf{b}^*\right| \\ &= o\left(K^{-(p+1)}\right), \quad \text{as } K \longrightarrow \infty,\end{aligned} \tag{10}$$

where

$$b_a\left(x\right) = -\frac{f^{(p+1)}\left(x\right)}{(p+1)!}\sum_{k=1}^{K} I\left(\kappa_{k-1} \le x < \kappa_k\right)\mathrm{Br}_{p+1}\left(\frac{x-\kappa_{k-1}}{K^{-1}}\right), \tag{11}$$

$I(a < x < b)$ is the indicator function of an interval (a,b), and $\mathrm{Br}_p(x)$ is the pth Bernoulli polynomial (see Zhou et al. [28]). It can be easily shown that $b_a(x) = O(1)$ as $K \to \infty$.

The penalized spline estimator can be written as

$$\begin{aligned}\widehat{f}\left(x\right) &= \mathbf{B}(x)^T\left(Z^TZ\right)^{-1}Z^T\mathbf{Y} \\ &\quad - \frac{\lambda}{n}\mathbf{B}(x)^T\Lambda_n^{-1}D_m^TD_mG_n^{-1}\left(\frac{Z}{n}\right)^T\mathbf{Y}.\end{aligned} \tag{12}$$

The first term of the right hand side of (12) is equal to the regression spline estimator denoted by $\widehat{f}_{\mathrm{rs}}(x)$. The asymptotics for the regression spline estimator has been developed by Zhou et al. [28] and can be expressed as

$$\begin{aligned}E\left[\widehat{f}_{\mathrm{rs}}\left(x\right)\right] &= f\left(x\right) + K^{-(p+1)}b_a\left(x\right) + o\left(K^{-(p+1)}\right), \\ V\left[\widehat{f}_{\mathrm{rs}}\left(x\right)\right] &= \frac{\sigma^2}{n}\mathbf{B}(x)^TG_n^{-1}\mathbf{B}\left(x\right)\left\{1 + o\left(1\right)\right\} = O\left(\frac{K}{n}\right).\end{aligned} \tag{13}$$

From Theorem 2(a) of Claeskens et al. [18], we have

$$\begin{aligned}E\left[\widehat{f}\left(x\right)\right] - E\left[\widehat{f}_{\mathrm{rs}}\left(x\right)\right] &= -\frac{\lambda}{n}\mathbf{B}(x)^T\Lambda_n^{-1}D_m^TD_m\mathbf{b}^* \\ &\quad + o\left(\frac{\lambda K^{1-m}}{n}\right) = O\left(\frac{\lambda K^{1-m}}{n}\right), \\ V\left[\widehat{f}\left(x\right)\right] &= V\left[\widehat{f}_{\mathrm{rs}}\left(x\right)\right] - \frac{\lambda K}{n}\frac{C\left(x \mid n,K,\lambda\right)}{K} + o\left(\frac{\lambda K^2}{n^2}\right),\end{aligned} \tag{14}$$

where

$$C\left(x \mid n,K,\lambda\right) = \frac{2\sigma^2}{n}\mathbf{B}(x)^T\Lambda_n^{-1}D_m^TD_mG_n^{-1}\mathbf{B}\left(x\right) \tag{15}$$

is the covariance of $\widehat{f}_{\mathrm{rs}}(x)$ and the second term of the right hand side of (12). The variance of the second term of (12) can be shown to be negligible (see the appendix). The following theorem leads to λ controlling the trade-off between the squared bias and variance of the penalized spline estimator.

Theorem 1. *The covariance $C(x \mid n,K,\lambda)/K$ in (15) is positive. Furthermore, as $n \to \infty$, $C(x \mid n,K,\lambda) = C(x \mid n,K,0)(1 + o(1))$ and $C(x \mid n,K,0)/K = O(K/n)$.*

From the asymptotic form of $E[\widehat{f}(x)]$ and $V[\widehat{f}(z)]$ and Theorem 1, we see that, for small λ, the bias of $\widehat{f}(x)$ is small and the variance becomes large. On the other hand, the large λ indicates the bias of $\widehat{f}(x)$ increases and the variance decreases. From Theorem 1, the MISE of $\widehat{f}(x)$ can be expressed as

$$\begin{aligned}\int_a^b E\left[\left\{\widehat{f}(x)-f(x)\right\}^2\right]dx &= \int_a^b \left\{K^{-(p+1)}b_a(x)-\frac{\lambda}{n}\mathbf{B}(x)^T\Lambda_n^{-1}D_m^TD_m\mathbf{b}^*\right\}^2 dx \\ &+\frac{\sigma^2}{n}\int_a^b \mathbf{B}(x)^TG_n\mathbf{B}(x)\,dx-\frac{\lambda K}{n}\frac{\int_a^b C(x\mid n,K,0)\,dx}{K} \\ &+r_{1n}(K)+r_{2n}(K,\lambda),\end{aligned} \tag{16}$$

where $r_{1n}(K)$ and $r_{2n}(K,\lambda)$ are of negligible order, respectively, compared to the regression spline and penalized spline of the second term of the right hand side of (12). Actually we have $r_{1n}(K)=o(K/n)$ and $r_{2n}(K,\lambda)=o(\lambda^2K^{2(1-m)}/n^2)$. The MISE of $\widehat{f}(x)$ is asymptotically quadratic and a global minimum exists. Let

$$\begin{aligned}\mathrm{MISE}(\lambda) &= \frac{\lambda^2}{n^2}\int_a^b\left\{\mathbf{B}(x)^T\Lambda_n^{-1}D_m^TD_m\mathbf{b}^*\right\}^2dx \\ &-\frac{\lambda}{n}\int_a^b\left\{2\frac{b_a(x)\,\mathbf{B}(x)^T\Lambda_n^{-1}D_m^TD_m\mathbf{b}^*}{K^{p+1}}\right. \\ &\left.+C(x\mid n,K,0)\right\}dx.\end{aligned} \tag{17}$$

And let λ_{opt} be the minimizer of MISE(λ). We suggest the use of λ_{opt} as the optimal smoothing parameter,

$$\lambda_{\mathrm{opt}}=\frac{n}{2}\frac{\int_a^b D_{2n}\left(x\mid f^{(p+1)},\mathbf{b}^*\right)dx}{\int_a^b D_{1n}(x\mid\mathbf{b}^*)\,dx}, \tag{18}$$

where

$$\begin{aligned}D_{1n}(x\mid\mathbf{b}^*) &= \left\{\mathbf{B}(x)^TG_n^{-1}D_m^TD_m\mathbf{b}^*\right\}^2, \\ D_{2n}\left(x\mid f^{(p+1)},\mathbf{b}^*\right) &= 2\frac{b_a(x)\,\mathbf{B}(x)^TG_n^{-1}D_m^TD_m\mathbf{b}^*}{K^{p+1}} \\ &+C(x\mid n,K,0).\end{aligned} \tag{19}$$

However it is easy to see that MISE(λ) and λ_{opt} contain an unknown function and parameters, and hence these must be estimated. We construct the estimator of λ_{opt} by using the consistent estimator of $f^{(p+1)}$ and $\mathbf{b}^*$. We can use the penalized spline estimator and its derivative as the pilot estimator of $f^{(p+1)}$ and $\mathbf{b}^*$. If we then use another smoothing parameter λ_0, it should be chosen appropriately. Therefore we use the regression spline estimator as the pilot estimator of $f^{(p+1)}(x)$ and $\mathbf{b}^*$. First we establish $\widetilde{\mathbf{b}}=(Z^TZ)^{-1}Z^TY$. Next we construct the pilot estimator with $f^{(p+1)}(x)$ by using the $(p+2)$th degree B-spline basis. Let $\mathbf{B}^{[p]}(x)=\left(B_{-p}^{[p]}(x)\ \cdots\ B_{K-1}^{[p]}(x)\right)^T$. Using the fundamental property of the B-spline function, $f^{(m)}(x)$ can be written as $\widetilde{f}^{(m)}(x)=B^{[p-m]}(x)^TD_m\mathbf{b}^*$ asymptotically. Hence the regression spline estimator $\widetilde{f}^{(p+1)}$ can be constructed as $\widetilde{f}^{(p+1)}(x)=\mathbf{B}^{[p+1]}(x)^TD_{p+1}\mathbf{b}^{(p+2)}$, where $\mathbf{b}^{(p+2)}=((Z^{[p+2]})^TZ^{[p+2]})^{-1}(Z^{[p+2]})^T\mathbf{Y}$ and $Z^{[p+2]}=(B_{-p+j-1}^{[p+2]}(x_i))_{ij}$. Since the regression spline estimator tends to be oscillatory with a higher order pth spline function, the fewer knots are used to construct $\widetilde{f}^{(p+1)}(x)$. The variance σ^2 included in $C(x\mid n,K,0)$ is estimated via

$$\widehat{\sigma}^2=\frac{1}{n-(K+p)}\sum_{i=1}^n\left\{Y_i-\mathbf{B}(x_i)^T\widetilde{\mathbf{b}}\right\}^2. \tag{20}$$

Using the above pilot estimators,

$$\widehat{\lambda}_{\mathrm{opt}}=\frac{n}{2}\frac{\sum_{j=1}^J D_2\left(z_j\mid\widetilde{f}^{(p+1)},\widetilde{\mathbf{b}}\right)}{\sum_{j=1}^J D_1\left(z_j\mid\widetilde{\mathbf{b}}\right)}, \tag{21}$$

with some finite grid points $\{z_j\}_1^J$ on $[a,b]$.

Consequently the final penalized spline estimator is defined as

$$\widehat{f}(x)=\mathbf{B}(x)^T\widehat{\mathbf{b}}_{\mathrm{opt}}, \tag{22}$$

where

$$\widehat{\mathbf{b}}_{\mathrm{opt}}=\left(Z^TZ+\widehat{\lambda}_{\mathrm{opt}}D_m^TD_m\right)^{-1}Z^T\mathbf{Y}. \tag{23}$$

It is known that the optimal order of K of the penalized spline estimator is the same as that of the regression spline estimator, $K=O(n^{1/(2p+3)})$ (see Kauermann et al. [19] and Zhou et al. [28]). Using this, we show the asymptotic property of λ_{opt} in following theorem.

Theorem 2. *Let $f\in C^{p+1}$. Suppose that $K=o(n^{1/2})$ and $\lambda=o(n/K^{1-m})$. Then λ_{opt} given in (21) exists, and $\lambda_{opt}=O(nK^{m-p-2}+K^{2m})$ as $n\to\infty$. Furthermore $K=O(n^{1/(2p+3)})$ leads to the optimal order*

$$\lambda_{\mathrm{opt}}=O\left(n^{(p+m+1)/(2p+3)}\right)+O\left(n^{2m/(2p+3)}\right) \tag{24}$$

and the rate of convergence of MISE of $\widehat{f}(x)$ becomes $O(n^{-2(p+1)/(2p+3)})$.

The proof of Theorem 2 is given in the appendix. At the end of this section, we give a few remarks.

Remark 3. The asymptotic order of the squared bias and variance of the penalized splines are $O(K^{-2(p+1)})+O(\lambda^2K^{2(1-m)}/n^2)$ and $O(K/n)$, respectively. Therefore under $K=(n^{1/(2p+3)})$ and $\lambda=O(n^{(p+m+1)/(2p+3)})$, the optimal rate of convergence of MISE of the penalized splines is $O(n^{-2(p+1)/(2p+3)})$. From Theorem 2, we see that the asymptotic order of λ_{opt} yields the optimal rate of convergence.

Remark 4. O'Sullivan [1] used $\mu \int_a^b \{s^{(m)}(x)\}^2 dx$ as the penalty term, where μ is the smoothing parameter. When equidistant knots are used, the penalty $\int_a^b \{s^{(m)}(x)\}^2 dx$ can be expressed as $K^{2m}\mathbf{b}^T D_m^T R D_m \mathbf{b}$, where $R = (\int_a^b B_i^{[p-m]}(x) B_j^{[p-m]}(x))_{ij}$. The penalty $\mathbf{b}^T D_m^T D_m \mathbf{b}$ provided by Eilers and Marx [2] can be seen as the simple version of $K^{2m}\mathbf{b}^T D_m^T R D_m \mathbf{b}$ by replacing R with $K^{-1}I$ and $\lambda = \mu K^{2m-1}$, where I is the identity matrix. So the order m of the difference matrix D_m controls the smoothness of the pth B-spline function $s(x)$, and hence m should be set such that $m \leq p$ to give a theoretical justification although the penalized spline estimator can be also calculated for $m > p$. Actually, $(p, m) = (3, 2)$ is often used by many authors.

Remark 5. The penalized spline regression is often considered as the mixed model representation (see Ruppert et al. [4]). In this frame work, we use the pth truncated spline model

$$s(x) = \beta_0 + \beta_1 x + \cdots + \beta_p x^p + \sum_{j=1}^{K} u_k (x - \kappa_k)_+^p, \tag{25}$$

where $(x - a)_+ = \max\{x - a, 0\}$ and the β's are unknown parameters. Each u_k is independently distributed as $u_k \sim N(0, \sigma_u^2)$, where σ_u^2 is an unknown variance parameter of u_k. The penalized spline fit of $f(x)$ is defined as the estimated BLUP (see Robinson [29]). The smoothing parameter in the ordinal spline regression model corresponds to σ^2/σ_u^2 in the spline mixed model. Since the σ^2 and σ_u^2 are estimated using the ML or REML method, we do not need to choose the smoothing parameter via a grid search. It is known that the estimated BLUP fit is linked to the penalized spline estimator (9) with $m = p + 1$ (see Kauermann et al. [19]). Hence the estimated BLUP tends to have a theoretically underfit (see Remark 4).

Remark 6. From Lyapunov's theorem, the asymptotic normality of the penalized spline estimator $\widehat{f}(x)$ with $\widehat{\lambda}_{\text{opt}}$ can be derived under the same assumption as Theorem 2 and some additional mild conditions. Although the proof is omitted, it is straightforward since $\widehat{\lambda}_{\text{opt}}$ is the consistent estimator of λ_{opt} and the asymptotic order of λ_{opt} satisfies Lyapunov's condition.

4. Extension to Additive Models

We extend the direct method to the regression model with multidimensional explanatory variables. In particular, we consider additive models in this section. For the dataset $\{(Y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ with 1-dimensional response Y_i and d-dimensional explanatory $\mathbf{x}_i = (x_{i1}, \ldots, x_{id})$, the additive models are connected via unknown regression functions f_j $(j = 1, \ldots, d)$ and the mean parameter μ such that

$$Y_i = \mu + f_1(x_{i1}) + \cdots + f_d(x_{id}) + \varepsilon_i. \tag{26}$$

We assume that x_{ij} is located on an interval $[a_j, b_j]$ and $f_1, \ldots, f_d$ are normalized as

$$\int_{a_j}^{b_j} f_j(u)\, du = 0, \quad j = 1, \ldots, d, \tag{27}$$

to ensure the identifiability of f_j. Then the intercept μ is typically estimated via

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i. \tag{28}$$

Hence we replace Y_i with $Y_i - \widehat{\mu}$ in (26) and set $\mu = 0$, redefining the additive model as

$$Y_i = f_1(x_{i1}) + \cdots + f_d(x_{id}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{29}$$

where each Y_i is centered. We aim to estimate f_j via a penalized spline method. Let

$$s_j(x) = \sum_{k=-p}^{K_j - 1} B_{j,k}(x)\, b_{j,k}, \quad j = 1, \ldots, d, \tag{30}$$

be the B-spline model, where $B_{j,k}(x)$ is the pth B-spline basis function with knots $\{\kappa_{j,k} \mid k = -p, \ldots, K_j + p + 1, \kappa_{j,0} = a_j, \kappa_{j,K_j+1} = b_j, j = 1, \ldots, d\}$ and $b_{j,k}$'s are unknown parameters. We consider the B-spline additive models

$$Y_i = s_1(x_{i1}) + \cdots + s_d(x_{id}) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{31}$$

and estimate s_j and $b_{j,k}$. For $j = 1, \ldots, d$, the penalized spline estimator $\widehat{\mathbf{b}}_j = \left(\widehat{b}_{j,-p} \;\cdots\; \widehat{b}_{j,K-1}\right)^T$ of $\mathbf{b}_j = \left(b_{j,-p} \;\cdots\; b_{j,K-1}\right)^T$ is defined as

$$\begin{bmatrix} \widehat{\mathbf{b}}_1 \\ \vdots \\ \widehat{\mathbf{b}}_d \end{bmatrix} = \underset{\mathbf{b}_1, \ldots, \mathbf{b}_d}{\operatorname{argmin}} \sum_{i=1}^{n} \left\{ Y_i - \sum_{j=1}^{d} \sum_{k=-p}^{K_j} B_{j,k}(x_i)\, b_{j,k} \right\}^2 + \sum_{j=1}^{d} \lambda_j \mathbf{b}_j^T D_{jm}^T D_{jm} \mathbf{b}_j, \tag{32}$$

where λ_j are the smoothing parameters and D_{jm} is the mth order difference matrix with size $(K_j + p - m) \times (K_j + p)$ for $j = 1, \ldots, d$. By using $\widehat{\mathbf{b}}_j$, the penalized spline estimator of $f_j(x_j)$ is defined as

$$\widehat{f}_j(x_j) = \sum_{k=-p}^{K_j - 1} B_{j,k}(x_j)\, \widehat{b}_{j,k}, \quad j = 1, \ldots, d. \tag{33}$$

The asymptotics for $\widehat{f}_j(x_j)$ have been studied by Yoshida and Naito [22] who derived the asymptotic bias and variance to be

$$E\left[\widehat{f}_j\left(x_j\right)\right] = f_j\left(x_j\right) + K_j^{-(p+1)} b_{ja}\left(x_j\right)(1 + o(1)) + \frac{\lambda_j}{n}\mathbf{B}_j\left(x_j\right)^T \Lambda_{jn}^{-1} D_{jm}^T D_{jm}\mathbf{b}_j^* + o\left(\frac{\lambda_j K_j^{1-m}}{n}\right),$$

$$V\left[\widehat{f}_j\left(x_j\right)\right] = \frac{\sigma^2}{n}\mathbf{B}_j\left(x_j\right)^T G_{jn}^{-1}\mathbf{B}_j\left(x_j\right) - \frac{\lambda K}{n}\frac{\sigma^2 C_j\left(x_j \mid n, K_j\right)}{K} + o\left(\frac{\lambda_j K_j^2}{n^2}\right), \tag{34}$$

where $\mathbf{B}_j(x) = \left(B_{-p}(x) \cdots B_{K_j-1}(x)\right)^T$, $G_{jn} = n^{-1}Z_j^T Z_j$, $\Lambda_{jn} = G_{jn} + (\lambda_j/n)D_m^T D_m$, $Z_j = (B_{-p+k-1}(x_{ij}))_{ik}$, and $\mathbf{b}_j^*$ is the best L_∞ approximation of f_j,

$$b_{ja}(x) = -\frac{f_j^{(p+1)}(x)}{(p+1)!}\sum_{k=0}^{K_j-1} I\left(\kappa_{j,k} \le x < \kappa_{j,k+1}\right) \times B_{p+1}\left(\frac{x - \kappa_{j,k}}{K_j^{-1}}\right),$$

$$C_j\left(x_j \mid n, K_j\right) = \frac{2}{n}\mathbf{B}_j\left(x_j\right)^T G_{jn}^{-1} D_{jm}^T D_{jm} G_n^{-1}\mathbf{B}_j\left(x_j\right). \tag{35}$$

The above asymptotic bias and variance of $\widehat{f}_j(x_j)$ are similar to that of the penalized spline estimator in univariate regression with $\{(Y_i, x_{ij}) : i = 1, \ldots, n\}$. Furthermore the asymptotic normality of $\left[\widehat{f}_1(x_1) \cdots \widehat{f}_d(x_d)\right]^T$ has been shown by Yoshida and Naito [22]. From their paper, we find that $\widehat{f}_i(x_i)$ and $\widehat{f}_j(x_j)$ are then asymptotically independent for $i \neq j$. This indicates some theoretical justification to select λ_j in minimizing the MISE of $\widehat{f}_j$. Similar to the discussion in Section 3, the minimizer $\lambda_{j,\mathrm{opt}}$ of the MISE of $\widehat{f}_j(x_j)$ can be obtained for $j = 1, \ldots, d$,

$$\lambda_{j,\mathrm{opt}} = \frac{n}{2}\frac{\int_{a_j}^{b_j} L_{jn}\left(x \mid f_j^{(p+1)}, \mathbf{b}_j^*, \sigma^2\right) dx}{\int_{a_j}^{b_j} H_{jn}\left(x \mid \mathbf{b}_j^*\right) dx}, \tag{36}$$

where

$$H_{jn}\left(x \mid \mathbf{b}_j^*\right) = \left\{\mathbf{B}_j\left(x_j\right)^T G_{jn}^{-1} D_{jm}^T D_{jm}\mathbf{b}_j^*\right\}^2,$$

$$L_{jn}\left(x \mid f_j^{(p+1)}, \mathbf{b}_j^*, \sigma^2\right) = 2\frac{b_a(x)\,\mathbf{B}_j\left(x_j\right)^T G_{jn}^{-1} D_{jm}^T D_{jm}\mathbf{b}_j^*}{K_j^{p+1}} + \sigma^2 C_j\left(x \mid n, K_j\right). \tag{37}$$

Since f_j, $\mathbf{b}_j^*$, and σ^2 are unknown, they should be estimated. The pilot estimators of f_j, $\mathbf{b}_j^*$, and σ^2 are constructed based on the regression spline method. By using the pilot estimators $\widetilde{f}_j$, $\widetilde{\mathbf{b}}_j$ of f_j, $\mathbf{b}_j^*$ and the estimator of σ^2, we construct the estimator of $\lambda_{j,\mathrm{opt}}$:

$$\widehat{\lambda}_{j,\mathrm{opt}} = \frac{n}{2}\frac{\sum_{r=1}^R L_{jn}\left(z_r \mid \widetilde{f}_j^{(p+1)}, \widetilde{\mathbf{b}}_j, \widehat{\sigma}^2\right)}{\sum_{r=1}^R H_{jn}\left(z_r \mid \widetilde{\mathbf{b}}_j\right)}, \tag{38}$$

where $\{z_r\}_1^R$ is some finite grid point sequence on $[a_j, b_j]$. We obtain for $j = 1, \ldots, d$, the penalized spline estimator $\widehat{f}_j(x_j)$ of $f_j(x_j)$,

$$\widehat{f}_j\left(x_j\right) = \mathbf{B}_j\left(x_j\right)^T\widehat{\mathbf{b}}_{j,\mathrm{opt}}, \tag{39}$$

where $\widehat{\mathbf{b}}_{j,\mathrm{opt}}$ is the penalized spline estimator of $\mathbf{b}_j$ using $\widehat{\lambda}_{j,\mathrm{opt}}$. From Theorem 2 and the proof of Theorem 3.4 of Yoshida and Naito [22], the asymptotic normality of $\left[\widehat{f}_1(x_1) \cdots \widehat{f}_d(x_d)\right]^T$ using $(\lambda_{1,\mathrm{opt}}, \ldots, \lambda_{d,\mathrm{opt}})$ can be shown.

Remark 7. Since the true regression functions are normalized, the estimator $\widehat{f}_j$ should be also centered as

$$\widehat{f}_j\left(x_j\right) - \frac{1}{n}\sum_{i=1}^n \widehat{f}_j\left(x_{ij}\right). \tag{40}$$

Remark 8. The penalized spline estimator of $\mathbf{b}_1, \ldots, \mathbf{b}_d$ can be obtained using a backfitting algorithm (Hastie and Tibshirani [30]). The backfitting algorithm for the penalized splines in additive regression is detailed in Marx and Eilers [3] and Yoshida and Naito [21].

Remark 9. Although we focus on the nonparametric additive regression in this paper, the direct method can be also applied to the generalized additive models. However we omit this discussion because the procedure is similar to the case of the additive models discussed in this section.

Remark 10. The direct method is quite computationally efficient when compared to the grid search method in additive models. In grid searches, we prepare the candidate of λ_j. Let M_j be the set of all possible candidate grid value of λ_j for $j = 1, \ldots, d$. Then we need to compute the backfitting algorithm $\{M_1 \times \cdots \times M_d\}$ times. On the other hand, it is sufficient to perform the backfitting algorithm for only two steps for the pilot estimator and the final penalized spline estimator. Thus, compared with the conventional grid search method, the direct method can drastically reduce computation time.

5. Numerical Study

In this section, we investigate the finite sample performance of the proposed direct method in a Monte Carlo simulation. Let us first consider the univariate regression model (1) for the data $\{(Y_i, x_i) : i = 1, \ldots, n\}$. Then we use the three

Table 1: Results of sample MISE for $n = 50$ and $n = 200$. All entries for MISE are 10^2 times their actual values.

True	F1		F2		F3	
Sample size	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$
L-Direct	0.526	0.322	3.684	1.070	1.160	0.377
L-GCV	0.726	0.621	5.811	1.082	1.183	0.379
L-REML	2.270	1.544	9.966	3.520	1.283	0.873
Local linear	0.751	0.220	4.274	0.901	1.689	0.503
C-Direct	0.401	0.148	3.027	0.656	1.044	0.326
C-GCV	0.514	0.137	3.526	0.666	1.043	0.290
C-REML	1.326	0.732	8.246	4.213	3.241	0.835

Table 2: Results of sample MSE of smoothing parameter obtained by direct method, GCV, and REML for $n = 50$ and $n = 200$. All entries for MSE are 10^2 times their actual values.

True	F1		F2		F3	
Sample size	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$
L-Direct	0.331	0.193	0.113	0.043	0.025	0.037
L-GCV	0.842	0.342	0.445	0.101	0.072	0.043
L-REML	1.070	1.231	0.842	0.437	0.143	0.268
C-Direct	0.262	0.014	0.082	0.045	0.053	0.014
C-GCV	0.452	0.123	0.252	0.092	0.085	0.135
C-REML	0.855	0.224	0.426	0.152	0.093	0.463

types of true regression function $f(x) = \cos(\pi(x - 0.3))$, $f(x) = \phi((x - 0.8)/0.05)/\sqrt{0.05} - \phi((x - 0.2)/0.04)/\sqrt{0.04}$, and $f(x) = 2x^3 + 3\sin(2\pi(x-0.8)^3) + 3\exp[-(x-0.5)^2/0.1]$, which are labeled by F1, F2, and F3, respectively. Here $\phi(x)$ is the density function of the normal distribution. The explanatory x_i and error ε_i are independently generated from uniform distribution on $[0, 1]$ and $N(0, 0.5^2)$, respectively. We estimate each true regression function via the penalized spline method. We then use the linear and cubic B-spline bases with equidistant knots and the second order difference penalty. In addition we set $K = 5n^{2/5}$ equidistant knots and the smoothing parameter is determined by the direct method. The penalized spline estimator with the linear spline and cubic spline are denoted as L-Direct and C-Direct, respectively. For comparison with L-Direct, the same studies are also implemented for the penalized spline estimator with a linear spline and the smoothing parameter selected via GCV and restricted maximum likelihood method (REML) in mixed model representation, and the local polynomial estimator with normal kernel and Plug-in bandwidth (see Ruppert et al. [24]). In GCV, we set the candidate values of λ as $\{i/10 : i = 0, 1, \ldots, 99\}$. The above three estimators are denoted by L-GCV, L-REML, and local linear, respectively. Furthermore we compare C-Direct with C-GCV and C-REML, which are the penalized spline estimator with the cubic spline and the smoothing parameter determined by GCV and REML. Let

$$\text{sMISE} = \frac{1}{J}\sum_{j=1}^{J}\left[\frac{1}{R}\sum_{r=1}^{R}\left\{\widehat{f}_r\left(z_j\right) - f\left(z_j\right)\right\}^2\right] \tag{41}$$

be the sample MISE of any estimator $\widehat{f}(x)$ of $f(x)$, where $\widehat{f}_r(z)$ is $\widehat{f}(z)$ with rth replication and $z_j = j/J$. We calculate the sample MISE of the penalized spline estimator with the direct method, GCV, and REML and the local linear estimator. In this simulation, we use $J = 100$ and $R = 1000$. We have simulated $n = 50$ and 200.

The sMISE of all estimators for each model and n are given in Table 1. The penalized spline estimator using the direct method shows good performance in each setting. In comparison with other smoothing parameter methods, the direct method is a little better than the GCV as a whole. However for $n = 200$, C-GCV is better than C-Direct in F1 and F3 though its difference is very small. We see that the sMISE of C-Direct is smaller than local linear, whereas local linear behaves better than L-Direct in some case. In F2, C-Direct is the smallest of all estimators for $n = 50$ and 200. Although the performance totally seems to depend on situation in which data sets are generated, we believe that the proposed method is one of the efficient methods.

Next the difference between $\widehat{\lambda}_{\text{opt}}$ and λ_{opt} is investigated empirically. Let $\widehat{\lambda}_{\text{opt},r}$ be the $\widehat{\lambda}_{\text{opt}}$ with rth replications for $r = 1, \ldots, 1000$. Then we calculate the sample MSE of $\widehat{\lambda}_{\text{opt}}$:

$$\text{sMSE} = \frac{1}{R}\sum_{r=1}^{R}\left\{\widehat{\lambda}_{\text{opt},r} - \lambda_{\text{opt}}\right\}^2 \tag{42}$$

for F1, F2, and F3 and $n = 50$ and 200. To construct λ_{opt} for F1, F2, and F3, we use true $f^{(4)}(x)$ and σ^2 and an approximate $\mathbf{b}^*$. Here the approximate $\mathbf{b}^*$ means the sample average of $\widetilde{\mathbf{b}}$ with $n = 200$ and 1000 replications.

In Table 2, the sMSE of $\widehat{\lambda}_{\text{opt}}$ for each true function are described. For comparison, the sMSE of the smoothing

Table 3: Results of sample PSE for $n = 50$ and $n = 200$. All entries for PSE are 10^2 times their actual values.

True	F1		F2		F3	
Sample size	$n = 50$	$n = 200$	$n = 50$	$n = 200$	$n = 50$	$n = 200$
L-Direct	0.424	0.052	0.341	0.070	0.360	0.117
L-GCV	0.331	0.084	0.381	0.089	0.333	0.139
L-REML	0.642	0.124	0.831	1.002	0.733	0.173
Local linear	0.751	0.220	0.474	0.901	0.489	0.143
C-Direct	0.341	0.48	0.237	0.63	0.424	0.086
C-GCV	0.314	0.43	0.326	0.76	0.533	0.089
C-REML	0.863	1.672	0.456	1.21	1.043	0.125

parameter obtained via GCV and REML are calculated. The sMSE of the L-Direct and the C-Direct is small even in $n = 50$ for all true functions. Therefore it seems that the accuracy of $\widehat{\lambda}_{\text{opt}}$ is guaranteed. It indicates that the pilot estimator constructed via least squares method is not bad. The sMSE with the direct method are smaller than that with GCV and REML. This result is not surprising since GCV and REML are not concerned with λ_{opt}. However together with Table 1, it seems that the sample MSE of the smoothing parameter is reflected in the sample MISE of the estimator.

The proposed direct method was derived based on the MISE. On the other hand the GCV and the REML are obtained in context of minimizing prediction squared error (PSE) and prediction error. Hence we compare the sample PSE of the penalized spline estimator with the direct method, GCV and REML, and the local linear estimator. Since the prediction error is almost the same as MISE (see Section 4 of Ruppert et al. [4]), we omit the comparison of the prediction error. Let

$$\text{sPSE} = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{R}\sum_{r=1}^{R}\left\{y_{ir} - \widehat{f}_r\left(x_i\right)\right\}^2\right] \tag{43}$$

be the sample PSE for any estimator $\widehat{f}(x)$, where y_{ir} $(r = 1, \ldots, R)$ is independently generated from $Y_i \sim N(f(x_i), (0.5)^2)$ for all i.

In Table 3, the modified sPSE, $|\text{sPSE} - (0.5)^2|$, of all estimators for each model and n are described. In Remark 11, we discuss the modified sPSE. From the result, we can confirm that the direct method has good predictability. GCV can be regarded as the estimator of sPSE. Therefore in some case, sPSE with GCV is smaller than that with the direct method. It seems that the cause is the accuracy of the estimates of variance (see Remark 11). However its difference is very small.

We admit that the computational time (in second) taken to obtain $\widehat{f}(x)$ for F1, $p = 3$, and $n = 200$. The fits with the direct method, GCV and REML took 0.04, 1.22, and 0.34. Although the difference is small, the computational time of the direct method was faster than that of GCV and REML.

Next we confirm the behavior of the penalized spline estimator with the direct method in the additive model. For the data $\{(Y_i, x_{i1}, x_{i2}, x_{i3}) : i = 1, \ldots, n\}$, we assume the additive model with true functions $f_1(x_1) = \sin(2\pi x_1)$, $f_2(x_2) = \phi((x_2 - 0.5)/0.2)$, and $f_3(x_3) = 0.4\phi((x_3 - 0.1)/0.2) + 0.6\phi((x_2 - 0.8)/0.2)$. The error is similar to the first simulation. The design points (x_{i1}, x_{i2}, x_{i3}) are generated by

$$\begin{bmatrix} x_{i1} \\ x_{i2} \\ x_{i3} \end{bmatrix} = \begin{bmatrix} \left(1+\rho+\rho^2\right)^{-1} & 0 & 0 \\ 0 & \left(1+2\rho\right)^{-1} & 0 \\ 0 & 0 & \left(1+\rho+\rho^2\right)^{-1} \end{bmatrix} \times \begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \begin{bmatrix} z_{i1} \\ z_{i2} \\ z_{i3} \end{bmatrix}, \tag{44}$$

where z_{ij} $(i = 1, \ldots, n,\ j = 1, 2, 3)$ are generated independently from the uniform distribution on $[0, 1]$. In this simulation, we adopt $\rho = 0$ and $\rho = 0.5$. We then corrected to satisfy

$$\int_0^1 f_j(z)\, dz = 0, \quad j = 1, 2, 3 \tag{45}$$

in each ρ. We construct the penalized spline estimator via the backfitting algorithm with the cubic spline and second order difference penalty. The number of equidistant knots is $K_j = 5n^{2/5}$ and the smoothing parameters are determined using the direct method. The pilot estimator to construct $\widehat{\lambda}_{j,\text{opt}}$ is the regression spline estimator with fifth spline and $K_j = n^{2/5}$. We calculate the sample MISE of each $\widehat{f}_j$ for $n = 200$ and 1000 Monte Carlo iterations. Then we calculate the sample MISE of $\widehat{f}_1$, $\widehat{f}_2$, and $\widehat{f}_3$. In order to compare with the direct method, we also conduct the same simulation with GCV and REML. In GCV, we set the candidate values of λ_j as $\{i/10 : i = 0, 1, \ldots, 9\}$ for $j = 1, 2, 3$. Table 4 summarizes the sample MISE of $\widehat{f}_j$ $(j = 1, 2, 3)$ denoted by MISEj for direct method, GCV, and REML. The penalized spline estimator with the direct method performs like that with the GCV in both the uncorrelated and correlated design cases. For $\rho = 0$, the behaviors of MISE1, MISE2, and MISE3 with the direct method are similar. On the other hand, for GCV, the MISE1 is slightly larger than MISE2 and MISE3. The direct method leads to an efficient estimate of all covariates. On the whole, the direct method is better than REML. From above, we believe that the direct method is preferable in practice.

Table 4: Results of sample MISE of the penalized spline estimator. For each direct method, GCV and ρ, MISE1, MISE2, and MISE3 are the sample MISE of $\widehat{f}_1$, $\widehat{f}_2$, and $\widehat{f}_3$, respectively. All entries for MISE are 10^2 times their actual values.

Method	$\rho = 0$			$\rho = 0.5$		
	MISE1	MISE2	MISE3	MISE1	MISE2	MISE3
Direct	0.281	0.209	0.205	0.916	0.795	0.972
GCV	0.687	0.237	0.390	0.929	0.857	1.112
REML	1.204	0.825	0.621	2.104	1.832	2.263

Table 5: Results of sample MSE of the selected smoothing parameter via direct method, GCV, and REML for $\rho = 0$ and $\rho = 0.5$. All entries for MSE are 10^2 times their actual values.

	$\rho = 0$			$\rho = 0.5$		
True	f_1	f_2	f_3	f_1	f_2	f_3
Direct	0.124	0.042	0.082	0.312	0.105	0.428
GCV	1.315	0.485	0.213	0.547	0.352	0.741
REML	2.531	1.237	2.656	1.043	0.846	1.588

Table 6: Results of sample PSE of the penalized spline estimator. All entries for PSE are 10^2 times their actual values.

$n = 200$	$\rho = 0$	$\rho = 0.5$
Direct	0.281	0.429
GCV	0.387	0.467
REML	1.204	0.925

To confirm the consistency of the direct method, the sample MSE of $\widehat{\lambda}_{j,\text{opt}}$ is calculated the same manner as that given in univariate regression case. For comparison, the sample MSE of GCV and REML are obtained. Here the true $\lambda_{j,\text{opt}}$ $(j = 1, 2, 3)$ is defined in the same way as that for univariate case. In Table 5, the results for each f_1, f_2, and f_3, and $\rho = 0$ and 0.5 are shown. We see from the results that the behavior of the direct method is good. The sample MSE of the direct method is smaller than that of GCV and REML for all j. For the random design $\rho = 0.5$, such tendency can be also seen.

Table 6 shows the sPSE of $\widehat{f}(\mathbf{x}_i) = \widehat{f}_1(x_{i1}) + \widehat{f}_2(x_{i2}) + \widehat{f}_3(x_{i3})$ with each smoothing parameter selection method. We see from result that the efficiency of the direct method is guaranteed in the context of prediction accuracy.

Finally we show the computational time (in second) required to construct the penalized spline estimator with each method. The computational times with the direct method, GCV and REML are 11.87 s, 126.43 s, and 43.5 s, respectively. We see that the direct method is more efficient than other methods in context of the computation (see Remark 11). All of computations in the simulation were done using a software R and a computer with 3.40 GHz CPU and 24.0 GB memory. Though this is only few examples, the direct method can be seen as one of the good methods to select the smoothing parameter.

Remark 11. We calculate $|\text{sPSE} - (0.5)^2|$ as the criteria to confirm the prediction squared error. The ordinal PSE is defined as

$$\begin{aligned}\text{PSE} &= \frac{1}{n}\sum_{i=1}^{n} E\left[\left\{Y_i - \widehat{f}(x_i)\right\}^2\right] \\ &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n} E\left[\left\{\widehat{f}(x_i) - f(x_i)\right\}^2\right],\end{aligned} \tag{46}$$

where Y_i's are the test data. The second term of the PSE is similar to MISE. So it can be said that the sample PSE evaluates the accuracy of the variance part and MISE part. To see the detail of the difference of the sample PSE between the direct method and other method, we calculated $|\text{sPSE} - \sigma^2|$ in this section.

6. Discussion

In this paper, we proposed a new direct method for determining the smoothing parameter for a penalized spline estimator in a regression problem. The direct method is based on minimizing MISE of the penalized spline estimator. We studied the asymptotic property of the direct method. The asymptotic normality of the penalized spline estimator using $\widehat{\lambda}_{\text{opt}}$ is theoretically guaranteed when the consistent estimator is used as the pilot estimator to obtain $\widehat{\lambda}_{\text{opt}}$. In numerical study, for the additive model, the computational cost of this direct method is dramatically reduced when compared to grid search methods such as GCV. Furthermore we find that the performance of the direct method is better than or at least similar to that of other methods.

The direct method will be developed for other regression models such as varying-coefficient models, Cox proportional hazards models, single-index models, and others if the asymptotic bias and variance of the penalized spline

estimator are derived. It is not limited to the mean regression; it can be applied to quantile regression. Actually, Yoshida [31] has presented the asymptotic bias and variance of the penalized spline estimator in univariate quantile regressions. Furthermore, it is seen that improving the direct method is important for various situations and datasets. In particular, the development of the determination of locally adaptive λ is an interesting avenue of further research.

Appendix

We describe the technical details. For the matrix $A = (a_{ij})_{ij}$, $\|A\|_\infty = \max_{ij}\{|a_{ij}|\}$. First, to prove Theorems 1 and 2, we introduce the fundamental property of penalized splines in the following lemma.

Lemma A.1. *Let $A = (a_{ij})_{ij}$ be $(K + p)$ matrix. Suppose that $K = o(n^{1/2})$, $\lambda = o(nK^{1-m})$, and $\|A\|_\infty = O(1)$. Then, $\|AG_n^{-1}\|_\infty = O(K)$ and $\|A\Lambda_n^{-1}\|_\infty = O(K)$.*

The proof of Lemma A.1 is shown in Zhou et al. [28] and Claeskens et al. [18]. The repeated use of Lemma A.1 yields that the asymptotic order of the variance of the second term of (12) is $O(\lambda^2(K/n)^3)$. Actually, the asymptotic order of the variance of the second term of (12) can be calculated as

$$\left(\frac{\lambda}{n}\right)^2 \frac{\sigma^2}{n}\mathbf{B}(x)^T\Lambda_n^{-1}D_m^TD_mG_n^{-1}D_m^TD_m\Lambda_n^{-1}\mathbf{B}(x) = O\left(\frac{\lambda^2}{n^3}K^3\right). \tag{A.1}$$

When $K = o(n^{1/2})$ and $\lambda = O(n/K^{1-m})$, $\lambda^2(K/n)^3 = o(\lambda(K/n)^2)$ holds.

Proof of Theorem 1. We write

$$C(x \mid n, K, \lambda) = \frac{2\sigma^2}{n}\mathbf{B}(x)^TG_n^{-1}D_m^TD_mG_n^{-1}\mathbf{B}(x) + \frac{2\sigma^2}{n}\mathbf{B}(x)^T\left\{\Lambda_n^{-1} - G_n^{-1}\right\}D_m^TD_mG_n^{-1}\mathbf{B}(x). \tag{A.2}$$

Since

$$\Lambda_n^{-1} - G_n^{-1} = \Lambda_n^{-1}\left\{I - \Lambda_nG_n^{-1}\right\} = \frac{\lambda}{n}\Lambda_n^{-1}D_m^TD_mG_n^{-1}, \tag{A.3}$$

we have $\|\Lambda_n^{-1} - G_n^{-1}\|_\infty = O(\lambda K^2/n)$ and hence the second term of right hand side of (A.2) can be shown $O((K/n)^2(\lambda K/n))$. From Lemma A.1, it is easy to derive that $C(x \mid n, K, 0) = O(K(K/n))$. Consequently we have

$$\frac{C(x \mid n, K, \lambda)}{K} = \frac{C(x \mid n, K, 0)}{K} + O\left(\left(\frac{K}{n}\right)^2\left(\frac{\lambda}{n}\right)\right) = \frac{C(x \mid n, K, 0)}{K} + o\left(\frac{K}{n}\right) \tag{A.4}$$

and this leads to Theorem 1. □

Proof of Theorem 2. It is sufficient to prove that $\int_a^b D_{1n}(x \mid \mathbf{b}^*)dx = O(K^{2(1-m)})$ and

$$\int_a^b D_{2n}\left(x \mid f^{(p+1)}, \mathbf{b}^*\right)dx = O\left(K^{-(p+m)}\right) + O\left(\frac{K^2}{n}\right). \tag{A.5}$$

Since

$$\left|\int_a^b \left\{\mathbf{B}(x)^TG_n^{-1}D_m^TD_m\mathbf{b}^*\right\}^2 dx\right| \le \sup_{x\in[a,b]}\left[\left\{\mathbf{B}(x)^TG_n^{-1}D_m^TD_m\mathbf{b}^*\right\}^2\right](a-b), \tag{A.6}$$

we have

$$\int_a^b D_{1n}(x \mid \mathbf{b}^*)dx = O\left(K^{2(1-m)}\right) \tag{A.7}$$

by using Lemma A.1 and the fact that $\|D_m\mathbf{b}_*\|_\infty = O(K^{-m})$ (see Yoshida [31]).

From the property of B-spline function, for $(K+p)$ square matrix A, there exists $C > 0$ such that

$$\int_a^b \mathbf{B}(x)^TA\mathbf{B}(x)\,dx \le \|A\|_\infty\int_a^b \mathbf{B}(x)^T\mathbf{B}(x)\,dx \le \|A\|_\infty C. \tag{A.8}$$

In other words, the asymptotic order of $\int_a^b \mathbf{B}(x)^TA\mathbf{B}(x)dx$ and that of $\|A\|_\infty$ are the same. Let $A_n = G_n^{-1}D_m^TD_mG_n^{-1}$. Then, since $\|A_n\|_\infty = K^2$, we obtain

$$\int_a^b C(x \mid n, K, 0)\,dx = O\left(\frac{K^2}{n}\right). \tag{A.9}$$

Furthermore because $\sup_{a\in[a.b]}\{b_a(x)\} = O(1)$, we have

$$K^{-(p+1)}\left|\int_a^b b_a(x)\,\mathbf{B}(x)^TG_n^{-1}D_m^TD_m\mathbf{b}^*dx\right| = O\left(K^{-(p+m)}\right), \tag{A.10}$$

and hence it can be shown that

$$\int_a^b D_{2n}\left(x \mid f^{(p+1)}, \mathbf{b}^*\right)dx = O\left(K^{-(p+m)}\right) + O\left(n^{-1}K^2\right). \tag{A.11}$$

Together with (A.7) and (A.11), $\lambda_{\text{opt}} = O(nK^{m-p-2} + K^{2m})$ can be obtained. Then rate of convergence of the penalized spline estimator with λ_{opt} is $O(n^{-2(p+2)}/(2p+3))$ when $K = O(n^{1/(2p+3)})$, which is detailed in Yoshida and Naito [22]. □

Conflict of Interests

The author declares that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

The author also would like to thank two anonymous referees for their careful reading and comments to improve the paper.

References

[1] F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statistical Science*, vol. 1, no. 4, pp. 502–527, 1986.

[2] P. H. C. Eilers and B. D. Marx, "Flexible smoothing with *B*-splines and penalties," *Statistical Science*, vol. 11, no. 2, pp. 89–121, 1996, With comments and a rejoinder by the authors.

[3] B. D. Marx and P. H. C. Eilers, "Direct generalized additive modeling with penalized likelihood," *Computational Statistics and Data Analysis*, vol. 28, no. 2, pp. 193–209, 1998.

[4] D. Ruppert, M. P. Wand, and R. J. Carroll, *Semiparametric Regression*, Cambridge University Press, Cambridge, UK, 2003.

[5] T. Krivobokova, "Smoothing parameter selection in two frameworks for penalized splines," *Journal of the Royal Statistical Society B*, vol. 75, no. 4, pp. 725–741, 2013.

[6] P. T. Reiss and R. T. Ogden, "Smoothing parameter selection for a class of semiparametric linear models," *Journal of the Royal Statistical Society B*, vol. 71, no. 2, pp. 505–523, 2009.

[7] S. N. Wood, "Modelling and smoothing parameter estimation with multiple quadratic penalties," *Journal of the Royal Statistical Society B*, vol. 62, no. 2, pp. 413–428, 2000.

[8] S. N. Wood, "Stable and efficient multiple smoothing parameter estimation for generalized additive models," *Journal of the American Statistical Association*, vol. 99, no. 467, pp. 673–686, 2004.

[9] S. N. Wood, "Fast stable direct fitting and smoothness selection for generalized additive models," *Journal of the Royal Statistical Society B*, vol. 70, no. 3, pp. 495–518, 2008.

[10] X. Lin and D. Zhang, "Inference in generalized additive mixed models by using smoothing splines," *Journal of the Royal Statistical Society B*, vol. 61, no. 2, pp. 381–400, 1999.

[11] M. P. Wand, "Smoothing and mixed models," *Computational Statistics*, vol. 18, no. 2, pp. 223–249, 2003.

[12] L. Fahrmeir, T. Kneib, and S. Lang, "Penalized structured additive regression for space-time data: a Bayesian perspective," *Statistica Sinica*, vol. 14, no. 3, pp. 731–761, 2004.

[13] L. Fahrmeir and T. Kneib, "Propriety of posteriors in structured additive regression models: theory and empirical evidence," *Journal of Statistical Planning and Inference*, vol. 139, no. 3, pp. 843–859, 2009.

[14] F. Heinzl, L. Fahrmeir, and T. Kneib, "Additive mixed models with Dirichlet process mixture and P-spline priors," *Advances in Statistical Analysis*, vol. 96, no. 1, pp. 47–68, 2012.

[15] G. Kauermann, "A note on smoothing parameter selection for penalized spline smoothing," *Journal of Statistical Planning and Inference*, vol. 127, no. 1-2, pp. 53–69, 2005.

[16] P. Hall and J. D. Opsomer, "Theory for penalised spline regression," *Biometrika*, vol. 92, no. 1, pp. 105–118, 2005.

[17] Y. Li and D. Ruppert, "On the asymptotics of penalized splines," *Biometrika*, vol. 95, no. 2, pp. 415–436, 2008.

[18] G. Claeskens, T. Krivobokova, and J. D. Opsomer, "Asymptotic properties of penalized spline estimators," *Biometrika*, vol. 96, no. 3, pp. 529–544, 2009.

[19] G. Kauermann, T. Krivobokova, and L. Fahrmeir, "Some asymptotic results on generalized penalized spline smoothing," *Journal of the Royal Statistical Society B*, vol. 71, no. 2, pp. 487–503, 2009.

[20] X. Wang, J. Shen, and D. Ruppert, "On the asymptotics of penalized spline smoothing," *Electronic Journal of Statistics*, vol. 5, pp. 1–17, 2011.

[21] T. Yoshida and K. Naito, "Asymptotics for penalized additive *B*-spline regression," *Journal of the Japan Statistical Society*, vol. 42, no. 1, pp. 81–107, 2012.

[22] T. Yoshida and K. Naito, "Asymptotics for penalized splines in generalized additive models," *Journal of Nonparametric Statistics*, vol. 26, no. 2, pp. 269–289, 2014.

[23] L. Xiao, Y. Li, and D. Ruppert, "Fast bivariate *P*-splines: the sandwich smoother," *Journal of the Royal Statistical Society B*, vol. 75, no. 3, pp. 577–599, 2013.

[24] D. Ruppert, S. J. Sheather, and M. P. Wand, "An effective bandwidth selector for local least squares regression," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1257–1270, 1995.

[25] M. P. Wand and M. C. Jones, *Kernel Smoothing*, Chapman & Hall, London, UK, 1995.

[26] C. de Boor, *A Practical Guide to Splines*, Springer, New York, NY, USA, 2001.

[27] D. Ruppert, "Selecting the number of knots for penalized splines," *Journal of Computational and Graphical Statistics*, vol. 11, no. 4, pp. 735–757, 2002.

[28] S. Zhou, X. Shen, and D. A. Wolfe, "Local asymptotics for regression splines and confidence regions," *The Annals of Statistics*, vol. 26, no. 5, pp. 1760–1782, 1998.

[29] G. K. Robinson, "That BLUP is a good thing: the estimation of random effects," *Statistical Science*, vol. 6, no. 1, pp. 15–32, 1991.

[30] T. J. Hastie and R. J. Tibshirani, *Generalized Additive Models*, Chapman & Hall, London, UKI, 1990.

[31] T. Yoshida, "Asymptotics for penalized spline estimators in quantile regression," *Communications in Statistics-Theory and Methods*, 2013.