Electronic Journal of Statistics
Vol. 8 (2014) 159–183
ISSN: 1935-7524
DOI: 10.1214/14-EJS879

Asymptotic optimality of a multivariate version of the generalized cross validation in adaptive smoothing splines

Heeyoung Kim[*]

Department of Industrial & Systems Engineering
KAIST (Korea Advanced Institute of Science and Technology)
e-mail: heeyoungkim@kaist.ac.kr

and

Xiaoming Huo

H. Milton Stewart School of Industrial & Systems Engineering
Georgia Institute of Technology
e-mail: xiaoming@isye.gatech.edu

Abstract: We consider an adaptive smoothing spline with a piecewise-constant penalty function $\lambda(x)$, in which a univariate smoothing parameter λ in the classic smoothing spline is converted into an adaptive multivariate parameter $\boldsymbol{\lambda}$. Choosing the optimal value of $\boldsymbol{\lambda}$ is critical for obtaining desirable estimates. We propose to choose $\boldsymbol{\lambda}$ by minimizing a multivariate version of the generalized cross validation function; the resulting estimator is shown to be consistent and asymptotically optimal under some general conditions—i.e., the counterparts of the nice asymptotic properties of the generalized cross validation in the ordinary smoothing spline are still provable. This provides theoretical justification of adopting the multivariate version of the generalized cross validation principle in adaptive smoothing splines.

MSC 2010 subject classifications: Primary 62G20; secondary 62G08.
Keywords and phrases: Adaptive smoothing splines, asymptotic optimality, generalized cross validation.

Received December 2013.

1. Introduction

We consider the problem of estimating an unknown function $f(\cdot)$ given observations

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

where the design points x_i follow a strictly positive continuous density function and ϵ_i are independent random noise with mean 0 and unknown variance σ^2. Without loss of generality, we assume that the domain of f is $[0, 1]$. Smoothing spline is one of the most popular methods for estimating f. Let $f^{(m)}$ denote the

[*] Corresponding author.

mth derivative of f. The smoothing spline estimator is the unique solution of the following problem

$$\min_{f\in W_2^m[0,1]} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda \int_0^1 \{f^{(m)}(x)\}^2 dx, \tag{2}$$

where $W_2^m[0,1]$ is the mth order Sobolev space defined as $\{f : f^{(j)}$ is absolutely continuous for $j = 0, \dots, m-1,$ and $f^{(m)} \in L_2[0,1]\}$ where $L_2[0,1]$ is the space of squared integrable functions, and λ is a smoothing parameter that controls the trade-off between smoothness of the estimated function (the second term) and the goodness of fit (the first term). Large values of λ produce smoother functions while smaller values produce more wiggly functions. For the automatic choice of λ, many procedures have been proposed, including cross validation (CV) (Stone, 1974), generalized cross validation (GCV) (Craven and Wahba, 1979), and Mallow's C_p (Mallows, 1973).

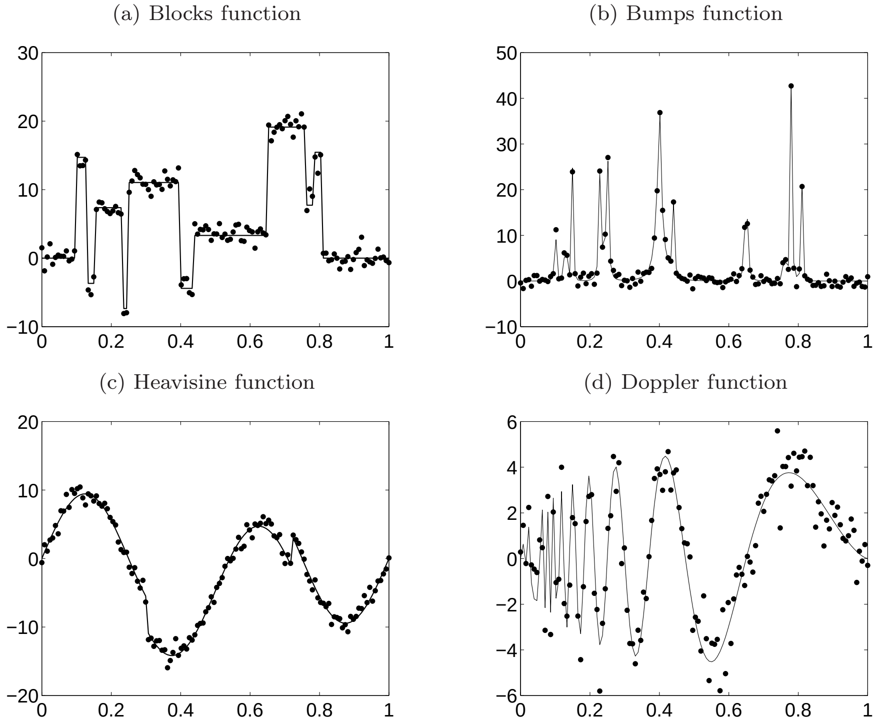
Even though smoothing splines have been shown to perform well in many examples, if the underlying function is spatially nonhomogeneous, traditional smoothing splines will fail to capture the varying degrees of smoothness properly. In practice, there are various types of functions with varying smoothness, and four popular scenarios in Donoho and Johnstone (1995) are illustrated in Fig. 1. The functions in Fig. 1 change rapidly in some regions while being smooth in others. If we choose the global smoothing parameter to be relatively small, the resulting spline estimate will describe the function well in the regions of large variations, however it will under-smooth in other regions. On the other hand, if the global smoothing parameter is chosen to be relatively large, then the estimated function will be over-smoothed in the regions of large variations. This indicates that in fitting functions with varying roughness, using a global smoothing parameter is not sufficient.

To resolve such a problem and attain more flexible estimation of the function, there have been attempts to allow for the smoothing parameter to vary adaptively with x (Abramovich and Steinberg, 1996; Pintore, Speckman and Holmes, 2006; Storlie, Bondell and Reich, 2010; Liu and Guo, 2010; Kim and Huo, 2012; Wang, Du and Shen, 2013). Instead of (2), the following minimization problem has been considered in the framework of smoothing splines:

$$\min_{f\in W_2^m[0,1]} \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \int_0^1 \lambda(x)\{f^{(m)}(x)\}^2 dx, \tag{3}$$

where $\lambda(x)$ is a *variable* smoothing parameter—a function of x.

In the framework in (3), one popular approach in deriving the solution is to discretize $\lambda(x)$. That is, $\lambda(x)$ is approximated by a step function, i.e., a piecewise constant function. Pintore, Speckman and Holmes (2006) assumed an equal-size piecewise structure for $\lambda(x)$. The number of jumps and the jump locations need to be prespecified. Then the step function is estimated by minimizing the multivariate version of the generalized cross validation. Liu and Guo (2010) extended the work of Pintore, Speckman and Holmes (2006). They also assume a step function for $\lambda(x)$, but the segmentation is data-driven. The number of jumps and the


(a) Blocks function

(b) Bumps function

(c) Heavisine function

(d) Doppler function

Fig 1. *Illustration of four widely-studied functions with varying degrees of smoothness: (a) Blocks function, (b) Bumps function, (c) Heavisine function, (d) Doppler function. Solid curve is the true function and dots are noisy observations.*

jump locations are chosen based on the structure of data. Then the step function is estimated by maximizing the generalized likelihood. In these methods, the optimal choice of the multivariate smoothing parameter and the associated asymptotic properties have not been studied. Wang, Du and Shen (2013) develop a general framework for asymptotic analysis of adaptive smoothing splines with the use of the Green's function. However, they require some highly technical mathematical knowledge, and the assumptions made in their paper to establish theoretical results seem to be very strong, e.g., one assumption is that the underlying true function is $2m$-times continuously differentiable.

With the assumption that $\lambda(x)$ is approximated by a step function, we study an optimal choice for $\lambda(x)$. We consider a discretized version of $\lambda(x)$ with knots s_i, $i = 1, \ldots, k$: $\lambda(x) \equiv \eta_i$ for $x \in [s_i, s_{i+1})$, where s_i satisfy $0 = s_0 < s_1 < \cdots < s_k < s_{k+1} = 1$ and $s_i \in \{x_1, \ldots, x_n\}$. Then we need to estimate $\eta = (\eta_1, \eta_2, \ldots, \eta_k)$ whose dimension is k. For the optimal choice of η, we propose to use the multivariate version of the generalized cross validation (mGCV). We show that under some moderate conditions, if we choose η by minimizing the mGCV, then the resulting estimate is *consistent* in the sense that the true loss tends to zero as the sample size goes to infinity, and is *asymptotically optimal* in the sense that it achieves the smallest possible loss in probability when the sample size goes to infinity. Our theoretical analysis depends only on simple linear algebra and calculus. In approximating $\lambda(x)$ by a step function, we allow the discretization of $\lambda(x)$ to be as flexible as possible: our theoretical results cover all possible cases of step functions assumed for $\lambda(x)$ with flexible step size and step number, regardless of the location of $s_1, \ldots, s_k$ and the number of k.

For the ordinary smoothing splines (i.e., with univariate λ), the optimal choice of λ has been extensively studied. In particular, it has been shown that if one chooses λ via GCV, the resulting estimate has nice asymptotic properties (Craven and Wahba, 1979; Li, 1985, 1986). However, for the adaptive smoothing spline, similar study on the optimal choice of $\lambda(x)$ (or its discretized version—multivariate smoothing parameter) has not appeared. The main contribution of this paper is to show that the adaptive smoothing spline estimator with the mGCV choice of the multivariate smoothing parameter has the same asymptotic properties as the ones established for the ordinary smoothing splines—consistency and asymptotic optimality. This paper focuses on the theoretical study of the mGCV. Nevertheless, we present some numerical study of the mGCV in Section 4 to show the practical advantage of the mGCV.

It is worth mentioning another popular approach to solve (3). Instead of considering a step function, there have been attempts to assume a particular continuously varying penalty function for $\lambda(x)$. Abramovich and Steinberg (1996) assumed that $\lambda(x)$ is proportional to $(f^{(2)}(x))^{-2}$, and Storlie, Bondell and Reich (2010) assumed that $\lambda(x)$ is proportional to $(|f^{(2)}(x)| + \delta)^{-2\gamma}$ that allows more flexibility due to two more tuning parameters δ and γ. Kim and Huo (2012) derived an asymptotically optimal local penalty function for $\lambda(x)$.

The rest of the paper is organized as follows. In Section 2, we review ordinary smoothing splines and the justification of GCV. In Section 3, we study asymptotic properties of the mGCV in choosing the multivariate smoothing parameter in adaptive smoothing splines. In Section 4, we show the practical effectiveness of the mGCV via simulations. We conclude in Section 5.

2. A review of ordinary smoothing splines and GCV

We review ordinary smoothing splines and the justification of GCV in Section 2.1 and Section 2.2, respectively.

2.1. A review of ordinary smoothing splines

We briefly review ordinary smoothing splines. For more details, we refer to Green and Silverman (1994) and Eubank (1999). Throughout this paper, we focus on the cubic smoothing splines (i.e., with $m = 2$ in (2)) which are the most commonly used splines in practice. Using the cubic smoothing splines, f is estimated by minimizing the following objective function:

$$J(\lambda; f) = \sum_{i=1}^{n} [y_i - f(x_i)]^2 + \lambda \int_0^1 \{f^{(2)}(x)\}^2 dx. \tag{4}$$

Let $\mathbf{y} = (y_1, \ldots, y_n)^T$, $\mathbf{f} = (f(x_1), \ldots, f(x_n))^T$, $\delta_i = f^{(2)}(x_i), i = 2, \ldots, n-1$, $\delta_1 = \delta_n = 0$, and $h = x_{i+1} - x_i, 1 \le i \le n-1$. (The last equation indicates that we restrict ourselves to the case of equally spaced samples—more general case is approachable, however not described here.) Using these notations, the objective function in (4) can be restated as

$$J(\lambda; \mathbf{f}) \quad = \quad (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \lambda \delta^T \mathbf{M} \delta, \tag{5}$$

where $\delta = (\delta_2, \ldots, \delta_{n-1})^T$ and $\mathbf{M}$ is defined to be the $(n-2) \times (n-2)$ matrix with elements m_{ij}, given by $m_{ii} = \frac{2h}{3}$ for $i = 1, \ldots, n-2$; $m_{i,i+1} = m_{i+1,i} = \frac{h}{6}$ for $i = 1, \ldots, n-3$; zeros elsewhere. Using the fact that $\mathbf{M}\delta = \mathbf{Q}\mathbf{f}$, one can find the minimizer of $J(\lambda; \mathbf{f})$ as follows:

$$\hat{\mathbf{f}}(\lambda) = \left(\mathbf{I} + \lambda \mathbf{Q}^T \mathbf{M}^{-1} \mathbf{Q}\right)^{-1} \mathbf{y}, \tag{6}$$

where $\mathbf{Q}$ is defined to be the $(n-2) \times n$ matrix with elements q_{ij}, given by $q_{ii} = q_{i,i+2} = \frac{1}{h}$ and $q_{i,i+1} = \frac{-2}{h}$ for $i = 1, \ldots, n-2$, and zeros elsewhere.

2.2. A review of generalized cross validation

In (6), the choice of λ is an important issue, and many procedures have been proposed for the optimal choice of λ. One of the most popular procedures is the generalized cross validation (GCV) (Craven and Wahba, 1979). GCV selects λ by minimizing

$$\mathrm{GCV}_n(\lambda) = \frac{n^{-1}\|\mathbf{y} - \hat{\mathbf{f}}(\lambda)\|^2}{[n^{-1}\mathrm{tr}(\mathbf{I} - \mathbf{A}(\lambda))]^2}, \tag{7}$$

where $\mathbf{A}(\lambda)$ is the smoothing matrix that satisfies $\hat{\mathbf{f}} = \mathbf{A}(\lambda)\mathbf{y}$ (i.e., $\mathbf{A}(\lambda) = (\mathbf{I} + \lambda \mathbf{Q}^T \mathbf{M}^{-1} \mathbf{Q})^{-1}$ in (6)), $\|\cdot\|$ indicates the Euclidean norm that will be used throughout the paper, and 'tr' denotes trace. In the remainder of this subsection, we justify the adoption of GCV: if we choose λ by minimizing GCV, the resulting estimate minimizes the true loss for estimating f with $\hat{f}(\lambda)$ defined by

$$L_n(\lambda) \quad = \quad n^{-1}\|\mathbf{f} - \hat{\mathbf{f}}(\lambda)\|^2. \tag{8}$$

In the following, we use $\mathbf{A}_n(\lambda)$ (instead of $\mathbf{A}(\lambda)$) to integrate the sample size n. Similarly, $\mathbf{f}_n$ denote $\mathbf{f}$ when the sample size is n.

Theorem 1 (Li (1985))**.** *Consider the following Stein's estimate $\tilde{\mathbf{f}}_n(\lambda)$, the associated Stein's unbiased risk estimator ($SURE_n(\lambda)$), and the loss $\tilde{L}_n(\lambda)$ while estimating $\mathbf{f}_n$ by $\tilde{\mathbf{f}}_n(\lambda)$:*

$$\tilde{\mathbf{f}}_n(\lambda) = \mathbf{y}_n - \sigma^2 \frac{tr(\mathbf{I} - \mathbf{A}_n(\lambda))}{\|(\mathbf{I} - \mathbf{A}_n(\lambda))\mathbf{y}_n\|^2}(\mathbf{I} - \mathbf{A}_n(\lambda))\mathbf{y}_n,$$

$$SURE_n(\lambda) = \sigma^2 - \sigma^4 \frac{[n^{-1} tr(\mathbf{I} - \mathbf{A}_n(\lambda))]^2}{n^{-1}\|(\mathbf{I} - \mathbf{A}_n(\lambda))\mathbf{y}_n\|^2},$$

$$\tilde{L}_n(\lambda) = n^{-1}\|\tilde{\mathbf{f}}_n(\lambda) - \mathbf{f}_n\|^2.$$

Under the following conditions:

(C.1) *The 4th moment of ϵ_i's are upper bounded by a constant, where ϵ_i are random noise in (1),*

(C.2) *There exists a constant K, such that $\forall a > 0$,*

$$\sup_{x \in \mathbb{R}} P\{x - a \le \epsilon_i \le x + a\} \le K \cdot a, \quad \text{for } \forall i,$$

for any $\delta > 0$, we have

$$\sup_{\mathbf{f}_n \in \mathbb{R}^n} P\left\{ \sup_{\lambda \in \mathbb{R}^+} \left| SURE_n(\lambda) - \tilde{L}_n(\lambda) \right| > \delta \right\} \to 0.$$

Theorem 1 demonstrates that $SURE_n(\lambda)$ is a uniformly consistent estimator of $\tilde{L}_n(\lambda)$. Also note that minimizing the GCV function in (7) is equivalent to minimizing the $SURE_n(\lambda)$. In Theorem 1, the conditions **(C.1)** and **(C.2)** can be replaced by the following condition **(C.3)**: ϵ_i are i.i.d. $N(0, \sigma^2)$. That is, **(C.3)** implies **(C.1)** and **(C.2)**.

The asymptotic equivalence between $\tilde{\mathbf{f}}_n(\lambda)$ and $\hat{\mathbf{f}}_n(\lambda)$ are known as in the following theorem due to Li (1986).

Theorem 2 (Li (1986))**.** *For any sequence λ_n such that $\hat{\mathbf{f}}_n(\lambda_n)$ is consistent in the sense that*

$$n^{-1} \|\mathbf{f}_n - \hat{\mathbf{f}}_n(\lambda_n)\|^2 \to 0, \tag{9}$$

and

$$(n^{-1} tr \mathbf{A}_n(\lambda_n))^2 / n^{-1} tr \mathbf{A}_n^2(\lambda_n) \to 0, \tag{10}$$

under the condition that

$$\inf_{\lambda \ge 0} \quad n \cdot \mathbb{E} L_n(\lambda) \to \infty, \tag{11}$$

$\tilde{\mathbf{f}}_n(\lambda_n)$ and $\hat{\mathbf{f}}_n(\lambda_n)$ are asymptotically indistinguishable in the sense that

$$n^{-1} \|\tilde{\mathbf{f}}_n(\lambda_n) - \hat{\mathbf{f}}_n(\lambda_n)\|^2 / L_n(\lambda_n) \to 0.$$

Let $\hat{\lambda}_G$ denote the value of λ chosen by minimizing GCV in (7). It is known (Li, 1986) that if f is not a polynomial of degree 2 or less, (11) holds, and that if x_i are equispaced and ϵ_i are i.i.d. $N(0, \sigma^2)$, (9) and (10) hold with $\lambda_n = \hat{\lambda}_G$. Under these general conditions, therefore, $\tilde{\mathbf{f}}_n(\hat{\lambda}_G)$ and $\hat{\mathbf{f}}_n(\hat{\lambda}_G)$ are asymptotically equivalent due to Theorem 2. Together with Theorem 1, this demonstrates that if we choose λ by minimizing GCV, the resulting smoothing spline estimate $\hat{\mathbf{f}}_n(\hat{\lambda}_G)$ minimizes the true loss $L_n(\lambda)$ in (8).

3. Adaptive smoothing splines and mGCV

We describe an optimization approach to achieve desirable spatial adaptation in the framework of (3). This section is organized as follows. We introduce a penalization method for spatial adaptation in Section 3.1. The key idea is to turn a univariate λ in Section 2.1 to a function of the location, $\lambda(x)$, which is subsequently discretized to a multivariate smoothing parameter. For the choice of the multivariate smoothing parameter, a multivariate version of the GCV (mGCV) is suggested in Section 3.2. The consistency and asymptotic optimality of the mGCV are shown in Section 3.3 and Section 3.4, respectively.

3.1. A strategy to achieve spatial adaptivity

The essence of achieving *spatial adaptivity* is to utilize $\lambda(x)$, which is a function of location x, instead of constant λ. We assume that $\lambda(x)$ is absolutely continuous nearly everywhere except for a set of points whose measure is zero. As described in Section 1, we consider a discretized version of $\lambda(x)$: $\lambda(x) \equiv \eta_i$ for $x \in [s_i, s_{i+1}), i = 1, \dots, k$, $0 = s_0 < s_1 < \cdots < s_k < s_{k+1} = 1$, and $s_i \in \{x_1, \dots, x_n\}$. Then we need to estimate $\eta = (\eta_1, \eta_2, \dots, \eta_k)$ whose dimension is k. The objective of this paper is to provide theoretical justification of the mGCV for choosing η by proving that the mGCV choice of η is asymptotically optimal.

For our theoretical results to cover all possible cases of the step function assumed for $\lambda(x)$ with flexible step size and step number, we allow the discretization of $\lambda(x)$ to be as flexible as possible. By adopting a new sequence $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{n-1})$, we consider the most general case of the discretization: we assume that $\lambda(x) \equiv \lambda_i$ for $x \in [x_i, x_{i+1})$, $1 \le i \le n-1$ and $x_1 < x_2 < \cdots < x_n$. Then estimating η is equivalent to estimating $\boldsymbol{\lambda}_\eta$ which is a special case of $\boldsymbol{\lambda}$ and is defined as

$$\boldsymbol{\lambda}_\eta = (\underbrace{\eta_1, \dots, \eta_1}_{n_1}, \underbrace{\eta_2, \dots, \eta_2}_{n_2}, \dots, \underbrace{\eta_k, \dots, \eta_k}_{n_k}), \tag{12}$$

where n_i is the number of design points included in $[s_i, s_{i+1})$, $i = 1, \dots, k$, and $n_1 + \cdots + n_k = n - 1$. To establish the asymptotic optimality of the mGCV for all possible cases of the step function, it suffices to prove the asymptotic optimality for the most general case, i.e., $\boldsymbol{\lambda}$. For this reason, $\boldsymbol{\lambda}$ will be considered rather than η in our theoretical analysis throughout this paper. Then our established theoretical results cover all possible cases of the step function for $\lambda(x)$, including an equal-size piecewise constant penalty function with a few steps, which is considered in Pintore, Speckman and Holmes (2006). We have the following theorem.

Theorem 3. *Let* $\hat{\mathbf{f}}(\boldsymbol{\lambda})$ *denote the solution to* (3) *given* $\boldsymbol{\lambda}$. *We have* $\hat{\mathbf{f}}(\boldsymbol{\lambda}) \approx (\mathbf{I} + \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\boldsymbol{\lambda})\mathbf{M}^{-1}\mathbf{Q})^{-1}\mathbf{y}$ *as* $n \to \infty$, *where* $h, \mathbf{Q}, \mathbf{M}$ *were previously defined, and* $\mathbf{M}(\boldsymbol{\lambda}) \in \mathbb{R}^{(n-2)\times(n-2)}$ *satisfies* $\mathbf{M}(\boldsymbol{\lambda})_{ii} = \lambda_i + \lambda_{i+1}$, $1 \le i \le n-2$, $\mathbf{M}(\boldsymbol{\lambda})_{i,i+1} = \mathbf{M}(\boldsymbol{\lambda})_{i+1,i} = \frac{\lambda_{i+1}}{2}$, $1 \le i \le n-3$, *and zeros elsewhere.*

See Appendix A for the proof of Theorem 3. We will propose to use the mGCV for the optimal choice of $\boldsymbol{\lambda}$ in Section 3.2.

3.2. Multivariate version of GCV (mGCV)

For the estimator $\hat{\mathbf{f}}(\boldsymbol{\lambda})$, the choice of the multivariate smoothing parameter $\boldsymbol{\lambda}$ is important for appropriate spatial adaptation. We suggest to use the *multivariate version of generalized cross validation (mGCV)* defined as

$$\mathrm{mGCV}_n(\boldsymbol{\lambda}) = \frac{n^{-1}\|\mathbf{y} - \hat{\mathbf{f}}(\boldsymbol{\lambda})\|^2}{[n^{-1}\mathrm{tr}(\mathbf{I} - \mathbf{S}(\boldsymbol{\lambda}))]^2}, \tag{13}$$

where $\mathbf{S}(\boldsymbol{\lambda})$ denotes the smoothing matrix which satisfies $\hat{\mathbf{f}}(\boldsymbol{\lambda}) = \mathbf{S}(\boldsymbol{\lambda})\mathbf{y}$.

Parallel to Theorem 1, we establish the uniform consistency of $SURE_n(\boldsymbol{\lambda})$ in the following theorem. Together with the fact that minimizing mGCV is equivalent to minimizing $SURE_n(\boldsymbol{\lambda})$, the following theorem gives a justification of the mGCV. In the following, we use $\mathbf{S}_n(\boldsymbol{\lambda})$ (instead of $\mathbf{S}(\boldsymbol{\lambda})$) to take into account the sample size n.

Theorem 4. *Consider the following Stein estimate $\tilde{\mathbf{f}}_n(\boldsymbol{\lambda})$, the associated unbiased risk estimate ($SURE_n(\boldsymbol{\lambda})$), and the true loss $\tilde{L}_n(\boldsymbol{\lambda})$ while estimating $\mathbf{f}_n$ by $\tilde{\mathbf{f}}_n(\boldsymbol{\lambda})$:*

$$\tilde{\mathbf{f}}_n(\boldsymbol{\lambda}) = \mathbf{y}_n - \sigma^2 \frac{tr(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}))}{\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}))\mathbf{y}_n\|^2}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}))\mathbf{y}_n,$$

$$SURE_n(\boldsymbol{\lambda}) = \sigma^2 - \sigma^4 \frac{[n^{-1}tr(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}))]^2}{n^{-1}\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}))\mathbf{y}_n\|^2}, \tag{14}$$

and

$$\tilde{L}_n(\boldsymbol{\lambda}) = n^{-1}\|\mathbf{f}_n - \tilde{\mathbf{f}}_n(\boldsymbol{\lambda})\|^2.$$

Then $SURE_n(\boldsymbol{\lambda})$ is a uniformly consistent estimate of $\tilde{L}_n(\boldsymbol{\lambda})$ over $\mathbf{f}_n$ and $\boldsymbol{\lambda}$: For any $\delta > 0$,

$$\sup_{\mathbf{f}_n \in \mathbb{R}^n} P\left\{\sup_{\boldsymbol{\lambda} \in \mathbb{R}^{n-1}} |SURE_n(\boldsymbol{\lambda}) - \tilde{L}_n(\boldsymbol{\lambda})| > \delta\right\} \to 0.$$

Proof of the above theorem is in Appendix B.

Theorem 4 establishes the uniform consistency between $SURE_n(\boldsymbol{\lambda})$ and $\tilde{L}_n(\boldsymbol{\lambda})$ which involves $\tilde{\mathbf{f}}_n(\boldsymbol{\lambda})$. To consider the original loss $L_n(\boldsymbol{\lambda})$ for estimating $\mathbf{f}_n$ with $\hat{\mathbf{f}}_n(\boldsymbol{\lambda})$, we establish the asymptotic equivalence between $\tilde{\mathbf{f}}_n(\boldsymbol{\lambda})$ and $\hat{\mathbf{f}}_n(\boldsymbol{\lambda})$ in the following theorem.

Theorem 5. *For any $\hat{\boldsymbol{\lambda}}$ such that*

$$L_n(\hat{\boldsymbol{\lambda}}) \to 0, \tag{15}$$

and

$$\frac{(n^{-1}tr\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))^2}{n^{-1}tr\mathbf{S}_n^2(\hat{\boldsymbol{\lambda}}))} \to 0, \tag{16}$$

under the following condition,

(A.1) $\inf_{\boldsymbol{\lambda} \in \mathbb{R}_+^{n-1}} \quad n \cdot \mathbb{E}L_n(\boldsymbol{\lambda}) \to \infty,$

we have

$$\frac{|SURE_n(\hat{\boldsymbol{\lambda}}) - \tilde{L}_n(\hat{\boldsymbol{\lambda}}) - n^{-1}\|\boldsymbol{\epsilon}_n\|^2 + \sigma^2|}{L_n(\hat{\boldsymbol{\lambda}})} \to 0, \tag{17}$$

and

$$\frac{n^{-1}\|\tilde{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}) - \hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}})\|^2}{L_n(\hat{\boldsymbol{\lambda}})} \to 0. \tag{18}$$

Proof of the above theorem is in Appendix C.

(A.1) states that the optimal rate of convergence of $\mathbb{E}L_n(\boldsymbol{\lambda})$ to zero must be slower than n^{-1}. For **(A.1)**, in the typical polynomial spline smoothing problems, $\inf_{\lambda>0}\mathbb{E}L_n(\lambda)$ tends to zero at the rate $n^{-1+\delta}$ for some small constant $\delta > 0$ except if the underlying function is the sampled values of a low order polynomial (Wahba, 1985). In our framework, we need to study when **(A.1)** holds—this is an open problem.

Let $\hat{\boldsymbol{\lambda}}_{mG}$ denote the value of $\boldsymbol{\lambda}$ chosen by minimizing the mGCV function in (13). Using Theorem 5, we can show that under certain conditions, $\tilde{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$ and $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$ are asymptotically equivalent, which will be established in Theorem 9 in Section 3.3.

3.3. Consistency of mGCV

We say that $\hat{\mathbf{f}}_n(\boldsymbol{\lambda})$ is consistent if $L_n(\boldsymbol{\lambda}) \to 0$ as $n \to \infty$, where $L_n(\boldsymbol{\lambda})$ is the loss while estimating $\mathbf{f}_n$ by $\hat{\mathbf{f}}_n(\boldsymbol{\lambda})$, i.e., $L_n(\boldsymbol{\lambda}) = n^{-1}\|\mathbf{f}_n - \hat{\mathbf{f}}_n(\boldsymbol{\lambda})\|^2$. In this section, we show that if we choose $\boldsymbol{\lambda}$ via mGCV, then the resulting $\hat{\mathbf{f}}$ is consistent. To establish the consistency of mGCV, we need the following two conditions:

(A.2) Recall that as $n \to \infty$, we have $\mathbf{S}_n(\boldsymbol{\lambda}) \approx (\mathbf{I} + \boldsymbol{\Sigma}_n(\boldsymbol{\lambda}))^{-1}$, where $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda}) = \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\boldsymbol{\lambda})\mathbf{M}^{-1}\mathbf{Q}$. Let $0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_n$ be the eigenvalues of $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda})$. For any m such that $m/n \to 0$, we have

$$\frac{\left(n^{-1}\sum_{i=m+1}^n \tau_i^{-1}\right)^2}{n^{-1}\sum_{i=m+1}^n \tau_i^{-2}} \to 0, \text{ as } n \to \infty,$$

(A.3) There exists $\boldsymbol{\lambda}_n$, such that $L_n(\boldsymbol{\lambda}_n) \to 0$.

The following theorem establishes the consistency of mGCV.

Theorem 6 (Consistency)**.** *Under **(A.2)** and **(A.3)**,* $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$*, where* $\hat{\boldsymbol{\lambda}}_{mG}$ *is the mGCV choice, is consistent, i.e.,* $L_n(\hat{\boldsymbol{\lambda}}_{mG}) \to 0$.

Proof of the above theorem is in Appendix D.

The above **(A.2)** involves eigenvalues, and seems hard to verify. The following theorem provides a simple sufficient condition for it.

Theorem 7. *For the estimator* $\hat{\mathbf{f}}(\boldsymbol{\lambda})$*, if*

$$\max(\boldsymbol{\lambda}_n)/\min(\boldsymbol{\lambda}_n) < \textit{Constant, as } n \to \infty,$$

where $\max(\boldsymbol{\lambda}_n)$ *and* $\min(\boldsymbol{\lambda}_n)$ *denote the maximal and minimal values among* λ_i*,* $1 \le i \le n-1$*, then condition **(A.2)** holds.*

Proof of the above theorem is in Appendix E.

For **(A.3)**, it is known (Li, 1985, Theorem 5.5) that such $\boldsymbol{\lambda}_n$ exists for the nonadaptive case if x_i's are equispaced and ϵ_i's are i.i.d. $N(0, \sigma^2)$. Since the nonadaptive smoothing spline is a special case of the spatially adaptive smoothing spline, under the same conditions, **(A.3)** holds for the spatially adaptive smoothing splines too. Then we have the following corollary, together with Theorem 7:

Corollary 8. *If x_i's are equispaced and ϵ_i's are i.i.d. $N(0, \sigma^2)$, $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$ is consistent, provided that $\max(\hat{\boldsymbol{\lambda}}_{mG}) / \min(\hat{\boldsymbol{\lambda}}_{mG}) <$ Constant as $n \to \infty$.*

In the next theorem, using Theorem 5 and Theorem 6, we show that under certain conditions, $\tilde{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$ and $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$ are asymptotically indistinguishable.

Theorem 9. *Under **(A.1)**, **(A.2)**, and **(A.3)**,*

$$\frac{n^{-1} \|\tilde{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG}) - \hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})\|^2}{L_n(\hat{\boldsymbol{\lambda}}_{mG})} \to 0,$$

where $\hat{\boldsymbol{\lambda}}_{mG}$ is the mGCV choice.

Proof of the above theorem is in Appendix F. Together with Theorem 4, Theorem 9 demonstrates that if we choose $\boldsymbol{\lambda}$ by minimizing mGCV, the resulting estimate $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$ asymptotically minimizes the true loss $L_n(\boldsymbol{\lambda})$, which is a direct measure for estimating $\mathbf{f}_n$ by $\hat{\mathbf{f}}_n(\boldsymbol{\lambda})$.

3.4. Asymptotic optimality of mGCV

In a series of historic papers (Li, 1985, 1986; Girard, 1991), *asymptotic optimality* has been established for the GCV. In this section, we establish the asymptotic optimality of the mGCV in the same sense as in Li (1986). The asymptotic optimality in adaptive smoothing splines is defined as follows:

$$\frac{L_n(\hat{\boldsymbol{\lambda}}_{mG})}{\inf_{\boldsymbol{\lambda} \in \mathbb{R}_+^{n-1}} L_n(\boldsymbol{\lambda})} \to 1, \quad \text{in probability,} \tag{19}$$

where $\hat{\boldsymbol{\lambda}}_{mG}$ is the minimizer of the mGCV function in (13). Under certain conditions, the mGCV method of selecting $\boldsymbol{\lambda}$ is asymptotically optimal as indicated by the following theorem.

Theorem 10 (Asymptotic Optimality)**.** *Under **(A.1)**, **(A.2)**, and **(A.3)**, $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}_{mG})$, where $\hat{\boldsymbol{\lambda}}_{mG}$ is the mGCV choice, is asymptotically optimal as in* (19).

Proof of the above theorem is in Appendix G.

Theorem 10 states that under certain conditions, the mGCV choice, $\hat{\boldsymbol{\lambda}}_{mG}$, and the optimal value of $\boldsymbol{\lambda}$ behave the same for sufficiently large n in terms of the corresponding values of the loss. It also says that $L_n(\hat{\boldsymbol{\lambda}}_{mG})$ will tend toward the minimal loss as $n \to \infty$.

Recall that $\boldsymbol{\lambda}$ has the dimension of the sample size minus one. Although this $\boldsymbol{\lambda}$ allows the most flexible adaptation to varying roughness, it may not be computationally efficient. For more efficient computation, it is generally recommended to reduce the dimension of $\boldsymbol{\lambda}$ by assuming a step function for $\lambda(x)$ with the number of jumps much less than the sample size, such as $\boldsymbol{\lambda}_\eta$ in (12). As a special case of Theorem 10, it can be easily shown that if we choose $\boldsymbol{\lambda}_\eta$ by minimizing the mGCV, the asymptotic optimality still holds. Under the following conditions **(A.1′)**, **(A.2′)**, and **(A.3′)**:

(A.1′) $\inf_\eta \in \mathbb{R}_+^k \quad n \cdot \mathbb{E} L_n(\boldsymbol{\lambda}_\eta) \to \infty$,

(A.2′) Recall that as $n \to \infty$, we have $\mathbf{S}_n(\boldsymbol{\lambda}_\eta) \approx (\mathbf{I} + \boldsymbol{\Sigma}_n(\boldsymbol{\lambda}_\eta))^{-1}$, where $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda}_\eta) = \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\boldsymbol{\lambda}_\eta)\mathbf{M}^{-1}\mathbf{Q}$. Let $0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_n$ be the eigenvalues of $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda}_\eta)$. For any m such that $m/n \to 0$, we have

$$\frac{\left(n^{-1}\sum_{i=m+1}^n \tau_i^{-1}\right)^2}{n^{-1}\sum_{i=m+1}^n \tau_i^{-2}} \to 0, \text{ as } n \to \infty,$$

(A.3′) There exists $\boldsymbol{\lambda}_{\eta,n}$ such that $L_n(\boldsymbol{\lambda}_{\eta,n}) \to 0$,

we have

$$\frac{L_n(\hat{\boldsymbol{\lambda}}_{\eta,mG})}{\inf_{\eta\in\mathbb{R}_+^k} L_n(\boldsymbol{\lambda}_\eta)} \to 1, \quad \text{in probability},$$

where $\hat{\boldsymbol{\lambda}}_{\eta,mG}$ is the value of $\boldsymbol{\lambda}_\eta$ chosen by minimizing the mGCV function in (13) with $\boldsymbol{\lambda}$ replaced by $\boldsymbol{\lambda}_\eta$.

4. Numerical study

The main objective of this paper is to provide the theoretical justification for employing the mGCV in adaptive smoothing splines. Nevertheless, we present some numerical study of the mGCV to show its practical effectiveness. We suggest a simple segment-based search algorithm to estimate a piecewise constant penalty function using the mGCV, which has proved to work well in practice:

1. Assume the initial $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_{n-1}) = (\lambda, \dots, \lambda)$, where λ is chosen via the GCV.
2. Given $\boldsymbol{\lambda}$ from the step 1, define a sequence $\tilde{\boldsymbol{\lambda}}^1(i_0, i_1, \alpha) = (\tilde{\lambda}_1^1, \dots, \tilde{\lambda}_{n-1}^1)$ where we impose the following: if $i_0 \le i \le i_1$, $\tilde{\lambda}_i^1 = \lambda_i + \alpha$; and $\tilde{\lambda}_i^1 = \lambda_i$, otherwise. We find i_0, i_1, α, such that the mGCV$(\tilde{\boldsymbol{\lambda}}^1)$ is minimized. This step can be done via an extensive search.
3. Given $\tilde{\boldsymbol{\lambda}}^1$ from the step 2, define a sequence $\tilde{\boldsymbol{\lambda}}^2(\beta)$ such that for $1 \le i \le n-1$, $\tilde{\lambda}_i^2 = \beta \cdot \tilde{\lambda}_i^1$ where $\beta > 0$. We find β such that the mGCV$(\tilde{\boldsymbol{\lambda}}^2)$ is minimized. This step can be done via an extensive search. We declare the convergence and terminate, if β is close enough to 1 or the newly obtained minimum value of the mGCV is larger than the one at the previous iteration. Otherwise, bring $\tilde{\boldsymbol{\lambda}}^2$ back to the step 2 with $\boldsymbol{\lambda}$ replaced by $\tilde{\boldsymbol{\lambda}}^2$. The final $\tilde{\boldsymbol{\lambda}}^2$ is our estimate of the multivariate smoothing parameter.

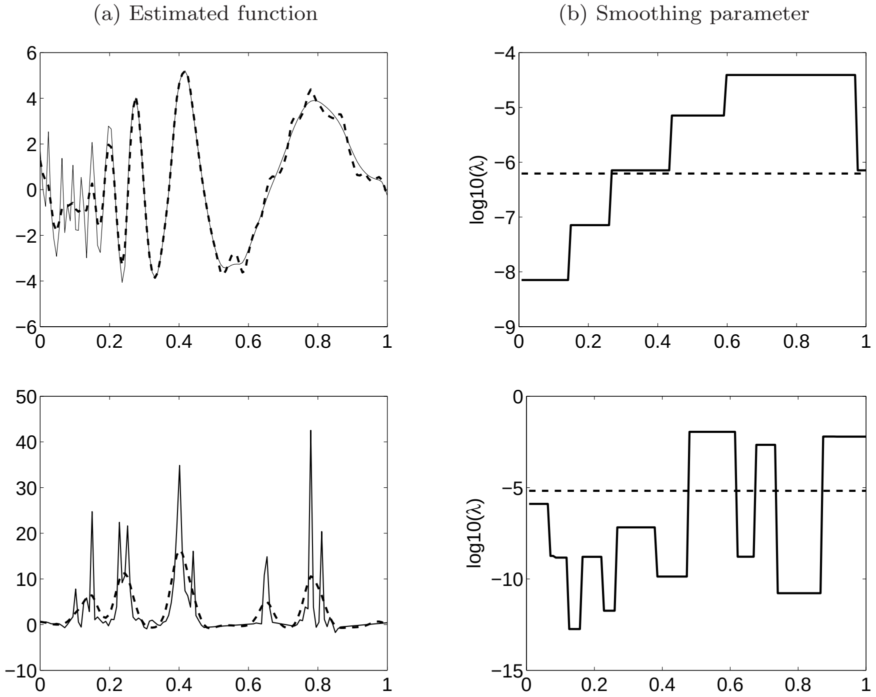


Fig 2. *Illustration of our numerical results using the Doppler and Bumps functions (top and bottom rows, respectively). In (a), the estimated function using our method (solid) is shown, together with the result from the classic smoothing spline (dashed). In (b), our multivariate smoothing parameter (solid) is shown, together with the global smoothing parameter from the standard GCV (dashed).*

Even though the asymptotic optimality of the mGCV holds for the most general case of the piecewise constant penalty function, assuming a small number of pieces is recommended in practice for more efficient computation. In our numerical study, we force the number of pieces to be small by specifying the step size of the above search algorithm to be large enough. The suggested simple algorithm only guarantees a local minimum of the mGCV function. Nevertheless, promising numerical results have been obtained in our simulation study: we found that our method outperforms or performs comparably with other competitive methods. As an alternative, well-known nonlinear optimization algorithms (e.g., Nelder-Mead, quasi-Newton, and conjugate gradient methods) can be used to minimize the mGCV function. However, researching on the best numerical strategy is beyond the scope of this paper.

In Fig. 2, we illustrate our numerical results using the two popular examples of the Doppler and Bumps functions introduced in Fig. 1. For both examples, we consider 128 data points sampled regularly on $[0, 1]$ and the signal-to-noise ratio of 7. Each row shows the estimated function in (a) and the estimated smoothing parameter in log scale in (b). Our results are shown in a solid line, together with the results from the standard smoothing spline in a dashed line. For the choice of the smoothing parameter in the standard smoothing spline, we use the GCV. We observe that our estimated functions in (a) show much better agreement with the true functions than the standard smoothing splines do. We also observe that our estimated multivariate smoothing parameter in (b) is spatially adaptive: it has relatively small values in the regions of large local variations, and large values in the regions of small local variations.

Based on repeated simulations, we further verify the performance of our method using the Doppler and Bumps examples via the following four ways:

- Comparison with the traditional smoothing splines (denoted as SS): This comparison will show the advantage of adopting the adaptive smoothing parameter over the global smoothing parameter. The global smoothing parameter for SS is chosen via the GCV.
- Comparison with the spatially adaptive smoothing splines (denoted as SASS) in Pintore, Speckman and Holmes (2006): SASS assumes an equal-size piecewise constant penalty function. 5 and 10 pieces are assumed for the Doppler and Bumps functions, respectively. This comparison will show the advantage of adopting more flexible structure on the piecewise constant penalty function rather than the equal-size pieces.
- Comparison with the locally optimal adaptive smoothing splines (denoted as LOASS) in Kim and Huo (2012): LOASS assumes an asymptotically optimal local penalty function that is continuously varying. This comparison will show the advantage of adopting a piecewise constant penalty function rather than the continuously varying penalty function.
- Comparison with the Wavelet shrinkage method in Donoho and Johnstone (1995): Wavelet shrinkage has emerged to be a powerful nonparametric smoothing method. We choose the Symmlet wavelets with 8 vanishing moments, and the coarsest level is set to be 4. This comparison will show the effectiveness of our spline-based method for fitting functions with varying roughness.

We compare the mean squared error (MSE) for fitting the two functions. The MSE is defined as MSE $= n^{-1}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2$. For each function, we consider 128, 256, 512 data points sampled regularly on $[0, 1]$ and the signal-to-noise ratio of 7. For each example, we run 100 experiments, and then take the averaged MSE as a performance measure. The results are summarized in Table 1: the averaged MSE (in bold-face if it is the smallest one) with the standard error (in parentheses) are reported. The results from the competitive methods are reproduced from Kim and Huo (2012). For our method to be computationally efficient, we intentionally made the number of pieces in the estimated piecewise constant penalty function to be small by setting the step size of the search algorithm large enough. As a result, the average of the number of pieces in our estimated penalty function was obtained as 5 and 8 for the Doppler and Bumps functions, respectively. In Table 1, we observe that our method outperforms or performs competitively with other methods.

Table 1
The five competing methods are compared using the Doppler and Bumps functions. The averaged MSE based on 100 simulations are shown. The values in the parentheses are standard errors. In each case, the smallest MSE is indicated in bold-face

n=128	Doppler	Bumps
SS	2.98 (0.32)	27.11 (0.55)
SASS	2.58 (0.42)	29.50 (0.41)
LOASS	1.46 (0.21)	17.92 (0.95)
Wavelets	1.91 (0.21)	3.51 (0.28)
Our method	**0.61** (0.14)	**0.89** (0.15)
n=256	Doppler	Bumps
SS	1.23 (0.15)	4.35 (0.45)
SASS	0.85 (0.25)	4.84 (0.50)
LOASS	0.78 (0.07)	**0.79** (0.07)
Wavelets	1.17 (0.13)	3.32 (0.19)
Our method	**0.39** (0.07)	0.87 (0.12)
n=512	Doppler	Bumps
SS	0.58 (0.04)	1.19 (0.14)
SASS	0.56 (0.05)	1.15 (0.21)
LOASS	0.56 (0.04)	**0.75** (0.07)
Wavelets	0.81 (0.07)	2.62 (0.14)
Our method	**0.35** (0.05)	0.85 (0.12)

5. Conclusion

The asymptotic optimality of the generalized cross validation (GCV) is well known in smoothing splines. The multivariate version of the GCV (mGCV) is more flexible in practice, and has been used to achieve spatial adaptivity in smoothing splines. However, little is known about its theoretical property. We show that the mGCV also has the asymptotic optimality, under general conditions that are comparable as those conditions in the case of GCV. Our analysis provides theoretical justification for employing the mGCV in choosing the multivariate smoothing parameter in spatially adaptive smoothing splines.

Appendix A: Proof of Theorem 3

We have

$$\int_{x_i}^{x_{i+1}} \{f^{(2)}(x)\}^2 dx = \frac{h}{3}\left(\delta_i^+\right)^2 + \frac{h}{3}\left(\delta_{i+1}^-\right)^2 + \frac{h}{3}\delta_i^+\delta_{i+1}^-, \quad i = 1, \dots, n-1, \tag{20}$$

where δ_i^+ is the second derivative of f from the right at x_i, and δ_{i+1}^- is the second derivative of f from the left at x_{i+1}, that is, $\delta_i^+ = f^{(2)}(x_i^+)$ and $\delta_i^- = f^{(2)}(x_i^-)$, where $x_i^+ = \lim_{a\to 0, a>0}(x_i + a)$ and $x_i^- = \lim_{a\to 0, a<0}(x_i + a)$, and $h = x_{i+1} - x_i, 1 \le i \le n-1$. Note that $\delta_i^- \neq \delta_i^+$ for $\lambda_{i-1} \neq \lambda_i$: $|\delta_i^+ - \delta_i^-|$ is constrained through the quantity $|\{\lambda(x_i^-) - \lambda(x_i^+)\}/\{\lambda(x_i^-)\lambda(x_i^+)\}|$ (Pintore, Speckman and

Holmes, 2006). Recall that $\lambda(x) \equiv \lambda_i$ for $x \in [x_i, x_{i+1}], 1 \le i \le n-1$. Then as $n \to \infty$, we have $\lambda(x_i^-) \approx \lambda(x_i^+)$, and consequently, $\delta_i^+ \approx \delta_i^-$. Then as $n \to \infty$, (20) can be approximated as

$$\int_{x_i}^{x_{i+1}} \{f^{(2)}(x)\}^2 dx \approx \frac{h}{3}\delta_i^2 + \frac{h}{3}\delta_{i+1}^2 + \frac{h}{3}\delta_i\delta_{i+1}, \quad i = 1, \ldots, n-1, \tag{21}$$

where $\delta_i = f^{(2)}(x_i), i = 2, \ldots, n-1$, $\delta_1 = \delta_n = 0$. Using (21), as $n \to \infty$, we have

$$\begin{aligned} J^\star(\boldsymbol{\lambda}; \mathbf{f}) &= (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \int_0^1 \lambda(x)\{f^{(2)}(x)\}^2 dx \\ &\approx (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \frac{h}{3}\sum_{i=1}^{n-1} \lambda_i(\delta_i^2 + \delta_{i+1}^2 + \delta_i\delta_{i+1}) \\ &= (\mathbf{y} - \mathbf{f})^T(\mathbf{y} - \mathbf{f}) + \frac{h}{3}\delta^T \mathbf{M}(\boldsymbol{\lambda})\delta, \end{aligned} \tag{22}$$

where $\delta = (\delta_2, \ldots, \delta_{n-1})^T$ and $\mathbf{M}(\boldsymbol{\lambda}) \in \mathbb{R}^{(n-2)\times(n-2)}$ satisfies $\mathbf{M}(\boldsymbol{\lambda})_{ii} = \lambda_i + \lambda_{i+1}$, $1 \le i \le n-2$, $\mathbf{M}(\boldsymbol{\lambda})_{i,i+1} = \mathbf{M}(\boldsymbol{\lambda})_{i+1,i} = \frac{\lambda_{i+1}}{2}$, $1 \le i \le n-3$, and zeros elsewhere. Pintore, Speckman and Holmes (2006) showed that with the assumption of step function on $\lambda(x)$, the solution $\hat{\mathbf{f}}(\boldsymbol{\lambda})$ satisfies all conditions, but the continuity of the second derivative of f, to be a natural cubic spline. However, since $\delta_i^+ \approx \delta_i^-$ as $n \to \infty$, we can take advantage of the fact that $\mathbf{M}\delta \approx \mathbf{Q}\mathbf{f}$ as $n \to \infty$. (Note that the identity $\mathbf{M}\delta = \mathbf{Q}\mathbf{f}$ holds if and only if $f(x)$ is a natural cubic spline.) Then, as $n \to \infty$, using the fact that $\mathbf{M}\delta \approx \mathbf{Q}\mathbf{f}$ and by considering the first order condition, from (22), $\hat{\mathbf{f}}(\boldsymbol{\lambda})$ can be approximated as

$$\hat{\mathbf{f}}(\boldsymbol{\lambda}) \approx \left(\mathbf{I} + \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\boldsymbol{\lambda})\mathbf{M}^{-1}\mathbf{Q}\right)^{-1}\mathbf{y}.$$

Appendix B: Proof of Theorem 4

To prove Theorem 4, we extend the proof of Theorem 1 in Li (1985). Note that $\boldsymbol{\lambda}$ can be factorized as $\lambda_1(1, \lambda_2/\lambda_1, \ldots, \lambda_{n-1}/\lambda_1)$ where λ_1 is the first element of $\boldsymbol{\lambda}$, and let $\tilde{\boldsymbol{\lambda}}$ denote $(1, \lambda_2/\lambda_1, \ldots, \lambda_{n-1}/\lambda_1)$. Due to Theorem 3, as $n \to \infty$, we have $\mathbf{S}_n(\boldsymbol{\lambda}) \approx (\mathbf{I} + \boldsymbol{\Sigma}_n(\boldsymbol{\lambda}))^{-1}$ where $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda}) = \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\boldsymbol{\lambda})\mathbf{M}^{-1}\mathbf{Q}$. Also note that $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda}) = \lambda_1\boldsymbol{\Sigma}_n(\tilde{\boldsymbol{\lambda}})$. Let $0 \le \tilde{\tau}_1 \le \tilde{\tau}_2 \le \cdots \le \tilde{\tau}_n$ be the eigenvalues of $\boldsymbol{\Sigma}_n(\tilde{\boldsymbol{\lambda}})$. As in Li (1985), the key to prove the above theorem is to establish the following three inequalities: There exist $\delta_1, \delta_2 > 0$ and $a_n \to 0$ such that

$$P\left\{\inf_{\lambda_1 \ge 0}\inf_{\tilde{\boldsymbol{\lambda}} \ge 0} \lambda_1^2 \sum_{i=1}^n (f(x_i) + \epsilon_i)^2(\lambda_1 + \tilde{\tau}_i)^{-2}/Q_n(\lambda_1) \le a_n\right\} \le \delta_2/2,$$

$$P\left\{\sup_{\lambda_1 \ge 0}\sup_{\tilde{\boldsymbol{\lambda}} \ge 0} \frac{\lambda_1^2}{nQ_n(\lambda_1)} \sum_{i=1}^n (\lambda_1 + \tilde{\tau}_i)^{-1}\left|\sum_{i=1}^n \frac{\epsilon_i^2 - \sigma^2}{\lambda_1 + \tilde{\tau}_i}\right| \ge a_n\delta_1\right\} \le \delta_2/2,$$

$$P\left\{\sup_{\lambda_1\geq 0}\sup_{\tilde{\boldsymbol{\lambda}}\geq 0}\frac{\lambda_1^2}{nQ_n(\lambda_1)}\sum_{i=1}^{n}(\lambda_1+\tilde{\tau}_i)^{-1}\left|\sum_{i=1}^{n}\frac{f(x_i)\epsilon_i}{\lambda_1+\tilde{\tau}_i}\right|\geq a_n\delta_1\right\}\leq \delta_2/2,$$

where $Q_n(\lambda_1)=\lambda_1^2\sum_{i=1}^{n}(\lambda_1+\tilde{\tau}_i)^{-2}(f(x_i)^2+\sigma^2)$. A careful scrutiny at the proof in Li (1985) revealed that the choice of δ_1, δ_2, and a_n does not depend on the values of $\tilde{\tau}_i$, so the argument in Li (1985) is sufficient to establish Theorem 4. To avoid repeating a lengthy discussion, we omit a detailed proof here.

Appendix C: Proof of Theorem 5

We first prove (17). Let $\mathbf{A}_n(\hat{\boldsymbol{\lambda}})=\mathbf{I}_n-\mathbf{S}_n(\hat{\boldsymbol{\lambda}})$. Rewrite (17) as

$$\begin{aligned}
& 2\left|\frac{\sigma^2\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}})}{n\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2}\langle\boldsymbol{\epsilon}_n,\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\rangle-\frac{\sigma^4(\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}}))^2}{n\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2}-n^1\|\boldsymbol{\epsilon}_n\|^2+\sigma^2\right|/L_n(\hat{\boldsymbol{\lambda}})\\
\leq\;& 2\sigma^2\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\left|\langle\boldsymbol{\epsilon}_n,\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{f}_n\rangle\right|/n\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2L_n(\hat{\boldsymbol{\lambda}}) \qquad (23)\\
& +2\sigma^2\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\left|\langle\boldsymbol{\epsilon}_n,\mathbf{S}_n(\hat{\boldsymbol{\lambda}})\boldsymbol{\epsilon}_n\rangle-\sigma^2\mathrm{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}})\right|/n\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2L_n(\hat{\boldsymbol{\lambda}}) \qquad (24)\\
& +2\left|\left(\frac{\sigma^2\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}})}{\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2}-1\right)(\sigma^2-n^{-1}\|\boldsymbol{\epsilon}_n\|^2)\right|/L_n(\hat{\boldsymbol{\lambda}}). \qquad (25)
\end{aligned}$$

It suffices to show (23), (24), (25) tend to 0. Note that (15) is equivalent to $n^{-1}\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2\to\sigma^2$. For (23) and (24), using (15), it suffices to show

$$\sup_{\boldsymbol{\lambda}>0}\left|n^{-1}\langle\boldsymbol{\epsilon}_n,\mathbf{A}_n(\boldsymbol{\lambda})\mathbf{f}_n\rangle\right|/\mathbb{E}L_n(\boldsymbol{\lambda})\to 0, \qquad (26)$$

$$\sup_{\boldsymbol{\lambda}>0}n^{-1}\left|\sigma^2\mathrm{tr}\mathbf{S}_n(\boldsymbol{\lambda})-\langle\boldsymbol{\epsilon}_n,\mathbf{S}_n(\boldsymbol{\lambda})\boldsymbol{\epsilon}_n\rangle\right|/\mathbb{E}L_n(\boldsymbol{\lambda})\to 0, \qquad (27)$$

and

$$\sup_{\boldsymbol{\lambda}>0}\left|L_n(\boldsymbol{\lambda})/\mathbb{E}L_n(\boldsymbol{\lambda})-1\right|\to 0. \qquad (28)$$

For (26), given any $\delta>0$, we have by Chebychev inequality that

$$\begin{aligned}
& P\{\sup_{\boldsymbol{\lambda}>0}\left|n^{-1}\langle\boldsymbol{\epsilon}_n,\mathbf{A}_n(\boldsymbol{\lambda})\mathbf{f}_n\rangle\right|/\mathbb{E}L_n(\boldsymbol{\lambda})>\delta\}\\
\leq\;& \delta^{-2}n^{-2}\mathbb{E}\|\boldsymbol{\epsilon}_n\|^2\|\mathbf{A}_n(\boldsymbol{\lambda})\mathbf{f}_n\|^2(\mathbb{E}L_n(\boldsymbol{\lambda}))^{-2}. \qquad (29)
\end{aligned}$$

Using the fact that $n^{-1}\|\mathbf{A}_n(\boldsymbol{\lambda})\mathbf{f}_n\|^2\leq\mathbb{E}L_n(\boldsymbol{\lambda})$ and **(A.1)**, (29) goes to 0. For (27), by using the fact that $\mathbb{E}(\sigma^2\mathrm{tr}\mathbf{S}_n(\boldsymbol{\lambda})-\langle\boldsymbol{\epsilon}_n,\mathbf{S}_n(\boldsymbol{\lambda})\boldsymbol{\epsilon}_n\rangle)^2\leq 2\sigma^4\mathrm{tr}\mathbf{S}_n^2(\boldsymbol{\lambda})$ and $\sigma^2n^{-1}\mathrm{tr}\mathbf{S}_n^2(\boldsymbol{\lambda})\leq\mathbb{E}L_n(\boldsymbol{\lambda})$, it can be proved in a similar way to prove (26). For (28), it suffices to show the following

$$\sup_{\boldsymbol{\lambda}>0}n^{-1}\left|\langle\mathbf{A}_n(\boldsymbol{\lambda})\mathbf{f}_n,\mathbf{S}_n(\boldsymbol{\lambda})\boldsymbol{\epsilon}_n\rangle\right|/\mathbb{E}L_n(\boldsymbol{\lambda})\to 0, \qquad (30)$$

and

$$\sup_{\boldsymbol{\lambda}>0}n^{-1}\left|\|\mathbf{S}_n(\boldsymbol{\lambda})\boldsymbol{\epsilon}_n\|^2-\sigma^2\mathrm{tr}\mathbf{S}_n^2(\boldsymbol{\lambda})\right|/\mathbb{E}L_n(\boldsymbol{\lambda})\to 0. \qquad (31)$$

Notice that (30) and (31) can be proved in the same way as (26) and (27), respectively. Now only (25) remains to be proved. To prove (25), it suffices to show

$$\left|\sigma^2 n^{-1}\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}}) - n^{-1}\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2\right| \left|\sigma^2 - n^{-1}\|\boldsymbol{\epsilon}_n\|^2\right| / L_n(\hat{\boldsymbol{\lambda}}) \to 0.$$

Since $\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2 = \|\boldsymbol{\epsilon}_n\|^2 + 2\langle\boldsymbol{\epsilon}_n, \mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{f}_n\rangle - 2\langle\boldsymbol{\epsilon}_n, \mathbf{S}_n(\hat{\boldsymbol{\lambda}})\boldsymbol{\epsilon}_n\rangle + \|\mathbf{f}_n - \hat{\mathbf{f}}_n\|^2$, we have

$$\begin{aligned}
&\left|\sigma^2 n^{-1}\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}}) - n^{-1}\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2\right| \\
\leq\ &\left|\sigma^2 - \frac{1}{n}\|\boldsymbol{\epsilon}_n\|^2\right| + L_n(\hat{\boldsymbol{\lambda}}) + \frac{2}{n}\left|\langle\boldsymbol{\epsilon}_n, \mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{f}_n\rangle\right| \\
&+ \frac{2}{n}\left|\langle\boldsymbol{\epsilon}_n, \mathbf{S}_n(\hat{\boldsymbol{\lambda}})\boldsymbol{\epsilon}_n\rangle - \sigma^2\mathrm{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}})\right| + n^{-1}\sigma^2\mathrm{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}}).
\end{aligned}$$

Then using (26), (27), and (15), it suffices to show

$$(\sigma^2 - n^{-1}\|\boldsymbol{\epsilon}_n\|^2)^2 / L_n(\hat{\boldsymbol{\lambda}}) \to 0, \tag{32}$$

and

$$(n^{-1}\mathrm{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))\left|\sigma^2 - n^{-1}\|\boldsymbol{\epsilon}_n\|^2\right| / L_n(\hat{\boldsymbol{\lambda}}) \to 0. \tag{33}$$

By **(A.1)**, (28), and the central limit theorem, we have (32). Using (32), (28), **(A.1)**, and the fact that $(n^{-1}\mathrm{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))^2 \leq \mathbb{E}L_n(\hat{\boldsymbol{\lambda}})$, we have (33). Hence, we complete the proof of (17). Now it remains to prove (18). The numerator of (18) can be rewritten as

$$\begin{aligned}
n^{-1}\|\tilde{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}}) - \hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}})\|^2 &= \left(\frac{n^{-1}\sigma^2\mathrm{tr}\mathbf{A}_n(\hat{\boldsymbol{\lambda}})}{n^{-1}\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2} - 1\right)^2 n^{-1}\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2 \\
= \quad &\frac{\left[(\sigma^2 - \frac{\|\boldsymbol{\epsilon}_n\|^2}{n}) - L_n(\hat{\boldsymbol{\lambda}}) - \frac{2\langle\boldsymbol{\epsilon}_n, \mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{f}_n\rangle}{n} + \frac{2\langle\boldsymbol{\epsilon}_n, \mathbf{S}_n(\hat{\boldsymbol{\lambda}})\boldsymbol{\epsilon}_n\rangle}{n} - \frac{\sigma^2\mathrm{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}})}{n}\right]^2}{n^{-1}\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2}.
\end{aligned}$$

Since $n^{-1}\|\mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{y}_n\|^2 \to \sigma^2$, to prove (18), it suffices to show the following:

$$(\sigma^2 - n^{-1}\|\boldsymbol{\epsilon}_n\|^2)^2 / L_n(\hat{\boldsymbol{\lambda}}) \to 0, \tag{34}$$

$$\left(n^{-1}\langle\boldsymbol{\epsilon}_n, \mathbf{A}_n(\hat{\boldsymbol{\lambda}})\mathbf{f}_n\rangle\right)^2 / L_n(\hat{\boldsymbol{\lambda}}) \to 0, \tag{35}$$

$$(n^{-1}\mathrm{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))^2 / L_n(\hat{\boldsymbol{\lambda}}) \to 0, \tag{36}$$

and

$$\left(n^{-1}\langle\boldsymbol{\epsilon}_n, \mathbf{S}_n(\hat{\boldsymbol{\lambda}})\boldsymbol{\epsilon}_n\rangle\right)^2 / L_n(\hat{\boldsymbol{\lambda}}) \to 0. \tag{37}$$

Notice that (34) is the same as (32). Using (26), (35) can be easily proven. (36) follows from (16) and (28). Finally, (37) follows from (27) and (36).

Appendix D: Proof of Theorem 6

Recall the Stein estimates ($\tilde{\mathbf{f}}_n$), the associated unbiased risk estimate ($SURE_n$), and the true loss ($\tilde{L}_n$) defined in Theorem 4. Note in Theorem 4 that $SURE_n(\boldsymbol{\lambda})$ is a uniformly consistent estimate of $\tilde{L}_n(\boldsymbol{\lambda})$ over $\mathbf{f}_n$ and $\boldsymbol{\lambda}$. Also note that by comparing (13) and (14), $\hat{\boldsymbol{\lambda}}_{mG}$ also minimizes $SURE_n(\boldsymbol{\lambda})$. We first need the following Lemmas 11 and 12 to establish the upcoming Lemma 13.

Lemma 11. *Under **(A.3)**, we have $\tilde{L}_n(\boldsymbol{\lambda}_n) \to 0$, where $\boldsymbol{\lambda}_n$ satisfies **(A.3)**.*

Proof. We have

$$
\begin{aligned}
\tilde{L}_n(\boldsymbol{\lambda}_n) &= \frac{1}{n}\|\tilde{\mathbf{f}}_n(\boldsymbol{\lambda}_n) - \mathbf{f}_n\|^2 \\
= \quad & \frac{1}{n}\left\|\boldsymbol{\epsilon}_n - \sigma^2 \frac{\mathrm{tr}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))}{\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\|^2}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\right\|^2 \\
\leq \quad & \frac{1}{n}\left(1 - \frac{\sigma^2 \mathrm{tr}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))}{\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\|^2}\right)^2 \|\boldsymbol{\epsilon}_n\|^2 \\
& + \frac{2}{n}\left|1 - \frac{\sigma^2 \mathrm{tr}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))}{\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\|^2}\right| \left|\frac{\sigma^2 \mathrm{tr}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))}{\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\|^2}\right| \|\boldsymbol{\epsilon}_n\| \|\mathbf{f}_n - \mathbf{S}_n(\boldsymbol{\lambda}_n)\mathbf{y}\| \\
& + \frac{1}{n}\left(\sigma^2 \frac{\mathrm{tr}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))}{\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\|^2}\right)^2 \|\mathbf{f}_n - \mathbf{S}_n(\boldsymbol{\lambda}_n)\mathbf{y}\|^2.
\end{aligned}
$$

It suffices to show that

$$
\frac{\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\|^2}{\mathrm{tr}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))} \to \sigma^2. \tag{38}
$$

Note that from the fact $\sigma^2(n^{-1}\mathrm{tr}\mathbf{S}_n(\boldsymbol{\lambda}_n))^2 \leq \sigma^2 n^{-1}\mathrm{tr}\mathbf{S}_n^2(\boldsymbol{\lambda}_n) \leq \mathbb{E}L_n(\boldsymbol{\lambda}_n) \to 0$, we have $n^{-1}\mathrm{tr}\mathbf{S}_n(\boldsymbol{\lambda}_n) \to 0$. Thus, in the denominator of (38), we have

$$
n^{-1}\mathrm{tr}(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n)) = 1 - n^{-1}\mathrm{tr}\mathbf{S}_n(\boldsymbol{\lambda}_n) \to 1. \tag{39}
$$

We also have, for the numerator of (38),

$$
n^{-1}\|(\mathbf{I} - \mathbf{S}_n(\boldsymbol{\lambda}_n))\mathbf{y}_n\|^2 \leq n^{-1}\|\boldsymbol{\epsilon}_n\| + n^{-1}\|\mathbf{f}_n - \hat{\mathbf{f}}_n\|^2 + 2n^{-1}|\langle \boldsymbol{\epsilon}_n, \mathbf{f}_n - \hat{\mathbf{f}}_n\rangle| \to \sigma^2, \tag{40}
$$

by the fact $n^{-1}\|\boldsymbol{\epsilon}_n\|^2 \to \sigma^2$, **(A.3)**, and the Cauchy-Schwartz inequality. Finally, (38) follows from (39) and (40). □

Lemma 12. *Under **(A.3)**, we have $\tilde{L}_n(\hat{\boldsymbol{\lambda}}_{mG}) \to 0$.*

Proof. From the uniform consistency of $SURE_n(\boldsymbol{\lambda})$, together with the fact that $\hat{\boldsymbol{\lambda}}_{mG}$ also minimizes $SURE_n(\boldsymbol{\lambda})$, we have $\tilde{L}_n(\hat{\boldsymbol{\lambda}}_{mG}) = SURE_n(\hat{\boldsymbol{\lambda}}_{mG}) + o_p(1) \leq SURE_n(\boldsymbol{\lambda}_n) + o_p(1) = \tilde{L}_n(\boldsymbol{\lambda}_n) + o_p(1) = o_p(1)$, where the last equality follows from Lemma 11. □

Once again using the fact of the uniform consistency of $SURE_n(\boldsymbol{\lambda})$, from the result of Lemma 12, we have $SURE_n(\hat{\boldsymbol{\lambda}}_{mG}) \to 0$. Equivalently, we have the following Lemma.

Lemma 13. *Under **(A.3)**, we have* $mGCV_n(\hat{\boldsymbol{\lambda}}_{mG}) \to \sigma^2$.

We also need the following Lemmas 14, 15 and 16.

Lemma 14. *If* $\epsilon_i s$ *are i.i.d.* $N(0, \sigma^2)$,

$$\lim_{n\to\infty} P\left\{ \frac{\|(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))\mathbf{y}_n\|^2}{\|(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))\mathbf{f}_n\|^2 + \sigma^2 tr(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))^2} \le 1-\delta \right\} = 0, \; \textit{for any } \delta > 0. \tag{41}$$

Proof. It can be proved similarly as in Li (1985): See the proof of Lemma 5.2 in Li (1985)[pp.1374–1376]. The key is to upper bound five terms in (7.3.8) in Li (1985) with a small quantity $\epsilon/5$. Note that as $n \to \infty$, our smoothing matrix $\mathbf{S}(\boldsymbol{\lambda})$ has the same canonical form of (4.9) in Li (1985) with λ_i replaced by τ_i. A careful check of the proof in Li (1985) reveals that the same argument applies for arbitrary λ_i, hence the above lemma can be established accordingly. □

Lemma 15. *For any sequence* $\hat{\boldsymbol{\lambda}}$ *such that*

$$mGCV_n(\hat{\boldsymbol{\lambda}}) \to \sigma^2, \tag{42}$$

*under **(A.2)**, we have* $n^{-1} tr\mathbf{S}_n(\hat{\boldsymbol{\lambda}}) \to 0$.

Proof. Using (42) and (41), we have $[n^{-1}\text{tr}(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))]^2 \ge [n^{-1}\text{tr}(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))^2](1-o_p(1))$. Then with the fact that $[n^{-1}\text{tr}(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))]^2 \le n^{-1}\text{tr}(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))^2$, we have the following:

$$\frac{[n^{-1}\text{tr}(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))]^2}{n^{-1}\text{tr}(\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}}))^2} \to 1. \tag{43}$$

Recall that as $n \to \infty$, we have $\mathbf{S}_n(\hat{\boldsymbol{\lambda}}) \approx (\mathbf{I} + \boldsymbol{\Sigma}_n(\hat{\boldsymbol{\lambda}}))^{-1}$, where $\boldsymbol{\Sigma}_n(\hat{\boldsymbol{\lambda}}) = \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\hat{\boldsymbol{\lambda}})\mathbf{M}^{-1}\mathbf{Q}$, and the eigenvalues of $\boldsymbol{\Sigma}_n(\hat{\boldsymbol{\lambda}})$ are $0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_n$. It is clear that the eigenvalues of $\mathbf{I}-\mathbf{S}_n(\hat{\boldsymbol{\lambda}})$ are $\tau_i(1+\tau_i)^{-1}$. Similarly as in Li (1985), let τ be the random variable taking values τ_i with probability n^{-1} for each i. Then (43) means $(\mathbb{E}\tau(1+\tau)^{-1})^2/\mathbb{E}\tau^2(1+\tau)^{-2} \to 1$, which implies

$$\tau(1+\tau)^{-1}/\mathbb{E}\tau(1+\tau)^{-1} \to 1 \text{ in probability.} \tag{44}$$

It is known (Girard, 1991) that **(A.2)** can be replaced with the following weaker condition **(A.2′)**: There exist constants p and q, $0 < p < q < 1$ such that $\limsup \tau_{[pn]}/\tau_{[qn]} < 1$, where $[x]$ denotes the greatest integer less than or equal to x. Since (44) implies that both $\tau_{[pn]}(1+\tau_{[pn]})^{-1}$ and $\tau_{[qn]}(1+\tau_{[qn]})^{-1}$ tend to $\mathbb{E}\tau(1+\tau)^{-1}$, we have $\mathbb{E}\tau(1+\tau)^{-1} \to 1$ due to **(A.2′)**. It is clear that $\mathbb{E}\tau(1+\tau)^{-1} \to 1$ implies $n^{-1}\text{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}}) \to 0$. □

Lemma 16. *For any sequence* $\hat{\boldsymbol{\lambda}}$ *such that* $mGCV_n(\hat{\boldsymbol{\lambda}}) \to \sigma^2$, $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}})$ *is consistent if and only if* $n^{-1} tr\mathbf{S}_n(\hat{\boldsymbol{\lambda}}) \to 0$.

Proof. If $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}})$ is consistent, $L_n(\hat{\boldsymbol{\lambda}}) \to 0$, and hence, $n^{-1}\|\mathbf{y}_n - \hat{\mathbf{f}}_n(\boldsymbol{\lambda})\|^2 \to \sigma^2$, since $n^{-1}\|\boldsymbol{\epsilon}_n\| \to \sigma^2$. Then, from the fact that $\text{mGCV}_n(\hat{\boldsymbol{\lambda}}) = n^{-1}\|\mathbf{y}_n - \hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}})\|_2^2/[n^{-1}\text{tr}(\mathbf{I} - \mathbf{S}_n(\hat{\boldsymbol{\lambda}}))]^2 \to \sigma^2$, we have $[n^{-1}\text{tr}(\mathbf{I} - \mathbf{S}_n(\hat{\boldsymbol{\lambda}}))]^2 \to 1$, and thus, $n^{-1}\text{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}}) \to 0$. Conversely, if $n^{-1}\text{tr}\mathbf{S}_n(\hat{\boldsymbol{\lambda}}) \to 0$, since $mGCV_n(\hat{\boldsymbol{\lambda}}) \to \sigma^2$, we have $n^{-1}\|\mathbf{y}_n - \hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}})\|^2 \to \sigma^2$. Then, with the fact that $n^{-1}\|\boldsymbol{\epsilon}_n\| \to \sigma^2$, we have $L_n(\hat{\boldsymbol{\lambda}}) \to 0$, which implies $\hat{\mathbf{f}}_n(\hat{\boldsymbol{\lambda}})$ is consistent. □

From Lemmas 13, 15, and 16, Theorem 6 is proved.

Appendix E: Proof of Theorem 7

Since we assume $n \to \infty$, we have $\hat{\mathbf{f}}(\boldsymbol{\lambda}) \approx (\mathbf{I} + \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\boldsymbol{\lambda})\mathbf{M}^{-1}\mathbf{Q})^{-1}\mathbf{y}$ as $n \to \infty$ provided in Theorem 3. We start with some basic linear algebra facts. It can be shown that for matrix $\Phi = \mathbf{M}^{-1}\mathbf{Q}$, we have $\Phi_{(1,2,\ldots,n-2)\times(2,3,\ldots,n-1)} = 2\mathbf{I} - 6\mathbf{M}^{-1}$. The above can be verified by multiplying $\mathbf{M}$ on both sides; it also indicates that we only need to consider the eigenvalues of the matrix $(2\mathbf{I} - 6\mathbf{M}^{-1})^T\mathbf{M}(\boldsymbol{\lambda})(2\mathbf{I} - 6\mathbf{M}^{-1})$, because one can then apply the interlacing theorem to describe the eigenvalues of $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$. Formally, we have

$$(2\mathbf{I} - 6\mathbf{M}^{-1})^T\mathbf{M}(\boldsymbol{\lambda})(2\mathbf{I} - 6\mathbf{M}^{-1}) = \boldsymbol{\Sigma}(\boldsymbol{\lambda})_{(2,\ldots,n-1)\times(2,\ldots,n-1)}. \tag{45}$$

That is, the left side is a principal submatrix of $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$. It is known (Craven and Wahba, 1979) that $\mathbf{M}^{-1} = \boldsymbol{\Gamma}^T\mathbf{D}_1\boldsymbol{\Gamma}$, where $\mathbf{D}_1 = \text{diag}\{1/(2 + \cos\frac{i\pi}{n})\}, i = 1, 2, \ldots, n-2$, and $\Gamma_{jk} = \sqrt{\frac{2}{n}}\sin\frac{jk\pi}{n}, 1 \le j, k \le n-2$. Noticing that $\boldsymbol{\Gamma}$ is orthogonal, we can easily derive the following: $2\mathbf{I} - 6\mathbf{M}^{-1} = \boldsymbol{\Gamma}^T\ \mathbf{D}_2\boldsymbol{\Gamma}$, where $\mathbf{D}_2 = \text{diag}\{\frac{2\cos\frac{i\pi}{n} - 2}{2 + \cos\frac{i\pi}{n}}\}, 1 \le i \le n-2$. Bringing this into (45), we have $(2\mathbf{I} - 6\mathbf{M}^{-1})^T\mathbf{M}(\boldsymbol{\lambda})(2\mathbf{I} - 6\mathbf{M}^{-1}) = \boldsymbol{\Gamma}^T\mathbf{D}_2\boldsymbol{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\boldsymbol{\Gamma}^T\mathbf{D}_2\boldsymbol{\Gamma}$, which has the same eigenvalues as $\mathbf{D}_2\boldsymbol{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\boldsymbol{\Gamma}^T\mathbf{D}_2$.

We need the following lemma.

Lemma 17. *Suppose matrix $\mathbf{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{A}$ is symmetric positive definite. Suppose the eigenvalues of $\mathbf{A}$ are $\gamma_1(\mathbf{A}) \le \gamma_2(\mathbf{A}) \le \cdots \le \gamma_n(\mathbf{A})$. Let $\mathbf{D}$ be an diagonal matrix $\mathbf{D} = diag\{d_1, d_2, \ldots, d_n\}$, where $0 \le d_1 \le d_2 \le \cdots \le d_n$. Let $\tau_1 \le \tau_2 \le \cdots \le \tau_n$ be the eigenvalues of matrix $\mathbf{DAD}$. We have $\gamma_1(\mathbf{A})d_i^2 \le \tau_i \le \gamma_n(\mathbf{A})d_i^2$.*

Proof. According to the minimax theorem, we have

$$\tau_i = \sup_{\Omega_i}\inf_{x\in\Omega_i}\frac{x^T\mathbf{DAD}x^T}{x^Tx},$$

where Ω_i is an i-dimensional linear subspace of $\mathbb{R}^n(\Omega_i \subset \mathbb{R}^n)$ and x is an n-dimensional vector. We can easily establish the following:

$$\gamma_1(\mathbf{A})x^T\mathbf{D}^2x \le x^T\mathbf{DAD}x \le \gamma_n(\mathbf{A})x^T\mathbf{D}^2x.$$

Applying operator $\sup_{\Omega_i} \inf_{x\in\Omega_i}$ on all the three terms above, we have

$$\gamma_1(\mathbf{A}) \sup_{\Omega_i} \inf_{x\in\Omega_i} \frac{x^T\mathbf{D}^2 x}{x^T x} \le \tau_i \le \gamma_n(\mathbf{A}) \sup_{\Omega_i} \inf_{x\in\Omega_i} \frac{x^T\mathbf{D}^2 x}{x^T x}.$$

Notice that the first term above is $\gamma_1(\mathbf{A})d_i^2$ and the last term is $\gamma_n(\mathbf{A})d_i^2$. □

Due to the above lemma, if $\tau_1 \le \tau_2 \le \cdots \le \tau_{n-2}$ are the eigenvalues of $\mathbf{D}_2\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T\mathbf{D}_2$, we have

$$\gamma_{\min}\left(\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T\right) \frac{4(1-\cos\frac{i\pi}{n})^2}{(2+\cos\frac{i\pi}{n})^2} \le \tau_i \le \gamma_{\max}\left(\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T\right) \frac{4(1-\cos\frac{i\pi}{n})^2}{(2+\cos\frac{i\pi}{n})^2},$$

where $\gamma_{\min}(\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T)$ and $\gamma_{\max}(\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T)$ are the minimum and maximum eigenvalues of $\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T$. Recall $\mathbf{\Gamma}$ is orthogonal, so we have $\gamma_{\min}(\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T) = \gamma_{\min}(\mathbf{M}(\boldsymbol{\lambda}))$ and $\gamma_{\max}(\mathbf{\Gamma}\mathbf{M}(\boldsymbol{\lambda})\mathbf{\Gamma}^T) = \gamma_{\max}(\mathbf{M}(\boldsymbol{\lambda}))$. We will need the following lemma to examine the coefficient of variations condition.

Lemma 18. *If there are two functions* $0 < C_1(n) < C_2(n)$ *such that*

$$\frac{C_2(n)}{C_1(n)} < c \text{ where } c > 0 \text{ is a constant,}$$

and

$$C_1(n)\cdot\frac{\left(1-\cos\frac{i\pi}{n}\right)^2}{(2+\cos\frac{i\pi}{n})^2} \le \tau_i \le C_2(n)\cdot\frac{\left(1-\cos\frac{i\pi}{n}\right)^2}{(2+\cos\frac{i\pi}{n})^2},$$

then we have, for any m *satisfying* $\frac{m}{n} \to 0$,

$$\frac{\left(\sum_m^n \tau_i^{-1}/n\right)^2}{\sum_m^n \tau_i^{-2}/n} \to 0 \quad as \quad n \to \infty.$$

Proof. For the numerator, we have

$$\begin{aligned} \frac{1}{n}\sum_m^n \frac{(2+\cos\frac{i\pi}{n})^2}{\left(1-\cos\frac{i\pi}{n}\right)^2} &\le \frac{1}{n}\sum_m^n \frac{9}{\left(1-\cos\frac{i\pi}{n}\right)^2} \approx \int_{\frac{m}{n}}^1 \frac{9}{(1-\cos x\pi)^2}dx \\ = & -\frac{(-2+\cos\frac{m\pi}{n})\cot\frac{m\pi}{2n}(\csc\frac{m\pi}{2n})^2}{2\pi} = A_1. \end{aligned}$$

For the denominator, we have

$$\begin{aligned} \frac{1}{n}\sum_m^n \frac{\left(2+\cos\frac{i\pi}{n}\right)^4}{\left(1-\cos\frac{i\pi}{n}\right)^4} &\ge \frac{1}{n}\sum_m^n \frac{1}{\left(1-\cos\frac{i\pi}{n}\right)^4} \approx \int_{\frac{m}{n}}^1 \frac{1}{(1-\cos x\pi)^4}dx \\ = & -\frac{(-32+29\cos\frac{m\pi}{n} - 8\cos\frac{2m\pi}{n} + \cos\frac{3m\pi}{n})\cot\frac{m\pi}{2n}(\csc\frac{m\pi}{2n})^6}{560\pi} = A_2. \end{aligned}$$

Now we put everything together:

$$\begin{aligned}\frac{\left(\sum_m^n \frac{1}{\tau_i}/n\right)^2}{\sum_m^n \frac{1}{\tau_i^2}/n} &\leq \frac{A_1^2 C_1^{-2}(n)}{A_2 C_2^{-2}(n)} \\ &= -\frac{C_2^2(n)}{C_1^2(n)} \frac{420 \cos\frac{m\pi}{2n}(-2+\cos\frac{m\pi}{n})^2 \sin\frac{m\pi}{2n}}{\pi(-32+29\cos\frac{m\pi}{n} - 8\cos\frac{2m\pi}{n} + \cos\frac{3m\pi}{n})} \to 0. \qquad \square\end{aligned}$$

Combining the discussion after Lemma 17 and Lemma 18, if we can find a sufficient condition of $\mathbf{M}(\boldsymbol{\lambda})$ such that $\gamma_{\max}(\mathbf{M}(\boldsymbol{\lambda}))/\gamma_{\min}(\mathbf{M}(\boldsymbol{\lambda}))$ is upper bounded by a constant, then we find a sufficient condition for **(A.2)**. We have the following lemma.

Lemma 19. *Let $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum among $\lambda_k's$, $1 \leq k \leq n-1$. We have*

$$\frac{\gamma_{\max}(\mathbf{M}(\boldsymbol{\lambda}))}{\gamma_{\min}(\mathbf{M}(\boldsymbol{\lambda}))} \leq 3 \cdot \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Proof. Let $\mathbf{M}(\boldsymbol{\lambda})_{ij}$ be the (i,j) entry of $\mathbf{M}(\boldsymbol{\lambda})$, and γ denote an eigenvalue of $\mathbf{M}(\boldsymbol{\lambda})$. Due to the Gershgorin circle theorem (Horn and Johnson, 1985), we have $|\gamma - \mathbf{M}(\boldsymbol{\lambda})_{ii}| \leq \sum_{j\neq i} |\mathbf{M}(\boldsymbol{\lambda})_{ij}|$, $1 \leq i \leq n-2$. Consequently, we have

$$|\gamma| \geq |\mathbf{M}(\boldsymbol{\lambda})_{ii}| - |\gamma - \mathbf{M}(\boldsymbol{\lambda})_{ii}| \geq |\mathbf{M}(\boldsymbol{\lambda})_{ii}| - \sum_{j\neq i} |\mathbf{M}(\boldsymbol{\lambda})_{ij}|. \tag{46}$$

By recalling the structure of $\mathbf{M}(\boldsymbol{\lambda})$, we have that

$$|\mathbf{M}(\boldsymbol{\lambda})_{11}| - \sum_{j\neq 1} |\mathbf{M}(\boldsymbol{\lambda})_{1j}| = \lambda_1 + \frac{1}{2}\lambda_2, \tag{47}$$

$$|\mathbf{M}(\boldsymbol{\lambda})_{n-2,n-2}| - \sum_{j\neq n-2} |\mathbf{M}(\boldsymbol{\lambda})_{n-2,j}| = \frac{1}{2}\lambda_{n-2} + \lambda_{n-1}, \text{ and} \tag{48}$$

$$|\mathbf{M}(\boldsymbol{\lambda})_{ii}| - \sum_{j\neq i} |\mathbf{M}(\boldsymbol{\lambda})_{ij}| = \frac{1}{2}\lambda_i + \frac{1}{2}\lambda_{i+1}, \quad 2 \leq i \leq n-3. \tag{49}$$

Combining (46) through (49), we have $\gamma_{\min}(\mathbf{M}(\boldsymbol{\lambda})) \geq \lambda_{\min}$. On the other hand, we have $|\gamma| - |\mathbf{M}(\boldsymbol{\lambda})_{ii}| \leq |\gamma - \mathbf{M}(\boldsymbol{\lambda})_{ii}| \leq \sum_{j\neq i} |\mathbf{M}(\boldsymbol{\lambda})_{ij}|$. Consequently, we have

$$|\gamma| \leq |\mathbf{M}(\boldsymbol{\lambda})_{ii}| + \sum_{j\neq i} |\mathbf{M}(\boldsymbol{\lambda})_{ij}|. \tag{50}$$

On the other hand, we have

$$|\mathbf{M}(\boldsymbol{\lambda})_{11}| + \sum_{j\neq 1} |\mathbf{M}(\boldsymbol{\lambda})_{1j}| = \lambda_1 + \frac{3}{2}\lambda_2, \tag{51}$$

$$|\mathbf{M}(\boldsymbol{\lambda})_{n-2,n-2}| + \sum_{j\neq n-2} |\mathbf{M}(\boldsymbol{\lambda})_{n-2,j}| = \frac{3}{2}\lambda_{n-2} + \lambda_{n-1}, \text{ and} \tag{52}$$

$$|\mathbf{M}(\boldsymbol{\lambda})_{ii}| + \sum_{j\neq i} |\mathbf{M}(\boldsymbol{\lambda})_{ij}| = \frac{3}{2}\lambda_i + \frac{3}{2}\lambda_{i+1}, \quad 2 \leq i \leq n-3. \tag{53}$$

From (50) through (53), we have $\gamma_{\max}(\mathbf{M}(\boldsymbol{\lambda})) \le 3\lambda_{\max}$. From all the above, we prove the lemma. □

The above lemma demonstrates that if we have $\lambda_{\max}/\lambda_{\min}$ upper bounded by a constant, then the **(A.2)** type condition is true for the terms on both sides of (45). Note they are the $(n-2)\times(n-2)$ principal submatrix of $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$. The relation of the eigenvalues of a principal submatrix and the eigenvalues of the original matrix is known due to the interlacing theorem by Cauchy (Stewart, 1977). Cauchy's interlacing theorem states that if we let τ_i' be the eigenvalues of $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$, then $\tau_i' \le \tau_i \le \tau_{i+1}'$. In essence, they have nearly identical behavior. It is not hard to show that when the condition in Lemma 19 is met, the condition **(A.2)** is satisfied for $\boldsymbol{\Sigma}(\boldsymbol{\lambda})$ as well. We skip some details and claim that Theorem 7 is established.

Appendix F: Proof of Theorem 9

For the proof, we use Theorem 5. Note that in Theorem 5, besides **(A.1)**, we need two conditions: (15) and (16). Under the conditions **(A.2)** and **(A.3)**, (15) is true with $\hat{\boldsymbol{\lambda}}_{mG}$ due to Theorem 6. The following lemma states that under **(A.1)** and **(A.2)**, (15) implies (16); therefore we prove Theorem 9.

Lemma 20. *Under **(A.1)** and **(A.2)**, for any $\boldsymbol{\lambda}$ such that*

$$L_n(\boldsymbol{\lambda}) \to 0, \tag{54}$$

we have

$$\frac{(n^{-1}tr\mathbf{S}_n(\boldsymbol{\lambda}))^2}{n^{-1}tr\mathbf{S}_n^2(\boldsymbol{\lambda})} \to 0, \tag{55}$$

Proof. Recall that as $n \to \infty$, we have $\mathbf{S}_n(\boldsymbol{\lambda}) \approx (\mathbf{I} + \boldsymbol{\Sigma}_n(\boldsymbol{\lambda}))^{-1}$, where $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda}) = \frac{h}{3}\mathbf{Q}^T\mathbf{M}^{-1}\mathbf{M}(\boldsymbol{\lambda})\mathbf{M}^{-1}\mathbf{Q}$, and the eigenvalues of $\boldsymbol{\Sigma}_n(\boldsymbol{\lambda})$ are $0 \le \tau_1 \le \tau_2 \le \cdots \le \tau_n$. We have

$$\frac{(n^{-1}\mathrm{tr}\mathbf{S}_n(\boldsymbol{\lambda}))^2}{n^{-1}\mathrm{tr}\mathbf{S}_n^2(\boldsymbol{\lambda})} = \frac{\left(n^{-1}\sum_{i=1}^n (1+\tau_i)^{-1}\right)^2}{n^{-1}\sum_{i=1}^n (1+\tau_i)^{-2}}.$$

Define $m = i$ such that $\tau_i \le 1 \le \tau_{i+1}$. Then we have

$$\sum_{i=1}^n (1+\tau_i)^{-1} \le m + \sum_{i=m+1}^n \tau_i^{-1},$$

and

$$\sum_{i=1}^n (1+\tau_i)^{-2} \ge \frac{1}{4}\left(m + \sum_{i=m+1}^n \tau_i^{-2}\right). \tag{56}$$

To show (55), therefore, it suffices to show

$$\frac{\left(\frac{m}{n} + \frac{1}{n}\sum_{i=m+1}^n \tau_i^{-1}\right)^2}{\frac{m}{n} + \frac{1}{n}\sum_{i=m+1}^n \tau_i^{-2}} \to 0. \tag{57}$$

On the other hand, from (54), we have $\mathbb{E}L_n(\boldsymbol{\lambda}) \to 0$ due to (28), and hence $n^{-1}\mathrm{tr}\mathbf{S}_n^2(\boldsymbol{\lambda}) \to 0$. Thus $m/n \to 0$ due to (56). Bringing this to (57), we have (55) under **(A.2)**. □

Appendix G: Proof of Theorem 10

It is clear that (54) is equivalent to (15). Then if we have any $\hat{\boldsymbol{\lambda}}$ such that $L_n(\hat{\boldsymbol{\lambda}}) \to 0$, by Lemma 20, we can utilize Theorem 5. Under **(A.3)**, this holds for $\hat{\boldsymbol{\lambda}} = \boldsymbol{\lambda}_n^*$, where $\boldsymbol{\lambda}_n^*$ is the minimizer of $L_n(\boldsymbol{\lambda})$. Thus we have

$$SURE_n(\boldsymbol{\lambda}_n^*) - n^{-1}\|\boldsymbol{\epsilon}_n\|^2 + \sigma^2 = L_n(\boldsymbol{\lambda}_n^*)(1 + o_p(1)).$$

On the other hand, by Theorem 6, this also holds for $\hat{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}_{mG}$. Therefore we have

$$SURE_n(\hat{\boldsymbol{\lambda}}_{mG}) - n^{-1}\|\boldsymbol{\epsilon}_n\|^2 + \sigma^2 = L_n(\hat{\boldsymbol{\lambda}}_{mG})(1 + o_p(1)).$$

From the fact that $SURE_n(\hat{\boldsymbol{\lambda}}_{mG}) \leq SURE_n(\boldsymbol{\lambda}_n^*)$ and $L_n(\boldsymbol{\lambda}_n^*) \leq L_n(\hat{\boldsymbol{\lambda}}_{mG})$, we have $L_n(\hat{\boldsymbol{\lambda}}_{mG})/L_n(\boldsymbol{\lambda}_n^*) \to 1$.

References

Abramovich, F. and Steinberg, D. M. (1996). Improved inference in nonparametric regression using L_k-smoothing splines. *Journal of Statistical Planning and Inference* **49** 327–341. MR1381163

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31** 377–403. MR0516581

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* **90** 1200–1224. MR1379464

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York. MR1680784

Girard, D. A. (1991). Asymptotic optimality of the fast randomized versions of GCV and C_L in ridge regression and regularization. *The Annals of Statistics* **19** 1950–1963. MR1135158

Green, P. J. and Silverman, B. W. (1994). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. *Monographs on Statistics and Applied Probability* **58**. Chapman & Hall, New York, NY. MR1270012

Horn, R. and Johnson, C. R. (1985). *Matrix Analysis*. Cambridge University Press, Cambridge. MR0832183

Kim, H. and Huo, X. (2012). Locally optimal adaptive smoothing splines. *Journal of Nonparametric Statistics* **24** 665–680. MR2968895

Li, K. C. (1985). From Stein's unbised risk estimates to the method of generalized cross validation. *The Annals of Statistics* **13** 1352–1377. MR0811497

Li, K. C. (1986). Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing. *The Annals of Statistics* **14** 1101–1112. MR0856808

Liu, Z. and Guo, W. (2010). Data driven adaptive spline smoothing. *Statistica Sinica* **20** 1143–1163. MR2730177

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

Pintore, A., Speckman, P. and Holmes, C. C. (2006). Spatially adaptive smoothing splines. *Biometrika* **93** 113–125. MR2277744

Stewart, G. W. (1977). Perturbation bounds for the QR factorization of a matrix. *SIAM Journal on Numerical Analysis* **14** 509–518. MR0436566

Stone, M. (1974). Cross-validation and multinomial prediction. *Biometrika* **61** 509–515. MR0415896

Storlie, C. B., Bondell, H. D. and Reich, B. J. (2010). A locally adaptive penalty for estimation of functions with varying roughness. *To appear in Journal of Computational and Graphical Statistics*. MR2732493

Wahba, G. (1985). A comparison of GCV and GML for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics* **13** 1378–1402. MR0811498

Wang, X., Du, P. and Shen, J. (2013). Smoothing splines with varying smoothing parameter. *Biometrika*. To appear.