Probab. Theory Relat. Fields 127, 153–176 (2003)
Digital Object Identifier (DOI) 10.1007/s00440-003-0277-z

S. C. Kou

On the efficiency of selection criteria in spline regression

Received: 23 July 2003 / Revised version: 26 March 2003 /
Published online: 4 July 2003 –

Abstract. This paper concerns the cubic smoothing spline approach to nonparametric regression. After first deriving sharp asymptotic formulas for the eigenvalues of the smoothing matrix, the paper uses these formulas to investigate the efficiency of different selection criteria for choosing the smoothing parameter. Special attention is paid to the generalized maximum likelihood (GML), C_p and extended exponential (EE) criteria and their marginal Bayesian interpretation. It is shown that (a) when the Bayesian model that motivates GML is true, using C_p to estimate the smoothing parameter would result in a loss of efficiency with a factor of 10/3, proving and strengthening a conjecture proposed in Stein (1990); (b) when the data indeed come from the C_p density, using GML would result in a loss of efficiency of ∞; (c) the loss of efficiency of the EE criterion is at most 1.543 when the data are sampled from its consistent density family. The paper not only studies equally spaced observations (the setting of Stein, 1990), but also investigates general sampling scheme of the design points, and shows that the efficiency results remain the same in both cases.

1. Introduction

Given n observed data points $\{(x_i, y_i)\}_{i=1}^n$ in the plane, regression models postulate that at the design points $x_1, x_2, \ldots, x_n$, the observations y_i satisfy

$$y_i = f(x_i) + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1.1}$$

where $\varepsilon_1, \varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed with zero mean. The goal of regression is to estimate the underlying function $f(x)$. In this paper we consider the spline smoothing approach to nonparametric regression, where, to make the theoretical statements clear, throughout this paper we assume that the design points are distinct and ordered such that $x_1 < x_2 < \cdots < x_n$. For general nonparametric methods, see, for example, Fan and Gijbels (1996), and the review papers of Fan (2000) and Hall (2001).

S. C. Kou: Department of Statistics, Science Center 6th Floor, Harvard University, Cambridge, MA 02138, USA. e-mail: kou@stat.harvard.edu.

This work is supported in part by NSF grant DMS-0204674 and Harvard University Clark-Cooke Fund.

Mathematics Subject Classification (2000): Primary: 62G08; Secondary: 62G20

Key words or phrases: Smoothing splines – Extended exponential criterion – Cp – Generalized maximum likelihood – Eigenvalue – Robustness – Sampling scheme

A cubic smoothing spline minimizes the penalized least square criterion

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int f''(t)^2 dt,$$

over the class of all functions f for which f and f' are absolutely continuous and f'' is square-integrable. The *smoothing parameter*, λ, balances the fidelity to the data and the roughness of the curve. This idea of trading-off the faithfulness to the data against the smoothness of the curve dates back to Whittaker (1923) and Schoenberg (1964a, 1964b), who coined the name "spline functions". Through important early developments such as Reinsch (1967), Kimeldorf and Wahba (1970), Demmler and Reinsch (1975), Wahba (1975), Craven and Wahba (1979), and Utreras (1979), nowadays spline smoothing has become a standard statistical technique, widely used in many scientific disciplines. For references, see, for example, Silverman (1985), Eubank (1988), Wahba (1990), Härdle (1990), Hastie and Tibshirani (1990), Rosenblatt (1991), Green and Silverman (1994), Simonoff (1996), and Bowman and Azzalini (1997).

It is well-known that smoothing splines are linear smoothers, meaning that the cubic smoothing spline estimate $\hat{\mathbf{f}}_\lambda$ of $\mathbf{f} = (f_1, f_2, \ldots, f_n)' = (f(x_1), f(x_2), \ldots, f(x_n))'$ can be written as

$$\hat{\mathbf{f}}_\lambda = A_\lambda \mathbf{y}, \tag{1.2}$$

where $\hat{\mathbf{f}}_\lambda = (\hat{f}_{\lambda 1}, \ldots, \hat{f}_{\lambda n})'$, $\mathbf{y} = (y_1, \ldots, y_n)'$. Furthermore, A_λ, the smoothing matrix, which does not depend on $\mathbf{y}$, has eigen decomposition $A_\lambda = U\mathbf{a}_\lambda U'$ with U an $n \times n$ orthonormal matrix *not* depending on λ, and $\mathbf{a}_\lambda = \mathrm{diag}(a_{\lambda i})$, a diagonal matrix whose ith diagonal element has the form

$$a_{\lambda i} = 1/(1 + \lambda k_i) \quad i = 1, 2, \ldots, n, \tag{1.3}$$

where the sequence $\mathbf{k} = (k_1, k_2, \ldots, k_n)$ is solely determined by $\mathbf{x} = (x_1, \ldots, x_n)$ and satisfies

$$0 = k_1 = k_2 < k_3 < \cdots < k_n. \tag{1.4}$$

The first two columns of the eigen matrix U represent linear functions of $\mathbf{x}$, while the other columns of U behave much like trigonometric functions of increasing frequency. Intuitively, cubic smoothing splines achieve the goal of smoothing by maintaining the part of $\mathbf{y}$ linear in $\mathbf{x}$ while shrinking the parts of $\mathbf{y}$ that are of higher frequencies towards zero; the higher the frequency, the stronger the shrinkage. The eigenvalue $a_{\lambda i}$ controls the amount of shrinkage. For more details on the eigen matrix U, see Utreras (1988) and Eubank (1999).

One object of this paper is to study bounds and asymptotic properties of the eigenvalues $a_{\lambda i}$. There are two reasons why they are interesting.

(a) Because of the shrinking mechanism of smoothing splines, which is controlled by $a_{\lambda i}$, obtaining a sharp result of the eigenvalues would help statisticians have a better understanding of the way smoothing splines work.

(b) To use smoothing splines in practice, it is necessary to choose a value for λ, the smoothing parameter. Two popular families of selection criteria are: (i) C_p-type criteria including C_p (Mallows, 1973), generalized cross validation (GCV) (Craven and Wahba, 1979) and Akaike Information Criterion (AIC) (Akaika, 1974), and (ii) empirical Bayes type criteria such as generalized maximum likelihood (GML) (Wecker and Ansley, 1983, Wahba, 1985). Kou and Efron (2002) introduces another selection criterion: the extended exponential (EE) criterion. For a problem at hand, since there are several ways to choose λ, one natural question is then: How do different selection criteria compare with each other? Efron (2001) and Kou and Efron (2002) geometrically compare C_p with GML in a small-sample, non-asymptotic setting, and show that the big variability of C_p is closely related to its geometric instability. In addition, Kou and Efron (2002) show that, under the small-sample setting, the EE criterion effectively combines the strength of C_p and GML. These small-sample results are quite applicable in general; however in many contexts, the large sample properties of selection criteria are also of significant interests. In this case, knowing the asymptotic behavior of the eigenvalues $a_{\lambda i}$ becomes indispensable.

There is a large amount of literature studying the large sample properties of selection criteria; for references, see, for example, Wahba (1985), Härdle, Hall and Marron (1988), Hall and Johnstone (1992), Li (1986, 1987), Speckman (1983, 1985), Kneip (1994) and Speckman and Sun (2001). Most of the literature is written in a frequentist framework. This paper, parallel to these frequentist developments, focuses the large sample investigation of the selection criteria on their Bayesian properties.

We first derive sharp asymptotic formulas for the eigenvalues $a_{\lambda i}$; then use these formulas to study the Bayesian efficiency of different selection criteria for choosing the smoothing parameter. In particular we obtain the following results:

- In a collection of papers, Wahba (1977a, 1977b, 1985) develops some theoretical results regarding the behavior of C_p (GCV) and GML and suggested that C_p (GCV) and GML would perform similarly under the Bayesian model for spline smoothing and that C_p (GCV) would perform better than GML in the frequentist case. Recently Speckman and Sun (2001) reinvestigate the frequentist setting and illustrate that C_p (GCV) does not necessarily performs better than GML. The current paper extends Wahba's conjecture for the Bayesian model setting; we show that under the Bayesian model using C_p, instead of GML, to estimate the smoothing parameter would encounter a loss of efficiency of factor 10/3.
- Stein (1990) also compares the performance of C_p (actually GCV) and GML under the Bayesian model that motivates GML, and conjectures, in the case of equally spaced observations, that when the Bayesian model that motivates GML is true, the use of C_p to estimate λ would result in a loss of efficiency with a factor of 10/3. In this paper, obtaining the sharp asymptotic expressions on $a_{\lambda i}$ enables us to rigorously prove the conjecture. Furthermore, we show that the conjecture holds true even if the data are not equally spaced, thus strengthening the result's applicability.

- Efron (2001) gives C_p a marginal Bayesian interpretation. This paper, under *both* equally and unequally spaced observations, also investigates what would happen to GML, if the data indeed are sampled from the C_p marginal density. The interesting result is that the relative efficiency of C_p versus GML in this case is ∞, namely, asymptotically C_p will do better than GML, if the data actually come from the C_p marginal density.
- Kou and Efron (2002), in addition to introducing the extended exponential (EE) criterion, show that the EE criterion has a marginal Bayesian interpretation. This paper also studies the Bayesian properties of the EE criterion and shows that the maximum loss of efficiency of the EE criterion is 1.543 when the data are sampled from its consistent density family, in certain sense suggesting the robustness of the EE criterion.

In the study, we consider not only the case of $(x_1, x_2, \ldots, x_n)$ being equally spaced, but also the case that they are drawn from a distribution on an interval $[\alpha, \beta]$. Interestingly, all the efficiency results remain the same no matter what the sampling scheme is, hence suggesting the general applicability of the results in the current paper.

The paper is organized as follows. Section 2 derives bounds on the eigenvalues and investigates their asymptotic properties in the case of equally spaced observations. After an overview of the C_p, GML and EE selection criteria and their marginal Bayesian interpretation in Section 3, the results of Section 2 are used to study the large sample properties of these three selection criteria in Section 4. The study focuses on the relative Bayesian efficiency of these selection criteria. Section 5 considers general sampling scheme of $(x_1, x_2, \ldots, x_n)$, and shows that essentially all the results established in the previous sections remain the same. The paper concludes in Section 6 with some discussion.

2. Bounds and limiting behavior of the eigenvalues for equally spaced observations

In this section, we study the properties of the eigenvalues of the smoothing matrix A_λ in the case of equally spaced data. That is, the observations are such that $x_{i+1} - x_i = \delta, i = 1, 2, \ldots, n-1$ for some δ. When $\delta = n^{-1}$, $(x_1, \ldots, x_n)$ are n equally spaced points in the $[x_1, x_1 + 1]$ interval, a standard setting in considering nonparametric regression problems. In this section we consider more generally $\delta = n^{-\rho}$, for $0 < \rho \leq 1$. Compared with $\delta = n^{-1}$, taking $0 < \rho < 1$ has the feature that as $n \to \infty$, not only the observations become denser, but also the range $x_n - x_1$ becomes larger. Section 5 considers the situation that $(x_1, x_2, \ldots, x_n)$ are sampled from a distribution on an interval $[\alpha, \beta]$. It is worth pointing out that the results in this section are not special cases of those in Section 5, the main reason being that we consider not only $\delta = n^{-1}$, but more generally $\delta = n^{-\rho}$, which is not covered by the setting of Section 5 (see Remark 4).

For $\delta = n^{-1}$, several authors have studied the problem of approximating the eigenvalues $a_{\lambda i}$. Utreras (1980, 1981) proposes an approximation by relating the eigenvalues of the spline matrix to those of the differential operator that one encounters in classical mechanics when describing the vibration of a rod with free ends.

Silverman (1984) suggests approximating $a_{\lambda i}$ by taking $k_i = \pi^4(i-1.5)^4/n$ for $i \geq 3$ in (1.3). Both approximations are shown to be accurate in the sense of approximating the trace of the smoothing matrix, $\mathrm{tr}(A_\lambda) = \sum_{i=1}^n a_{\lambda i}$. Speckman (1983, 1985) and Nussbaum (1985) provide further approximation for $a_{\lambda i}$.

For general δ, Culpin (1986) gives a semi-explicit formula for the eigenvalues and eigenvectors of A_λ. In particular, for $i \geq 3$, the k_i sequence in (1.3) has the form

$$k_i = 12\delta^{-3}\frac{(1-\cos\theta_i)^2}{2+\cos\theta_i}, \tag{2.1}$$

where θ_i $(3 \leq i \leq n)$ are the $n-2$ distinct roots in $(0,\pi)$ of the equation $F(\theta)=0$ given by

$$F(\theta) = \cos(n\theta-\omega) - 2e^{-n\phi}\cos\omega + e^{-2n\phi}\cos(n\theta+\omega) = 0, \tag{2.2}$$

where

$$\phi = \phi(\theta) = \cosh^{-1}((5-2\cos\theta)/(2+\cos\theta)), \tag{2.3}$$

$$\omega = \omega(\theta) = \sin^{-1}(2(1-\cos\theta)/(5+\cos\theta)) \tag{2.4}$$

The above formulas are quite useful in our study, because (i) we want to consider general δ, not only $\delta = n^{-1}$, (ii) in order to describe the limiting behavior of the eigenvalues, as well as to compare different selection criteria, only considering the trace of the smoothing matrix A_λ is not enough.

The following proposition, giving a tight bound on k_i, is the starting point of our study.

Proposition 2.1. *The k_i sequence in (1.3) is bounded by*

$$L_i \leq k_i \leq U_i \quad \textit{for } i \geq 3$$

where

$$L_i = \frac{12\delta^{-3}(1-\cos\frac{i-2}{n}\pi)^2}{2+\cos\frac{i-2}{n}\pi},\ U_i = \frac{12\delta^{-3}(1-\cos\frac{i-1}{n}\pi)^2}{2+\cos\frac{i-1}{n}\pi}. \tag{2.5}$$

Proof. See the appendix.

Working with trigonometric function sometimes is not very convenient. The bounds (2.5) can be simplified.

Corollary 2.2. *The k_i sequence is bounded above and below by*

$$k_i^- \leq k_i \leq k_i^+,\ i \geq 3 \tag{2.6}$$

where

$$k_i^- = \frac{1}{n^4\delta^3}(i-2)^4\pi^4\left(1-\frac{(i-2)^2\pi^2}{18n^2}\right), \quad k_i^+ = \frac{1}{n^4\delta^3}(i-1)^4\pi^4. \tag{2.7}$$

Subsequently the eigenvalues $a_{\lambda i}$ $(i \geq 3)$ are bounded by

$$1/(1+\lambda k_i^+) \leq a_{\lambda i} \leq 1/(1+\lambda k_i^-). \tag{2.8}$$

Proof. The result is a simple follow-up of the fact

$$\frac{\theta^4}{12} - \frac{\theta^6}{216} \leq \frac{(1-\cos\theta)^2}{2+\cos\theta} \leq \frac{\theta^4}{12}, \text{ for all } \theta \in (0,\pi). \qquad \square$$

Remark 1. When $\delta = n^{-1}$, note that $k_i^- \approx \frac{1}{n}(i-2)^4\pi^4$, $k_i^+ = \frac{1}{n}(i-1)^4\pi^4$ for moderate i and large n. The corollary, hence, suggests the effectiveness of Silverman's (1984) approximation.

The bounds (2.6) and (2.8) are very useful to study the limiting behavior of the eigenvalues — they serve as the basic building block of our large sample investigation. The following theorem, built upon (2.8), provides a general result regarding the asymptotic behavior of the eigenvalues in the case of equally spaced observations.

Theorem 2.3. *Suppose* $\frac{n^4\delta^3}{\lambda} \to \infty$, $\delta^{-3}\lambda \to \infty$, *then for any real numbers* $r > \frac{1}{4}$ *and* $s > -\frac{1}{4}$,

$$\sum_{i=3}^{n} a_{\lambda i}^r (1-a_{\lambda i})^s = \frac{1}{4\pi} B(r-\frac{1}{4}, s+\frac{1}{4})(\frac{n^4\delta^3}{\lambda})^{1/4} + o\left((\frac{n^4\delta^3}{\lambda})^{1/4}\right),$$

where the beta function $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$.

Proof. See the appendix.

This theorem covers a variety of situations. For $\delta = n^{-\rho}$, $0 < \rho \leq 1$, it says that as long as λ is of $o(n^{4-3\rho})$ and $n^{3\rho}\lambda \to \infty$, the summation $\sum_i a_{\lambda i}^r (1-a_{\lambda i})^s$ goes to infinity at the same order $O((n^{4-3\rho}/\lambda)^{1/4})$ for all values of $r > 1/4$ and $s > -1/4$. A special case of the theorem concerning the limiting behavior of the trace of the smoothing matrix for $\delta = n^{-1}$ is readily available by taking r to be 1, and s to be 0.

Corollary 2.4. *Suppose* $\frac{n}{\lambda} \to \infty$, $n^3\lambda \to \infty$, *then* $\operatorname{tr}(A_\lambda) = \sum_{i=1}^n a_{\lambda i} = 2^{-3/2}(\frac{n}{\lambda})^{1/4} + o\left((\frac{n}{\lambda})^{1/4}\right)$.

This corollary echoes the well-known results of Utreras (1980, 1981) and Silverman (1984). The following theorem, covering the case of $s < -\frac{1}{4}$, complements Theorem 2.3.

Theorem 2.5. *Suppose* $\frac{n^4\delta^3}{\lambda} \to \infty$, $\delta^{-3}\lambda \to \infty$, *then for any real numbers* $r > \frac{1}{4}$ *and* $s < -\frac{1}{4}$,

$$\sum_{i=3}^{n} a_{\lambda i}^r (1-a_{\lambda i})^s = O\left((\frac{n^4\delta^3}{\lambda})^{-s}\right).$$

Proof. See the appendix.

The bounds (2.6) and (2.8), together with Theorems 2.3 and 2.5, depict the finite sample as well as the large sample properties of the eigenvalues $a_{\lambda i}$: The bounds (2.6) and (2.8) are valid for all values of $i \geq 3$ and n, while Theorems 2.3 and 2.5 describe the limiting behavior of $a_{\lambda i}$. In Section 4 we will see the utility of these results by applying them to analyze the Bayesian properties of selection criteria for choosing the smoothing parameter λ.

3. Selecting the smoothing parameter

3.1. The C_p, GML, and EE criteria

Because the smoothing parameter λ controls the trade-off between the closeness to the data and the roughness of the curve, in practice the use of a smoothing spline requires the choice of λ. In this Section, we review three selection criteria for choosing the smoothing parameter: the C_p criterion (Mallows, 1973), the generalized maximum likelihood (GML) criterion (Wecker and Ansley, 1983, Wahba, 1985), and the extended exponential (EE) criterion (Kou, 2001, Kou and Efron, 2002). We start by strengthening (1.1) to the normal sampling model $\mathbf{y} \sim N(\mathbf{f}, \sigma^2 I)$, where σ^2 is assumed known.

The C_p selection criterion chooses λ as the minimizer of the C_p statistic $C_p(\lambda) = \| \mathbf{y} - \hat{\mathbf{f}}_\lambda \|^2 + 2\sigma^2 \mathrm{tr}(A_\lambda) - n\sigma^2$ with $\hat{\mathbf{f}}_\lambda$ as in (1.2). It is based on the fact that the C_p statistic is an unbiased estimate of the total estimation error: $E\{C_p(\lambda)\} = E \| \hat{\mathbf{f}}_\lambda - \mathbf{f} \|^2$. Let

$$\mathbf{z} = U'\mathbf{y}/\sigma, \quad \mathbf{g} = U'\mathbf{f}/\sigma, \quad \hat{\mathbf{g}}_\lambda = U'\hat{\mathbf{f}}_\lambda/\sigma \tag{3.1}$$

where, recall, U is the orthogonal matrix consisting of the eigenvectors of A_λ. Then the C_p statistic can be succinctly expressed as $C_p(\lambda) = \sigma^2 \sum_{i=1}^n (b_{\lambda i}^2 z_i^2 - 2b_{\lambda i} + 1)$ with $b_{\lambda i} = 1 - a_{\lambda i}$, and consequently the C_p criterion chooses λ by

$$\hat{\lambda}^{C_p} = \arg\min_\lambda \sum_i (b_{\lambda i}^2 z_i^2 - 2b_{\lambda i}). \tag{3.2}$$

The transformation (3.1) also provides $\mathbf{z} \sim N(\mathbf{g}, I)$ and $\hat{\mathbf{g}}_\lambda = \mathbf{a}_\lambda \mathbf{z}$.

The GML criterion is another method to choose λ. It is an empirical Bayes estimator. Suppose in addition to the normal model $\mathbf{z} \sim N(\mathbf{g}, I)$, one puts a Gaussian prior on the curve $\mathbf{g} \sim N(\mathbf{0}, \mathbf{c}_\lambda)$, where $\mathbf{c}_\lambda$ is a diagonal matrix with the iith entry $c_{\lambda i} = a_{\lambda i}/(1 - a_{\lambda i}) = a_{\lambda i}/b_{\lambda i}$. Then by Bayes theorem,

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{1}/\mathbf{b}_\lambda), \quad \mathbf{g}|\mathbf{z} \sim N(\mathbf{a}_\lambda \mathbf{z}, \mathbf{a}_\lambda), \tag{3.3}$$

where $\mathbf{1}/\mathbf{b}_\lambda$ denotes the diagonal matrix with the ith diagonal element being $1/b_{\lambda i} = 1/(1 - a_{\lambda i})$. The second relationship in (3.3) gives a Bayesian justification for using the linear smoother (expressed in terms of $\mathbf{z}$ and $\mathbf{g}$) $\hat{\mathbf{g}}_\lambda = \mathbf{a}_\lambda \mathbf{z}$ — it is the posterior mean of $\mathbf{g}$ given the observation $\mathbf{z}$. The first relationship motivates the GML choice for λ,

$$\hat{\lambda}^{GML} = \text{MLE of } \lambda \text{ based on } \mathbf{z} \sim N(\mathbf{0}, \mathbf{1}/\mathbf{b}_\lambda).$$

Under the distribution $\mathbf{z} \sim N(\mathbf{0}, \mathbf{1}/\mathbf{b}_\lambda)$, $\mathbf{z}^2$ is the minimal sufficient statistic for λ with the joint density $d_\lambda(\mathbf{z}^2)$ given by

$$d_\lambda(\mathbf{z}^2) = \exp(-\frac{1}{2} \sum_i (b_{\lambda i} z_i^2 - \log b_{\lambda i}))/\prod_i \sqrt{2\pi z_i^2}. \tag{3.4}$$

The GML estimate, thus, can be written as

$$\hat{\lambda}^{GML} = \arg\max_{\lambda} d_{\lambda}(\mathbf{z}^2) = \arg\min_{\lambda} \sum_i (b_{\lambda i} z_i^2 - \log b_{\lambda i}). \tag{3.5}$$

The extended exponential (EE) selection criterion, studied in Kou (2001) and Kou and Efron (2002), provides a third way to choose the smoothing parameter. It is motivated by the idea of combining the strength of C_p and GML while mitigating their weaknesses, since in practice the C_p selected smoothing parameter tends to be highly variable, whereas the GML criterion has a serious problem with large bias (for details, see Kou, 2001, and Kou and Efron, 2002). Expressed in terms of $\mathbf{z}$, the EE criterion selects the smoothing parameter λ according to

$$\hat{\lambda}^{EE} = \arg\min_{\lambda} \sum_i [\xi b_{\lambda i} {z_i}^{4/3} - 3 b_{\lambda i}^{1/3}], \tag{3.6}$$

where the constant $\xi = \frac{\sqrt{\pi}}{2^{2/3}\Gamma(7/6)} = 1.203$. Kou and Efron (2002) explain the construction of the EE criterion from a geometric point of view and show, through a finite-sample non-asymptotic analysis, that the EE criterion effectually combines the strength of C_p and GML [see Kou and Efron (2002) for details].

An interesting fact about the three criteria (C_p, GML and EE) is that they have a unified structure. Let $p \geq 1$, $q \geq 1$ be two fixed constants. Define the function

$$l_{\lambda}^{(p,q)}(\mathbf{u}) = \begin{cases} \sum_i [(c_q b_{\lambda i}^{1/q})^p u_i - \frac{p}{p-1}((c_q b_{\lambda i}^{1/q})^{p-1} - 1)] & \text{if } p > 1 \\ \sum_i (c_q b_{\lambda i}^{1/q} u_i - \log b_{\lambda i}^{1/q}) & \text{if } p = 1 \end{cases}$$

where

$$b_{\lambda i} = 1 - a_{\lambda i}, \quad c_q = \frac{\sqrt{\pi}}{2^{1/q}\Gamma(1/2 + 1/q)}, \tag{3.7}$$

and a corresponding selection criterion

$$\begin{aligned} \hat{\lambda}^{(p,q)} &= \arg\min_{\lambda} \{l_{\lambda}^{(p,q)}(\mathbf{z}^{2/q})\} \\ &= \begin{cases} \arg\min_{\lambda} \sum_i [c_q b_{\lambda i}^{p/q} z_i^{2/q} - \frac{p}{p-1} b_{\lambda i}^{(p-1)/q}] & \text{if } p > 1 \\ \arg\min_{\lambda} \sum_i (c_q b_{\lambda i}^{1/q} z_i^{2/q} - \log b_{\lambda i}^{1/q}) & \text{if } p = 1 \end{cases}. \end{aligned} \tag{3.8}$$

Then it is easy to verify that

(i) $l_{\lambda}^{(p,q)}(\cdot) \to l_{\lambda}^{(1,q)}(\cdot)$ as $p \to 1$;

(ii) taking $p = 1, q = 1$ gives the GML criterion; $p = 2, q = 1$ gives the C_p criterion; $p = q = \frac{3}{2}$ gives the EE criterion.

The class (3.8), therefore, unites the three criteria in a continuous fashion.

3.2. The marginal Bayesian interpretation of the C_p, GML and EE criteria

GML is motivated from the Bayesian framework (3.3). Actually, like GML, the C_p and EE criteria also have a marginal Bayesian interpretation [see Efron (2001) and Kou (2001) for further background]. In general, every member of (3.8) has a marginal Bayesian interpretation. Suppose instead of the GML marginal density (3.4), one assumes that $\mathbf{u} = \mathbf{z}^{2/q}$ comes from

$$\mathbf{u} = \mathbf{z}^{2/q} \sim \exp(-C_0 l_\lambda^{(p,q)}) d_0^{(p,q)}(\mathbf{u}), \tag{3.9}$$

Then the MLE of (3.9) yields the (p, q)-criterion (3.8). Three special cases of (3.9) that are of particular interest to us, are the GML marginal density (3.4), the C_p marginal density

$$\mathbf{z}^2 \sim \exp[-C_0 \sum_i (b_{\lambda i}^2 z_i^2 - 2b_{\lambda i})] d_0^{C_p}(\mathbf{z}^2) \tag{3.10}$$

and the EE marginal density

$$u_i = z_i^{4/3} \overset{\text{ind.}}{\sim} \exp[-C_0(\xi^{3/2} b_{\lambda i} u_i - 3\xi^{1/2} b_{\lambda i}^{1/3})] d_0^{EE}(u_i). \tag{3.11}$$

(The properties of $d_0^{C_p}(\cdot)$ and $d_0^{EE}(\cdot)$ will be discussed shortly.) The density (3.9) forms an exponential family, which means it can be written as

$$\mathbf{u} = \mathbf{z}^{2/q} \sim \exp(\boldsymbol{\eta}_\lambda' \mathbf{u} - \psi_\lambda) d_0^{(p,q)}(\mathbf{u}),$$

where $d_0^{(p,q)}(\mathbf{u})$ is the normalizing (carrier) density, $\boldsymbol{\eta}_\lambda = -C_0(c_q \mathbf{b}_\lambda^{1/q})^p$ is the natural parameter vector and ψ_λ is the cumulant generating function given by

$$\psi_\lambda = \psi(\boldsymbol{\eta}_\lambda) = \begin{cases} -C_0 \sum_i \frac{p}{p-1} (c_q b_{\lambda i}^{1/q})^{p-1} & \text{if } p > 1 \\ -C_0 \sum_i \log b_{\lambda i}^{1/q} & \text{if } p = 1 \end{cases}. \tag{3.12}$$

Because of the one-to-one correspondence between a density and its cumulant generating function, (3.12) in fact completely determines the distribution. For instance, (i) the C_p marginal density (3.10) is inverse Gaussian, which is first shown in Efron (2001); (ii) in the EE marginal distribution (3.11), the carrier density $d_0^{EE}(u_i)$ follows a positive stable law with order 1/3, and consequently, its exponential tilt gives the EE marginal density. [For reference concerning stable laws, see Feller (1971).] The cumulant generating function (3.12) furthermore determines the mean $\boldsymbol{\mu}_\lambda^{(p,q)} = E_{p,q}\{\mathbf{u}\}$ and the covariance matrix $V_\lambda^{(p,q)}$ of the density (3.9) in a simple way: for all $p \geq 1, q \geq 1$,

$$\boldsymbol{\mu}_\lambda^{(p,q)} = \frac{\partial \psi(\boldsymbol{\eta}_\lambda)}{\partial \boldsymbol{\eta}_\lambda} = 1/(c_q \mathbf{b}_\lambda^{1/q}) \tag{3.13}$$

$$V_\lambda^{(p,q)} = \frac{\partial^2 \psi(\boldsymbol{\eta}_\lambda)}{\partial \boldsymbol{\eta}_\lambda^2} = \operatorname{diag}\left(\frac{1}{C_0 p} (c_q \mathbf{b}_\lambda^{1/q})^{-(p+1)}\right). \tag{3.14}$$

Notice that the right hand side of (3.13) only depends on the value of q, which reveals an interesting feature about the (p, q) marginal density (3.9): different (p, q)-densities sharing the same q value have the same expectation. We conclude this section with another important property of the density family (3.9): Fisher consistency. Suppose $\mathbf{u} = \mathbf{z}^{2/q}$ happens to take on the value of its expectation $\boldsymbol{\mu}_{\lambda_0}^{(p,q)}$ for some λ_0, then the (p, q)-criterion would choose λ_0 to be the smoothing parameter. That is

$$\hat{\lambda}^{(p,q)}\left(1/(c_q \mathbf{b}_{\lambda_0}^{1/q})\right) = \lambda_0. \tag{3.15}$$

This property can be easily seen by observing that $\frac{\partial}{\partial\lambda} l^{(p,q)}\left(1/(c_q \mathbf{b}_{\lambda}^{1/q})\right)\Big|_{\lambda=\lambda_0} = 0$.

4. Marginal Bayesian efficiency

The marginal Bayesian interpretation says that the GML, C_p, and EE criteria are MLE's under the densities (3.5), (3.10) and (3.11) respectively. It follows that if the data indeed come from these distributions, then using GML, C_p, and EE respectively would be most efficient. Now an interesting question is: What would happen to a specific criterion, if the data, instead of coming from the density of which it is MLE, actually follow some other distribution?

In this section, we use our earlier results on the eigenvalues to study this problem, namely the relative Bayesian efficiency. Suppose the data come from the marginal density corresponding to criterion (p_1, q) [see (3.8)], but the (p_2, q)-criterion is used to estimate the smoothing parameter. Then because we are not using the MLE $\hat{\lambda}^{(p_1,q)}$ of the (p_1, q)-density, the loss of efficiency is expected. Investigating the loss of efficiency under various situations would help us understand and compare the robustness of different selection criteria.

Theorem 4.1. *Under the (p_1, q)-density, the estimator $\hat{\lambda}^{(p_2,q)}$ satisfies*

$$\left[\frac{q^2 c_q^{1-p_1}\lambda^2}{C_0 p_1}\frac{\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p_2-p_1-1)/q}}{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q})^2}\right]^{-1/2}(\hat{\lambda}^{(p_2,q)} - \lambda) \Longrightarrow N(0,1), \text{ as } n \to \infty.$$

In particular, the asymptotic variance of $\hat{\lambda}^{(p_2,q)}$ under the (p_1, q)-density is

$$\mathrm{var}_{p_1,q}(\hat{\lambda}^{(p_2,q)}) \approx \frac{q^2 c_q^{1-p_1}\lambda^2}{C_0 p_1}\frac{\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p_2-p_1-1)/q}}{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q})^2}. \tag{4.1}$$

We give a heuristic proof here, which itself reveals some interesting features about the (p, q)-density family. For a rigorous proof, see the appendix.

Heuristic Proof. Since the (p_2, q)-criterion chooses $\hat{\lambda}^{(p_2,q)}$ as the minimizer of $l_{\lambda}^{(p_2,q)}(\mathbf{u})$ where $\mathbf{u} = \mathbf{z}^{2/q}$, it must satisfy the normal equation

$$\frac{\partial}{\partial\lambda}\left[l_{\lambda}^{(p_2,q)}(\mathbf{u})\right]\Big|_{\lambda=\hat{\lambda}^{(p_2,q)}} = 0.$$

Applying the implicit function theorem, we can calculate the delta-influence of u_i on $\hat{\lambda}^{(p_2,q)}$

$$\frac{\partial \hat{\lambda}^{(p_2,q)}}{\partial u_i} = -\left[(\frac{\partial^2}{\partial \lambda^2} l_\lambda(\mathbf{u}))^{-1} \frac{\partial^2}{\partial u_i \partial \lambda} l_\lambda(\mathbf{u}) \right]\Big|_{\lambda=\hat{\lambda}^{(p_2,q)}},$$

which, by some simple but tedious algebra, is

$$\frac{\partial \hat{\lambda}^{(p_2,q)}}{\partial u_i} = -\left[\lambda (Q_\lambda(\mathbf{u}))^{-1} a_{\lambda i} (c_q b_{\lambda i}^{1/q})^{p_2}\right]\Big|_{\lambda=\hat{\lambda}^{(p_2,q)}}$$

where $Q_\lambda(\mathbf{u}) = \sum_i a_{\lambda i} (c_q b_{\lambda i}^{1/q})^{p_2-1} \left\{ \frac{1}{q} a_{\lambda i} + [(1+\frac{p_2}{q}) a_{\lambda i} - 2](c_q b_{\lambda i}^{1/q} u_i - 1) \right\}$.

Now taking a first order Taylor expansion on $\hat{\lambda}^{(p_2,q)}$ around λ, the smoothing parameter under the Bayesian model that generates the data, we obtain the approximation

$$\begin{aligned}\hat{\lambda}^{(p_2,q)} - \lambda &\approx \sum_i \frac{\partial \hat{\lambda}^{(p_2,q)}}{\partial u_i}\Big|_{\mathbf{u}=1/(c_q \mathbf{b}_{\lambda i}^{1/q})} (u_i - \frac{1}{c_q b_{\lambda i}^{1/q}}) \\ &= -\frac{q c_q \lambda}{\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q}} \sum_i a_{\lambda i} b_{\lambda i}^{p_2/q} (u_i - \frac{1}{c_q b_{\lambda i}^{1/q}}). \qquad (4.2)\end{aligned}$$

Note that in deriving this approximation, we have used the property of Fisher consistency (3.15).

Since the data are sampled from the (p_1, q)-density, (3.13) and (3.14) say that each individual u_i independently has mean $1/(c_q b_{\lambda i}^{1/q})$, and variance $\mathrm{var}_{p_1,q}(u_i) = \frac{1}{C_0 p_1} (c_q b_{\lambda i}^{1/q})^{-(p_1+1)}$. It follows that the right hand side of (4.2) has mean 0 and variance

$$\frac{q^2 c_q^{1-p_1} \lambda^2}{C_0 p_1} \frac{\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p_2-p_1-1)/q}}{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q})^2}$$

So from the approximation, the variance of $\hat{\lambda}^{(p_2,q)}$ under the (p_1, q)-density is

$$\mathrm{var}_{p_1,q}(\hat{\lambda}^{(p_2,q)}) \approx \frac{q^2 c_q^{1-p_1} \lambda^2}{C_0 p_1} \frac{\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p_2-p_1-1)/q}}{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q})^2}.$$

and

$$\left[\frac{q^2 c_q^{1-p_1} \lambda^2}{C_0 p_1} \frac{\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p_2-p_1-1)/q}}{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q})^2} \right]^{-1/2} (\hat{\lambda}^{(p_2,q)} - \lambda) \Longrightarrow N(0,1), \text{ as } n \to \infty.$$

□

An important application of this theorem is

Corollary 4.2. *Under the (p_1, q)-density, the asymptotic efficiency ratio of $\hat{\lambda}^{(p_2,q)}$ relative to the MLE $\hat{\lambda}^{(p_1,q)}$ is given by*

$$E(p_1, q; p_2, q) = \lim_{n\to\infty} \frac{\mathrm{var}_{p_1,q}(\hat{\lambda}^{(p_2,q)})}{\mathrm{var}_{p_1,q}(\hat{\lambda}^{(p_1,q)})} = \lim_{n\to\infty} \frac{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p_2-p_1-1)/q})(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_1-1)/q})}{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q})^2}. \quad (4.3)$$

As an example, consider the case that the data come from the GML density (where $p = 1, q = 1$), but C_p ($p = 2, q = 1$) is used to estimate the smoothing parameter. Then (4.3) tells us that the relative efficiency ratio

$$E(1, 1; 2, 1) = \lim_{n\to\infty} \frac{(\sum_i a_{\lambda i}^2 b_{\lambda i}^2)(\sum_i a_{\lambda i}^2)}{(\sum_i a_{\lambda i}^2 b_{\lambda i})^2}.$$

When $(x_1, x_2, \ldots x_n)$ are equally spaced with $\delta = x_{i+1} - x_i = n^{-\rho}$, $0 < \rho \le 1$, applying the result of Theorem 2.3 gives

$$E(1, 1; 2, 1) = \frac{B(\frac{7}{4}, \frac{9}{4})B(\frac{7}{4}, \frac{1}{4})}{B(\frac{7}{4}, \frac{5}{4})^2} = \frac{10}{3}. \quad (4.4)$$

Remark 2. Wahba (1985) compares GML and C_p (GCV) under the Bayesian setting and conjectures that they perform similarly. Stein (1990) further considers the above problem — for equally spaced $(x_1, x_2, \ldots x_n)$ how much the loss of efficiency would be when C_p (GCV) is used and in fact the GML Bayesian model is true. He conjectured the above ratio of $10/3$. Theorem 2.3 enables us to rigorously prove the conjecture.

The result of (4.4) can be strengthened (by using Theorem 2.3) to

$$\max_{1\le p\le 2} E(p, 1; 2, 1) = \max_{1\le p\le 2} \frac{B(\frac{7}{4}, \frac{13}{4} - p)B(\frac{7}{4}, p - \frac{3}{4})}{B(\frac{7}{4}, \frac{5}{4})^2} = \frac{10}{3} \quad (4.5)$$

In other words, the use of C_p under the $(p, 1)$ density ($1 \le p \le 2$) would encounter a maximum loss of efficiency of $10/3$, which actually occurs at the GML density.

The C_p criterion also has a marginal Bayesian interpretation (3.10). It is, thus, interesting to ask the question the other way round — what would happen if the data are sampled from the C_p density but GML is used to estimate the smoothing parameter? By (4.3), the relative efficiency is equal to

$$E(2, 1; 1, 1) = \lim_{n\to\infty} \frac{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{-1})(\sum_i a_{\lambda i}^2 b_{\lambda i})}{(\sum_i a_{\lambda i}^2)^2}.$$

For $\delta = x_{i+1} - x_i = n^{-\rho}$, $0 < \rho \le 1$, the summations $\sum_i a_{\lambda i}^2 b_{\lambda i}$ and $\sum_i a_{\lambda i}^2$ have the asymptotic order $O(n^{(4-3\rho)/4})$ according to Theorem 2.3, while $\sum_i a_{\lambda i}^2 b_{\lambda i}^{-1}$ is

of the order $O(n^{4-3\rho})$ by Theorem 2.5. We, hence, have an interesting result

$$E(2,1;1,1) = \lim_{n\to\infty} \frac{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{-1})(\sum_i a_{\lambda i}^2 b_{\lambda i})}{(\sum_i a_{\lambda i}^2)^2} = \infty. \tag{4.6}$$

That is, if the data indeed come from the C_p marginal density, then using GML asymptotically would not work at all.

A careful reader might notice that in the derivation of (4.3), we assume that the density from which the data come and the criterion used to choose the smoothing parameter share the same q value. This assumption is needed for studying the relative efficiency. The reason hinges on the fact that in order for the estimator $\hat{\lambda}^{(p_2,q_2)}$ to be consistent under density-(p_1, q_1), we must have $q_1 = q_2$, which can be seen from the Taylor approximation (4.2), as it relies on two crucial properties: (i) $\hat{\lambda}^{(p_2,q)}\left(1/(c_q \mathbf{b}_\lambda^{1/q})\right) = \lambda$; and (ii) under the (p_1, q_1)-density, $1/(c_q \mathbf{b}_\lambda^{1/q})$ is the expectation of $\mathbf{u}$.

If $q_1 \neq q_2$, although we can still expand $\hat{\lambda}^{(p_2,q_2)}(\mathbf{u})$ around $\mathbf{u} = 1/(c_{q_2} \mathbf{b}_\lambda^{1/q_2})$, however since $1/(c_{q_2} \mathbf{b}_\lambda^{1/q_2})$ is no longer the expectation of $\mathbf{u}$ under the (p_1, q_1)-density, $\hat{\lambda}^{(p_2,q_2)}$ could not be consistent. GML and Cp, both having $q = 1$, provide an example of this mutual consistency as we have seen. C_p and EE provide a different example — using C_p on data sampled from EE density would not give a consistent estimate for λ; conversely, using EE on data sampled from the C_p density would not be consistent either.

Parallel to (4.6) and (4.5), we have the following result regarding the maximum loss of efficiency of the EE criterion (where $p = \frac{3}{2}$, $q = \frac{3}{2}$) in the case of equally spaced observations

$$\begin{aligned}
\max_{1\le p\le 2} E(p, \frac{3}{2}; \frac{3}{2}, \frac{3}{2}) &= \max_{1\le p\le 2} \lim_{n\to\infty} \frac{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(4-2p)/3})(\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p-2)/3})}{(\sum_i a_{\lambda i}^2 b_{\lambda i}^{1/3})^2} \\
&= \max_{1\le p\le 2} \frac{B(\frac{7}{4}, \frac{19}{12} - \frac{2}{3}p) B(\frac{7}{4}, \frac{2}{3}p - \frac{5}{12})}{B(\frac{7}{4}, \frac{7}{12})^2} \\
&= 1.543
\end{aligned} \tag{4.7}$$

Comparing (4.7) with (4.6) and (4.5), it can be seen that the when each estimator (GML, C_p, EE) is used on data sampled from its own consistent density family, the maximum loss of efficiency of the EE criterion is the much smaller 1.543, compared with 10/3 of C_p and ∞ of GML.

Remark 3. Efron (2001) investigates the selection criteria family $\{\hat{\lambda}^{(p,1)} : p \geq 1\}$, where it is shown that having $p > 2$ would result in a very unstable criterion. For this reason, in (4.5) and (4.7) we confine our attention on $1 \leq p \leq 2$.

5. General sampling schemes of the design points

In the study so far we have focused on equally spaced design points $(x_1, x_2, \ldots, x_n)$. In this section, we will study the case of unequally spaced design points to complement the results of Sections 2 and 4. More specifically, suppose $(x_1, x_2, \ldots, x_n)$

are drawn from a continuous distribution on an interval $[\alpha, \beta]$ such that

$$x_i = G^{-1}(\frac{2i-1}{2n}), \tag{5.1}$$

where G is the c. d. f. of the distribution. Let $p(x)$ be the density function of G. For regularity purposes, suppose also that the constant

$$C^* = \int_\alpha^\beta p^{1/4}(x)dx \tag{5.2}$$

is finite and positive. With this setting, one naturally wonders to what extent the previous results remain valid. Quite interestingly, we shall see that all the early results stay essentially intact. In particular, we first note that Theorem 4.1 and Corollary 4.2 do not require $(x_1, x_2, \ldots, x_n)$ to be equally spaced, and hence are carried through under (5.1).

To establish the counterparts of Theorems 2.3 and 2.5 for the setting of (5.1), we need the following handy result of Speckman (1983, 1985): Let l_n and u_n be two sequence (depending on n) such that $l_n \to \infty$, $u_n = o(n^{2/5})$, then for $l_n \le i \le u_n$,

$$k_i = (\frac{\pi}{C^*})^4 \frac{i^4}{n}(1+\epsilon_n), \tag{5.3}$$

where $\lim_{n\to\infty} \epsilon_n = 0$. Using this result, we have the parallel results of Theorems 2.3 and 2.5, whose proofs are deferred to the Appendix.

Theorem 5.1. *Under the sampling scheme (5.1), suppose both* $\lambda \log n \to \infty$, *and* $\frac{n^{1-\varepsilon}}{\lambda} \to \infty$ *for some* $\varepsilon > 0$, *then for all real numbers* $r > \frac{5}{4}$ *and* $s > -\frac{1}{4}$,

$$\sum_{i=3}^{n} a_{\lambda i}^r (1-a_{\lambda i})^s = \frac{C^*}{4\pi} B(r-\frac{1}{4}, s+\frac{1}{4})(\frac{n}{\lambda})^{1/4} + o\left((\frac{n}{\lambda})^{1/4}\right),$$

where the constant C^* *determined by the sampling distribution is given by (5.2).*

Theorem 5.2. *For the sampling scheme (5.1), suppose both* $\lambda \log n \to \infty$, *and* $\frac{n^{1-\varepsilon}}{\lambda} \to \infty$ *for some* $\varepsilon > 0$, *then for all* $r > \frac{1}{4}$ *and* $s < -\frac{1}{4}$,

$$\sum_{i=3}^{n} a_{\lambda i}^r (1-a_{\lambda i})^s \ge O\left((\frac{n^4\delta^3}{\lambda})^{-s-\eta}\right), \text{ for any } \eta > 0.$$

Remark 4. Theorems 5.1 and 5.2, instead of containing Theorems 2.3 and 2.5 as special cases, complement them in that (i) Theorems 5.1 and 5.2 require stronger conditions, and (ii) Theorems 2.3 and 2.5 work on the general setting of $\delta = x_{i+1} - x_i = n^{-\rho}$ for $0 < \rho \le 1$, which is not covered by Theorems 5.1 and 5.2 unless $\rho = 1$.

Applying Theorems 5.1 and 5.2 to Corollary 4.2 and noting that the constant C^* cancels out in the ratios, we have

$$\begin{aligned} E(1,1;2,1) &= \max_{1\le p\le 2} E(p,1;2,1) = \frac{10}{3} \\ E(2,1;1,1) &= \infty \\ \max_{1\le p\le 2} E(p,\frac{3}{2};\frac{3}{2},\frac{3}{2}) &= 1.543. \end{aligned} \tag{5.4}$$

It is noteworthy that the relative efficiency ratios remain the same no matter how the design points $(x_1, x_2, \ldots, x_n)$ are placed, making our results quite general in nature.

6. Discussion

This paper studies the Bayesian large sample properties of the selection criteria, where, in addition to proving and strengthening Stein's (1990) conjecture to general sampling schemes, we obtain the interesting results (4.6) and (5.4), which say that GML can asymptotically perform poorly when its underlying Bayesian structure is violated. This result parallels that of Speckman and Sun (2001), where the performances of C_p (GCV) and GML are compared under the frequentist setting. We also consider the EE criterion and show that its maximum loss of efficiency is 1.543, which parallels the small sample analysis of Kou and Efron (2002) and suggests the robustness of the EE criterion. It is worth noting that in the study we consider both equally spaced and unequally spaced observation, and thus the results obtained are of general applicability.

In this paper, working on cubic smoothing splines, the efficiency of different selection criteria, in terms of estimating the smoothing parameter, is considered. Several authors (Speckman, 1983 and 1985, and Stein, 1993) have studied higher and flexible order smoothing splines. We expect, at the expense of more complicated calculation, the conclusion of the present paper could be qualitatively extended to these cases as well, since the well established results (Wahba, 1990, Speckman, 1983 and 1985, and Stein, 1990 and 1993) indicate that in the sense of general properties cubic smoothing splines are representative. Kou and Efron (2002) also study the relationship between estimating the curve and estimating the smoothing parameter, and suggest that there is a second order connection between the two and that comparing different selection criteria based on their performance of estimating the smoothing parameter is more sensitive. Therefore in terms of estimating the curve, we expect that under the GML Bayesian model, GML works better than C_p to the second order, which agrees with the general conclusion of Stein (1990), and that to the second order the EE criterion would behave more robustly than GML and C_p as (4.7) and (5.4) indicate.

Appendix: Detailed proofs

Proof of Proposition 2.1. We first note that both $\phi(\theta)$ and $\omega(\theta)$ defined in (2.3) and (2.4) are strictly increasing functions of θ for $\theta \in (0, \pi)$. In addition, $\phi(\theta) > 0$

and $\omega(\theta) \in (0, \pi/2)$ for $\theta \in (0, \pi)$. Taking $\theta = \frac{i-2}{n}\pi$ in (2.2) gives

$$F(\frac{i-2}{n}\pi) = (-1)^i[1-(-1)^i e^{-n\phi_i}]^2 \cos\omega_i$$

where $\phi_i = \phi(\frac{i-2}{n}\pi)$, and $\omega_i = \omega(\frac{i-2}{n}\pi)$. Similarly, letting $\theta = \frac{i-1}{n}\pi$ yields

$$F(\frac{i-1}{n}\pi) = (-1)^{i+1}[1+(-1)^i e^{-n\phi_{i+1}}]^2 \cos\omega_{i+1}$$

where $\phi_{i+1} = \phi(\frac{i-1}{n}\pi)$, and $\omega_{i+1} = \omega(\frac{i-1}{n}\pi)$. Since $\phi_i > 0$ and $\omega_i \in (0, \pi/2)$ for all $i \geq 3$, we have

$$F(\frac{i-2}{n}\pi) \cdot F(\frac{i-1}{n}\pi) < 0, \text{ for } i \geq 3.$$

This gives $\frac{i-2}{n}\pi \leq \theta_i \leq \frac{i-1}{n}\pi$. The monotone relationship between k_i and θ_i further provides

$$\frac{12\delta^{-3}(1-\cos\frac{i-2}{n}\pi)^2}{2+\cos\frac{i-2}{n}\pi} \leq k_i \leq \frac{12\delta^{-3}(1-\cos\frac{i-1}{n}\pi)^2}{2+\cos\frac{i-1}{n}\pi}, \text{ for } i \geq 3. \qquad \square$$

Proof of Theorem 2.3. Denote $b_{\lambda i} = 1 - a_{\lambda i}$ for convenience. We shall prove the theorem by considering two cases.

Case I: $-\frac{1}{4} < s \leq 0$. In this case, since $a_{\lambda i}^r(1-a_{\lambda i})^s = (\frac{1}{1+\lambda k_i})^r(\frac{\lambda k_i}{1+\lambda k_i})^s$ is a decreasing function of k_i, it follows that

$$\sum_{i=3}^n a_{\lambda i}^r b_{\lambda i}^s \leq \sum_{i=3}^n (\frac{1}{1+\lambda k_i^-})^r(\frac{\lambda k_i^-}{1+\lambda k_i^-})^s \leq \int_2^n (\frac{1}{1+\lambda k_x^-})^r(\frac{\lambda k_x^-}{1+\lambda k_x^-})^s dx, \tag{A.1}$$

where $k_x^- = \frac{1}{n^4\delta^3}(x-2)^4\pi^4(1-\frac{(x-2)^2\pi^2}{18n^2})$. The second inequality is an easy follow-up of a typical graphical argument. Denoting $f_{r,s}(\lambda, k) = (\frac{1}{1+\lambda k})^r(\frac{\lambda k}{1+\lambda k})^s$, and changing the variable in the integral to $t = (\frac{n^4\delta^3}{\lambda})^{-1/4}(x-2)\pi$, the last term of (A.1) becomes

$$\frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_0^{(n^4\delta^3/\lambda)^{-1/4}(n-2)\pi} f_{r,s}(1, t^4(1-\frac{t^2\delta^{3/2}}{18\lambda^{1/2}}))dt. \tag{A.2}$$

Since $r > \frac{1}{4}$, $-\frac{1}{4} < s \leq 0$, $\delta^{-3}\lambda \to \infty$, by the dominated convergence theorem (A.2) converges to $\frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4}\int_0^\infty(\frac{1}{1+t^4})^r(\frac{t^4}{1+t^4})^s dt$, as $n \to \infty$. Similarly

$$\begin{aligned}
\sum_{i=3}^n a_{\lambda i}^r b_{\lambda i}^s &\geq \sum_{i=3}^n f_{r,s}(\lambda, k_i^+) \geq \int_3^{n+1} f_{r,s}(\lambda, \frac{1}{n^4\delta^3}(x-1)^4\pi^4)dx \\
&= \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_{2\pi(n^4\delta^3/\lambda)^{-1/4}}^{\pi(\delta^{-3}\lambda)^{1/4}} (\frac{1}{1+t^4})^r(\frac{t^4}{1+t^4})^s dt \\
&\to \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_0^\infty (\frac{1}{1+t^4})^r(\frac{t^4}{1+t^4})^s dt.
\end{aligned} \tag{A.3}$$

Combining (A.1), (A.2) and (A.3), and noting that $\int_0^\infty (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt = \frac{1}{4} B(r - \frac{1}{4}, s+\frac{1}{4})$ [see formula (3.251.11) of Gradshteyn and Ryzhik (1994)] yield the desired result.

Case II: $s > 0$. Break the sum $\sum_{i=3}^n a_{\lambda i}^r b_{\lambda i}^s$ into three parts

$$\sum_{i=3}^{n} a_{\lambda i}^r b_{\lambda i}^s = \sum_{\{i: k_i^+ < s/(r\lambda)\}} + \sum_{\{i: k_i^- \le s/(r\lambda) \le k_i^+\}} + \sum_{\{i: k_i^- > s/(r\lambda)\}} . \tag{A.4}$$

Note that the index sets $\{i : k_i^+ < s/(r\lambda)\} = \{i : i < 1 + \frac{1}{\pi}\alpha_n\}$ and $\{i : k_i^- > s/(r\lambda)\} = \{i : i > 2 + \frac{1}{\pi}\beta_n\}$, where $\alpha_n = (\frac{s}{r}\frac{n^4\delta^3}{\lambda})^{1/4}$, β_n is the solution of the equation $\beta^4(1 - \frac{\beta^2}{18n^2}) = \frac{s}{r}\frac{n^4\delta^3}{\lambda}$, and $\lim_{n\to\infty} \frac{\beta_n}{\alpha_n} = 1$. It then follows that $\left|\{i : k_i^- \le s/(r\lambda) \le k_i^+\}\right| = o\left((\frac{n^4\delta^3}{\lambda})^{1/4}\right)$, which implies

$$\sum_{\{i: k_i^- \le s/(r\lambda) \le k_i^+\}} a_{\lambda i}^r b_{\lambda i}^s = o\left((\frac{n^4\delta^3}{\lambda})^{1/4}\right). \tag{A.5}$$

On the set $\{k < s/(r\lambda)\}$, $f_{r,s}(\lambda, k)$ is an increasing function of k. So

$$\sum_{\{i: k_i^+ < s/(r\lambda)\}} f_{r,s}(\lambda, k_i^-) \le \sum_{\{i: k_i^+ < s/(r\lambda)\}} a_{\lambda i}^r b_{\lambda i}^s \le \sum_{\{i: k_i^+ < s/(r\lambda)\}} f_{r,s}(\lambda, k_i^+).$$

But by dominated convergence theorem, the upper bound

$$\begin{aligned}
\sum_{\{i: k_i^+ < s/(r\lambda)\}} f_{r,s}(\lambda, k_i^+) &\le \int_3^{2+\frac{1}{\pi}\alpha_n} f_{r,s}(\lambda, \frac{1}{n^4\delta^3}(x-1)^4\pi^4) dx \\
&= \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_{2\pi(n^4\delta^3/\lambda)^{-1/4}}^{\pi(n^4\delta^3/\lambda)^{-1/4}+(s/r)^{1/4}} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt \\
&\to \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_0^{(s/r)^{1/4}} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt
\end{aligned}$$

and the lower bound

$$\begin{aligned}
\sum_{\{i: k_i^+ < s/(r\lambda)\}} f_{r,s}(\lambda, k_i^-) &\ge \int_2^{1+\frac{1}{\pi}\alpha_n} f_{r,s}(\lambda, k_x^-) dx \\
&\to \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_0^{(s/r)^{1/4}} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt,
\end{aligned}$$

where $k_x^- = \frac{1}{n^4\delta^3}(x-2)^4\pi^4(1 - \frac{(x-2)^2\pi^2}{18n^2})$. We thus have

$$\sum_{\{i: k_i^+ < s/(r\lambda)\}} a_{\lambda i}^r b_{\lambda i}^s = \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_0^{(s/r)^{1/4}} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt + o\left((\frac{n^4\delta^3}{\lambda})^{1/4}\right). \tag{A.6}$$

Identical treatment of $\sum_{\{i:k_i^- > s/(r\lambda)\}} a_{\lambda i}^r b_{\lambda i}^s$ yields

$$\sum_{\{i:k_i^- > s/(r\lambda)\}} a_{\lambda i}^r b_{\lambda i}^s = \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_{(s/r)^{1/4}}^{\infty} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt + o\left((\frac{n^4\delta^3}{\lambda})^{1/4}\right) \tag{A.7}$$

Combining (A.5), (A.6) and (A.7) yields

$$\begin{aligned}\sum_{i=3}^{n} a_{\lambda i}^r b_{\lambda i}^s &= \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_{0}^{\infty} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt + o\left((\frac{n^4\delta^3}{\lambda})^{1/4}\right) \\ &= \frac{1}{4\pi} B(r-\frac{1}{4}, s+\frac{1}{4})(\frac{n^4\delta^3}{\lambda})^{1/4} + o\left((\frac{n^4\delta^3}{\lambda})^{1/4}\right). \qquad \square\end{aligned}$$

Proof of Theorem 2.5. Again let $f_{r,s}(\lambda, k) = (\frac{1}{1+\lambda k})^r (\frac{\lambda k}{1+\lambda k})^s$, and $b_{\lambda i} = 1 - a_{\lambda i}$. Since $f_{r,s}(\lambda, k)$ is a decreasing function of k,

$$\sum_{i=4}^{n} a_{\lambda i}^r b_{\lambda i}^s \le \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_{(n^4\delta^3/\lambda)^{-1/4}\pi}^{(n^4\delta^3/\lambda)^{-1/4}(n-2)\pi} f_{r,s}(1, x^4(1 - \frac{x^2\delta^{3/2}}{18\lambda^{1/2}})dx. \tag{A.8}$$

Break the above integral into two parts $\int_{(n^4\delta^3/\lambda)^{-1/4}\pi}^{1}$ and $\int_{1}^{(n^4\delta^3/\lambda)^{-1/4}(n-2)\pi}$. The second part converges to a constant as $n \to \infty$. Let $\alpha_n = 1 - \frac{\delta^{3/2}}{18\lambda^{1/2}}$, the first part is less than

$$\int_{(n^4\delta^3/\lambda)^{-1/4}\pi}^{1} f_{r,s}(1, \alpha_n x^4)dx = \alpha_n^{-1/4} \int_{\pi(n^4\delta^3/\lambda)^{-1/4}\alpha_n^{1/4}}^{\alpha_n^{1/4}} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt,$$

which is $O\left((\frac{n^4\delta^3}{\lambda})^{-s-\frac{1}{4}}\right)$ by L'Hôpital's rule. This together with (A.8) implies $\sum_{i=4}^{n} a_{\lambda i}^r b_{\lambda i}^s \le O\left((\frac{n^4\delta^3}{\lambda})^{-s}\right)$. Note that $a_{\lambda 3}^r b_{\lambda 3}^s = (\frac{1}{1+\lambda k_3})^r (\frac{\lambda k_3}{1+\lambda k_3})^s$ is $O\left((\frac{n^4\delta^3}{\lambda})^{-s}\right)$. Therefore

$$\sum_{i=3}^{n} a_{\lambda i}^r b_{\lambda i}^s = a_{\lambda 3}^r b_{\lambda 3}^s + \sum_{i=4}^{n} a_{\lambda i}^r b_{\lambda i}^s \le O\left((\frac{n^4\delta^3}{\lambda})^{-s}\right).$$

For the lower bound, similar to (A.8), we have

$$\sum_{i=3}^{n} a_{\lambda i}^r b_{\lambda i}^s \ge \frac{1}{\pi}(\frac{n^4\delta^3}{\lambda})^{1/4} \int_{2\pi(n^4\delta^3/\lambda)^{-1/4}}^{(\delta^{-3}\lambda)^{1/4}\pi} f_{r,s}(1, x^4)dx = O\left((\frac{n^4\delta^3}{\lambda})^{-s}\right). \qquad \square$$

Proof of Theorem 4.1. $\hat{\lambda}^{(p_2,q)}$ satisfies the normal equation $\frac{\partial}{\partial\lambda}\,[l_\lambda^{(p_2,q)}(\mathbf{u})]\Big|_{\lambda=\hat{\lambda}^{(p_2,q)}} = 0$, which (by simple algebra) is

$$\left\{\frac{p_2}{q\lambda}\sum_i a_{\lambda i}(c_q b_{\lambda i}^{1/q})^{p_2-1}(c_q b_{\lambda i}^{1/q}u_i - 1)\right\}\Bigg|_{\lambda=\hat{\lambda}^{(p_2,q)}} = 0. \tag{A.9}$$

Letting $s_i(\lambda) = c_q a_{\lambda i} b_{\lambda i}^{p_2/q}$ and $t_i(\lambda) = -a_{\lambda i} b_{\lambda i}^{(p_2-1)/q}$, (A.9) is equivalent to $\sum_i\{s_i(\hat{\lambda}^{(p_2,q)})u_i + t_i(\hat{\lambda}^{(p_2,q)})\} = 0$. Applying a first order Taylor expansion around λ on a coordinate by coordinate basis gives

$$\begin{aligned} 0 &= \sum_i\{s_i(\hat{\lambda}^{(p_2,q)})u_i + t_i(\hat{\lambda}^{(p_2,q)})\} \\ &= \sum_i\{s_i(\lambda)u_i + t_i(\lambda)\} + (\hat{\lambda}^{(p_2,q)} - \lambda)\sum_i\{s_i'(\lambda_i^*)u_i + t_i'(\lambda_i^*)\}, \end{aligned} \tag{A.10}$$

where for each i, λ_i^* lies between $\hat{\lambda}^{(p_2,q)}$ and λ. Note that $\sum_i\{s_i(\lambda)u_i + t_i(\lambda)\}$ is a sum of independent mean-zero random variables. It follows from Lyaponouv's theorem [Billingsley (1995), page 362],

$$\left[\frac{c_q^{1-p_1}}{C_0 p_1}\sum_i a_{\lambda i}^2 b_{\lambda i}^{(2p_2-p_1-1)/q}\right]^{-1/2}\sum_i\{s_i(\lambda)u_i + t_i(\lambda)\} \Longrightarrow N(0,1).$$

Therefore by (A.10) we only need to show that

$$\frac{\sum_i\{s_i'(\lambda_i^*)u_i + t_i'(\lambda_i^*)\}}{\frac{1}{q\lambda}\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q}} \to 1 \text{ in probability,}$$

Denote $T_n = \frac{1}{q\lambda}\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q}$. Note that $s_i'(\lambda) = \frac{c_q}{\lambda}a_{\lambda i}b_{\lambda i}^{p_2/q}(\frac{p_2}{q}a_{\lambda i} - b_{\lambda i})$, $t_i'(\lambda) = -\frac{1}{\lambda}a_{\lambda i}b_{\lambda i}^{(p_2-1)/q}(b_{\lambda i} - \frac{p_2-1}{q}a_{\lambda i})$, and

$$\begin{aligned} &\sum_i\{s_i'(\lambda_i^*)u_i + t_i'(\lambda_i^*)\} - T_n \\ &\quad= \sum_i\{(s_i'(\lambda_i^*) - s_i'(\lambda))u_i + (t_i'(\lambda_i^*) - t_i'(\lambda))\} + \sum_i\{s_i'(\lambda)(u_i - \frac{1}{c_q b_{\lambda i}^{1/q}})\} \\ &\quad= \text{Term A} + \text{Term B}. \end{aligned}$$

We need to show that both $\frac{\text{Term A}}{Tn} \to 0$ and $\frac{\text{Term B}}{Tn} \to 0$ in probability. First consider $\frac{\text{Term A}}{Tn}$. For any $\varepsilon > 0, \theta > 0$,

$$\begin{aligned} P\left(\left|\frac{\text{Term A}}{Tn}\right| > \varepsilon\right) \le\ & P\left(\left|\frac{\text{Term A}}{Tn}\right| > \varepsilon, \left|\hat{\lambda}^{(p_2,q)} - \lambda\right| \le \theta\right) \\ &+ P\left(\left|\hat{\lambda}^{(p_2,q)} - \lambda\right| > \theta\right). \end{aligned}$$

By simple (but tedious) calculation, it can be shown that there exists a constant M such that

$$|s_i''(\tilde{\lambda})| \le M c_q a_{\lambda i} b_{\lambda i}^{p_2/q},\ |t_i''(\tilde{\lambda})| \le M a_{\lambda i} b_{\lambda i}^{(p_2-1)/q} \text{ for all } |\tilde{\lambda}-\lambda| \le \theta.$$

Therefore by the mean value theorem, on the event $\{\left|\hat{\lambda}^{(p_2,q)}-\lambda\right| \le \theta\}$

$$|\text{Term A}| \le M\theta \sum_i (c_q a_{\lambda i} b_{\lambda i}^{p_2/q} u_i + a_{\lambda i} b_{\lambda i}^{(p_2-1)/q}). \tag{A.11}$$

But using the result of Theorem 2.3, the expectation

$$E\{\frac{1}{T_n} M\theta \sum_i (c_q a_{\lambda i} b_{\lambda i}^{p_2/q} u_i + a_{\lambda i} b_{\lambda i}^{(p_2-1)/q})\} = \theta \cdot (2q\lambda M \frac{\sum_i a_{\lambda i} b_{\lambda i}^{p_2/q} u_i}{\sum_i a_{\lambda i}^2 b_{\lambda i}^{(p_2-1)/q}}) \le C\theta,$$

for some constant C, when n is sufficiently large. Thus we have from (A.11)

$$\begin{aligned} P(\left|\frac{\text{Term A}}{Tn}\right| > \varepsilon) &\le P\left(\frac{1}{T_n} M\theta \sum_i (c_q a_{\lambda i} b_{\lambda i}^{p_2/q} u_i + a_{\lambda i} b_{\lambda i}^{(p_2-1)/q}) > \varepsilon\right) \\ &\quad + P(\left|\hat{\lambda}^{(p_2,q)}-\lambda\right| > \theta) \\ &\le \frac{1}{\varepsilon} E\{\frac{1}{T_n} M\theta \sum_i (c_q a_{\lambda i} b_{\lambda i}^{p_2/q} u_i + a_{\lambda i} b_{\lambda i}^{(p_2-1)/q})\} \\ &\quad + P(\left|\hat{\lambda}^{(p_2,q)}-\lambda\right| > \theta) \\ &\le \frac{1}{\varepsilon} C\theta + P(\left|\hat{\lambda}^{(p_2,q)}-\lambda\right| > \theta). \end{aligned}$$

Let $n \to \infty$,

$$\lim_{n\to\infty} \sup P(\left|\frac{\text{Term A}}{Tn}\right| > \varepsilon) \le \frac{1}{\varepsilon} C\theta + \lim_{n\to\infty} P(\left|\hat{\lambda}^{(p_2,q)}-\lambda\right| > \theta) = \frac{1}{\varepsilon} C\theta.$$

The last equality follows by the Fisher consistency of $\hat{\lambda}^{(p_2,q)}$, which is stronger than the ordinary consistency. Now sending $\theta \downarrow 0$ yields $\frac{\text{Term A}}{Tn} \to 0$ in probability. To prove $\frac{\text{Term B}}{Tn} \to 0$ in probability, note that

$$\begin{aligned} P(\frac{\text{Term B}}{Tn} > \varepsilon) &= P(\sum_i s_i'(\lambda)(u_i - \frac{1}{c_q b_{\lambda i}^{1/q}}) > T_n\varepsilon) \\ &\le E\{\exp\left(C_0\theta[\sum_i s_i'(\lambda)(u_i - \frac{1}{c_q b_{\lambda i}^{1/q}}) - T_n\varepsilon]\right)\},\ \forall \theta > 0 \\ &= e^{-C_0\theta\varepsilon T_n} \left(\prod_i e^{-C_0\theta s_i'(\lambda)/(c_q b_{\lambda i}^{1/q})}\right) \prod_i E e^{C_0\theta s_i'(\lambda) u_i}. \end{aligned}$$

The moment generating function of u_i, according to (3.12), is

$$Ee^{\theta u_i} = \exp(C_0 \frac{p_1}{p_1 - 1}\{(c_q b_{\lambda i}^{1/q})^{p_1 - 1} - ((c_q b_{\lambda i}^{1/q})^{p_1} - \frac{\theta}{C_0})^{\frac{p_1-1}{p_1}}\}).$$

Therefore

$$\begin{aligned}
\log P(\frac{\text{Term B}}{Tn} > \varepsilon) &\le -C_0\theta\varepsilon T_n - C_0\theta \sum_i s_i'(\lambda)/(c_q b_{\lambda i}^{1/q}) + \\
&\quad + \sum_i C_0 \frac{p_1}{p_1 - 1}\{(c_q b_{\lambda i}^{1/q})^{p_1 - 1} \\
&\quad -((c_q b_{\lambda i}^{1/q})^{p_1} - \theta s_i'(\lambda))^{\frac{p_1-1}{p_1}}\} \\
&\le -C_0[\theta\varepsilon T_n + \theta \sum_i s_i'(\lambda)/(c_q b_{\lambda i}^{1/q}) \\
&\quad - \frac{p_1}{p_1 - 1}\theta^{\frac{p_1-1}{p_1}} \sum_i |s_i'(\lambda)|^{\frac{p_1-1}{p_1}}]. \qquad \text{(A.12)}
\end{aligned}$$

Theorem 2.3 says that $\varepsilon T_n + \sum_i s_i'(\lambda)/(c_q b_{\lambda i}^{1/q}) = O((\frac{n^4\delta^3}{\lambda})^{1/4}) > 0$ and $\sum_i |s_i'(\lambda)|^{\frac{p_1-1}{p_1}} = O((\frac{n^4\delta^3}{\lambda})^{1/4})$. So for sufficiently large θ, the right hand side of (A.12) is $O((\frac{n^4\delta^3}{\lambda})^{1/4})$ and negative. It follows that $P(\frac{\text{Term B}}{Tn} > \varepsilon) \to 0$, as $n \to \infty$. Applying similar method on the inequality

$$P(\frac{\text{Term B}}{Tn} < -\varepsilon) \le E\{\exp\left(-C_0\theta[\sum_i s_i'(\lambda)(u_i - \frac{1}{c_q b_{\lambda i}^{1/q}}) + T_n\varepsilon]\right)\}, \ \forall\theta > 0,$$

we can show that $P(\frac{\text{Term B}}{Tn} < -\varepsilon) \to 0$, which finally yields $\frac{\text{Term B}}{Tn} \to 0$ in probability. □

Proof of Theorem 5.1. Break the sum $\sum_{i=3}^n a_{\lambda i}^r b_{\lambda i}^s$ into three parts $\sum_{i=3}^n a_{\lambda i}^r b_{\lambda i}^s = \sum_{i=3}^{l_n} + \sum_{l_n}^{u_n} + \sum_{u_n}^n$. Consider the second term first. Using (5.3) provides

$$\begin{aligned}
\sum_{i=l_n}^{u_n} a_{\lambda i}^r b_{\lambda i}^s &= \sum_{l_n}^{u_n} (\frac{1}{1 + \lambda k_i})^r (\frac{\lambda k_i}{1 + \lambda k_i})^s \\
&= \sum_{i=l_n}^{u_n} (\frac{1}{1 + \lambda(\frac{\pi}{C^*})^4 \frac{i^4}{n}(1 + \epsilon_n)})^r (\frac{\lambda(\frac{\pi}{C^*})^4 \frac{i^4}{n}(1 + \epsilon_n)}{1 + \lambda(\frac{\pi}{C^*})^4 \frac{i^4}{n}(1 + \epsilon_n)})^s
\end{aligned}$$

Applying identical arguments as in the proof of Theorem 2.3, it can be seen that $\sum_{l_n}^{u_n} a_{\lambda i}^r b_{\lambda i}^s$ is asymptotically equivalent to $\int_{l_n}^{u_n} (\frac{1}{1+\lambda(\frac{\pi}{C^*})^4 \frac{x^4}{n}(1+\epsilon_n)})^r$

$\times (\frac{\lambda(\frac{\pi}{C^*})^4 \frac{x^4}{n}(1+\epsilon_n)}{1+\lambda(\frac{\pi}{C^*})^4 \frac{x^4}{n}(1+\epsilon_n)})^s dx$, which, after a change of variable $t = [\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} x$, becomes

$$[\frac{n}{\lambda(1+\epsilon_n)}]^{1/4} \frac{C^*}{\pi} \int_{[\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} l_n}^{[\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} u_n} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt. \tag{A.13}$$

Taking $l_n = o(n^{\varepsilon/4})$, $n^{2/5-\varepsilon/4} < u_n = o(n^{2/5})$, the conditions of Theorem 5.1 together with dominated convergence theorem implies that as $n \to \infty$, (A.13) is equivalent to $\frac{C^*}{4\pi} B(r-\frac{1}{4}, s+\frac{1}{4})(\frac{n}{\lambda})^{1/4}$.

Next consider $\sum_{i=3}^{l_n} a_{\lambda i}^r b_{\lambda i}^s$, it is easily seen (from the proof of Theorem 2.3) that for $i \le l_n < n^{\varepsilon/4}$, $a_{\lambda i}^r b_{\lambda i}^s$ is a monotone function of i (for all given $s > -1/4$). Therefore, $\sum_{i=3}^{l_n} a_{\lambda i}^r b_{\lambda i}^s \le l_n \max(a_{\lambda 3}^r b_{\lambda 3}^s, a_{\lambda l_n}^r b_{\lambda l_n}^s)$. Since $a_{\lambda 3}^r b_{\lambda 3}^s = (\frac{1}{1+(\lambda/n)(nk_3)})^r \times (\frac{(\lambda/n)(nk_3)}{1+(\lambda/n)(nk_3)})^s$, using the result of Speckman (1985, Theorem 2.2 of the paper) that $\lim_{n\to\infty} nk_3 = const > 0$ gives $l_n a_{\lambda 3}^r b_{\lambda 3}^s = O(l_n(\frac{\lambda}{n})^s)$. Thus, taking l_n sufficiently small guarantees $l_n a_{\lambda 3}^r b_{\lambda 3}^s = o((\frac{n}{\lambda})^{1/4})$. As for $l_n a_{\lambda l_n}^r b_{\lambda l_n}^s$, applying (5.3) yields $l_n a_{\lambda l_n}^r b_{\lambda l_n}^s = l_n(\frac{1}{1+\frac{\lambda}{n}(\frac{\pi}{C^*}l_n)^4(1+\epsilon_n)})^r (\frac{\frac{\lambda}{n}(\frac{\pi}{C^*}l_n)^4(1+\epsilon_n)}{1+\frac{\lambda}{n}(\frac{\pi}{C^*}l_n)^4(1+\epsilon_n)})^s = O(l_n^{1+4s}(\frac{\lambda}{n})^s) = o((\frac{n}{\lambda})^{1/4})$ for l_n sufficiently small. Hence for all fixed $s > -\frac{1}{4}$, $\sum_{i=3}^{l_n} a_{\lambda i}^r b_{\lambda i}^s = o((\frac{n}{\lambda})^{1/4})$.

Finally consider $\sum_{i=u_n}^{n} a_{\lambda i}^r b_{\lambda i}^s$. Following the proof of Theorem 2.3, it can be seen that for $i \ge u_n$, $a_{\lambda i}^r b_{\lambda i}^s$ is a monotone decreasing function of i. It then follows that $\sum_{u_n}^{n} a_{\lambda i}^r b_{\lambda i}^s \le (n-u_n) a_{\lambda u_n}^r b_{\lambda u_n}^s \le n a_{\lambda u_n}^r b_{\lambda u_n}^s$, which, by (5.3), is $n(\frac{1}{1+\frac{\lambda}{n}(\frac{\pi}{C^*}u_n)^4(1+\epsilon_n)})^r (\frac{\frac{\lambda}{n}(\frac{\pi}{C^*}u_n)^4(1+\epsilon_n)}{1+\frac{\lambda}{n}(\frac{\pi}{C^*}u_n)^4(1+\epsilon_n)})^s = O(n \cdot (\frac{\lambda}{n}u_n^4)^{-r})$. For any fixed $r > 5/4$, because $\lambda \log n \to \infty$, taking $u_n = n^{2/5-\eta/4}$ yields $\sum_{u_n}^{n} a_{\lambda i}^r b_{\lambda i}^s \le O(n(n^{\frac{3}{5}-\eta}\lambda)^{-r}) = O\left((\frac{n}{\lambda})^{1/4}(n^{(\frac{3}{5}(r-\frac{5}{4})-\eta r)/(r-\frac{1}{4})}\lambda)^{\frac{1}{4}-r}\right)$, which is $o((\frac{n}{\lambda})^{1/4})$ if we take η to be sufficiently small, for example $\eta = \frac{3}{10r}(r-\frac{5}{4})$.

The proof is terminated by combining the three parts of the sum. □

Proof of Theorem 5.2. For $l_n \to \infty, u_n = n^{2/5-\varepsilon}$, $\sum_{i=3}^{n} a_{\lambda i}^r (1-a_{\lambda i})^s \ge \sum_{i=l_n}^{u_n} a_{\lambda i}^r b_{\lambda i}^s$, which is $\sum_{i=l_n}^{u_n} (\frac{1}{1+\lambda(\frac{\pi}{C^*})^4 \frac{i^4}{n}(1+\epsilon_n)})^r (\frac{\lambda(\frac{\pi}{C^*})^4 \frac{i^4}{n}(1+\epsilon_n)}{1+\lambda(\frac{\pi}{C^*})^4 \frac{i^4}{n}(1+\epsilon_n)})^s$ by (5.3). Since $a_{\lambda i}^r (1-a_{\lambda i})^s$ is monotone decreasing for i, a simple graphical argument gives

$$\begin{aligned}\sum_{i=l_n}^{u_n} a_{\lambda i}^r b_{\lambda i}^s &\ge \int_{l_n}^{u_n} (\frac{1}{1+\lambda(\frac{\pi}{C^*})^4 \frac{x^4}{n}(1+\epsilon_n)})^r (\frac{\lambda(\frac{\pi}{C^*})^4 \frac{x^4}{n}(1+\epsilon_n)}{1+\lambda(\frac{\pi}{C^*})^4 \frac{x^4}{n}(1+\epsilon_n)})^s dx \\ &= [\frac{n}{\lambda(1+\epsilon_n)}]^{1/4} \frac{C^*}{\pi} \int_{[\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} l_n}^{[\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} u_n} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt.\end{aligned}$$

Note that $\int_1^{[\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} u_n} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt \to \int_0^{\infty} (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt < \infty$; and that $\int_{[\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} l_n}^1 (\frac{1}{1+t^4})^r (\frac{t^4}{1+t^4})^s dt = O(\int_{[\frac{\lambda}{n}(1+\epsilon_n)]^{1/4} \frac{\pi}{C^*} l_n}^1 t^{4s} dt) = O((\frac{n}{\lambda})^{-s-\frac{1}{4}} l_n^{4s+1})$. The desired result follows by taking $l_n = (\frac{n}{\lambda})^{\frac{\eta}{-(4s+1)}}$. □

Acknowledgements. The author thanks Professors Bradley Efron, Iain Johnstone, Trevor Hastie, Rob Tibshirani and Zhiliang Ying for helpful discussion. The author is also grateful to the associate editor and the referees for many insightful suggestions.

References

Akaike, H.: A new look at statistical model identification. IEEE Trans. Auto. Cont. **AU-19**, 716–722 (1974)

Billingsley, P.: Probability and Measure, 3rd ed. Wiley, New York, 1995

Bowman, A., Azzalini, A.: Applied smoothing techniques for data analysis: the kernel approach with *S*-Plus illustrations. Oxford University Press, New York, 1997

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

Culpin, D.: Calculation of cubic smoothing splines for equally spaced data. Numer. Math. **48**, 627–638 (1986)

Demmler, A., Reinsch, C.: Oscillation matrices with spline smoothing. Numer. Math. **24**, 375–382 (1975)

Efron, B.: Selection criteria for scatterplot smoothers. Ann. Statist. **29**, 470–504 (2001)

Eubank, R.: Spline Smoothing and Nonparametric Regression. Marcel Dekker, New York, 1988

Eubank, R.: Nonparametric Regression and Spline Smoothing, 2nd ed, Marcel Dekker, New York, 1999

Fan, J.: Prospects of nonparametric modeling. J. Amer. Statist. Assoc. **95**, 1296–1300 (2000)

Fan, J., Gijbels, I.: Local Polynomial Modelling and Its Applications. Chapman and Hall, London, 1996

Feller, W.: An Introduction to Probability Theory and Its Applications, Vol. II. Wiley, New York, 1971

Gradshteyn, I., Ryzhik, I.: Table of Integrals, Series, and Products. Academic Press, Boston, 1994

Green, P., Silverman, B.: Nonparametric Regression and Generalized Linear Models. Chapman and Hall, London, 1994

Hall, P.: Biometrika century: nonparametrics. Biometrika **88**, 143–165 (2001)

Hall, P., Johnstone, I.: Empirical functionals and efficient smoothing parameter selection (with discussion). J. Roy. Statist. Soc. B **54**, 475–530 (1992)

Härdle, W.: Applied Nonparametric Regression. Cambridge University Press, Cambridge, 1990

Härdle, W., Hall, P., Marron, S.: How far are the optimally chosen smoothing parameters from their optimum? (with discussion.) J. Amer. Statist. Assoc. **83**, 86–101 (1988)

Hastie, T., Tibshirani, R.: Generalized Additive Models. Chapman and Hall, London, 1990

Kimeldorf, G., Wahba, G.: A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. Ann. Math. Statist. **41**, 495–502 (1970)

Kneip, A.: Ordered linear smoothers. Ann. Statist. **22**, 835–866 (1994)

Kou, S.C.: Extended exponential criterion: a new selection procedure for scatterplot smoothers. Ph. D. thesis, Stanford University, 2001

Kou, S.C., Efron, B.: Smoothers and the C_p, GML and EE criteria: A geometric approach. J. Amer. Statist. Assoc. **97**, 766–782 (2002)

Li, K.-C.: Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. Ann. Statist. **14**, 1101–1112 (1986)

Li, K.-C.: Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. Ann. Statist. **15**, 958–975 (1987)

Mallows, C.: Some comments on C_p. Technometrics **15**, 661–675 (1973)

Nussbaum, M: Spline smoothing in regression models and asymptotic efficiency in L_2. Ann. Statist. **13**, 984–997 (1985)

Rosenblatt, M.: Stochastic Curve Estimation. NSF-CBMS Regional Conference Series in Probability and Statistics, Volume 3. IMS, Hayward, 1991
Reinsch, C.: Smoothing by spline functions. Numer. Math. **10**, 177–183 (1967)
Schoenberg, I.: Spline functions and the problem of graduation. Proc. Nat. Acad. Sci. USA. **52**, 947–950 (1964a)
Schoenberg, I.: On interpolation by spline functions and its minimum properties. Internat. Ser. Numer. Anal. **5**, 109–129 (1964b)
Silverman, B.: A fast and efficient cross-validation method for smoothing parameter choice in spline regression. J. Amer. Statist. Assoc. **79**, 584–589 (1984)
Silverman, B.: Some aspects of the spline smoothing approach to nonparametric regression curve fitting (with discussion). J. Roy. Statist. Soc. B **47**, 1–52 (1985)
Simonoff, J.: Smoothing Methods in Statistics. Springer-Verlag, New York, 1996
Speckman, P.: Efficient nonparametric regression with cross-validated smoothing splines. Unpublished manuscript, 1983
Speckman, P.: Spline smoothing and optimal rates of convergence in nonparametric regression models. Ann. Statist. **13**, 970–983 (1985)
Speckman, P., Sun, D.: Asymptotic properties of smoothing parameter selection in spline regression. Preprint, 2001
Stein, M.: A comparison of generalized cross validation and modified maximum likelihood for estimating the parameters of a stochastic process. Ann. Statist. **18**, 1139–1157 (1990)
Stein, M.: Spline smoothing with an estimated order parameter. Ann. Statist. **21**, 1522–1544 (1993)
Utreras, F.: Cross-validation techniques for smoothing spline functions in one or two dimensions. In: Smoothing Techniques for Curve Estimation, (T. Gasser, M. Rosenblatt, ed.), Springer-Verlag, Heidelberg, 1979, pp. 196–232
Utreras, F.: Sur le choix du parametre d'ajustement dans le lissage par fonctions spline. Numer. Math. **34**, 15–28 (1980)
Utreras, F.: Optimal smoothing of noisy data using spline functions. SIAM J. Sci. and Statist. Comput. **2**, 349–362 (1981)
Utreras, F.: Boundary effects on convergence rates for Tikhonov regularization. J. Approx. Theor. **54**, 235–249 (1988)
Wahba, G.: Smoothing noisy data by spline functions. Numer. Math. **24**, 383–393 (1975)
Wahba, G.: Optimal smoothing of density estimates. In: Classification and Clustering (J. Van Ryzin, ed.), Academic Press, New York, 1977a, pp. 423–458
Wahba, G.: A survey of some smoothing problems and the method of generalized cross-validation for solving them. In: Applications of Statistics (P. R. Krishnaiah, ed.), North Holland, Amsterdam. 1977b, pp. 507–523
Wahba, G.: A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. Ann. Statist. **13**, 1378–1402 (1985)
Wahba, G.: Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics, 59. SIAM, Philadelphia, 1990
Wecker, W., Ansley, C.: The signal extraction approach to nonlinear regression and spline smoothing. J. Amer. Statist. Assoc. **78**, 81–89 (1983)
Whittaker, E.: On a new method of graduation. Proc. Edinburgh Math. Soc. **41**, 63–75 (1923)