A Non-iterative Optimization for Smoothness in Penalized Spline Regression

(Last Modified: November 26, 2009)

Hirokazu YANAGIHARA[1]

[1] *Department of Mathematics, Graduate School of Science, Hiroshima University*
1-3-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8626, Japan

Abstract

Typically, an optimal smoothing parameter in a penalized spline regression is determined by minimizing an information criterion, such as one of the C_p, CV and GCV criteria. Since an explicit solution to the minimization problem for an information criterion cannot be obtained, it is necessary to carry out an iterative procedure to search for the optimal smoothing parameter, i.e., a grid search method. In order to avoid such extra calculation, a non-iterative optimization method for smoothness in penalized spline regression is proposed using the formulation of generalized ridge regression. By conducting numerical simulations, we verify that our method has better performance than other methods which optimize the number of basis functions and the single smoothing parameter by means of the CV or GCV criteria.

MSC 2010 *subject classifications*: Primary 62G08; Secondary 62J07.
Key words: *B*-spline, Generalized ridge regression, Mallows' C_p criterion, Multiple smoothing parameters, Penalized least square estimation.

1. Introduction

In a penalized spline regression, the problem of optimal choice of the smoothing parameter λ ($\lambda \geq 0$) is important because λ controls the local variation along an estimated curve. Typically, an optimal value of λ is determined by minimizing the value of an information criterion (see e.g., Hastie & Tibshirani, 1990; Green & Silverman, 1994; Hastie,

[1] Corresponding author, E-mail: *yanagi@math.sci.hiroshima-u.ac.jp*

Tibshirani & Friedman, 2001; Konishi & Kitagawa, 2008). Information criteria to optimize λ can be roughly divided into two types; firstly, those that measure the goodness of fit of a model by the predictive Kullback-Leibler discrepancy, i.e., AIC (Eilers & Marx, 1996), GIC (Konishi & Kitagawa, 1996), $\mathrm{AIC_C}$ (Hurvich, Simonoff & Tsai, 1998) and SPIC (Imoto & Konishi, 2003), and secondly, those that measure the goodness of fit of a model using the mean square error (MSE) of prediction, i.e., the C_p criterion (Mallows, 1973, 1995; Hastie & Tibshirani, 1990), the cross-validation (CV) criterion (Stone, 1974) and the generalized cross validation (GCV) criterion (Craven & Wahba, 1979). Generally, an information criterion based on the MSE of prediction is widely used for optimizing the smoothing parameter. It is well known that an explicit solution to the minimization problem for the information criterion cannot be obtained. Hence, we need to carry out an iterative procedure to search for the optimal smoothing parameter, i.e., we need to apply a grid search method. Since it is also necessary to optimize the number of basis functions by minimizing the information criterion, for each set of basis functions, we have to repeat the optimization processes for λ. Consequently, considerable computation is required until a final estimated curve is obtained. In order to avoid such an inconvenient optimization, we propose a non-iterative method for optimizing a smoothing parameter.

For generalized ridge regression as proposed by Hoerl and Kennard (1970), it is known that multiple ridge parameters minimizing the C_p criterion can be obtained as closed forms (see e.g., Lawless, 1981; Walker & Page, 2001). Thus, we expect that our objective can be achieved by applying the optimization method for generalized ridge regression to penalized spline regression. In penalized spline regression, unknown parameters are estimated by minimizing a penalized residual sum of squares (PRSS) having a roughness penalty which consists of a known penalty matrix $\boldsymbol{K}'\boldsymbol{K}$, where $\boldsymbol{K}$ is a row-full rank matrix. Unfortunately, the fact that $\boldsymbol{K}'\boldsymbol{K}$ is not an identity matrix prevents us from directly applying the optimization method for generalized ridge regression to optimization of the smoothing parameter. Therefore, we construct transformations of the unknown parameters and a matrix of basis functions by $\boldsymbol{K}_+^{-1}$ and $\boldsymbol{K}_+$, respectively, where $\boldsymbol{K}_+$ is a nonsingular matrix constructed from the singular-value decomposition of $\boldsymbol{K}$ and the identity matrix. Then a PRSS for a penalized spline regression is transformed into a PRSS for a partial ridge regression. By applying the optimization method for the generalized ridge regression to such a transformed model, we can obtain closed forms of optimal smoothing parameters minimizing the C_p criterion.

Multiple smoothing parameters are used in our new optimization method for smoothness. Although there are also several papers that have dealt with multiple smoothing parameters, e.g., Gu and Wahba (1991) and Wood (2000), an optimization method for multiple smoothing parameters turns out to be more difficult than one for a single smoothing parameter. Nevertheless, our optimization method is very simple because optimal values for the smoothing parameters are given by closed forms.

By using an asymptotic expansion of the information criterion, an approximate optimal smoothing parameter can also be derived as a closed form (e.g., Wand, 1999). However, it is well known that an approximate solution based on an asymptotic expansion departs from the true solution when the sample size is small or the number of basis functions is large. Moreover, there exists a serious problem in that the approximate optimal smoothing parameter does not become ∞ although the explicit solution may become ∞. In addition, since the information criterion is expanded around $\lambda = 0$, the approximate solution tends to a value near 0. We are able to avoid such disadvantages with our new optimized smoothing parameters, because we have solved the minimization problem for the information criterion exactly.

This paper is organized as follows: In Section 2, we clarify the relation between penalized spline regression and partial ridge regression. In Section 3, we propose a non-iterative optimization method for smoothing parameters minimizing the C_p criterion. Additionally, a C_p criterion for optimizing the number of basis functions is reconsidered. In Section 4, by conducting numerical simulations, we examine the accuracy of the estimated curve derived by our optimization method. Section 5 contains a discussion and our conclusion. Technical details are provided in the Appendix.

2. Correspondence with Partial Ridge Regression

Let $\{(x_i, y_i)|i = 1, \ldots, n\}$ be n observable data pairs, each consisting of an explanatory variable x and a response variable y. We consider the following non-linear regression model:

$$y_i = \mu(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mu(x)$ is an unknown smooth function. In this model, it is assumed that $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed according to a distribution with mean 0 and variance σ^2. Then, without loss of generality, we assume that $x_1 \leq \cdots \leq x_n$.

In penalized spline regression, $\mu(x)$ is expressed as a linear combination of m known basis functions $b_1(x), \ldots, b_m(x)$, i.e.,

$$\mu(x) = \sum_{j=1}^{m} \alpha_j b_j(x) = \boldsymbol{\alpha}'\boldsymbol{b}(x),$$

where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$ is an m-dimensional unknown parameter vector and $\boldsymbol{b}(x) = (b_1(x), \ldots, b_m(x))'$ is an m-dimensional known vector of basis functions.

Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$ be an n-dimensional vector of response variables and $\boldsymbol{B} = (\boldsymbol{b}(x_1), \ldots, \boldsymbol{b}(x_n))$ be an $n \times m$ matrix of basis function values. Then, the PRSS for the spline regression is defined as

$$\mathrm{RSS}(\boldsymbol{\alpha}|\lambda) = (\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha})'(\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\alpha}) + \lambda\boldsymbol{\alpha}'\boldsymbol{K}'\boldsymbol{K}\boldsymbol{\alpha}, \tag{2.2}$$

where λ is the smoothing parameter ($\lambda \geq 0$), and $\boldsymbol{K}'\boldsymbol{K}$ is a $m \times m$ known penalty matrix. Here, $\boldsymbol{K}$ is a $k \times m$ matrix of row-full rank k ($\leq m$). For example, for $\boldsymbol{K}$, the following ath-order difference matrix is commonly used:

$$(\boldsymbol{K})_{ij} = \begin{cases} (-1)^{j-i}\binom{a}{j-i}, & (i = 1, \ldots, m-a; j = i, \ldots, i+a) \\ 0, & (\text{otherwise}) \end{cases}, \tag{2.3}$$

where $\binom{a}{i}$ is the binomial coefficient, and $(\boldsymbol{A})_{ij}$ denotes the (i, j)th element of the matrix $\boldsymbol{A}$. The unknown parameter value $\boldsymbol{\alpha}$ is estimated by minimizing PRSS, i.e., the penalized least square estimator (PLSE) of $\boldsymbol{\alpha}$ is defined as

$$\hat{\boldsymbol{\alpha}}_\lambda = \arg\min_{\boldsymbol{\alpha}} \mathrm{RSS}(\boldsymbol{\alpha}|\lambda) = (\boldsymbol{B}'\boldsymbol{B} + \lambda\boldsymbol{K}'\boldsymbol{K})^{-1}\boldsymbol{B}'\boldsymbol{y}. \tag{2.4}$$

Then, the estimated curve $\hat{\mu}(x)$ is given by $\hat{\mu}(x) = \hat{\boldsymbol{\alpha}}_\lambda'\boldsymbol{b}(x)$.

Let us decompose $\boldsymbol{K}$ into $\boldsymbol{G}(\boldsymbol{L}, \boldsymbol{O}_{k,m-k})\boldsymbol{C}'$ by singular-value decomposition, where $\boldsymbol{L}$ is a $k \times k$ diagonal matrix, $\boldsymbol{O}_{k,m-k}$ is a $k \times (m-k)$ matrix of zeros, and $\boldsymbol{C}$ and $\boldsymbol{G}$ are $m \times m$ and $k \times k$ orthogonal matrices respectively, such that $\boldsymbol{K}'\boldsymbol{K}$ and $\boldsymbol{K}\boldsymbol{K}'$ become diagonal matrices, i.e., $\boldsymbol{C}'\boldsymbol{K}'\boldsymbol{K}\boldsymbol{C} = \mathrm{diag}(\boldsymbol{L}^2, \boldsymbol{O}_{m-k,m-k})$ and $\boldsymbol{G}'\boldsymbol{K}\boldsymbol{K}'\boldsymbol{G} = \boldsymbol{L}^2$,. Here, $\mathrm{diag}(\boldsymbol{A}_1, \boldsymbol{A}_2)$ denotes a block diagonal matrix when $\boldsymbol{A}_1$ and $\boldsymbol{A}_2$ are square matrices. We define the $m \times m$ nonsingular matrix $\boldsymbol{K}_+$ by $\boldsymbol{K}_+ = \boldsymbol{G}_+\boldsymbol{L}_+\boldsymbol{C}'$, where $\boldsymbol{L}_+ = \mathrm{diag}(\boldsymbol{L}, \boldsymbol{I}_{m-k})$ and $\boldsymbol{G}_+ = \mathrm{diag}(\boldsymbol{G}, \boldsymbol{I}_{m-k})$. It is easy to see that $\boldsymbol{K} = (\boldsymbol{I}_k, \boldsymbol{O}_{k,m-k})\boldsymbol{K}_+$. Hence, we obtain the relation

$$\boldsymbol{K}'\boldsymbol{K} = \boldsymbol{K}_+'\mathrm{diag}(\boldsymbol{I}_k, \boldsymbol{O}_{m-k,m-k})\boldsymbol{K}_+.$$

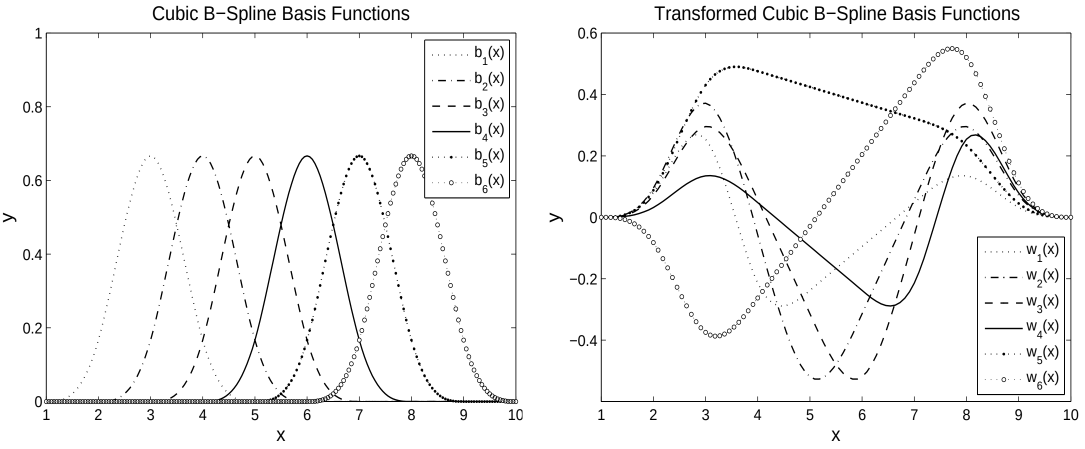

FIGURE 1. Cubic B-spline basis functions and these functions transformed by $\boldsymbol{K}_+^{-1}$

Then, we define an $n\times m$ matrix $\boldsymbol{W} = (\boldsymbol{W}_1, \boldsymbol{W}_2)$ by transforming $\boldsymbol{B}$ by $\boldsymbol{K}_+^{-1} = \boldsymbol{C}\boldsymbol{L}_+^{-1}\boldsymbol{G}'_+$, i.e., $\boldsymbol{W} = \boldsymbol{B}\boldsymbol{K}_+^{-1}$, where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are $n\times k$ and $n\times(m-k)$ matrices, respectively. This corresponds to the transformation of basis functions $\boldsymbol{b}(x)$ to $\boldsymbol{w}(x) = (w_1(x), \ldots, w_m(x))' = \boldsymbol{K}_+^{-1}\boldsymbol{b}(x)$. For instance, when we use cubic B-spline basis functions with $m = 6$ (knots $1, \ldots, 10$) and the second-order difference matrix in (2.3) as $\boldsymbol{K}$, basis functions are transformed as in Figure 1. Moreover, we can determine an m-dimensional known parameter vector $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2)'$ by transforming $\boldsymbol{\alpha}$ by $\boldsymbol{K}_+$, i.e., $\boldsymbol{\beta} = \boldsymbol{K}_+\boldsymbol{\alpha}$, where $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are k- and $(m-k)$-dimensional unknown parameter vectors, respectively. Notice that

$$\boldsymbol{B}\boldsymbol{\alpha} = \boldsymbol{B}\boldsymbol{K}_+^{-1}\boldsymbol{K}_+\boldsymbol{\alpha} = \boldsymbol{W}\boldsymbol{\beta},$$

and

$$\boldsymbol{\alpha}'\boldsymbol{K}'\boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{\alpha}'\boldsymbol{K}'_+\mathrm{diag}(\boldsymbol{I}_k, \boldsymbol{O}_{m-k,m-k})\boldsymbol{K}_+\boldsymbol{\alpha} = \boldsymbol{\beta}'_1\boldsymbol{\beta}_1.$$

Hence, we can rewrite $\mathrm{RSS}(\boldsymbol{\alpha}|\lambda)$ in (2.2) by $\boldsymbol{W}$ and $\boldsymbol{\beta}$ as

$$(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{W}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}'_1\boldsymbol{\beta}_1. \tag{2.5}$$

This result indicates that the PRSS for the penalized spline regression is equivalent to that for the partial ridge regression. Since $\boldsymbol{K}_+$ is a nonsingular matrix, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are in one-to-one correspondence. Therefore, we can obtain $\hat{\boldsymbol{\alpha}}_\lambda$ as $\hat{\boldsymbol{\alpha}}_\lambda = \boldsymbol{K}_+^{-1}\hat{\boldsymbol{\beta}}_\lambda$, where $\hat{\boldsymbol{\beta}}_\lambda$ is a minimizer of PRSS in (2.5). Moreover, an estimated curve can be derived by using $\hat{\boldsymbol{\beta}}_\lambda$ instead of $\hat{\boldsymbol{\alpha}}_\lambda$, because we have $\hat{\mu}(x) = \hat{\boldsymbol{\alpha}}'_\lambda\boldsymbol{b}(x) = \hat{\boldsymbol{\beta}}'_\lambda(\boldsymbol{K}_+^{-1})'\boldsymbol{K}'_+\boldsymbol{w}(x) = \hat{\boldsymbol{\beta}}'_\lambda\boldsymbol{w}(x)$.

Minimizing the PRSS for $\boldsymbol{\beta}$ in (2.5) yields

$$\hat{\boldsymbol{\beta}}_\lambda = \{\boldsymbol{W}'\boldsymbol{W} + \lambda \mathrm{diag}(\boldsymbol{I}_k, \boldsymbol{O}_{m-k,m-k})\}^{-1}\boldsymbol{W}'\boldsymbol{y}.$$

Let $\hat{\boldsymbol{\beta}}_\lambda$ be partitioned into $(\hat{\boldsymbol{\beta}}'_{\lambda,1}, \hat{\boldsymbol{\beta}}'_{\lambda,2})'$. Using a formula for inversion of a matrix in block form, $\hat{\boldsymbol{\beta}}_{\lambda,1}$ and $\hat{\boldsymbol{\beta}}_{\lambda,2}$ can be expressed as

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\lambda,1} &= \left\{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1 + \lambda\boldsymbol{I}_k\right\}^{-1} \boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{y}, \\
\hat{\boldsymbol{\beta}}_{\lambda,2} &= (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'(\boldsymbol{y} - \boldsymbol{W}_1\hat{\boldsymbol{\beta}}_{\lambda,1}) \qquad (2.6) \\
&= (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'\left[\boldsymbol{I}_n - \boldsymbol{W}_1\left\{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1 + \lambda\boldsymbol{I}_k\right\}^{-1} \boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\right]\boldsymbol{y},
\end{aligned}$$

where $\boldsymbol{P}_{\boldsymbol{W}}$ is the projection matrix to the subspace spanned by the columns of $\boldsymbol{W}$, i.e., $\boldsymbol{P}_{\boldsymbol{W}} = \boldsymbol{W}(\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'$. Then, the hat matrix $\boldsymbol{H}_\lambda$ becomes

$$\boldsymbol{H}_\lambda = \boldsymbol{P}_{\boldsymbol{W}_2} + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\left\{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1 + \lambda\boldsymbol{I}_k\right\}^{-1}\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}). \qquad (2.7)$$

Notice that $\lim_{\lambda\to\infty}\hat{\boldsymbol{\beta}}_{\lambda,1} = \boldsymbol{0}_k$ and $\lim_{\lambda\to\infty}\hat{\boldsymbol{\beta}}_{\lambda,2} = (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'\boldsymbol{y}$, where $\boldsymbol{0}_k$ is a k-dimensional vector of zeros. These limit values imply that $\mu(x)$ is estimated by multiple regression using only $\boldsymbol{W}_2$, i.e., that $\hat{\mu}(x) = \boldsymbol{y}'\boldsymbol{W}_2(\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{w}(x)$, when $\lambda \to \infty$. In Figure 1, $\boldsymbol{W}_2$ corresponds to the transformed basis functions $w_5(x)$ and $w_6(x)$. Notice that data exists from the fourth to the seventh knot. In the range where data exists, it seems that $w_5(x)$ and $w_6(x)$ are almost straight lines. Then, $\hat{\mu}(x) = \boldsymbol{y}'\boldsymbol{W}_2(\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{w}(x)$ becomes a straight line. Therefore, $\hat{\mu}(x)$ will approach a straight line when λ becomes large.

In order to guarantee the nonsingularity of $\boldsymbol{K}_+$, it was defined as $\boldsymbol{K}_+ = \boldsymbol{G}_+\boldsymbol{L}_+\boldsymbol{C}'$. On the other hand, for any $(m-k)\times(m-k)$ diagonal matrix $\boldsymbol{L}_0$ and orthogonal matrix $\boldsymbol{G}_0$, the following equation always holds:

$$\boldsymbol{K}'\boldsymbol{K} = \boldsymbol{C}\mathrm{diag}(\boldsymbol{L}, \boldsymbol{L}_0)\mathrm{diag}(\boldsymbol{G}', \boldsymbol{G}_0')\mathrm{diag}(\boldsymbol{I}_k, \boldsymbol{O}_{m-k,m-k})\mathrm{diag}(\boldsymbol{G}, \boldsymbol{G}_0)\mathrm{diag}(\boldsymbol{L}, \boldsymbol{L}_0)\boldsymbol{C}'.$$

This means that $\boldsymbol{K}_+$ cannot be determined uniquely. However, it seems that $\hat{\mu}(x)$ is invariant for any diagonal matrix $\boldsymbol{L}_0$ and orthogonal matrix $\boldsymbol{G}_0$, because $\boldsymbol{P}_{\boldsymbol{W}_2}$ is invariant for any diagonal matrix $\boldsymbol{L}_0$ and orthogonal matrix $\boldsymbol{G}_0$. Therefore in this paper, we take $\boldsymbol{L}_0$ and $\boldsymbol{G}_0$ to be identity matrices for simplicity.

3. New Smoothing Method

3.1. New PLSE

Let $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ be a $k \times k$ diagonal matrix of multiple smoothing parameters and $\boldsymbol{Q}$ be a $k \times k$ orthogonal matrix such that $\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1$ becomes a diagonal matrix, i.e., $\boldsymbol{Q}'\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\boldsymbol{Q} = \boldsymbol{D} = \mathrm{diag}(d_1, \ldots, d_k)$. By replacing $\lambda\boldsymbol{I}_k$ in (2.6) with $\boldsymbol{Q\Lambda Q}'$, the partial generalized ridge estimator of $\boldsymbol{\beta}$ is derived as

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda},1} &= \left\{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1 + \boldsymbol{Q\Lambda Q}'\right\}^{-1}\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{y}, \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda},2} &= (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'(\boldsymbol{y} - \boldsymbol{W}_1\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda},1}) \\ &= (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'\left[\boldsymbol{I}_n - \boldsymbol{W}_1\left\{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1 + \boldsymbol{Q\Lambda Q}'\right\}^{-1}\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\right]\boldsymbol{y}.\end{aligned} \tag{3.1}$$

Then, the hat matrix $\boldsymbol{H}_{\boldsymbol{\Lambda}}$ becomes

$$\boldsymbol{H}_{\boldsymbol{\Lambda}} = \boldsymbol{P}_{\boldsymbol{W}_2} + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\left\{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1 + \boldsymbol{Q\Lambda Q}'\right\}^{-1}\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}). \tag{3.2}$$

Let $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}} = (\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda},1}', \hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda},2}')'$. The estimated curve can be calculated by $\hat{\mu}(x) = \hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}'\boldsymbol{w}(x)$ instead of using the PLSE of $\boldsymbol{\alpha}$. On the other hand, the PLSE of $\boldsymbol{\alpha}$ can also be derived using $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$, i.e., $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}} = \boldsymbol{K}_+^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$. Notice that

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}} = \{\boldsymbol{W}'\boldsymbol{W} + \mathrm{diag}(\boldsymbol{Q\Lambda Q}', \boldsymbol{O}_{m-k,m-k})\}^{-1}\boldsymbol{W}'\boldsymbol{y}.$$

By using the formula for inversion of a matrix in block form and the relation $\boldsymbol{K} = (\boldsymbol{I}_k, \boldsymbol{O}_{k,m-k})\boldsymbol{K}_+$, after substituting $\boldsymbol{W} = \boldsymbol{B}\boldsymbol{K}_+^{-1}$ into the above $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$, the following relation is obtained:

$$\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}} = \boldsymbol{K}_+^{-1}\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}} = (\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{K}'\boldsymbol{Q\Lambda Q}'\boldsymbol{K})^{-1}\boldsymbol{B}'\boldsymbol{y}. \tag{3.3}$$

This PLSE $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}}$ in (3.3) is equivalent to the minimizer of the following PRSS for $\boldsymbol{\alpha}$:

$$(\boldsymbol{y} - \boldsymbol{B\alpha})'(\boldsymbol{y} - \boldsymbol{B\alpha}) + \boldsymbol{\alpha}'\boldsymbol{K}'\boldsymbol{Q\Lambda Q}'\boldsymbol{K\alpha}.$$

The value $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}}$ with $\boldsymbol{\Lambda} = \lambda\boldsymbol{I}_k$ directly corresponds with $\hat{\boldsymbol{\alpha}}_\lambda$ in (2.4). Therefore, we can view $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}}$ as an extended version of the ordinary PLSE of $\boldsymbol{\alpha}$.

3.2. C_p Criterion for Selecting Smoothing Parameters

Notice that $\boldsymbol{y} - \boldsymbol{W}\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}} = (\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{\Lambda}})\boldsymbol{y}$. By using this result and $\boldsymbol{H}_{\boldsymbol{\Lambda}}$, the C_p criterion for selecting $\boldsymbol{\Lambda}$ in $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$ can be defined by

$$C_p(\boldsymbol{\Lambda}|m) = \frac{1}{\hat{\sigma}^2}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{\Lambda}})^2\boldsymbol{y} + 2\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\Lambda}}), \tag{3.4}$$

where $\hat{\sigma}^2$ is an estimator of σ^2. Since equations $\boldsymbol{y} - \boldsymbol{B}\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}} = (\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{\Lambda}})\boldsymbol{y}$ and $\mathrm{tr}\{(\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{K}'\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}'\boldsymbol{K})^{-1}\boldsymbol{B}'\boldsymbol{B}\} = \mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\Lambda}})$ are satisfied, the C_p criterion for selecting $\boldsymbol{\Lambda}$ in $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}}$ also becomes the equation (3.4). For the C_p criterion, a consistent and more higher-order asymptotically unbiased estimator of σ^2 is appropriate for $\hat{\sigma}^2$. If the model which is used to estimate σ^2 includes the true model, consistency and unbiasedness of an estimator of σ^2 can be guaranteed. Thus, for the original C_p criterion, the full model is used to estimate σ^2, because the probability that the true model is included is greatest for the full model. However, a well-defined full model for the spline regression does not exist because we can freely determine the maximum number of basis functions. Hence, in the same way as in Yanagihara and Ohtaki (2004), we use a nonparametric estimator of σ^2 proposed by Gasser, Sroka and Jennen-Steinmetz (1986).

Suppose that $x_1 < \cdots < x_n$. Let $\boldsymbol{R}$ be an $n \times (n-2)$ matrix whose (i,j)th element is defined by

$$(\boldsymbol{R})_{ij} = \begin{cases} (x_{j+2} - x_{j+1})/(x_{j+2} - x_j) & (i = j; j = 1, \ldots, n-2) \\ -1 & (i = j+1; j = 1, \ldots, n-2) \\ (x_{j+1} - x_j)/(x_{j+2} - x_j) & (i = j+2; j = 1, \ldots, n-2) \\ 0 & (\text{otherwise}) \end{cases}, \tag{3.5}$$

and $\boldsymbol{S}$ be an $(n-2) \times (n-2)$ diagonal matrix whose jth diagonal element is equal to the jth diagonal element of $\boldsymbol{R}'\boldsymbol{R}$. Then, a nonparametric estimator of σ^2 is defined by

$$\hat{\sigma}^2 = \frac{1}{n-2}\boldsymbol{y}'\boldsymbol{R}\boldsymbol{S}^{-1}\boldsymbol{R}'\boldsymbol{y}. \tag{3.6}$$

Suppose that $|x_i - x_{i-1}| = O(n^{-1})$. From Gasser, Sroka and Jennen-Steinmetz (1986), we can see that $\hat{\sigma}^2$ is a consistent estimator of σ^2, and $E[\hat{\sigma}^2] = \sigma^2 + O(n^{-4})$ holds, when $\mu(x)$ is twice continuously differentiable and the fourth moment of ε_i exists. If $\mu(x)$ is once differentiable, the bias of $\hat{\sigma}^2$ becomes $O(n^{-2})$. If there are multiple measurements in $x_1, \ldots, x_n$, we have to modify the definition of $\boldsymbol{R}$ in (3.5). For such a modification, see Gasser, Sroka and Jennen-Steinmetz (1986). Although there are other nonparametric estimators of σ^2, e.g., Buckley, Eagleson and Silverman (1988), Hall, Kay and Titterington (1990), and Seifert, Gasser and Wolf (1993), we use (3.6) as an estimator of σ^2, because this provides the simplest nonparametric estimator of σ^2.

3.3. Optimization for Smoothing Parameters

Let $\boldsymbol{z} = (z_1, \ldots, z_k)'$ be a k-dimensional vector defined by

$$\boldsymbol{z} = \frac{1}{\hat{\sigma}} \boldsymbol{D}^{-1/2} \boldsymbol{Q}' \boldsymbol{W}_1' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}) \boldsymbol{y}. \tag{3.7}$$

From the Appendix A.1, it appears that $C_p(\boldsymbol{\Lambda}|m)$ in (3.4) can be rewritten as

$$C_p(\boldsymbol{\Lambda}|m) = \frac{1}{\hat{\sigma}^2} \boldsymbol{y}' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}) \boldsymbol{y} + 2(m-k) + 2 \sum_{j=1}^{k} f(\lambda_j | d_j, z_j^2), \tag{3.8}$$

where the function $f(\lambda_j | d_j, z_j^2)$ is given by

$$f(\lambda_j | d_j, z_j^2) = \frac{d_j (1 - z_j^2)}{d_j + \lambda_j} + \frac{d_j^2 z_j^2}{2(d_j + \lambda_j)^2}.$$

Notice that

$$\frac{\partial}{\partial \lambda_j} C_p(\boldsymbol{\Lambda}|m) = 2 \frac{\partial}{\partial \lambda_j} f(\lambda_j | d_j, z_j^2) = \frac{2 d_j \{(z_j^2 - 1)\lambda_j - d_j\}}{(d_j + \lambda_j)^3}.$$

Since $\lambda_j \geq 0$ and $d_j > 0$, the function $f(\lambda_j | d_j, z_j^2)$ becomes a minimum at $\lambda_j = d_j/(z_j^2 - 1)$ when $z_j^2 - 1 > 0$ and at $\lambda_j = \infty$ when $z_j^2 - 1 \leq 0$. Accordingly, the optimal λ_j minimizing $C_p(\boldsymbol{\Lambda}|m)$ is given by

$$\hat{\lambda}_j = \begin{cases} d_j/(z_j^2 - 1) & (z_j^2 > 1) \\ \infty & (z_j^2 \leq 1) \end{cases}. \tag{3.9}$$

Let $\boldsymbol{V}$ be a $k \times k$ diagonal matrix defined by

$$\boldsymbol{V} = \mathrm{diag}(v_1, \ldots, v_k), \;\; v_j = I(z_j^2 > 1) \left(1 - \frac{1}{z_j^2} \right), \;\; (j = 1, \ldots, k), \tag{3.10}$$

where $I(z_j^2 > 1)$ is an indicator function, i.e., $I(z_j^2 > 1) = 1$ if $z_j^2 > 1$ and $I(z_j^2 > 1) = 0$ if $z_j^2 \leq 1$. By using $\boldsymbol{V}$ and the relation

$$\boldsymbol{W}_1' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}) \boldsymbol{W}_1 + \boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}' = \boldsymbol{Q} (\boldsymbol{D} + \Lambda) \boldsymbol{Q}', \tag{3.11}$$

we have

$$\left\{ \boldsymbol{W}_1' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}) \boldsymbol{W}_1 + \boldsymbol{Q} \hat{\boldsymbol{\Lambda}} \boldsymbol{Q}' \right\}^{-1} = \boldsymbol{Q} \boldsymbol{V} \boldsymbol{D}^{-1} \boldsymbol{Q}',$$

where $\hat{\boldsymbol{\Lambda}} = \mathrm{diag}(\hat{\lambda}_1, \ldots, \hat{\lambda}_k)$. Therefore, after optimization of $\boldsymbol{\Lambda}$, the PLSE $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$ in (3.12) becomes

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}},1} &= \boldsymbol{Q} \boldsymbol{V} \boldsymbol{D}^{-1} \boldsymbol{Q}' \boldsymbol{W}_1' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}) \boldsymbol{y}, \\ \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}},2} &= (\boldsymbol{W}_2' \boldsymbol{W}_2)^{-1} \boldsymbol{W}_2' (\boldsymbol{y} - \boldsymbol{W}_1 \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}},1}) \\ &= (\boldsymbol{W}_2' \boldsymbol{W}_2)^{-1} \boldsymbol{W}_2' \left\{ \boldsymbol{I}_n - \boldsymbol{W}_1 \boldsymbol{Q} \boldsymbol{V} \boldsymbol{D}^{-1} \boldsymbol{Q}' \boldsymbol{W}_1' (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}) \right\} \boldsymbol{y}. \end{aligned} \tag{3.12}$$

The hat matrix after optimizing $\boldsymbol{\Lambda}$ is given by

$$\boldsymbol{H}_{\hat{\boldsymbol{\Lambda}}} = \boldsymbol{P}_{\boldsymbol{W}_2} + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\boldsymbol{Q}\boldsymbol{V}\boldsymbol{D}^{-1}\boldsymbol{Q}'\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}). \tag{3.13}$$

Let $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}}} = (\hat{\boldsymbol{\beta}}'_{1,\hat{\boldsymbol{\Lambda}}}, \hat{\boldsymbol{\beta}}'_{2,\hat{\boldsymbol{\Lambda}}})'$. We can calculate the PLSE of $\boldsymbol{\alpha}$ after optimizing $\boldsymbol{\Lambda}$ by $\hat{\boldsymbol{\alpha}}_{\hat{\boldsymbol{\Lambda}}} = \boldsymbol{K}_+^{-1}\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}}}$. However, the estimated curve $\hat{\mu}(x)$ after optimizing $\boldsymbol{\Lambda}$ is obtained by $\hat{\boldsymbol{\beta}}'_{\hat{\boldsymbol{\Lambda}}}\boldsymbol{w}(x)$ instead of using $\hat{\boldsymbol{\alpha}}'_{\hat{\boldsymbol{\Lambda}}}\boldsymbol{b}(x)$. Needless to say, $\hat{\boldsymbol{\beta}}'_{\hat{\boldsymbol{\Lambda}}}\boldsymbol{w}(x)$ is equivalent to $\hat{\boldsymbol{\alpha}}'_{\hat{\boldsymbol{\Lambda}}}\boldsymbol{b}(x)$.

3.4. Statistical Implications for the PLSE after Optimizing $\boldsymbol{\Lambda}$

Let $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_1', \hat{\boldsymbol{\beta}}_2')'$ be the ordinary least square estimator (LSE) of $\boldsymbol{\beta}$ given by

$$\begin{aligned}\hat{\boldsymbol{\beta}}_1 &= \{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\}^{-1}\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{y}, \\ \hat{\boldsymbol{\beta}}_2 &= (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'(\boldsymbol{y} - \boldsymbol{W}_1\hat{\boldsymbol{\beta}}_1) \\ &= (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'\left[\boldsymbol{I}_n - \boldsymbol{W}_1\{\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\}^{-1}\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\right]\boldsymbol{y}.\end{aligned} \tag{3.14}$$

Then, the optimized PLSE of $\boldsymbol{\beta}$ in (3.12) can be rewritten as

$$\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}},1} = \boldsymbol{Q}\boldsymbol{V}\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_1, \quad \hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}},2} = \hat{\boldsymbol{\beta}}_2 + (\boldsymbol{W}_2'\boldsymbol{W}_2)^{-1}\boldsymbol{W}_2'\boldsymbol{W}_1(\boldsymbol{I}_k - \boldsymbol{Q}\boldsymbol{V}\boldsymbol{Q}')\hat{\boldsymbol{\beta}}_1.$$

From the above results, we can see that $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}},1}$ is a shrinkage estimator of the ordinary LSE of $\boldsymbol{\beta}_1$, because $0 \le v_j \le 1$ is satisfied. A pseudospline proposed by Hastie (1996) is the same type of shrinkage estimator as $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}},1}$. For the pseudospline, the number of effective degrees of freedom determines whether or not v_j should be 0. With our method, the size of v_j is decided by z_j^2. We note that z_j^2 is a test statistics for $H_{0,j} : \gamma_j = 0$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_k)' = \boldsymbol{Q}'\boldsymbol{\beta}_1$. Since the null distribution of z_j is asymptotically distributed according to $N(0,1)$, we have $P(z_j^2 > 1) \approx 0.32$ under the assumption that the null hypothesis $H_{0,j}$ is true. Hence, if the null hypothesis $H_{0,j}$ is accepted at about the 0.32 level, v_j becomes 0. On the contrary, if the null hypothesis $H_{0,j}$ is rejected at about the 0.32 level, v_j becomes large as z_j^2 increases, and eventually approaches 1. This means that the shrinkage ratio is scaled by the size of z_j^2 when the null hypothesis is significant at about the 0.32 level.

3.5. The C_p Criterion for Selecting the Number of Basis Functions

For the ordinary optimization method, information criteria for selecting the smoothing parameter and the number of basis functions are generally used. Hence we will use the following information criterion for selecting m:

$$C_p(m) = C_p(\hat{\mathbf{\Lambda}}|m) = \frac{1}{\hat{\sigma}^2}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H}_{\hat{\mathbf{\Lambda}}})^2\boldsymbol{y} + 2\left\{m - k + \mathrm{tr}(\boldsymbol{V})\right\}. \tag{3.15}$$

Since $0 \leq \mathrm{tr}(\boldsymbol{V}) < k$, $-k \leq -k + \mathrm{tr}(\boldsymbol{V}) < 0$ is satisfied. Hence, $C_p(m)$ in (3.15) will tend to be small as m is increasing. This means that the penalty for increasing the number of smoothing parameters may not be evaluated using (3.15), because k increases as m increases.

The simplest technique to avoid such an under-evaluation problem is to remove $-k + \mathrm{tr}(\boldsymbol{V})$ from the C_p criterion for selecting m. Thus, we define the following C_p criterion:

$$C_p^{\#}(m) = \frac{1}{\hat{\sigma}^2}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H}_{\hat{\mathbf{\Lambda}}})^2\boldsymbol{y} + 2m. \tag{3.16}$$

Let $\boldsymbol{y}_+ = (\boldsymbol{y}', \boldsymbol{y}_0')'$ and $\boldsymbol{W}_+ = (\boldsymbol{W}', \boldsymbol{W}_0')'$, where $\boldsymbol{y}_0$ and $\boldsymbol{W}_0$ are observations defined by convention as $\boldsymbol{y}_0 = \boldsymbol{0}_k$ and $\boldsymbol{W}_0 = (\hat{\mathbf{\Lambda}}^{1/2}\boldsymbol{Q}', \boldsymbol{O}_{k,m-k})$. Then, we can derive $\hat{\boldsymbol{\beta}}_{\hat{\mathbf{\Lambda}}}$ in (3.12) as the ordinary LSE, i.e., $\hat{\boldsymbol{\beta}}_{\hat{\mathbf{\Lambda}}} = (\boldsymbol{W}_+'\boldsymbol{W}_+)^{-1}\boldsymbol{W}_+'\boldsymbol{y}_+$. The C_p criterion based on $\boldsymbol{y}_+$ and $\boldsymbol{W}_+$ becomes

$$\frac{1}{\hat{\sigma}^2}\boldsymbol{y}_+'(\boldsymbol{I}_{n+k} - \boldsymbol{P}_{\boldsymbol{W}_+})\boldsymbol{y}_+ + 2m. \tag{3.17}$$

Notice that

$$\boldsymbol{y}_+'(\boldsymbol{I}_{n+k} - \boldsymbol{P}_{\boldsymbol{W}_+})\boldsymbol{y}_+ = \boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H}_{\hat{\mathbf{\Lambda}}})^2\boldsymbol{y} + (\boldsymbol{y}_0 - \hat{\mathbf{\Lambda}}^{1/2}\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_{\hat{\mathbf{\Lambda}},1})'(\boldsymbol{y}_0 - \hat{\mathbf{\Lambda}}^{1/2}\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_{\hat{\mathbf{\Lambda}},1}).$$

The second term on the left hand side of this equation measures the discrepancy between $\boldsymbol{y}_0$ and its fitted value. However, it is meaningless to measure this discrepancy, because $\boldsymbol{y}_0$ is not an actual value. Since $\boldsymbol{y}_0$ is a vector of zeros, the second term should be removed from the discrepancy. The $C_p^{\#}(m)$ in (3.16) corresponds with the C_p criterion in (3.17), when we omit the term $(\boldsymbol{y}_0 - \hat{\mathbf{\Lambda}}^{1/2}\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_{\hat{\mathbf{\Lambda}},1})'(\boldsymbol{y}_0 - \hat{\mathbf{\Lambda}}^{1/2}\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_{\hat{\mathbf{\Lambda}},1})$.

On the other hand, the MSE of prediction is defined by

$$\frac{1}{\sigma^2}E_{\boldsymbol{y}}E_{\boldsymbol{u}}\left[(\boldsymbol{u} - \hat{\boldsymbol{\mu}})'(\boldsymbol{u} - \hat{\boldsymbol{\mu}})\right], \tag{3.18}$$

where $\boldsymbol{u}$ is an n-dimensional random vector which is independent of $\boldsymbol{y}$ and has the same distribution as $\boldsymbol{y}$, and $\hat{\boldsymbol{\mu}} = \boldsymbol{W}\hat{\boldsymbol{\beta}}_{\hat{\mathbf{\Lambda}}}$. The C_p criterion is defined as an estimator of (3.18). Let $\boldsymbol{a}_j$ be the jth row vector of $\boldsymbol{Q}'\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})$, i.e.,

$$\boldsymbol{Q}'\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}) = (\boldsymbol{a}_1, \ldots, \boldsymbol{a}_k)'. \tag{3.19}$$

Suppose that $\varepsilon_1, \ldots, \varepsilon_n \sim i.i.d.\ N(0, \sigma^2)$. Then, from the Appendix A.2, the C_p criterion as an estimator of (3.18) is defined by

$$C_p^*(m) = C_p(m) + 4 \sum_{j=1}^{k} I(z_j^2 > 1) \left\{ \frac{1}{z_j^2} - \frac{\boldsymbol{a}_j' \boldsymbol{R} \boldsymbol{S}^{-1} \boldsymbol{R}' \boldsymbol{y}}{(n-2) \boldsymbol{a}_j' \boldsymbol{y}} \right\}. \tag{3.20}$$

4. Numerical Studies

4.1. Real-Data Examples

In this subsection, we give demonstrations of scedastic smoothing for real-data examples. Two well known real datasets, i.e., the motorcycle dataset (Härdle, 1990) and the LIDAR dataset (Holst *et al.*, 1996), were used for demonstrations. The scedastic smoothings were carried out using the following three methods based on our proposed smoothing algorithm:

Method 1 (C_p): $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$ is used after optimizing smoothing parameters $\boldsymbol{\Lambda}$ and the number of basis functions by minimizing $C_p(\boldsymbol{\Lambda}|m)$ in (3.4) and $C_p(m)$ in (3.15), respectively.

Method 2 ($C_p^{\#}$): $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$ is used after optimizing smoothing parameters $\boldsymbol{\Lambda}$ and the number of basis functions by minimizing $C_p(\boldsymbol{\Lambda}|m)$ in (3.4) and $C_p^{\#}(m)$ in (3.16), respectively.

Method 3 (C_p^*): $\hat{\boldsymbol{\beta}}_{\boldsymbol{\Lambda}}$ is used after optimizing smoothing parameters $\boldsymbol{\Lambda}$ and the number of basis functions by minimizing $C_p(\boldsymbol{\Lambda}|m)$ in (3.4) and $C_p^*(m)$ in (3.20), respectively.

These three methods differ in the information criterion used for selecting the number of basis functions. In order to compare our new method with the ordinary method, we also performed scedastic smoothing using the following two methods:

Method 4 (CV): $\hat{\boldsymbol{\alpha}}_\lambda$ in (2.4) is used after optimizing the smoothing parameter λ and the number of basis functions by minimizing the usual CV criterion.

Method 5 (GCV): $\hat{\boldsymbol{\alpha}}_\lambda$ in (2.4) is used after optimizing the smoothing parameter λ and the number of basis functions by minimizing the usual GCV criterion.

Since the optimal λ minimizing the CV and GCV criteria under fixed m cannot be given as a closed form, we used Matlab's minimization function 'fminsearch' which set the initial value to 1 for searching for the optimal λ under fixed m.

Figures 2 and 3 show estimated curves optimized by each method. The result of the motorcycle data is given in Figure 2 and the result of the LIDAR data is given in Figure 3. We used cubic B-spline basis functions and the second-order difference matrix in (2.3) as $\boldsymbol{B}$ and $\boldsymbol{K}$, respectively. Moreover, an equidistant arrangement (see e.g., Yanagihara & Ohtaki, 2003) was used for knot-placement, i.e., $t_j = x_1 + (j-4)(x_n - x_1)/(m-3)$ $(j = 1, \ldots, m+4)$. The optimal number of basis functions was searched for in the range $m = 4, \ldots, 30$. In both figures, the lower right part shows the scatter plot of the values of the information criterion into which the optimized $\boldsymbol{\Lambda}$ or λ is substituted and the number of basis functions. The optimal number of basis functions chosen by each information criterion is also shown in the legend of the figure.

From both figures, we can see that our new optimization method will perform well if the number of basis functions can be optimized appropriately. $C_p(m)$ chose a larger number of basis functions as optimal than $C_p^{\#}(m)$ and $C_p^{*}(m)$. This result causes overfitting of the estimated curves obtained from Method 1. Hence, we can conclude that $C_p(m)$ is not suitable for choosing the number of basis functions. $C_p^{*}(m)$ chose a larger number of basis functions as optimal than did $C_p^{\#}(m)$. $C_p^{\#}(m)$ chose the same number of basis functions as resulted from the CV and GCV criteria in Methods 4 and 5. By comparing the estimated curve obtained by Method 2 with those obtained by Methods 4 and 5, it seems that our new smoothing method tends to choose slightly more flexible curves as optimal, compared to the ordinary smoothing method.

Please insert Figures 2 and 3 around here

4.2. Simulation Examinations

In this subsection, we study the performance of our new smoothing method by examining some simulations. Simulation data were generated from $y_i = \mu(x_i) + \sigma\delta_i$ $(i = 1, \ldots, n)$ with $n = 20, 50, 100$ and $\sigma = 0.5, 1.0, 2.0$. The following four functions were used as the true trend $\mu(x)$ (the shape of each trend is showed in Figure 4).

Trend 1 (Hastie, Tibshirani & Friedman, 2001): $\mu(x) = \dfrac{\sin\{12(x+0.2)\}}{x+0.2}$.

Trend 2 (partial linear trend): $\mu(x) = \begin{cases} -60(x-17/60)^2 + 16/15 & (x < 1/4) \\ 4x & (1/4 \le x < 3/4) \\ 80(x-29/40)^2 + 59/20 & (3/4 \le x) \end{cases}$.

Trend 3 (linear trend): $\mu(x) = 6x$.

Trend 4 (Wand, 2000): $\mu(x) = 8\{1.5\phi((x-0.35)/0.15) - \phi((x-0.8)/0.04)\}$, where $\phi(x)$ is the probability density function of the standard normal distribution.

Explanatory variables $x_1, \ldots, n$ were generated independently from the uniform distribution $U(0,1)$. Error variables $\delta_1, \ldots, \delta_n$ were generated independently by the following three models (κ_3 and κ_4 denote the skewness and kurtosis of the distribution, respectively):

Distribution 1 (Normal Distribution): $\delta_i \sim N(0,1)$ ($\kappa_3 = 0$ and $\kappa_4 = 0$).

Distribution 2 (Laplace Distribution): δ_i is generated from a Laplace distribution with mean 0 and standard deviation 1 ($\kappa_3 = 0$ and $\kappa_4 = 4.5$).

Distribution 3 (Skew-Laplace Distribution): δ_i is generated from a skew Laplace distribution with location parameter 0, dispersion parameter 1 and skew parameter 1 standardized to mean 3/4 and standard deviation $\sqrt{23}/4$ ($\kappa_3 \approx 1.06$ and $\kappa_4 \approx 4.88$).

The skew-Laplace distribution was proposed by Balakrishnan and Ambagaspitiya (1994) (for the probability density function, see e.g., Yanagihara & Yuan, 2005). By comparing with results derived from distribution 1 and the others, we can study the influence of non-normality in the smoothing method.

Please insert Figure 4 around here

We applied five smoothing methods in the previous subsection to the simulation data. Settings of $\boldsymbol{B}$, $\boldsymbol{K}$ and the knot-placement were the same as those in the previous subsection. The optimal number of basis functions was searched for in the ranges $m = 4, \ldots, 6$, $m = 4, \ldots, 16$ and $m = 4, \ldots, 26$ when $n = 20, 50, 100$, respectively. Then, we calculated the MSE of an estimated curve as

$$\frac{1}{100\sigma^2} \sum_{j=1}^{100} E\left[\{\hat{\mu}(\tau_j) - \mu(\tau_j)\}^2\right],$$

where $\tau_j = x_1 + (x_n - x_1)(j-1)/99$ ($j = 1, \ldots, 100$). The above MSE was evaluated by Monte Carlo simulation with 1,000 iterations.

Tables 1, 2, 3 and 4 show the MSE in the case of trends 1, 2, 3 and 4, respectively. From the tables, we can see that the MSE of our method becomes smaller than that of the ordinary method when the true trend is flexible. In particular, when the sample size was small, the difference appeared to be considerable. When the true trend is linear, the ordinary method was superior over our method. However, Method 2 was not so bad compared with the ordinary method, even when the true trend is linear. When the true trend is 2, the MSE of the ordinary method became large. The reason is that the ordinary method controls the roughness of estimated curve by the single smoothing parameter. Hence, the ordinary method brought the estimated curve to the straight line. Especially the MSE of Method 4 became very large. The CV criterion tends to choose the under-fitting model as the best model. This property brought the estimated curve close to the straight line further. When the true trend is 4, there was no difference among the MSE for each method so much. On average, Method 2 was better than the other methods from the viewpoint of the MSE. Moreover, we were not able to find a clear tendency for the influence of non-normality.

Please insert Tables 1, 2, 3 and 4 around here

5. Conclusion and Discussion

In this paper, we have proposed a non-iterative optimization method for smoothness in penalized spline regression. By clarifying the relation between the PRSS of penalized spline regression and the PRSS of partial ridge regression, the optimization method for generalized ridge regression was able to be applied to optimization of the penalized spline regression. After that, we can solve for explicit solutions to the minimization problem of the C_p criterion. Using the explicit solutions as optimal smoothing parameters yields a non-iterative optimization method. Additionally, we studied the C_p criterion for selecting the number of basis functions. From simulations, we verified that our smoothing method for selecting the number of basis functions by $C_p^{\#}(m)$ worked well.

From results with numerical studies, we found that $\hat{\mu}(x)$ consisting of $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}}}$ becomes flexible. On the other hand, $C_p^{\#}(m)$ tended to choose a smaller number of basis functions as optimal. Hence, it seems that a smoothing result based on $\hat{\boldsymbol{\beta}}_{\hat{\boldsymbol{\Lambda}}}$ will become appropriate if $C_p^{\#}(m)$ is used for selecting the number of basis functions.

Essentially, $\boldsymbol{G}_+$ in $\boldsymbol{K}_+$ is not necessary for our new smoothing method. The orthogonal matrix $\boldsymbol{G}_+$ is needed to satisfy only the second equality in equation (3.3). Hence, we can simplify our new method by removing $\boldsymbol{G}_+$ from the definition of $\boldsymbol{K}_+$ if we do not need to stick to the specific form of $\hat{\boldsymbol{\alpha}}_{\boldsymbol{\Lambda}}$ as in (3.3). Even if we omit $\boldsymbol{G}_+$, the smoothing result does not change.

In the simulations, we used cubic B-spline basis functions and the second-order difference matrix in (2.3) as the basis function and the penalty matrix, respectively. Of course, our method can be applied to other sets of basis functions and penalty matrices, e.g., Green and Silverman (1994), Eilers and Marx (1996) and Wand (1999). Moreover, extensive computation is not required, our method will be more suitable to apply to a model for which optimization requires many calculations, e.g., the generalized additive model (Hastie & Tibshirani, 1990), the mixed model (Wand, 2003) and the multivariate adaptive spline model (Zhang, 1997; 2004). Our smoothing method may therefore be useful in many situations.

Appendix

A.1. Derivation of the Equation (3.8)

From equations (3.2) and (3.11), we can see that

$$\begin{aligned}\boldsymbol{H}_{\boldsymbol{\Lambda}} &= \boldsymbol{P}_{\boldsymbol{W}_2} + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\boldsymbol{Q}(\boldsymbol{D}+\Lambda)^{-1}\boldsymbol{Q}'\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}),\\ \boldsymbol{H}_{\boldsymbol{\Lambda}}^2 &= \boldsymbol{P}_{\boldsymbol{W}_2} + (\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\boldsymbol{Q}(\boldsymbol{D}+\Lambda)^{-1}\boldsymbol{D}(\boldsymbol{D}+\Lambda)^{-1}\boldsymbol{Q}'\boldsymbol{W}_1'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2}).\end{aligned}$$

Hence, $\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\Lambda}})$ can be calculated as

$$\mathrm{tr}(\boldsymbol{H}_{\boldsymbol{\Lambda}}) = m - k + \sum_{j=1}^{k}\frac{d_j}{d_j+\lambda_j}. \tag{A.1}$$

By using $\boldsymbol{z}$ given by (3.7), the following equations are derived:

$$\begin{aligned}\frac{1}{\hat{\sigma}^2}\boldsymbol{y}'\boldsymbol{H}_{\boldsymbol{\Lambda}}\boldsymbol{y} &= \frac{1}{\hat{\sigma}^2}\boldsymbol{y}'\boldsymbol{P}_{\boldsymbol{W}_2}\boldsymbol{y} + \boldsymbol{z}'\boldsymbol{D}^{1/2}(\boldsymbol{D}+\Lambda)^{-1}\boldsymbol{D}^{1/2}\boldsymbol{z},\\ \frac{1}{\hat{\sigma}^2}\boldsymbol{y}'\boldsymbol{H}_{\boldsymbol{\Lambda}}^2\boldsymbol{y} &= \frac{1}{\hat{\sigma}^2}\boldsymbol{y}'\boldsymbol{P}_{\boldsymbol{W}_2}\boldsymbol{y} + \boldsymbol{z}'\boldsymbol{D}^{1/2}(\boldsymbol{D}+\Lambda)^{-1}\boldsymbol{D}(\boldsymbol{D}+\Lambda)^{-1}\boldsymbol{D}^{1/2}\boldsymbol{z}.\end{aligned}$$

These equations imply that

$$\frac{1}{\hat{\sigma}^2}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{H}_{\boldsymbol{\Lambda}})^2\boldsymbol{y} = \frac{1}{\hat{\sigma}^2}\boldsymbol{y}'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{y} - 2\sum_{j=1}^{k}\frac{d_j z_j^2}{d_j+\lambda_j} + \sum_{j=1}^{k}\frac{d_j^2 z_j^2}{(d_j+\lambda_j)^2}. \tag{A.2}$$

Consequently, substituting equations (A.1) and (A.2) into equation (3.4) yields equation (3.8).

A.2. Derivation of the Equation (3.20)

From Efron (2004), we can see the MSE of prediction in (3.18) can be rewritten as

$$\frac{1}{\sigma^2}\left\{E_{\boldsymbol{y}}\left[\boldsymbol{y}'(\boldsymbol{I}_n-\boldsymbol{H}_{\hat{\boldsymbol{\Lambda}}})^2\boldsymbol{y}\right]+2E_{\boldsymbol{y}}\left[(\boldsymbol{y}-\boldsymbol{\mu})'(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right]\right\},$$

where $\boldsymbol{\mu}=(\mu(x_1),\ldots,\mu(x_n))'$. Under the assumption that $\varepsilon_1,\ldots,\varepsilon_n \sim i.i.d.\ N(0,\sigma^2)$, we obtain the following equation from Efron (2004):

$$\frac{1}{\sigma^2}E_{\boldsymbol{y}}\left[(\boldsymbol{y}-\boldsymbol{\mu})'(\hat{\boldsymbol{\mu}}-\boldsymbol{\mu})\right]=\sum_{i=1}^{n}E_{\boldsymbol{y}}\left[\frac{\partial\hat{\mu}(x_i)}{\partial y_i}\right].$$

Hence, we can define a new C_p criterion for selecting m as

$$C_p^*(m)=\frac{1}{\hat{\sigma}^2}\boldsymbol{y}'(\boldsymbol{I}_n-\boldsymbol{H}_{\hat{\boldsymbol{\Lambda}}})^2\boldsymbol{y}+2\sum_{i=1}^{n}\frac{\partial\hat{\mu}(x_i)}{\partial y_i}. \tag{A.3}$$

Let $\dot{\boldsymbol{V}}_i$ be the $k\times k$ matrix defined by $\dot{\boldsymbol{V}}_i=\partial\boldsymbol{V}/\partial y_i$ $(i=1,\ldots,n)$, where $\boldsymbol{V}$ is given by (3.10). Using $\dot{\boldsymbol{V}}_i$, we have

$$\sum_{i=1}^{n}\frac{\partial\hat{\mu}(x_i)}{\partial y_i}=m-k+\mathrm{tr}(\boldsymbol{V})+\sum_{i=1}^{n}\boldsymbol{e}_i'(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\boldsymbol{Q}\dot{\boldsymbol{V}}_i\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_1, \tag{A.4}$$

where $\hat{\boldsymbol{\beta}}_1$ is the ordinary LSE of $\boldsymbol{\beta}_1$, which is given by (3.14), and $\boldsymbol{e}_i$ is an n-dimensional vector whose ith element is 1 and the others are 0. Notice that

$$\frac{\partial v_j}{\partial\boldsymbol{y}}=\frac{I(z_j^2>1)}{z_j^4}\frac{\partial z_j^2}{\partial\boldsymbol{y}},$$

where z_j is the jth element of $\boldsymbol{z}$ given by (3.7). Thus, the last term on the right hand side of equation (A.4) is calculated as

$$\sum_{i=1}^{n}\boldsymbol{e}_i'(\boldsymbol{I}_n-\boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\boldsymbol{Q}\dot{\boldsymbol{V}}_i\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_1=\sum_{j=1}^{k}\boldsymbol{a}_j'\left(\frac{\partial z_j^2}{\partial\boldsymbol{y}}\right)\frac{I(z_j^2>1)\boldsymbol{a}_j'\boldsymbol{y}}{z_j^4d_j}, \tag{A.5}$$

where $\boldsymbol{a}_j$ is given by (3.19). Recall that $z_j^2=(n-2)(\boldsymbol{a}_j'\boldsymbol{y})^2/(d_j\boldsymbol{y}'\boldsymbol{R}\boldsymbol{S}^{-1}\boldsymbol{R}'\boldsymbol{y})$. Continuing some tedious calculations produces the first derivative of z_j^2 with respect to $\boldsymbol{y}$ as

$$\frac{\partial z_j^2}{\partial\boldsymbol{y}}=\frac{2}{\hat{\sigma}^2}\left(\frac{\hat{\sigma}z_j}{\sqrt{d_j}}\boldsymbol{a}_j-\frac{z_j^2}{n-2}\boldsymbol{R}\boldsymbol{S}^{-1}\boldsymbol{R}'\boldsymbol{y}\right). \tag{A.6}$$

By using equation (A.6), equation (A.5) becomes

$$\sum_{i=1}^{n} \boldsymbol{e}_i'(\boldsymbol{I}_n - \boldsymbol{P}_{\boldsymbol{W}_2})\boldsymbol{W}_1\boldsymbol{Q}\dot{\boldsymbol{V}}_i\boldsymbol{Q}'\hat{\boldsymbol{\beta}}_1 = 2\sum_{j=1}^{k} I(z_j^2 > 1)\left\{\frac{1}{z_j^2} - \frac{\boldsymbol{a}_j'\boldsymbol{R}\boldsymbol{S}^{-1}\boldsymbol{R}'\boldsymbol{y}}{(n-2)\boldsymbol{a}_j'\boldsymbol{y}}\right\}. \qquad \text{(A.7)}$$

Hence, from equations (A.4) and (A.7), the second term on the right hand side of equation (A.3) becomes

$$\sum_{i=1}^{n} \frac{\partial\hat{\mu}(x_i)}{\partial y_i} = m - k + \mathrm{tr}(\boldsymbol{V}) + 2\sum_{j=1}^{k} I(z_j^2 > 1)\left\{\frac{1}{z_j^2} - \frac{\boldsymbol{a}_j'\boldsymbol{R}\boldsymbol{S}^{-1}\boldsymbol{R}'\boldsymbol{y}}{(n-2)\boldsymbol{a}_j'\boldsymbol{y}}\right\}. \qquad \text{(A.8)}$$

Consequently, substituting equation (A.8) into equation (A.3) yields equation (3.20).

Acknowledgments

The author thanks Dr. Hironori Fujisawa, The Institute of Statistical Mathematics, and Prof. Hirofumi Wakaki, Hiroshima University, for helpful comments on Stein's Lemma and the non-uniqueness of $\boldsymbol{K}_+$. This research was supported by the Ministry of Education, Science, Sports, and Culture, Grant-in-Aid for Young Scientists (B), #19700265, 2007–2010.

References

[1] Balakrishnan, N. & Ambagaspitiya, R. S. (1994). On skew Laplace distribution. *Technical Report, Department of Mathematics & Statistics, McMaster University*, Hamilton, Ontario, Canada.

[2] Buckley, M. J., Eagleson, G. K. & Silverman, B. W. (1988). The estimation of residual variance in nonparametric regression. *Biometrika*, **75**, 189–199.

[3] Craven, P. & Wahba, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. *Numer. Math.*, **31**, 377–403.

[4] Eilers, P. & Marx, B. (1996). Flexible smoothing with B-splines and penalties (with discussion). *Statist. Sci.*, **11**, 89-121.

[5] Efron, B. (2004). The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, **99**, 619–632.

[6] Gasser, T., Sroka, L. & Jennen-Steinmetz, C. (1986). Residual variance and residual pattern in nonlinear regression model. *Biometrika*, **73**, 625–633.

[7] Green, P. J. & Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*. Chapman & Hall/CRC.

[8] Gu, C. & Wahba, G. (1991). Minimizing GCV/GML scores with multiple smoothing parameters via the Newton method. *SIAM J. Sci. Statist. Comput.*, **12**, 383–398.

[9] Hall, P., Kay, J. W. & Titterington, D. M. (1990). Asymptotically optimal difference-based estimation of variance in nonparametric regression. *Biometrika*, **77**, 521–528.

[10] Härdle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press.

[11] Hastie, T. (1996). Pseudosplines. *J. Roy. Statist. Soc. Ser.* **B**, **58**, 379–396.

[12] Hastie, T. J. & Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman & Hall /CRC.

[13] Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*. Springer-Verlag, New York.

[14] Holst, U., Hössjer, O., Björklund, C., Ragnarson, P. & Edner, H. (1996). Locally weighted least squares kernel regression and statistical evaluation of LIDAR measurements. *Environmetrics*, **7**, 401–416.

[15] Hoerl, A. E. & Kennard, R. W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55–67.

[16] Hurvich, C. M., Simonoff, J. S. & Tsai, C.-L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion. *J. Roy. Statist. Soc. Ser.* **B**, **60**, 271–293.

[17] Imoto, S. & Konishi, S. (2003). Selection of smoothing parameters in B-spline nonparametric regression models using information criteria. *Ann. Inst. Statist. Math.*, **55**, 671–687.

[18] Konishi, S. & Kitagawa, G. (1996). Generalised information criteria in model selection. *Biometrika*, **83**, 875–890.

[19] Konishi, S. & Kitagawa, G. (2008). *Information Criteria and Statistical Modeling.* Springer Science+Business Media, LLC, New York.

[20] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[21] Mallows, C. L. (1995). More comments on C_p. *Technometrics*, **37**, 362–372.

[22] Lawless, J. F. (1981). Mean squared error properties of generalized ridge regression. *J. Amer. Statist. Assoc.*, **76**, 462–466.

[23] Seifert, B., Gasser, T. & Wolf, A. (1993). Nonparametric estimation of residual variance revisited. *Biometrika*, **80**, 373–383.

[24] Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser.* **B**, **36**, 111–147.

[25] Walker, S. G. & Page, C. J. (2001). Generalized ridge regression and a generalization of the C_p statistic. *J. Appl. Statist.*, **28**, 911–922.

[26] Wand, M. P. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika*, **86**, 936–940.

[27] Wand, M. P. (2000). A comparison of regression spline smoothing procedures. *Comput. Statist.*, **15**, 443–462.

[28] Wand, M. P. (2003). Smoothing and mixed models. *Comput. Statist.*, **18**, 223–249.

[29] Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. *J. R. Stat. Soc. Ser.* **B** *Stat. Methodol.*, **62**, 413–428.

[30] Yanagihara, H. & Ohtaki, M. (2003). Knot-placement to avoid over fitting in B-spline scedastic smoothing. *Comm. Statist. Simulation Comput.*, **32**, 771–785.

[31] Yanagihara, H. & Ohtaki, M. (2004). On avoidance of over-fitting in the B-spline non-parametric regression model. *Japanese J. Appl. Statist.*, **33**, 51–69 (in Japanese).

[32] Yanagihara, H. & Yuan, K.-H. (2005). Four improved statistics for contrasting means by correcting skewness and kurtosis. *British J. Math. Statist. Psych.*, **58**, 209–237.

[33] Zhang, H. (1997). Multivariate adaptive splines for analysis of longitudinal data. *J. Comput. Graph. Statist.*, **6**, 74–91.

[34] Zhang, H. (2004). Mixed effects multivariate adaptive splines model for the analysis of longitudinal and growth curve data. *Stat. Methods Med. Res.*, **13**, 63–82.

Table 1. The MSE of an estimated curve in the case of Trend 1

					MSE		
σ	n	Dist.	C_p	$C_p^{\#}$	C_p^{*}	CV	GCV
0.5	20	1	0.8484	0.8484	0.8484	1.7856	1.2256
		2	0.8476	0.8500	0.8512	1.7500	1.3152
		3	0.8416	0.8432	0.8436	1.6820	1.2508
	50	1	0.3680	0.2888	0.3120	0.2636	0.2592
		2	0.4120	0.3152	0.3524	0.2760	0.3012
		3	0.3856	0.2944	0.3368	0.2656	0.2840
	100	1	0.2164	0.1304	0.1556	0.1272	0.1244
		2	0.2224	0.1324	0.1632	0.1224	0.1248
		3	0.2216	0.1344	0.1660	0.1304	0.1304
1.0	20	1	0.4767	0.5205	0.5151	0.9414	0.7939
		2	0.4737	0.5187	0.5082	0.9514	0.8151
		3	0.4721	0.5155	0.5067	0.9317	0.7603
	50	1	0.3712	0.2394	0.2786	0.3467	0.3428
		2	0.3895	0.2560	0.3032	0.3521	0.3459
		3	0.3685	0.2469	0.2909	0.3373	0.3462
	100	1	0.2051	0.1192	0.1460	0.1153	0.1132
		2	0.2016	0.1142	0.1426	0.1089	0.1110
		3	0.1992	0.1172	0.1450	0.1107	0.1111
2.0	20	1	0.3187	0.3540	0.3409	0.4205	0.4114
		2	0.3186	0.3543	0.3422	0.4117	0.4077
		3	0.3102	0.3399	0.3303	0.4173	0.4032
	50	1	0.3470	0.2149	0.2616	0.2959	0.2952
		2	0.3552	0.2233	0.2840	0.2823	0.2893
		3	0.3409	0.2255	0.2774	0.2939	0.2978
	100	1	0.1957	0.0934	0.1319	0.0953	0.0935
		2	0.1972	0.0960	0.1337	0.0975	0.0978
		3	0.1948	0.0908	0.1306	0.0930	0.0921
Average			0.3741	0.3140	0.3370	0.4817	0.4127

Table 2. The MSE of an estimated curve in the case of Trend 2

					MSE		
σ	n	Dist.	C_p	$C_p^{\#}$	C_p^*	CV	GCV
0.5	20	1	0.3864	0.3860	0.3936	1.7588	0.4540
		2	0.3876	0.3820	0.3904	1.8372	0.4976
		3	0.3876	0.3828	0.3908	1.7980	0.4748
	50	1	0.2792	0.1828	0.2156	2.8612	0.3628
		2	0.2928	0.1924	0.2328	2.8588	0.4024
		3	0.2904	0.1948	0.2296	2.8456	0.3624
	100	1	0.1852	0.0960	0.1276	0.1040	0.0960
		2	0.1876	0.0976	0.1336	0.0996	0.0964
		3	0.1880	0.0988	0.1340	0.1032	0.0996
1.0	20	1	0.3122	0.2957	0.3030	0.5474	0.4192
		2	0.3080	0.2916	0.3017	0.5328	0.4040
		3	0.2962	0.2789	0.2888	0.5333	0.3931
	50	1	0.2772	0.1471	0.1948	0.6719	0.5100
		2	0.2842	0.1501	0.2050	0.6775	0.5075
		3	0.2858	0.1568	0.2069	0.6702	0.5052
	100	1	0.1675	0.0764	0.1083	0.0855	0.0858
		2	0.1691	0.0779	0.1120	0.0847	0.0849
		3	0.1739	0.0765	0.1132	0.0881	0.0853
2.0	20	1	0.2435	0.2272	0.2332	0.2437	0.2434
		2	0.2312	0.2147	0.2206	0.2414	0.2403
		3	0.2273	0.2144	0.2170	0.2352	0.2330
	50	1	0.2644	0.1234	0.1800	0.2044	0.2018
		2	0.2847	0.1295	0.2009	0.2062	0.2019
		3	0.2796	0.1303	0.1978	0.2088	0.2068
	100	1	0.1674	0.0619	0.1008	0.0919	0.0868
		2	0.1744	0.0632	0.1079	0.0888	0.0863
		3	0.1737	0.0626	0.1060	0.0900	0.0866
Average			0.2557	0.1775	0.2091	0.7322	0.2751

Table 3. The MSE of an estimated curve in the case of Trend 3

					MSE		
σ	n	Dist.	C_p	$C_p^{\#}$	C_p^{*}	CV	GCV
0.5	20	1	0.1656	0.1540	0.1548	0.1224	0.1228
		2	0.1612	0.1512	0.1508	0.1180	0.1272
		3	0.1704	0.1596	0.1584	0.1300	0.1336
	50	1	0.2304	0.0896	0.1540	0.0612	0.0596
		2	0.2496	0.0904	0.1664	0.0536	0.0548
		3	0.2280	0.0956	0.1592	0.0592	0.0664
	100	1	0.1464	0.0388	0.0860	0.0268	0.0272
		2	0.1512	0.0428	0.0892	0.0268	0.0272
		3	0.1484	0.0384	0.0844	0.0268	0.0268
1.0	20	1	0.1674	0.1566	0.1577	0.1287	0.1296
		2	0.1650	0.1550	0.1556	0.1303	0.1341
		3	0.1666	0.1559	0.1571	0.1279	0.1317
	50	1	0.2324	0.0901	0.1570	0.0590	0.0588
		2	0.2564	0.1063	0.1794	0.0577	0.0596
		3	0.2560	0.0997	0.1768	0.0580	0.0592
	100	1	0.1490	0.0424	0.0893	0.0279	0.0286
		2	0.1471	0.0418	0.0873	0.0273	0.0275
		3	0.1438	0.0412	0.0842	0.0269	0.0268
2.0	20	1	0.1650	0.1534	0.1560	0.1300	0.1266
		2	0.1711	0.1588	0.1612	0.1289	0.1312
		3	0.1701	0.1577	0.1598	0.1307	0.1346
	50	1	0.2488	0.0957	0.1704	0.0587	0.0605
		2	0.2536	0.0984	0.1772	0.0566	0.0581
		3	0.2511	0.1040	0.1710	0.0562	0.0607
	100	1	0.1504	0.0417	0.0873	0.0286	0.0279
		2	0.1457	0.0376	0.0838	0.0252	0.0251
		3	0.1383	0.0389	0.0775	0.0275	0.0275
Average			0.1863	0.0976	0.1367	0.0708	0.0724

Table 4. The MSE of an estimated curve in the case of Trend 4

			MSE				
σ	n	Dist.	C_p	$C_p^{\#}$	C_p^*	CV	GCV
0.5	20	1	1.6048	1.6996	1.6396	1.7984	1.6944
		2	1.5972	1.6840	1.6272	1.7808	1.6784
		3	1.6084	1.7080	1.6480	1.8220	1.7080
	50	1	0.5576	0.5572	0.5756	0.5292	0.4364
		2	0.5664	0.5664	0.5848	0.5352	0.4600
		3	0.5544	0.5608	0.5752	0.5328	0.4432
	100	1	0.2880	0.2512	0.2672	0.2168	0.2124
		2	0.2792	0.2464	0.2624	0.2088	0.2100
		3	0.2872	0.2548	0.2680	0.2148	0.2140
1.0	20	1	0.6317	0.6978	0.6630	0.9654	0.7308
		2	0.6202	0.6809	0.6489	0.8709	0.7236
		3	0.6291	0.6899	0.6590	0.8734	0.7308
	50	1	0.4264	0.3861	0.3941	0.3537	0.3463
		2	0.4567	0.4178	0.4255	0.3599	0.3757
		3	0.4209	0.3900	0.3962	0.3497	0.3505
	100	1	0.2489	0.2087	0.2236	0.1861	0.1823
		2	0.2493	0.2067	0.2239	0.1769	0.1802
		3	0.2453	0.2044	0.2191	0.1782	0.1785
2.0	20	1	0.3623	0.3796	0.3710	0.5088	0.4806
		2	0.3704	0.3878	0.3784	0.5055	0.4737
		3	0.3592	0.3774	0.3681	0.4869	0.4701
	50	1	0.3475	0.2600	0.2910	0.2109	0.2277
		2	0.3638	0.2712	0.3127	0.2131	0.2309
		3	0.3719	0.2678	0.3110	0.2162	0.2313
	100	1	0.2080	0.1457	0.1682	0.1328	0.1329
		2	0.2084	0.1444	0.1645	0.1308	0.1331
		3	0.2073	0.1426	0.1603	0.1282	0.1311
Average			0.5211	0.5106	0.5121	0.5365	0.4951

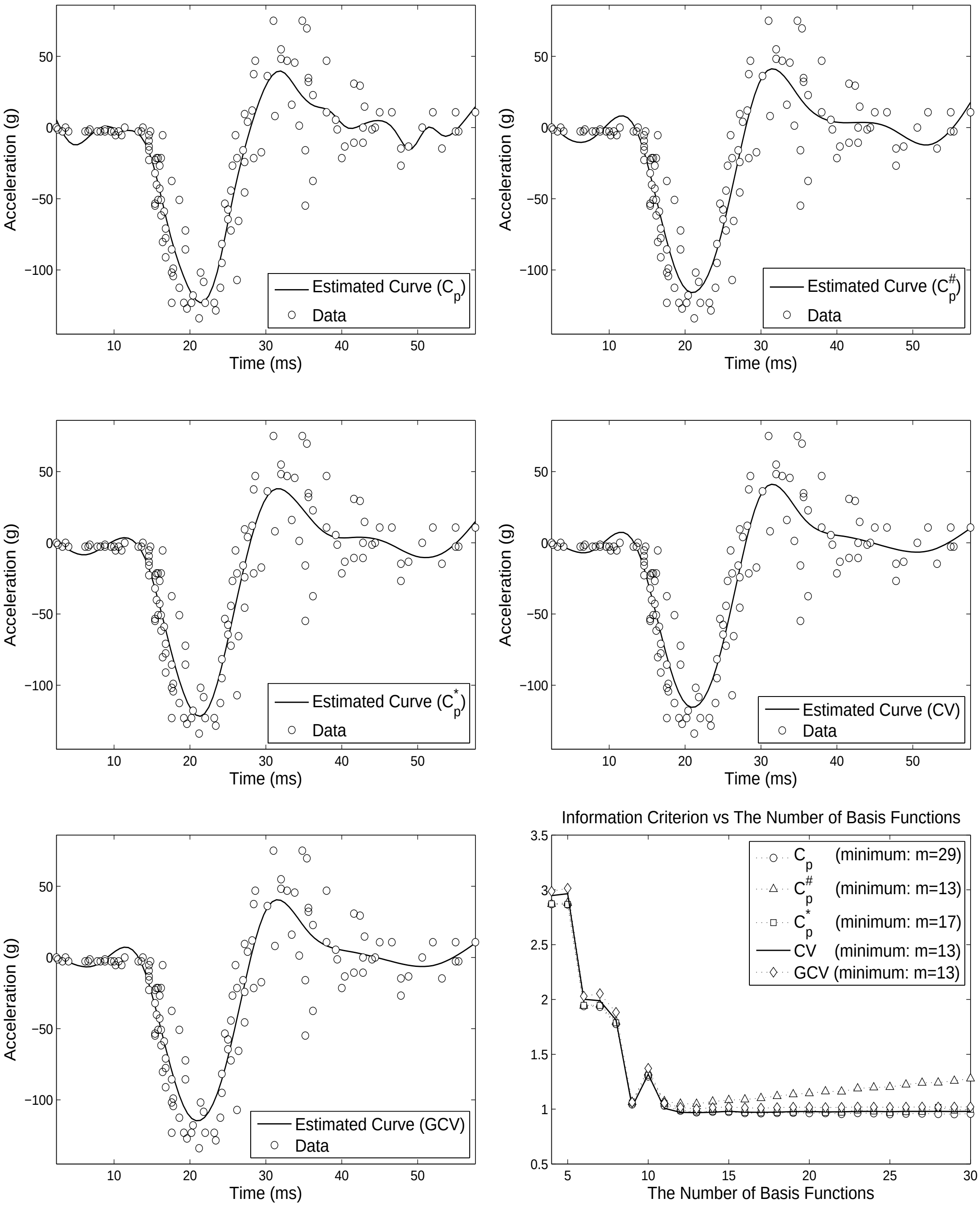

FIGURE 2. Estimated curves for motorcycle data

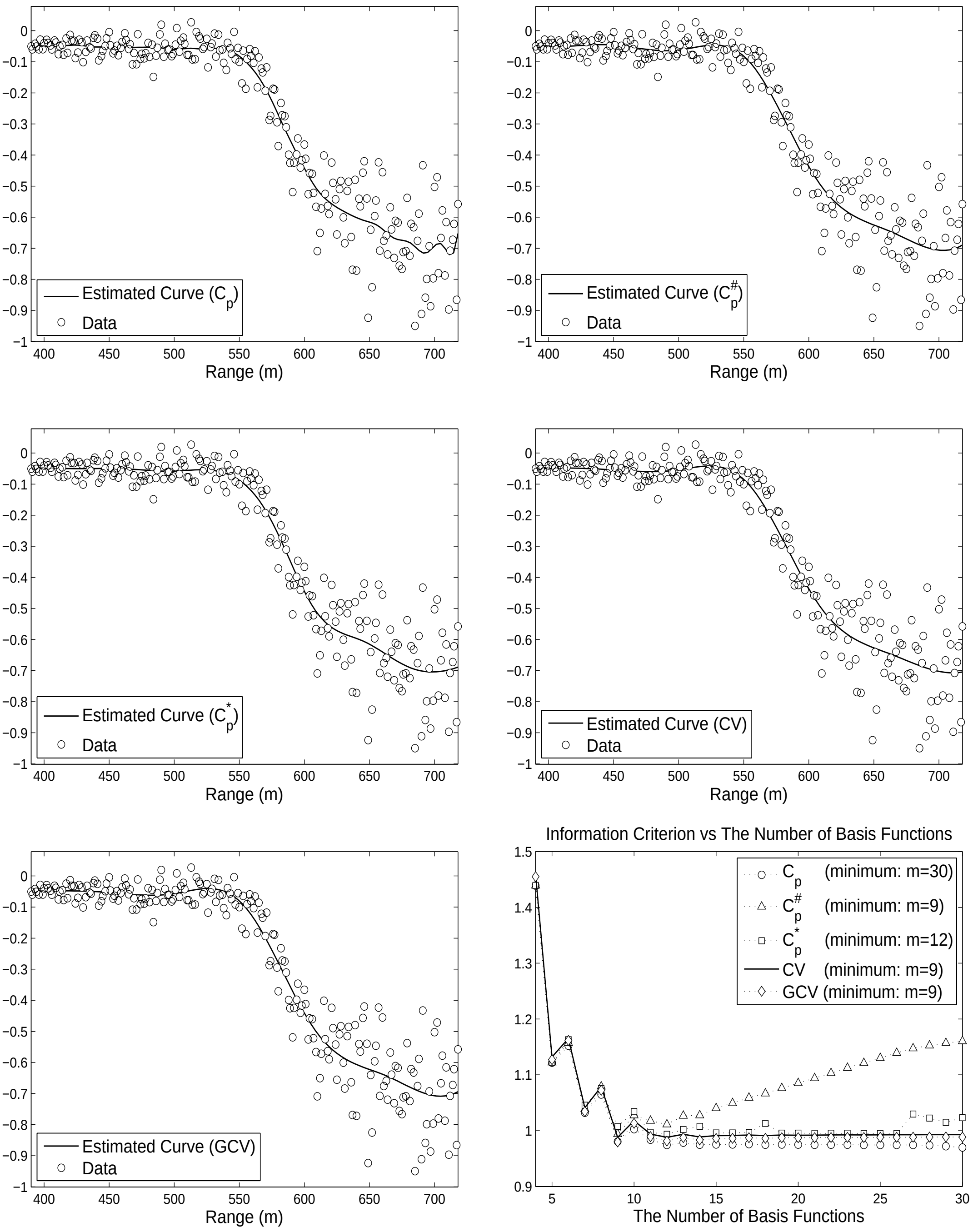

Figure 3. Estimated curves for LIDAR data

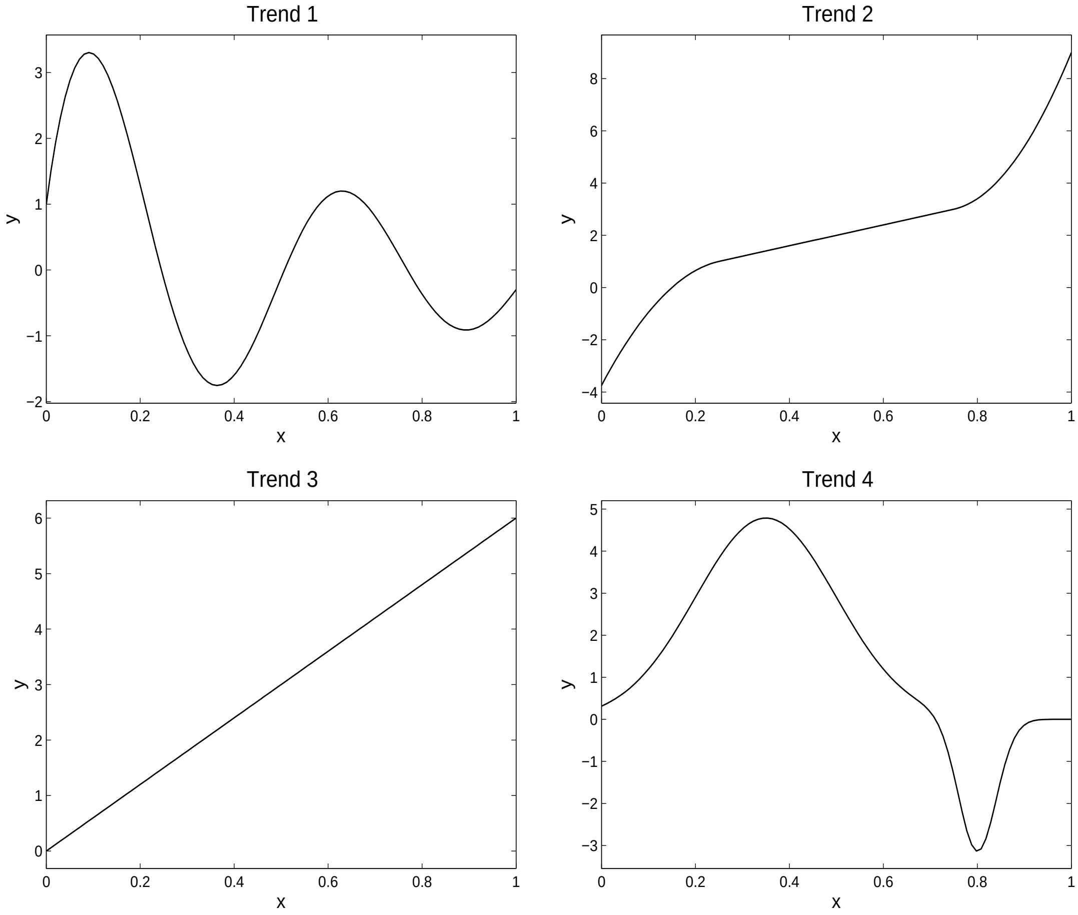

Figure 4. True trends for simulations