CARLETON UNIVERSITY

SCHOOL OF MATHEMATICS AND STATISTICS

HONOURS PROJECT

TITLE: Smoothing splines in nonparametric regression analysis

AUTHOR: Xiang Zhao

SUPERVISOR: Dr. Natalia Stepanova

DATE: April 30th, 2021

Abstract

In this project, we consider the smoothness penalty approach to find the smoothing spline estimator for the regression curve in nonparametric regression analysis, and investigate its mathematical and statistical properties. Also, we discuss the smoothing parameter selection criteria and examine the algorithms for smoothing spline computation. This project presents the theory of smoothing spline estimators in detail and is supplemented by the solutions to selected problems posed in the book Nonparametric Regression and Spline Smoothing, by Eubank [**2**].

Acknowledgement

Foremost, I would like to thank my supervisor, Dr. Natalia Stepanova, for her support, guidance, and most importantly the time she has given in these unprecedented situations. I am also grateful to my second reader, Dr. Mohamedou Ould Haye. I would also like to extend my thanks to my professors and the teaching assistants in the School of Mathematics and Statistics at Carleton University for everything they have done for me. Lastly, I would like to thank my friends and family for their love and support throughout my undergraduate journey.

Contents

1 Introduction **6**

1.1 Splines 8

1.2 Regression Splines 12

1.3 Smoothing Splines 13

2 Form of the Estimator **17**

2.1 PLS Estimator 17

2.2 The Demmler-Reinsch basis 22

3 Selection of smoothing parameter **33**

3.1 Prediction Risk 33

3.2 Cross-validation 37

4 Efficient Computation **41**

5 Large-Sample Properties **48**

6 Conclusion **65**

7 Appendix **66**

8 References **72**

9 R Code **74**

1 Introduction

In regression analysis, we build mathematical models to estimate the relationship between the independent variable (often referred to as 'predictor') and the dependent variable (often referred to as 'response'); these models may be used to draw statistical inferences and make predictions.

To begin, consider the case where the independent variables $t_i, i = 1, \ldots, n$, are predetermined values on the interval $[0, 1]$ and the dependent variables $Y_i, i = 1, \ldots, n$, are independent continuous random variables defined at different values of t_is. Now, for pairs $(t_i, Y_i), i = 1, \ldots, n$, assume that t_i and Y_i are related through:

$$Y_i = \mu(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \tag{1.1}$$

where ϵ_i are zero mean independent and identically distributed (iid) random variables with common variance σ^2, and $\mu : [0, 1] \to \mathbb{R}$ is the regression function which is unknown and needs to be estimated.

In a parametric regression framework, we often approximate such regression function by a straight line:

$$\mu(t) = \beta_0 + \beta_1 t$$

or, for a nonlinear regression function, by a polynomial of order m:

$$\mu(t) = \beta_0 + \beta_1 t + \cdots + \beta_{m-1} t^{m-1}. \tag{1.2}$$

That is, the parametric regression model assumes that the regression function $\mu(t)$ has a known explicit form with a finite number of unknown parameters.

These forms of regression functions may be deduced from scientific theories and therefore provide valid inferences in some fields, but such assumptions can be restrictive or even invalid for many applications where the relation between the variables is not clearly explained by some theories.

In contrast, the nonparametric model does not assume that the regression function has a particular parametric form, instead it only assumes that μ belongs to some infinite-dimensional collection of functions. The choice usually depends on the qualitative properties of μ such as the 'smoothness' of μ. The basic idea of nonparametric regression is to let the data speak for itself rather than make restrictive assumptions when there are little prior information available. That is, let the data decide the best form of the regression function μ. It is also important to note that, in nonparametric regression analysis, we do not assume a particular form for μ, but this does not mean we cannot estimate μ by some linear combination of basis functions.

Typically, it is only assumed that μ belongs to some function space. There are many choices for the function space, depending on the prior beliefs of the smoothness of a regression function μ. One reasonable choice is all continuous functions on the interval $[0, 1]$ (or any other finite interval $[a, b]$; there is no loss of generality since we can always find a continuous mapping function to scale the interval), or, more generally, all functions that has m continuous derivatives on $[0, 1]$. We will focus on the mth order *Sobolev space* denoted by $W_2^m[0, 1]$ and defined as follows:

$$W_2^m[0,1] = \{f : [0,1] \to \mathbb{R} \mid f, f', \ldots, f^{(m-1)} \text{ are absolutely continuous,} \int_0^1 (f^{(m)}(t))^2 dt < \infty\}. \tag{1.3}$$

In connection with the nonparametric regression model (1.1), the standard measure of goodness of fit is the average residual sum of squares (RSS):

$$\frac{1}{n} RSS(\mu) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(t_i))^2. \tag{1.4}$$

If $\mu \in W_2^m[0,1]$, then a direct fit to minimize (1.4) will be the dependent vector $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ resulting RSS to be 0 and the regression function μ to be an interpolation of data points. Interpolation is typically not the goal of regression analysis, instead the goal is to come up with some function that represents the process that could have generated the data. Therefore some penalty term of smoothness needs to be added. A natural measure of smoothness in space $W_2^m[0,1]$ is $\int_0^1 (\mu^{(m)}(t))^2 dt$. (We will comment on this in detail below.) Thus, the overall assessment of the regression function could be provided by the *Penalized Least Squares* (PLS), which is defined as

$$\frac{1}{n} \sum_{i=1}^{n} (Y_i - \mu(t_i))^2 + \lambda \int_0^1 (\mu^{(m)}(t))^2 dt, \tag{1.5}$$

for some $\lambda > 0$. As we will see later on, the minimizer to (1.5) is the smoothing spline.

1.1 Splines

Out of many techniques used in nonparametric regression analysis, splines are widely used. In general, a *spline* is a special function defined piecewisely

by polynomials. More formally, let $0 < t_1 < \cdots < t_n < 1$ be fixed points called *knots* with $t_0 = 0$ and $t_{n+1} = 1$. A *spline* of order r (or degree $r-1$) with knots at $(t_1, \ldots, t_n)$ is any real-valued function S defined on $[0,1]$ of the form

$$S(t) = \sum_{j=0}^{r-1} \theta_j t^j + \sum_{j=1}^{n} \eta_j (t - t_j)_+^{r-1}, \tag{1.6}$$

where

$$(x)_+^r = \begin{cases} x^r, & x \geq 0, \\ 0, & x < 0, \end{cases}$$

and $\theta_0, \ldots, \theta_{r-1}, \eta_1, \ldots, \eta_k$ are some coefficients. This definition is equivalent to saying that:

1) S is a piece-wise polynomial of order r on any subintervals $[t_i, t_{i+1}]$; hence by (1.6),

$$S(t) = \begin{cases} \sum_{j=0}^{r-1} \theta_j t^j, & 0 \leq t < t_1, \\ \sum_{j=0}^{r-1} \theta_j t^j + \eta_1 (t - t_1)^{r-1}, & t_1 \leq t < t_2, \\ \vdots & \\ \sum_{j=0}^{r-1} \theta_j t^j + \sum_{j=1}^{n} \eta_j (t - t_j)^{r-1}, & t_n < t \leq 1, \end{cases} \tag{1.7}$$

2) S has $r-2$ continuous derivative, namely, $S^{(j)}$ for $j = 1, \ldots, r-2$ with

$$S'(t_i^-) = S'(t_i^+), \ldots, S^{(r-2)}(t_i^-) = S^{(r-2)}(t_i^+), \quad \text{for} \quad i = 1, \ldots, n-1,$$

where $S'(t_i^-) = \lim_{x \nearrow t_i^-} S'(x)$, that is, the limit as x approaches t_i from the left side. And $S'(t_i^+) = \lim_{x \searrow t_i^+} S'(x)$, which means the limit as x approaches t_i from the right side.

3) S has a discontinuous $(r-1)$st derivative with jumps at knots $t_1, \ldots, t_n$, since it follows from (1.6) that

$$S^{(r-1)}(t_i^-) \neq S^{(r-1)}(t_i^+), \quad \text{for} \quad i = 1, \ldots, n-1,$$

and we can see that the $(r-1)$st derivative at t_1 satisfies

$$(r-1)!\,\theta_{r-1} = S^{(r-1)}(t_1^-) \neq S^{(r-1)}(t_1^+) = (r-1)!\,(\theta_{r-1} + \eta_1);$$

the other cases are treated analogously.

Now, consider a spline of even order $r = 2m$. A *spline* of even order $2m$ is a *natural spline* of order $2m$ with knots at $(t_1, \ldots, t_n)$ if in addition to 1), 2) and 3), it also satisfies the *natural boundary conditions*

4) $S^{(j)}(0) = S^{(j)}(1) = 0, \quad j = m, \ldots, 2m-1.$

The natural boundary conditions imply that S is a polynomial of order m (or degree $m-1$) outside of $[t_1, t_n]$. Specifically, using (1.6),

$$S(t) = \begin{cases} \sum_{j=0}^{m-1} \theta_j t^j, & 0 \leq t < t_1, \\ \sum_{j=0}^{m-1} \theta_j t^j + \sum_{j=1}^{n} \eta_j (t - t_j)^{m-1}, & t_n < t \leq 1. \end{cases} \tag{1.8}$$

Natural splines are particular important for our study because, as will be demonstrated later on, the natural spline estimator of μ tends to display a smoother behavior at the boundaries of the observation interval than spline estimator.

From the above definitions, we can also deduce that the set of splines of order r with knots $(t_1, \ldots, t_n)$ has dimension $n + r$, since the basis functions $1, t, \ldots, t^{r-1}, (t - t_1)_+^{r-1}, \ldots, (t - t_n)_+^{r-1}$ are linearly independent. And the

subset of natural splines has dimension n. Indeed, initially there are $n + 2m$ basis functions, but from (1.7) and (1.8) we have:

$$\theta_m = \cdots = \theta_{2m-1} = 0, \quad \text{for} \quad 0 \leq t < t_1 \quad \text{and} \quad t_n < t \leq 1,$$

that is, all functions that have the order greater than m outside of the interval $[t_1, t_n]$ must be excluded from the basis functions. These additional constraints reduce the number of basis functions by $r = 2m$, giving that the dimension of the set of natural splines is n. This is an important property: the dimension of the set of natural splines of order r with knots $(t_1, \ldots, t_n)$ only depends on the number of knots.

Example 1: Cubic Splines.

Cubic splines (or splines of order $2m = 4$) are commonly used in applications. One of the reasons is that cubic spline overcomes the overfitting problem, as shown in Figure 1 taken from page 18 of [**11**].

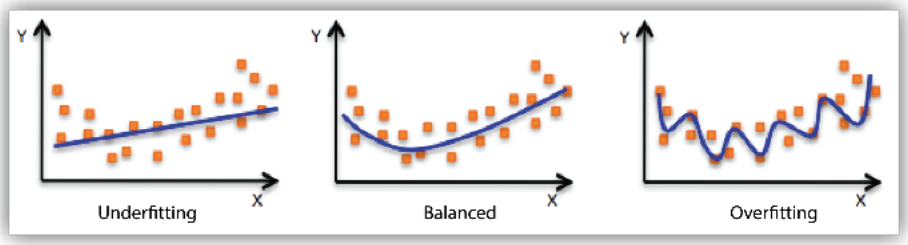


Figure 1: Overfitting can occur since higher order spline allows more flexibility and will follow the data points more closely.

It is worth noting, there might be situations where higher order spline fit the data better than cubic splines. But often for data that contains noise,

higher order splines could mistakenly take noise as signal, thus, as a rule of thumb, we recommend not to go higher than cubic splines in applications. The explicit form of such spline is:

$$S(t) = \theta_0 + \theta_1 t + \theta_2 t^2 + \theta_3 t^3 + \eta_1 (t - t_1)_+^3 + \cdots + \eta_n (t - t_n)_+^3, \quad t \in [0, 1].$$

1.2 Regression Splines

Return to the nonparametric regression model (1.1). For pairs $(t_i, Y_i), i = 1, \ldots, n$, we can estimate the regression function μ by spline of order r with chosen knots $(\xi_1, \ldots, \xi_m)$ with $0 \leq \xi_1 < \cdots < \xi_m \leq 1$. The change of notation here is due to the special design of regression spline in which the knots do not necessarily coincide with the design points. Rewriting equation (1.6) as

$$S(t) = \sum_{j=1}^{m+r} \beta_j x_j(t),$$

where the functions $x_j, j = 1, \ldots, m + r$, form the truncated power basis and are defined as follows:

$$x_1(t) = 1, \quad x_2(t) = t, \quad x_3(t) = t^2, \ldots, x_r = t^{r-1},$$
$$x_{r+1}(t) = (t - \xi_1)_+^{r-1}, \ldots, x_{r+m}(t) = (t - \xi_n)_+^{r-1},$$

we can estimate the coefficients $\beta_j, j = 1, \ldots, m + r$, by the *least squares* (LS) method. This method prescribes to minimize the following function:

$$\sum_{i=1}^{n} (Y_i - S(t_i))^2 = \sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{m+r} \beta_j x_j(t) \right)^2 \tag{1.9}$$

with respect to $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{m+r})^\top \in \mathbb{R}^{m+r}$.

Let us define the matrix $\boldsymbol{X} = \{x_j(t_i)\}_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, m+r$, and the vector $\boldsymbol{Y}$ as follows:

$$\boldsymbol{X} = \begin{bmatrix} x_1(t_1) & \ldots & x_{m+r}(t_1) \\ \vdots & \ddots & \vdots \\ x_1(t_n) & \ldots & x_{m+r}(t_n) \end{bmatrix}, \quad \boldsymbol{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}.$$

Then, in matrix notation, (1.9) takes the form: $(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})$. Taking the derivative with respect to (w.r.t) $\boldsymbol{\beta}$ and setting it equals to 0, we obtain the solution $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_{m+r})^\top$ in the form

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}.$$

It should be mentioned that it is necessary to have the knots different than the data points, so there is no a column of zeros in the matrix $\boldsymbol{X}$, in order for $\boldsymbol{X}^\top \boldsymbol{X}$ to be invertible.

The estimated regression function is then as follows:

$$\hat{\mu}(t) = \sum_{j=1}^{m+r} \hat{\beta}_j x_j(t), \quad t \in [0, 1].$$

The regression spline method can work well provided we choose good knots $\xi_1, \ldots, \xi_m$. One way to obtain these knots is by visually check the data scatter plot, and look for the 'turning points'. But, in general, choosing knots is tricky and it may be preferable to use some data driven methods. This is the beauty of smoothing splines, we don't have to choose the knots.

1.3 Smoothing Splines

Smoothing spline techniques are interesting because they places the knots at each data point and overcome the problem of overfitting by integrating a

smoothness penalty into the measure of goodness of fit. Now, recall regression model (1.1) and assume that $\mu \in W_2^m[0,1]$. Taylor's theorem implies that for $i = 1, \ldots, n$, (see Theorem 3.4 on page 121 of [**2**])

$$\mu(t_i) = \sum_{j=0}^{m-1} \frac{\mu^{(j)}(0)}{j!}(t_i - 0)^j + \int_0^1 \frac{(t_i - \xi)_+^{m-1}}{(m-1)!}\mu^{(m)}(\xi)d\xi.$$

Note that

$$\int_0^1 (t-\xi)_+^{m-1} d\xi = \int_0^t (t-\xi)^{m-1} d\xi = \frac{t^m}{m}.$$

The polynomial model (1.1) – (1.2) ignores the remainder term

$$\int_0^1 \frac{(t_i - \xi)_+^{m-1}}{(m-1)!}\mu^{(m)}(\xi)d\xi,$$

while model (1.1) with $\mu \in W_2^m[0,1]$ takes this term into consideration and let the data decide how large the remainder should be. Indeed, from the Cauchy-Schwartz inequality we have

$$\left(\int_0^1 \frac{(t_i - \xi)_+^{m-1}}{(m-1)!}\mu^{(m)}(\xi)d\xi\right)^2 \leq \frac{\int_0^1 (\mu^{(m)}(\xi))^2 d\xi}{[(m-1)!]^2}\frac{t_i^{2m-1}}{(2m-1)},$$

and, by taking the maximum value of t_i, we obtain

$$\max_{1\leq i\leq n}\left(\int_0^1 \frac{(t_i - \xi)_+^{m-1}}{(m-1)!}\mu^{(m)}(\xi)d\xi\right)^2 \leq \frac{\int_0^1 (\mu^{(m)}(\xi))^2 d\xi}{[(m-1)!]^2(2m-1)}.$$

This result implies that if we knew that

$$\int_0^1 (\mu^{(m)}(t))^2 dt \leq \rho \quad \text{for some} \quad \rho > 0, \tag{1.10}$$

then we would have a bound on how far apart is μ from a polynomial model. We could add this information in the estimation process, by minimizing the

average RSS as in (1.4) with the constraint (1.10) using the Lagrange multiplier method, that is, by minimizing

$$\frac{1}{n}RSS(\mu) + \lambda\left(\int_0^1 (\mu^{(m)}(t))^2 dt - \rho\right). \tag{1.11}$$

In the view of Theorem 1 below, this constrained minimization problem is equivalent to minimizing the PLS as defined in (1.5):

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(t_i))^2 + \lambda\int_0^1 (\mu^{(m)}(t))^2 dt.$$

This result was proved in [**8**].

Theorem 1 (Theorem 5.1 on page 229 of [**2**])**.** *Assume that $n \geq m$ and let $\hat{\mu}(\cdot;\rho)$ be the minimizer of (1.11) in $W_2^m[0,1]$. Let $\hat{\mu}_\lambda$ denote the minimizer of (1.5) in $W_2^m[0,1]$. Then there is a computable constant ρ_0 such that the sets $\{\hat{\mu}(\cdot;\rho)\colon 0 \leq \rho \leq \rho_0\}$ and $\{\hat{\mu}_\lambda(\cdot)\colon 0 \leq \lambda \leq \infty\}$ are identical in that for any value of λ there is a unique ρ such that $\hat{\mu}_\lambda(\cdot) = \hat{\mu}(\cdot;\rho)$ and conversely. If $\rho \leq \rho_0$, then $\int_0^1(\hat{\mu}(\cdot;\rho)^{(m)}(t))^2dt = \rho$.*

The minimizer of (1.5), denoted by $\hat{\mu}_\lambda$, is a smoothing spline that overcomes the problem of knot selection and for a given value of λ the data will decide the distance $\hat{\mu}_\lambda$ departs from the polynomial model.

Another way to look at PLS is to think of $\int_0^1(\mu^{(m)}(t))^2dt$ as a measure of smoothness. Looking at the penalty term in (1.5), we would want μ to have a small mth derivative, that is, a less 'wiggly' μ, provided it also has good fit to the data. The integration part in (1.5) gives an overall measure and the square part accounts for the negative derivatives. Smoothing spline takes data points as knots, and by adding the penalty term that accounts for smoothness, it allows the data to decide the suitable regression function.

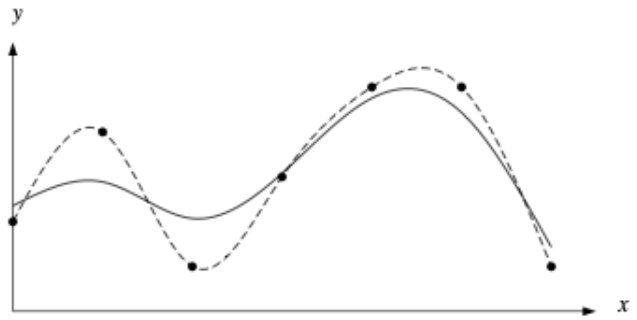
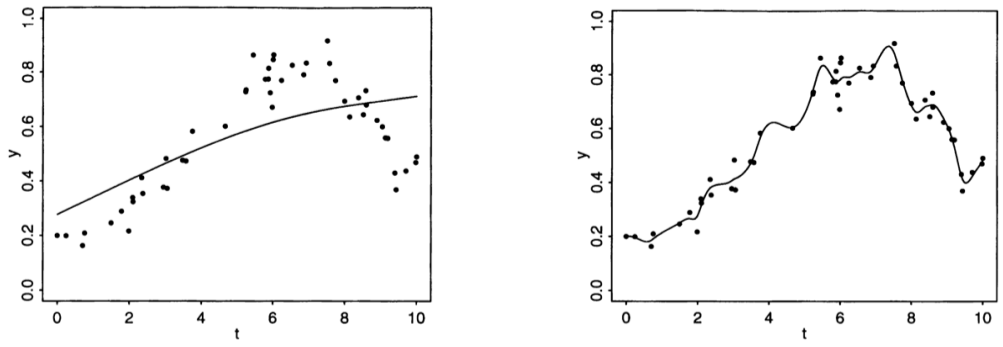


Figure 2: The dashed line is a cubic spline, the solid line is the cubic smoothing spline. Notice that the cubic splines interpolates the data while the smoothing spline do not fit the data exactly and has less up and down movement, this is because of the added smoothness penalty (see Figure 3 on page 307 of [**7**]).

The parameter λ is often refereed to as the smoothing parameter. It governs the trade off between the smoothness and goodness of fit. If λ is large, then the main component of (1.5) will be the penalty term, hence the minimizer $\hat{\mu}_\lambda$ will have little curvature. When $\lambda = \infty$, $\hat{\mu}_\lambda$ is the mth order polynomial, since the infinity term will be forced to 0 and $\hat{\mu}_\lambda$ will approach the linear regression fit. And if λ is small, then the main component of (1.5) will be the RSS, and the curve will track the data closely. When $\lambda = 0$, $\hat{\mu}_\lambda$ interpolates the data points. Figure 3 taken from pages 6 and 7 of [**4**] illustrates this point.



(a) Large λ (b) Small λ

Figure 3: Effects of smoothing parameter

2 Form of the Estimator

2.1 PLS Estimator

In this section, we obtain the explicit expression for the smoothing spline estimator as a minimizer of (1.5) assuming the value of λ is predetermined. This is an infinite-dimensional minimization problem over all $\mu \in W_2^m[0,1]$. Remarkably, such minimizer is unique and is a natural spline of order $2m$ with knots at $t_1, \ldots, t_n$. Here are the theorem and its proof in the cubic case ($2m = 4$) as stated on page 17 of [**4**].

Theorem 2 . *If $\tilde{\mu}$ is any twice differentiable function on the interval $[0,1]$, then there exists a natural cubic spline μ with knots at $t_1, \ldots, t_n$ such that $\tilde{\mu}(t_i) = \mu(t_i)$ for $i = 1, \ldots, n$ and*

$$\int_0^1 (\mu''(t))^2 \, dt \leq \int_0^1 (\tilde{\mu}''(t))^2 \, dt,$$

giving,

$$\frac{1}{n}RSS(\mu) + \int_0^1 \left(\mu''(t)\right)^2 dt \leq \frac{1}{n}RSS(\tilde{\mu}) + \int_0^1 \left(\tilde{\mu}''(t)\right)^2 dt.$$

Proof. First, for any $\tilde{\mu}(t_i) = Y_i, i = 1, \ldots, n$, we can always find a natural cubic spline μ with knots at $t_1, \ldots, t_n$ that also satisfies $\mu(t_i) = Y_i, i = 1, \ldots, n$. This is by the design of natural splines. Now, let

$$h(x) = \tilde{\mu}(t) - \mu(t)$$

and consider,

$$\begin{aligned}
\int_0^1 \mu''(t)h''(t)dt &= \mu''(t)h'(t)|_0^1 - \int_0^1 \mu'''(t)h'(t)dt \\
&= -\int_{t_1}^{t_n} \mu'''(t)h'(t)dt \\
&= -\sum_{j=1}^{n-1} \mu'''(t)h(t)|_{t_j}^{t_{j+1}} + \int_{t_1}^{t_n} \mu^{(4)}(t)h'(t)dt \\
&= -\sum_{j=1}^{n-1} c_j(h(t_{j+1}) - h(t_j)) \\
&= 0,
\end{aligned}$$

where c_j denotes a constant function $\mu'''(t)$ for $t \in (t_j, t_{j+1})$.

In the first equality, we used integration by parts, in the second equality we used the natural boundary conditions $\mu''(0) = \mu''(1) = 0$ and the fact that $\mu'''(t) = 0$ for interval outside $[t_1, t_n]$. Using integration by parts again, we obtain the third equality. In the fourth equality, we used the fact that $\mu'''(t)$ is a constant and $\mu^{(4)}(t) = 0$. Finally, $h(t_{j+1}) - h(t_j) = \tilde{\mu}(t_{j+1}) - \mu(t_{j+1}) -$

$\tilde{\mu}(t_j) + \mu(t_j) = 0$. Next, we have that

$$\begin{aligned}
\int_0^1 \left(\tilde{\mu}''(t)\right)^2 dt &= \int_0^1 \left(\mu''(t) + h''(t)\right)^2 dt \\
&= \int_0^1 \left(\mu''(t)\right)^2 dt + \int_0^1 \left(h''(t)\right)^2 dt + 2\int_0^1 \mu''(t)h''(t)dt \\
&= \int_0^1 \left(\mu''(t)\right)^2 dt + \int_0^1 \left(h''(t)\right)^2 dt \\
&\geq \int_0^1 \left(\mu''(t)\right)^2 dt.
\end{aligned}$$

Equality holds if and only if $\int_0^1 (h^{''}(t))^2 dt = 0$ or $h'' = 0$. This implies that h must be linear on $[0, 1]$, and $h(t_j) = 0, j = 1, \ldots, n$, which is equivalent to $h = 0$. In other words, the equality holds only when $\tilde{\mu}$ and μ are the same function. Thus, in the cubic spline case, the minimizer to the PLS criterion (1.5) is a unique natural cubic spline with knots at the data inputs. □

This means that, unless μ itself is a natural cubic spline, we can always find a natural cubic spline which attains a smaller value of PLS criterion (1.5) for $m = 2$, and the minimizer must also be a natural cubic spline. This property also holds true in higher order cases, as stated in Theorem 6.6.8 on page 170 of [**6**]. Therefore, if $x_1, \ldots, x_n$ are the basis functions of the space of natural splines of order $2m$ with knots at $t_1, \ldots, t_n$, then the smoothing spline estimator $\hat{\mu}_\lambda$ is a natural spline of order $2m$ with knots at $t_1, \ldots, t_n$, which is obtained by minimizing the following w.r.t $\mu \in W_2^m[0, 1]$:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(t_i))^2 + \lambda \int_0^1 (\mu^{(m)}(t))^2 dt.$$

This optimization problem is equivalent to minimizing the following w.r.t

$b \in \mathbb{R}^n$:

$$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{n} b_j x_j(t_i)\right)^2 + \lambda \int_0^1 \left(\sum_{j=1}^{n} b_j x_j^{(m)}(t)\right)^2 dt. \tag{2.1}$$

Thus, the infinite-dimensional minimization problem is reduced to a finite-dimensional problem of minimization over n dimensional basis of natural splines, and the parameterized smoothing spline estimator can be expressed as follows:

$$\hat{\mu}_\lambda(t) = \sum_{j=1}^{n} \hat{b}_{\lambda j} x_j(t), \tag{2.2}$$

where $\hat{\boldsymbol{b}}_\lambda = (\hat{b}_{\lambda 1}, \ldots, \hat{b}_{\lambda n})^\top$ is an estimator of $\boldsymbol{b} = (b_1, \ldots, b_n)^\top$. Now, define $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$,

$$\boldsymbol{X} = \{x_j(t_i)\}_{i,j=1,\ldots,n} = \begin{bmatrix} x_1(t_1) & \ldots & x_n(t_1) \\ \vdots & \ddots & \vdots \\ x_1(t_n) & \ldots & x_n(t_n) \end{bmatrix},$$

$$\boldsymbol{H} = \left\{x_i^{(m)}(t)x_j^{(m)}(t)dt\right\}_{i,j=1,\ldots,n} = \begin{bmatrix} \left(x_1^{(m)}(t)\right)^2 & \ldots & x_1^{(m)}(t)x_n^{(m)}(t) \\ \vdots & \ddots & \vdots \\ x_n^{(m)}(t)x_1^{(m)}(t) & \ldots & \left(x_n^{(m)}(t)\right)^2 \end{bmatrix},$$

$$\boldsymbol{\Omega} = \left\{\int_0^1 x_i^{(m)}(t)x_j^{(m)}(t)dt\right\}_{i,j=1,\ldots,n} = \int_0^1 \boldsymbol{H}(t)dt. \tag{2.3}$$

Then, we can express (2.1) as

$$\begin{aligned} &\frac{1}{n}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}) + \lambda\int_0^1 \boldsymbol{b}^\top\boldsymbol{H}(t)\boldsymbol{b}\,dt \\ &= \frac{1}{n}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})^\top(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b}) + \lambda\boldsymbol{b}^\top\boldsymbol{\Omega}\boldsymbol{b} \\ &= \frac{1}{n}(\boldsymbol{Y}^\top\boldsymbol{Y} - 2\boldsymbol{b}^\top\boldsymbol{X}^\top\boldsymbol{Y} + \boldsymbol{b}^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{b}) + \lambda\boldsymbol{b}^\top\boldsymbol{\Omega}\boldsymbol{b}. \end{aligned} \tag{2.4}$$

Before we solve the minimization problem (2.4). Recall the following rules of differentiation: Let $\boldsymbol{u} = (u_1, \ldots, u_n)^\top$ and $\boldsymbol{v} = (v_1, \ldots, v_n)^\top$ be two vectors and let $A = (a_{ij})_{n\times n}$ be a symmetric $n \times n$ matrix. Then, using the notation

$$\frac{d}{d\boldsymbol{v}} = \left(\frac{d}{dv_1}, \ldots, \frac{d}{dv_n}\right)^\top,$$

we have

$$\frac{d}{d\boldsymbol{v}}(\boldsymbol{u}^\top \boldsymbol{v}) = \frac{d}{d\boldsymbol{v}}(\boldsymbol{v}^\top \boldsymbol{u}) = \boldsymbol{u}, \tag{2.5}$$

$$\frac{d}{d\boldsymbol{v}}(\boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{v}) = 2\boldsymbol{A}\boldsymbol{v}. \tag{2.6}$$

Now, using rules (2.5) and (2.6) and taking the derivative of (2.4) w.r.t $\boldsymbol{b}$, we obtain

$$\frac{1}{n}(-2\boldsymbol{X}^\top \boldsymbol{Y} + 2\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{b}) + 2\lambda\boldsymbol{\Omega}\boldsymbol{b}. \tag{2.7}$$

Then, setting (2.7) equal to 0, we get the solution in the form

$$\widehat{\boldsymbol{b}}_\lambda = (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top \boldsymbol{Y} \tag{2.8}$$

and the vector of fitted values of $\hat{\mu}_\lambda$ is

$$\widehat{\boldsymbol{\mu}}_\lambda = (\hat{\mu}_\lambda(t_1), \ldots, \hat{\mu}_\lambda(t_n))^\top = \boldsymbol{S}_\lambda \boldsymbol{Y}, \tag{2.9}$$

where $\boldsymbol{S}_\lambda$ is the smoothing matrix given by

$$\boldsymbol{S}_\lambda = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top. \tag{2.10}$$

Note: From Lemma 5.2 of [**2**], the matrix $\boldsymbol{X}$ has full rank n and thus is invertable.

2.2 The Demmler-Reinsch basis

To gain insights to how smoothing spline works, we will consider a special basis which will result in $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{\Omega}$ to be diagonalized. Consider the case where we estimate μ in model (1.1) by a natural spline of order 2 with uniform designs knots,

$$t_i = \frac{2i-1}{2n}, \quad i = 1, \ldots, n.$$

The estimator $\hat{\mu}_\lambda$ is a minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mu(t_i))^2 + \lambda \int_0^1 (\mu'(t))^2 dt,$$

which is a linear smoothing spline that has the parametric form (2.2).

Consider the basis $\{x_j, j = 1, \ldots, n\}$ of the space of natural splines of order 2 with knots at $t_1, \ldots, t_n$, which is defined as follows: $x_1(t) = 1$ and

$$x_{j+1}(t) = \begin{cases} \sqrt{2}\cos(j\pi t_1), & 0 \le t < t_1, \\ \sqrt{2}\cos(j\pi t_i) & \\ +\sqrt{2}\frac{t-t_i}{t_{i+1}-t_i}\left[\cos(j\pi t_{i+1}) - \cos(j\pi t_i)\right], & t_i \le t < t_{i+1}, \\ & i = 1, \ldots, n-1, \\ \sqrt{2}\cos(j\pi t_n), & t_n \le t \le 1, \end{cases}$$

for $j = 1, \ldots, n-1$. Each basis function is a natural linear spline that is constant outside of $[t_1, t_n]$ and linear over each subinterval $[t_i, t_{i+1}], i = 1, \ldots, n-1$. Notice that $x_j(t_i) = \sqrt{2}\cos((j-1)\pi t_i)$ and let the matrix $\boldsymbol{X}$ be as before:

$$\boldsymbol{X} = \{x_j(t_i)\}_{i,j=1,\ldots,n} = \begin{bmatrix} 1 & \sqrt{2}\cos(\pi t_1) & \ldots & \sqrt{2}\cos((n-1)\pi t_1) \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \sqrt{2}\cos(\pi t_n) & \ldots & \sqrt{2}\cos((n-1)\pi t_n) \end{bmatrix}.$$

Through lengthy calculations, we obtain that

$$\boldsymbol{X}^\top \boldsymbol{X} = \left\{ \sum_{k=1}^{n} x_i(t_k) x_j(t_k) \right\}_{i,j=1,\ldots,n}$$
$$= \begin{bmatrix} n & \sum_{k=1}^n \sqrt{2}\cos(\pi t_i) & \ldots & \sum_{k=1}^n \sqrt{2}\cos((n-1)\pi t_i) \\ \sum_{k=1}^n \sqrt{2}\cos(\pi t_i) & \sum_{k=1}^n 2\cos^2(\pi t_i) & \ldots & \sum_{k=1}^n 2\cos(\pi t_i)\cos((n-1)\pi t_i) \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{k=1}^n \sqrt{2}\cos((n-1)\pi t_i) & \sum_{k=1}^n 2\cos((n-1)\pi t_i)\cos(\pi t_i) & \ldots & \sum_{k=1}^n 2\cos^2((n-1)\pi t_i) \end{bmatrix}.$$

Lemma 1 and Lemma 2 below are important for establishing good properties of smoothing splines.

Lemma 1 (Lemma 3.4 on page 144 of [**2**]). *Let* $t_i = \frac{2i-1}{2n}, i = 1, \ldots, n$, *and define*

$$d_{jr} = \frac{2}{n} \sum_{i=1}^{n} \cos(j\pi t_i)\cos(r\pi t_i).$$

For $j = 1, \ldots, n-1$,

$$d_{jr} = \begin{cases} 1, & \text{if } r - j = 2kn \text{ or } r + j = 2kn \text{ for } k \text{ even}, \\ -1, & \text{if } r - j = 2kn \text{ or } r + j = 2kn \text{ for } k \text{ odd}, \\ 0, & \text{otherwise.} \end{cases}$$

Applying Lemma 1, whose proof is given in the Appendix, we can obtain

that $\boldsymbol{X}^\top \boldsymbol{X} = n\boldsymbol{I}$. Indeed, we have

$$\begin{aligned}
\boldsymbol{X}\boldsymbol{X}^\top &= \left\{ \sum_{k=1}^{n} x_k(t_i) x_k(t_j) \right\}_{i,j=1,\ldots,n} \\
&= \left\{ 1 + \sum_{k=1}^{n-1} 2\cos(k\pi t_i)\cos(k\pi t_j)) \right\}_{i,j=1,\ldots,n} \\
&= \left\{ 1 + \sum_{k=1}^{n-1} \cos(k\pi(t_i - t_j)) + \sum_{k=1}^{n-1} \cos(k\pi(t_i + t_j)) \right\}_{i,j=1,\ldots,n} \\
&= \left\{ 1 + \sum_{k=1}^{n-1} \cos\left(\frac{(i-j)\pi k}{n}\right) + \sum_{k=1}^{n-1} \cos\left(\frac{(i+j-1)\pi k}{n}\right) \right\}_{i,j=1,\ldots,n} \\
&= \mathrm{Diag}(n, \ldots, n).
\end{aligned}$$

The second equality in the above display is obtained by separating the first basis function $x_1(t) = 1$ from the rest. The third equality is obtained by using the cosine product identity: $\cos u \cos v = \frac{1}{2}[\cos(u-v) + \cos(u+v)]$. The last equality is obtained as follows: we use the identity

$$\sum_{k=1}^{n} \cos(kx) = \frac{\sin\left(\frac{nx}{2}\right)}{\sin\left(\frac{x}{2}\right)} \cos\left(\frac{(n+1)x}{2}\right), \quad \text{if} \quad \sin\left(\frac{x}{2}\right) \neq 0,$$

which gives

$$\sum_{k=1}^{n-1} \cos(kx) = \frac{\sin\left(\frac{nx}{2}\right)}{\sin\left(\frac{x}{2}\right)} \cos\left(\frac{(n+1)x}{2}\right) - \cos(nx).$$

Then, setting $x = \frac{(i-j)\pi}{n}$ and using the identity $\cos(u+v) = \cos u \cos v - \sin u \sin v$ results in the following:

$$\sum_{k=1}^{n-1}\cos\left(\frac{(i-j)\pi k}{n}\right) = \frac{\sin\left(\frac{i-j}{2}\pi\right)}{\sin\left(\frac{i-j}{2n}\pi\right)}\cos\left(\left(1+\frac{1}{n}\right)\frac{i-j}{2}\pi\right) - \cos((i-j)\pi)$$
$$= \frac{\sin\left(\frac{i-j}{2}\pi\right)}{\sin\left(\frac{i-j}{2n}\pi\right)}\cos\left(\frac{i-j}{2}\pi\right)\cos\left(\frac{i-j}{2n}\pi\right)$$
$$- \sin^2\left(\frac{i-j}{2}\pi\right) - \cos((i-j)\pi),$$

from which, recalling the identity $\sin^2 x = \frac{1}{2}(1-\cos(2x))$, we obtain

$$\sum_{k=1}^{n-1}\cos\left(\frac{(i-j)\pi k}{n}\right) = -\frac{1}{2}(1-\cos((i-j)\pi)) - \cos((i-j)\pi)$$
$$= -\frac{1}{2} - \frac{1}{2}\cos((i-j)\pi). \tag{2.11}$$

The second term can be handled similarly, that is,

$$\sum_{k=1}^{n-1}\cos\left(\frac{(i+j-1)\pi k}{n}\right) = -\frac{1}{2} - \frac{1}{2}\cos((i+j-1)\pi). \tag{2.12}$$

Then, summing (2.11) and (2.12) gives

$$\sum_{k=1}^{n-1}\cos\left(\frac{(i-j)\pi k}{n}\right) + \sum_{k=1}^{n-1}\cos\left(\frac{(i+j-1)\pi k}{n}\right)$$
$$= -1 - \frac{1}{2}[\cos((i-j)\pi) + \cos((i+j-1)\pi)], \tag{2.13}$$

and we can see that if $i-j$ is even then $i+j-1$ must be odd, and vice versa. Hence, the expression in (2.13) equals 0, for $i \neq j$. For the case of $i = j$, we have that

$$\sum_{k=1}^{n-1}\cos\left(\frac{(i-j)\pi k}{n}\right) + \sum_{k=1}^{n-1}\cos\left(\frac{(i+j-1)\pi k}{n}\right)$$
$$= n - 1 - \frac{1}{2} - \frac{1}{2}\cos((2i-1)\pi) = n-1,$$

for $i = 1, \ldots, n$. Therefore, the diagonal elements of $\boldsymbol{X}\boldsymbol{X}^\top$ are equal to n and the off-diagonal elements are all zeros.

The fact that $\boldsymbol{X}^\top\boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^\top = n\boldsymbol{I}$ implies that the columns of $\boldsymbol{X}$ form an orthogonal basis in $\mathbb{R}^n$ and hence the collection of functions $\{x_j, j = 1, \ldots, n\}$ forms a basis in the space of natural splines of order 2 with knots at $t_1, \ldots, t_n$. Now, let us look at the $\boldsymbol{\Omega}$ matrix defined by (2.3), which also turns out to be a diagonal matrix. To see this, we note that the first row and the first column are

$$\int_0^1 x_i'(t)x_1'(t)dt = \int_0^1 x_1'(t)x_j'(t)dt = 0, \quad i, j = 1, \ldots, n,$$

and the other elements of $\boldsymbol{\Omega}$ are

$$\begin{aligned}
&\int_0^1 x_{i+1}'(t)x_{j+1}'(t)dt \\
&= 2n^2 \int_0^1 \left[\cos\left(i\pi t_{i+1}\right) - \cos\left(i\pi t_i\right)\right]\left[\cos\left(j\pi t_{i+1}\right) - \cos\left(j\pi t_i\right)\right] dt \\
&= 2n^2 \sum_{k=1}^{n-1} \int_{t_k}^{t_{k+1}} \left[\cos\left(i\pi t_{k+1}\right) - \cos\left(i\pi t_k\right)\right]\left[\cos\left(j\pi t_{k+1}\right) - \cos\left(j\pi t_k\right)\right] dt \\
&= 2n \sum_{k=1}^{n-1} \left[\cos\left(i\pi t_{k+1}\right) - \cos\left(i\pi t_k\right)\right]\left[\cos\left(j\pi t_{k+1}\right) - \cos\left(j\pi t_k\right)\right] \\
&= 2n \sum_{k=1}^{n-1} \left[-2\sin\left(\frac{i\pi k}{n}\right)\sin\left(\frac{i\pi}{2n}\right)\right]\left[-2\sin\left(\frac{j\pi k}{n}\right)\sin\left(\frac{j\pi}{2n}\right)\right] \\
&= 8n \sin\left(\frac{i\pi}{2n}\right)\sin\left(\frac{j\pi}{2n}\right)\sum_{k=1}^{n} \sin\left(\frac{i\pi k}{n}\right)\sin\left(\frac{j\pi k}{n}\right). \qquad (2.14)
\end{aligned}$$

In the second equality of (2.14), we use the fact that $\frac{1}{t_{i+1}-t_i} = n$. Then, in the fourth equality, we use the fact $\int_{t_k}^{t_{k+1}} dt = \frac{1}{n}$. We next apply the cosine sum identity: $\cos u - \cos v = -2\sin(\frac{u+v}{2})\sin(\frac{u-v}{2})$, to obtain equality five.

To get the final equality in (2.14), we rearrange the summation, then use the fact $\sin(n\pi) = 0, \text{for } n \in \mathbb{Z}$, to extend the order of summation to apply further the following Lemma.

Lemma 2 (Lemma 3.5 on page 145 of [**2**]). *Let* $t_i = \frac{i}{n}, i = 1, \ldots, n$, *and define*

$$d_{jr} = \frac{2}{n} \sum_{i=1}^{n} \sin(j\pi t_i) \sin(r\pi t_i).$$

For $j = 1, \ldots, n$,

$$d_{jr} = \begin{cases} 1, & \text{if } r - j = 2kn, \\ -1, & \text{if } r - j = 2kn, \\ 0, & \text{otherwise.} \end{cases}$$

Applying Lemma 2, whose proof is given in the Appendix, to the right side of (2.14), we obtain the elements of $\mathbf{\Omega}$ in the form

$$\int_0^1 x'_{i+1}(t) x'_{j+1}(t) dt = \begin{cases} \left(2n \sin\left(\frac{j\pi}{2n}\right)\right)^2, & i = j, \\ 0, & i \neq j, \end{cases}$$

for $i, j = 1, \ldots, n-1$. Now, let

$$\gamma_j = \gamma_{j,n} = \left(2n \sin\left(\frac{j\pi}{2n}\right)\right)^2, \quad j = 1, \ldots, n-1.$$

The γ_j are the called the Demmler-Reinsch eigenvalues, and $x_j, j = 1, \ldots, n$, are called the Demmler-Reinsch basis functions.

If we set $\gamma_0 = 0$, we obtain the smoothing matrix in (2.10) of the form

$$\begin{aligned}\boldsymbol{S}_\lambda &= \boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top \\ &= \boldsymbol{X}(n\boldsymbol{I} + n\lambda\text{Diag}(\gamma_0, \gamma_1, \ldots, \gamma_{n-1}))^{-1}\boldsymbol{X}^\top \\ &= \frac{1}{n}\sum_{j=1}^{n}\frac{1}{1+\lambda\gamma_{j-1}}\boldsymbol{x}_j\boldsymbol{x}_j^\top,\end{aligned}$$

where $\boldsymbol{x}_j$ is the jth column of matrix $\boldsymbol{X}$, and the vector of the fitted values using smoothing spline is

$$\widehat{\boldsymbol{\mu}}_\lambda = (\hat{\mu}_\lambda(t_1), \ldots, \hat{\mu}_\lambda(t_n))^\top = \frac{1}{n}\sum_{j=1}^{n}\frac{\boldsymbol{x}_j^\top\boldsymbol{Y}}{1+\lambda\gamma_{j-1}}\boldsymbol{x}_j. \tag{2.15}$$

This expression says that the smoothing spline performs linear transformation on the response $\boldsymbol{Y}$ with respect to $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, and then shrinks each component by a factor $\frac{1}{1+\lambda\gamma_{j-1}}$. For a large value of λ, the shrink effect is more sever and the smoothing spline fit is closer to a linear fit, and for a smaller value of λ the fit tracks the data more closely. Figure 4(a) shows this effect.

The smoothing spline coefficients as defined in (2.8) are equal to

$$\begin{aligned}\widehat{\boldsymbol{b}}_\lambda &= (n\boldsymbol{I} + n\lambda\text{Diag}(\gamma_0, \gamma_1, \ldots, \gamma_{n-1}))^{-1}\boldsymbol{X}^\top\boldsymbol{Y} \\ &= \text{Diag}\left(\frac{1}{n}, \frac{1}{n+n\lambda\gamma_1}, \ldots, \frac{1}{n+n\lambda\gamma_{n-1}}\right)\boldsymbol{X}^\top\boldsymbol{Y} \\ &= \begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{\sqrt{2}\cos(\pi t_1)}{n+n\lambda\gamma_1} & \cdots & \frac{\sqrt{2}\cos(\pi t_n)}{n+n\lambda\gamma_1} \\ \vdots & \ddots & \vdots \\ \frac{\sqrt{2}\cos((n-1)\pi t_1)}{n+n\lambda\gamma_{n-1}} & \cdots & \frac{\sqrt{2}\cos((n-1)\pi t_n)}{n+n\lambda\gamma_{n-1}} \end{bmatrix}\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}.\end{aligned}$$

In particular, we obtain $\hat{b}_{\lambda 1} = \bar{\boldsymbol{Y}}$, and

$$\hat{b}_{\lambda j} = \frac{\sqrt{2}}{n} \sum_{i=1}^{n} Y_i \frac{1}{1+\lambda\gamma_{j-1}} \cos\left((j-1)\pi t_i\right) \quad \text{for} \quad j = 2, \ldots, n.$$

Thus, the smoothing spline for a given value of λ is given by

$$\begin{aligned}
\hat{\mu}_\lambda(t) &= \frac{1}{n}\sum_{i=1}^{n} Y_i + \sum_{j=2}^{n} \left(\frac{\sqrt{2}}{n} \sum_{i=1}^{n} Y_i \frac{1}{1+\lambda\gamma_{j-1}} \cos\left((j-1)\pi t_i\right) \right) x_j(t) \\
&= \frac{1}{n}\sum_{i=1}^{n} Y_i \left(1 + \sqrt{2} \sum_{j=1}^{n-1} \frac{\cos\left(j\pi t_i\right) x_{j+1}(t))}{1+\lambda\gamma_j} \right), \quad t \in [0,1].
\end{aligned}$$

This expression is similar to the one for $\widehat{\boldsymbol{\mu}}_\lambda$ in (2.15) . We can also see that this linear smoothing spline is similar to a kernel estimator

$$\hat{\mu}_\lambda(t) = \frac{1}{n}\sum_{i=1}^{n} K_n(t, t_i; \lambda) Y_i,$$

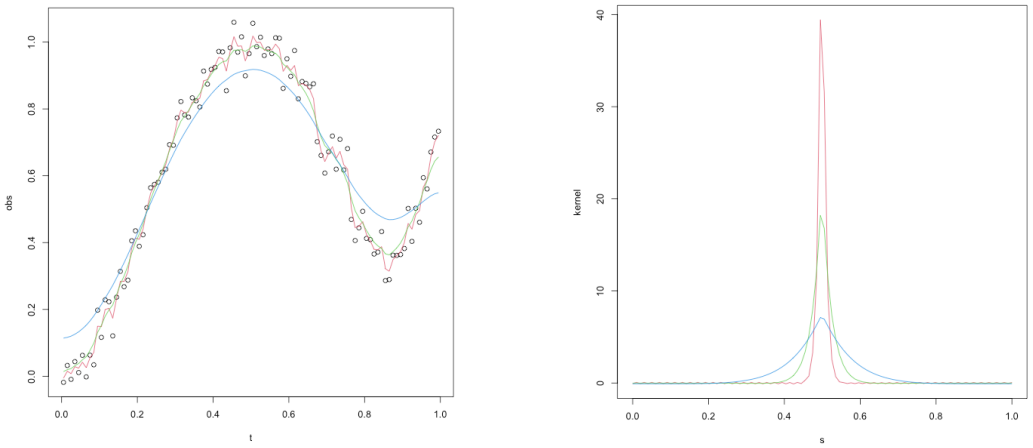
where

$$K_n(t, s; \lambda) = 1 + \sqrt{2} \sum_{j=1}^{n-1} \frac{\cos\left(j\pi s\right) x_{j+1}(t)}{1+\lambda\gamma_j}.$$

In Figure 4(b), we show the kernel weight function $K_n(t, t_i; \lambda)$ for different values of λ.

From Figure 4, the smaller value of $\lambda = 0.00006$ produces a slavish estimator, and the lager value of $\lambda = 0.006$ ignores many of the data features. The middle choice of $\lambda = 0.0006$ is visually more appropriate, it balances the goodness of fit and smoothness penalty.

Example 2 (Exercise 5.9.1 from [**2**])



(a) Smoothing spline fit

(b) Spline weight functions

Figure 4: (a) Data generated from model (1.1) using normal random errors with $\sigma = 0.05$ and regression function $\mu(t) = 16t^2(1-t)^2 + 32(t - 0.7)^3\mathbb{I}_{(t\geq 0.7)}(t)$, with $\mathbb{I}_A$ being the indicator function. The fits corresponding to different level of smoothing parameter (red: $\lambda = 0.00006$, green: $\lambda = 0.0006$ and blue: $\lambda = 0.006$). (b) Kernel weight function $K_n(0.5, s; \lambda)$, we can see that a fit that closely tracks the data points have a narrower kernel weight function.

There is no simple form of the Demmler-Reinsch basis functions for general m or nonuniform designs, but their property can be drawn from the $m = 1$ case. From [**1**], the basis $\{x_1, \ldots, x_n\}$ of the space natural spline of order $2m$ with knots at $(t_1, \ldots, t_n)$ may be chosen as follows:

1) $x_1, \ldots, x_m$ span the polynomial of order m for $m \leq n$,

2) $x_j, j = 1, \ldots, n$, has at least $j - 1$ sign changes in $(0, 1)$,

3) $\boldsymbol{X}^\top \boldsymbol{X} = \boldsymbol{X}\boldsymbol{X}^\top = n\boldsymbol{I}$,

4) $\boldsymbol{\Omega} = \text{Diag}(\underbrace{0, \ldots, 0}_{m}, \gamma_1, \ldots, \gamma_{n-m})$ for eigenvalues $0 < \gamma_1 \leq \ldots \leq \gamma_{n-m}$.

Using properties 1) to 4), we can modify the smoothing spline estimator in (2.2) as follows:

$$\hat{\mu}_\lambda(t) = \sum_{j=1}^{m} \hat{b}_{\lambda j} x_j(t) + \sum_{j=m+1}^{n} \frac{\hat{b}_{\lambda j}}{1 + \lambda_{j-m}} x_j(t) \tag{2.16}$$

where $\hat{b}_{\lambda j} = \frac{1}{n}\sum_{i=1}^{n} Y_i x_j(t_i), j = 1, \ldots, n$. Formula (2.16) reveals the two parts of smoothing splines: a projection part $\sum_{j=1}^{m} \hat{b}_{\lambda j} x_j$, which is a polynomial of order m fit to the data, and a departure term $\sum_{j=m+1}^{n} \frac{\hat{b}_{\lambda j}}{1+\lambda_{j-m}} x_j(t)$ that estimates the distance from a polynomial model. Hence, the choice of $\lambda = \infty$ will result in the smoothing spline estimator having only the projection part, and thus, gives the smoothest possible fit.

It is important to note that the smoothing spline estimator developed from criterion (1.5) is appropriate if each observation has the same variance. If heteroscedasticity exists, the following smoothing criterion should be considered:

$$\frac{1}{n}\sum_{i=1}^{n} \omega_i \left(Y_i - \mu(t_i)\right)^2 + \lambda \int_0^1 \left(\mu^{(m)}(t)\right)^2 dt \tag{2.17}$$

where

$$\omega_i = [\text{Var}(Y_i)]^{-1}, \quad \text{for} \quad i = 1, \ldots, n.$$

Let $\{x_1, \ldots, x_n\}$ be a basis of the space of natural spline of order $2m$ with knots at $t_1, \ldots, t_n$. Under the same assumptions on $\boldsymbol{X}$ and $\boldsymbol{\Omega}$ matrices as in Section 2.1, with $\boldsymbol{W} = \text{Diag}(\omega_1, \ldots, \omega_n)$. We can express (2.17) as

$$\frac{1}{n}(\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}))^\top (\boldsymbol{W}^{\frac{1}{2}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})) + \lambda \boldsymbol{b}^\top \boldsymbol{\Omega} \boldsymbol{b}.$$

Then if we take the derivative of this function w.r.t to $\boldsymbol{b}$, set it equal to zero and solve the equation, we obtain the minimizer of (2.17) as $\hat{\mu}_\lambda = \sum_{j=1}^{n} \hat{b}_{\lambda j} x_j$, where

$$\widehat{\boldsymbol{b}}_\lambda = (\hat{b}_{\lambda 1}, \ldots, \hat{b}_{\lambda n})^\top = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + n\lambda \boldsymbol{\Omega})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{Y}.$$

Example 3 (Exercise 5.9.4 from [**2**])

To illustrate the use of (2.17), let us look at the example of repeated observations at design points with constant variance. Suppose that our data set follows the model

$$Y_{ij} = \mu(t_i) + \epsilon_{ij}, \quad j = 1, \ldots, n_i, i = 1, \ldots, r.$$

Criterion (1.5) under such model can be expressed as

$$\frac{1}{n} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \mu(t_i))^2 + \lambda \int_0^1 \left(\mu^{(m)}(t)\right)^2 dt, \tag{2.18}$$

where $n = \sum_{i=1}^{r} n_i$.

Let $\bar{Y}_i = \sum_{j=1}^{n_i} Y_{ij}$. Using the identity $(Y_{ij} - \mu(t_i)) = (Y_{ij} - \bar{Y}_i) + (\bar{Y}_i - \mu(t_i))$, we obtain

$$\begin{aligned} \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \mu(t_i))^2 &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{r} \sum_{j=1}^{n_i} (\bar{Y}_i - \mu(t_i))^2 \\ &= \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{r} n_i (\bar{Y}_i - \mu(t_i))^2 \end{aligned}$$

with the cross product equal to 0. Therefore, relation (2.18) can be rewritten as

$$\frac{1}{n} \left\{ \sum_{i=1}^{r} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 + \sum_{i=1}^{r} n_i (\bar{Y}_i - \mu(t_i))^2 \right\} + \lambda \int_0^1 \left(\mu^{(m)}(t)\right)^2 dt \tag{2.19}$$

and the minimization of (2.19) w.r.t $\mu \in W_2^m[0,1]$ is equivalent to the minimization of

$$\frac{1}{n}\sum_{i=1}^{r} n_i(\bar{Y}_i - \mu(t_i))^2 + \lambda \int_0^1 \left(\mu^{(m)}(t)\right)^2 dt \tag{2.20}$$

w.r.t $\mu \in W_2^m[0,1]$. This is true because we are looking for μ that minimizes (2.19), and the term $\sum_{i=1}^{r}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2$ is free of μ, thus we can ignore it in the minimization problem. In fact, under the same assumption on $\boldsymbol{X}$ and $\boldsymbol{\Omega}$ matrices as in Section 2.1 with $\boldsymbol{W} = \mathrm{Diag}(n_1, \ldots, n_n)^\top$ and $\bar{\boldsymbol{Y}} = (\bar{Y}_1, \ldots, \bar{Y}_n)^\top$, we can rewrite (2.20) as

$$\frac{1}{n}(\boldsymbol{W}^{\frac{1}{2}}(\bar{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b}))^\top(\boldsymbol{W}^{\frac{1}{2}}(\bar{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{b})) + \lambda \boldsymbol{b}^\top \boldsymbol{\Omega} \boldsymbol{b},$$

then take the derivative w.r.t $\boldsymbol{b}$, set it equal to 0 and solve the observed equation. This yields the solution in the form

$$\widehat{\boldsymbol{b}}_\lambda = (\hat{b}_{\lambda 1}, \ldots, \hat{b}_{\lambda n})^\top = (\boldsymbol{X}^\top \boldsymbol{W} \boldsymbol{X} + n\lambda \boldsymbol{\Omega})^{-1} \boldsymbol{X}^\top \boldsymbol{W} \bar{\boldsymbol{Y}}.$$

3 Selection of smoothing parameter

3.1 Prediction Risk

In this section, we will introduce some criteria for selection of the smoothing parameter λ in (1.5). First, let us define the *loss function* as

$$L(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(\mu(t_i) - \hat{\mu}_\lambda(t_i))^2, \tag{3.1}$$

which is just a measure of distance between μ and $\hat{\mu}_\lambda$. The associated *risk function* is

$$R(\lambda) = \boldsymbol{E}(L(\lambda)) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}(\mu(t_i) - \hat{\mu}_\lambda(t_i))^2. \tag{3.2}$$

A value of λ that minimizes (3.1) will provide the best estimator of μ for the particular data set, and a value of λ that minimizes (3.2) can be viewed as a best estimator of μ in repeated sampling or best for prediction since only the expectation of the loss function is measured.

Another criterion for selecting λ is the *prediction risk*. Recall model (1.1) and suppose we have acquired n new observations: $\boldsymbol{Y}^* = (Y_1^*, \ldots, Y_n^*)^\top$ at the same t_is, and the model then is

$$Y_i^* = \mu(t_i) + \epsilon_i^*, \quad \text{for} \quad i = 1, \ldots, n,$$

where μ is the same as in model (1.1) with ϵ_i^*s are zero mean random variables that share a common variance σ^2, which are uncorrelated with each other and with all ϵ_i in model (1.1). From these additional observations, a natural measure of performance of an estimator $\hat{\mu}_\lambda$ is the *prediction risk* defined as

$$P(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{E}(Y_i^* - \hat{\mu}_\lambda(t_i))^2. \tag{3.3}$$

Using $(Y_i^* - \hat{\mu}_\lambda(t_i)) = (Y_i^* - \boldsymbol{E}(Y_i^*)) + (\boldsymbol{E}(Y_i^*) - \hat{\mu}_\lambda(t_i))$ and $\boldsymbol{E}(Y_i^*) = \mu(t_i)$, we obtain

$$P(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{E}(Y_i^* - \boldsymbol{E}(Y_i^*))^2 + \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{E}(\mu(t_i) - \hat{\mu}_\lambda(t_i))^2 = \sigma^2 + R(\lambda). \tag{3.4}$$

Hence, the value of λ that minimizes (3.2) will also minimize (3.3) and vice versa.

Unfortunately, criteria (3.1), (3.2), and (3.3) all require the knowledge of μ. Thus, in order to implement them in applications, an estimator of these criteria is needed for λ selection. We will focus on the estimation of the $P(\lambda)$ criterion.

One naive estimator of $\boldsymbol{E}(Y_i^* - \hat{\mu}_\lambda(t_i))^2$ is $\frac{1}{n}\sum_{i=1}^n (Y_i - \hat{\mu}_\lambda(t_i))^2$ which looks similar to

$$RSS(\hat{\mu}_\lambda) = \sum_{i=1}^{n} (Y_i - \hat{\mu}_\lambda(t_i))^2. \tag{3.5}$$

By (2.9), we can express (3.5) in the following form:

$$\begin{aligned} RSS(\hat{\mu}_\lambda) &= (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_\lambda)^\top (\boldsymbol{Y} - \widehat{\boldsymbol{\mu}}_\lambda) \\ &= (\boldsymbol{Y} - \boldsymbol{S}_\lambda \boldsymbol{Y})^\top (\boldsymbol{Y} - \boldsymbol{S}_\lambda \boldsymbol{Y}) \\ &= ((\boldsymbol{I} - \boldsymbol{S}_\lambda)\boldsymbol{Y})^\top ((\boldsymbol{I} - \boldsymbol{S}_\lambda)\boldsymbol{Y}) \\ &= \boldsymbol{Y}^\top (\boldsymbol{I} - \boldsymbol{S}_\lambda)^2 \boldsymbol{Y}. \end{aligned} \tag{3.6}$$

Next, we shall need the following result.

Lemma 3 (Lemma (2.2) form [**2**])**.** *Define the random variable*

$$Q = \boldsymbol{\epsilon}^\top \boldsymbol{B} \boldsymbol{\epsilon}$$

where $\boldsymbol{B}$ *is a symmetric* $n \times n$ *matrix, and* $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ *is a vector of random variables. If* $\boldsymbol{E}(\boldsymbol{\epsilon}) = 0$ *and* $\boldsymbol{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top) = \boldsymbol{\Sigma}$*, then*

$$\boldsymbol{E}(Q) = tr(\boldsymbol{\Sigma}\boldsymbol{B}),$$

where $tr(\boldsymbol{A})$ *denotes the trace of a square matrix* $\boldsymbol{A}$*.*

We can see that $\boldsymbol{S}_\lambda$ as specified by (2.10) is symmetric. Indeed,

$$\boldsymbol{S}_\lambda^\top = (\boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top)^\top = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top = \boldsymbol{S}_\lambda,$$

implying that $(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2$ is also symmetric. Thus, from Lemma 3,

$$\begin{aligned}
\boldsymbol{E}(RSS(\hat{\mu}_\lambda)) &= \boldsymbol{E}(\boldsymbol{Y}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{Y}) \\
&= \boldsymbol{E}((\boldsymbol{\mu}+\boldsymbol{\epsilon})^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2(\boldsymbol{\mu}+\boldsymbol{\epsilon})) \\
&= \boldsymbol{E}(\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\mu}+\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\epsilon} \\
&\quad +\boldsymbol{\epsilon}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\mu}+\boldsymbol{\epsilon}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\epsilon}) \\
&= \boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\mu}+\sigma^2 tr[(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2] \\
&= \boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\mu}+\sigma^2(n-2tr(\boldsymbol{S}_\lambda)+tr(\boldsymbol{S}_\lambda^2)).
\end{aligned} \tag{3.7}$$

Next, from (3.4), we obtain

$$\begin{aligned}
P(\lambda) = \sigma^2 + R(\lambda) &= \sigma^2 + \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{E}(\mu(t_i)-\hat{\mu}_\lambda(t_i))^2 \\
&= \sigma^2 + \frac{1}{n}\boldsymbol{E}\left[(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_\lambda)^\top(\boldsymbol{\mu}-\widehat{\boldsymbol{\mu}}_\lambda)\right] \\
&= \sigma^2 + \frac{1}{n}\boldsymbol{E}\left[(\boldsymbol{\mu}-\boldsymbol{S}_\lambda(\boldsymbol{\mu}+\boldsymbol{\epsilon}))^\top(\boldsymbol{\mu}-\boldsymbol{S}_\lambda(\boldsymbol{\mu}+\boldsymbol{\epsilon}))\right] \\
&= \sigma^2 + \frac{1}{n}\boldsymbol{E}\left[((\boldsymbol{I}-\boldsymbol{S}_\lambda)\boldsymbol{\mu}-\boldsymbol{S}_\lambda\boldsymbol{\epsilon})^\top((\boldsymbol{I}-\boldsymbol{S}_\lambda)\boldsymbol{\mu}-\boldsymbol{S}_\lambda\boldsymbol{\epsilon})\right] \\
&= \sigma^2 + \frac{1}{n}\boldsymbol{E}\left[\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\mu}+\boldsymbol{\epsilon}^\top\boldsymbol{S}_\lambda^2\boldsymbol{\epsilon}\right] \\
&= \sigma^2 + \frac{1}{n}(\boldsymbol{\mu}^\top(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2\boldsymbol{\mu}+\sigma^2 tr(\boldsymbol{S}_\lambda^2)).
\end{aligned} \tag{3.8}$$

Compare (3.7) and (3.8), we can see that

$$\boldsymbol{E}\left(\frac{1}{n}RSS(\hat{\mu}_\lambda)-\frac{2\sigma^2 tr(\boldsymbol{S}_\lambda)}{n}\right)=P(\lambda).$$

Therefore, assuming that σ^2 is known, the unbiased estimator for $P(\lambda)$ is

$$\hat{P}(\lambda)=\frac{1}{n}RSS(\hat{\mu}_\lambda)-\frac{2\sigma^2 tr(\boldsymbol{S}_\lambda)}{n}. \tag{3.9}$$

Intuitively, since $\hat{\mu}_\lambda$ was constructed from the data $Y_i, i=1,\ldots,n$, we would expect that $\hat{\mu}_\lambda$ would perform better in predicting Y_i than some future

value Y_i^*. Therefore, on average, $\frac{1}{n}RSS(\hat{\mu}_\lambda)$ is smaller than $P(\lambda)$: indeed, $\frac{1}{n}RSS(\hat{\mu}_\lambda)$ underestimate $P(\lambda)$ by a factor $\frac{2\sigma^2 tr(\boldsymbol{S}_\lambda)}{n}$.

Typically, σ^2 is unknown and one simple estimator for σ^2 is just $\hat{\sigma}^2 = \frac{1}{n-1}RSS(\hat{\mu}_\lambda)$. However, such estimator will be different for each model, namely, for different λ we choose, while the variance is assumed to remain the same. Therefore, an estimator of σ^2 that is independent from the model is desired. One strongly consistent estimator of σ^2 that is independent of model is proposed by Gasser, Sroka, and Jenner-Steinmetz (GSJS) in [**3**]:

$$\hat{\sigma}^2_{GSJS} = \frac{2}{3(n-2)}\sum_{i=2}^{n-1}\tilde{\epsilon_i}^2 \tag{3.10}$$

where

$$\tilde{\epsilon_i} = Y_i - \frac{Y_{i-1} + Y_{i+1}}{2}.$$

Hence, in practice, one way for smoothing parameter selection is by minimizing the following w.r.t λ:

$$\frac{1}{n}RSS(\hat{\mu}_\lambda) - \frac{2\hat{\sigma}^2_{GSJS}tr(\boldsymbol{S}_\lambda)}{n}.$$

3.2 Cross-validation

Cross-validation method provides another way to perform error estimation and smoothing parameter selection. The cross-validation criterion is given by

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{\mu}_{\lambda(i)}(t_i))^2$$

where $\hat{\mu}_{\lambda(i)}$ denotes the estimator based on all data except for the ith pair (t_i, Y_i). The following reduction is often used for efficient computation of

$CV(\lambda)$:

$$\hat{\mu}_{\lambda(i)}(t_i) = \frac{1}{1 - \boldsymbol{S}_{\lambda ii}}(\hat{\mu}_\lambda(t_i) - \boldsymbol{S}_{\lambda ii} Y_i) \tag{3.11}$$

where $\boldsymbol{S}_{\lambda ii}, i = 1, \ldots, n$, are the diagonal elements of (2.10). Result (3.11) is proved below.

Proof. Let $\boldsymbol{X}_{(i)}$ be the matrix obtained from removing the ith row $\mathbf{x}_i$ of $\boldsymbol{X}$ from (2.3) and $\boldsymbol{Y}_{(i)}$ be the vector with ith element Y_i removed from $\boldsymbol{Y}$. Then

$$\hat{\boldsymbol{b}}_{\lambda(i)} = (\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)} + n\lambda\boldsymbol{\Omega})^{-1} \boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)}.$$

Using $\sum_{i=1}^n \mathbf{x}_i^\top \mathbf{x}_i = \boldsymbol{X}^\top \boldsymbol{X}$, we obtain

$$\begin{aligned}
(\boldsymbol{X}_{(i)}^\top \boldsymbol{X}_{(i)} + n\lambda\boldsymbol{\Omega})^{-1} &= (\boldsymbol{X}^\top \boldsymbol{X} - \mathbf{x}_i^\top \mathbf{x}_i + n\lambda\boldsymbol{\Omega})^{-1} \\
&= (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \\
&\quad + \frac{(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top \mathbf{x}_i (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}}{1 - \mathbf{x}_i (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top}.
\end{aligned}$$

The second equality above is obtained by the Sherman–Morrison formula (see the Appendix for detail). Note that $\mathbf{x}_i (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top = \boldsymbol{S}_{\lambda ii}$. It is also true that $\boldsymbol{X}_{(i)}^\top \boldsymbol{Y}_{(i)} = \boldsymbol{X}^\top \boldsymbol{Y} - \mathbf{x}_i^\top Y_i$, and therefore

$$\begin{aligned}
\hat{\boldsymbol{b}}_{\lambda(i)} &= \left[(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} + \frac{(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top \mathbf{x}_i (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1}}{1 - \boldsymbol{S}_{\lambda ii}} \right] \\
&\quad \times (\boldsymbol{X}^\top \boldsymbol{Y} - \mathbf{x}_i^\top Y_i) \\
&= \hat{\boldsymbol{b}}_\lambda - (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top Y_i + \frac{(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top \mathbf{x}_i \hat{\boldsymbol{b}}_\lambda}{1 - \boldsymbol{S}_{\lambda ii}} \\
&\quad - \frac{(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top \boldsymbol{S}_{\lambda ii} Y_i}{1 - \boldsymbol{S}_{\lambda ii}} \\
&= \hat{\boldsymbol{b}}_\lambda - \frac{(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top}{1 - \boldsymbol{S}_{\lambda ii}} \left[Y_i (1 - \boldsymbol{S}_{\lambda ii}) - \mathbf{x}_i \hat{\boldsymbol{b}}_\lambda + \boldsymbol{S}_{\lambda ii} Y_i \right] \\
&= \hat{\boldsymbol{b}}_\lambda - \frac{(\boldsymbol{X}^\top \boldsymbol{X} + n\lambda\boldsymbol{\Omega})^{-1} \mathbf{x}_i^\top}{1 - \boldsymbol{S}_{\lambda ii}} \left[Y_i - \hat{\mu}_\lambda(t_i) \right],
\end{aligned}$$

giving

$$\mathbf{x}_i\hat{\boldsymbol{b}}_{\lambda(i)} = \mathbf{x}_i\hat{\boldsymbol{b}}_{\lambda} - \frac{\boldsymbol{S}_{\lambda ii}}{1-\boldsymbol{S}_{\lambda ii}}\left[Y_i - \hat{\mu}_{\lambda}(t_i)\right],$$
$$\hat{\mu}_{\lambda(i)}(t_i) = \frac{1}{1-\boldsymbol{S}_{\lambda ii}}(\hat{\mu}_{\lambda}(t_i) - \boldsymbol{S}_{\lambda ii}Y_i).$$

□

Hence, using (3.11), we obtain the following short-cut formula for computing $CV(\lambda)$ that allows us to avoid recomputing estimator after dropping out each observation:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{\mu}_{\lambda}(t_i)}{1-\boldsymbol{S}_{\lambda ii}}\right)^2. \tag{3.12}$$

Another useful method for λ selection is the generalized cross-validation (GCV). Assuming $tr(\boldsymbol{S}_{\lambda}) < n$, the GCV criterion is defined as

$$GCV(\lambda) = \frac{nRSS(\hat{\mu}_{\lambda})}{tr(I-S_{\lambda})^2}. \tag{3.13}$$

Exercise 5.9.7 from [**2**]. Derive the form of GCV criterion in the case of n_i replicate observations at each design points t_i all of which are to be weighted by a common factor w_i. Comment on the computational implications this criterion for smoothing with replicate data.

Solution: Suppose that our data follows the model

$$Y_{ij} = \mu(t_i) + \epsilon_{ij}, \quad j = 1, \ldots, n_i, i = 1, \ldots, r.$$

with corresponding sample means $\bar{Y}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} Y_{ij}$ and sample variances $S_i^2 = \frac{1}{n_i-1}\sum_{j=1}^{n_i}(Y_{ij} - \bar{Y}_i)^2, i = 1, \ldots, r$. This is a repeated sample problem

with non-constant variance. Hence, $\hat{\mu}_\lambda$ is obtained by minimizing

$$\frac{1}{n}\sum_{i=1}^{r} S_i^{-2}\sum_{j=1}^{n_i}(Y_{ij}-\mu(t_i))^2+\lambda\int_0^1(\mu^{(m)}(t))^2dt.$$

As we have seen in Exercise 5.9.4, this is equivalent as minimizing

$$\frac{1}{n}\sum_{i=1}^{r} S_i^{-2}n_i(\bar{Y}_i-\mu(t_i))^2+\lambda\int_0^1(\mu^{(m)}(t))^2dt,$$

and we have that the minimizer of above as $\hat{\mu}_\lambda=\sum_{j=1}^{n}\hat{b}_{\lambda j}x_j$, where

$$\widehat{\boldsymbol{b}}_\lambda=(\hat{b}_{\lambda 1},\ldots,\hat{b}_{\lambda n})^\top=(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}+n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\bar{\boldsymbol{Y}},$$

where $\boldsymbol{W}=\text{Diag}\left(\frac{n_1}{S_1^2},\ldots,\frac{n_n}{S_n^2}\right)^\top$ and $\bar{\boldsymbol{Y}}=(\bar{Y}_1,\ldots,\bar{Y}_n)^\top$. Hence, the vector of fitted values is

$$\widehat{\boldsymbol{\mu}}_\lambda=\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}+n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top\boldsymbol{W}\bar{\boldsymbol{Y}},$$

with the smoothing matrix

$$\boldsymbol{S}_\lambda=\boldsymbol{X}(\boldsymbol{X}^\top\boldsymbol{W}\boldsymbol{X}+n\lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^\top\boldsymbol{W}.$$

Thus, we can write the $GCV(\lambda)$ in (3.12) as

$$\begin{aligned}
GCV(\lambda)&=\frac{n\sum_{i=1}^{r}S_i^{-2}\sum_{j=1}^{n_i}(Y_{ij}-\hat{\mu}_\lambda(t_i))^2}{tr(\boldsymbol{I}-\boldsymbol{S}_\lambda)^2}\\
&=\frac{n\sum_{i=1}^{r}S_i^{-2}\sum_{j=1}^{n_i}(Y_{ij}-\bar{Y}_i)^2+(\bar{Y}_i-\hat{\mu}_\lambda(t_i))^2}{(n-tr(\boldsymbol{S}_\lambda))^2}\\
&=\frac{n(\sum_{i=1}^{r}(n_i-1)+\sum_{i=1}^{r}S_i^{-2}n_i(\bar{Y}_i-\hat{\mu}_\lambda(t_i))^2)}{n^2(1-n^{-1}tr(\boldsymbol{S}_\lambda))^2}\\
&=\frac{n-r+\sum_{i=1}^{r}S_i^{-2}n_i(\bar{Y}_i-\hat{\mu}_\lambda(t_i))^2}{n(1-n^{-1}tr(\boldsymbol{S}_\lambda))^2}.
\end{aligned}$$

The $GCV(\lambda)$ computation of repeated responses with non-constant variance converts the vector of responses into a mean vector and then evaluates the smoothing spline estimator based on the smoothing matrix for the obtained mean vector.

4 Efficient Computation

As we have seen in Sections 2 and 3, computing the smoothing spline estimator and selecting λ rely on computation of the smoothing matrix $\boldsymbol{S}_\lambda$. In this section, we consider an efficient way to compute $\boldsymbol{S}_\lambda$ in the cubic spline case, which can complete computation in $O(n)$ steps. To accomplish this, the natural spline basis function should be chosen to make $\boldsymbol{X}$ band-limited.

This method depends on two band-limited matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ defined as follows. Let $h_i = t_{i+1} - t_i, i = 1, \ldots, n-1$. Define $\boldsymbol{Q}$ to be an $n \times (n-2)$ matrix of the form

$$\boldsymbol{Q} = \{q_{ij}\}_{\substack{i=1,\ldots,n \\ j=2,\ldots,n-1}} = \begin{bmatrix} h_1^{-1} & 0 & \ldots & 0 \\ -h_1^{-1}-h_2^{-1} & h_2^{-1} & \ldots & 0 \\ h_2^{-1} & -h_2^{-1}-h_3^{-1} & \ldots & 0 \\ 0 & h_3^{-1} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & h_{n-2}^{-1} \\ 0 & 0 & \ldots & -h_{n-2}^{-1}-h_{n-1}^{-1} \\ 0 & 0 & \ldots & h_{n-1}^{-1} \end{bmatrix},$$

and $\boldsymbol{R}$ to be an $(n-2) \times (n-2)$ symmetric matrix of the form

$$\boldsymbol{R} = \{r_{ij}\}_{i,j=2,\ldots,n-1} = \begin{bmatrix} \frac{1}{3}(h_1+h_2) & \frac{1}{6}h_2 & 0 & \ldots & 0 \\ \frac{1}{6}h_2 & \frac{1}{3}(h_2+h_3) & \frac{1}{6}h_3 & \ldots & 0 \\ 0 & \frac{1}{6}h_3 & \frac{1}{3}(h_3+h_4) & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \ldots & \frac{1}{3}(h_{n-2}+h_{n-1}) \end{bmatrix}.$$

Note the columns of matrix $\boldsymbol{R}$ are linearly independent, hence $\boldsymbol{R}$ is invertable. Therefore, we can define matrix $\boldsymbol{K}$ by

$$\boldsymbol{K} = \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top.$$

Theorem 3 (Theorem 2.1 of [5]). *Suppose μ is a natural cubic spline with knots $0 < t_1 < \cdots < t_n < 1$, and define*

$$\mu_i = \mu(t_i) \text{ and } \delta_i = \mu''(t_i) \text{ for } i = 1, \ldots, n.$$

Then vectors $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$ *and* $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)^\top$ *specify a natural cubic spline if and only if*

$$\boldsymbol{Q}^\top \boldsymbol{\mu} = \boldsymbol{R}\boldsymbol{\delta},$$

if such condition is satisfied the roughness penalty can be expressed as

$$\int_0^1 (\mu''(t))^2 dt = \boldsymbol{\delta}^\top \boldsymbol{R}\boldsymbol{\delta} = \boldsymbol{\mu}^\top \boldsymbol{K}\boldsymbol{\mu}.$$

Proof. Since μ is a cubic function on the subintervals (t_i, t_{i+1}) for $i = 1, \ldots, n-1$, μ'' is linear on the subintervals, and it can be expressed as $\mu''(t) = at + b$ for $t_i \leq t \leq t_{i+1}$. Evaluating this function at t_i and t_{i+1} results in

$$\mu''(t) = \frac{(t - t_i)\delta_{i+1} + (t_{i+1} - t)\delta_i}{h_i}. \tag{4.1}$$

Differentiating both sides of (4.1), we obtain

$$\mu'''(t) = \frac{\delta_{i+1} - \delta_i}{h_i}. \tag{4.2}$$

Next, we substitute relations (4.1), (4.2), along with the initial conditions $\mu(t_i) = \mu_i$ and $\mu(t_{i+1}) = \mu_{i+1}$, into

$$\begin{aligned} \mu(t) &= at^3 + bt^2 + ct + d, \\ \mu''(t) &= 6at + 2b, \\ \mu'''(t) &= 6a. \end{aligned}$$

Through lengthy calculations, we obtain the following expression:

$$\begin{aligned} \mu(t) = &\frac{(t - t_i)\mu_{i+1} + (t_{i+1} - t)\mu_i}{h_i} \\ &- \frac{1}{6}(t - t_i)(t_{i+1} - t)\left\{\left(1 + \frac{t - t_i}{h_i}\right)\delta_{i+1} + \left(1 + \frac{t_{i+1} - t}{h_i}\right)\delta_i\right\}. \end{aligned}$$

Taking the derivative yields

$$\mu'(t) = \frac{\mu_{i+1} - \mu_i}{h_i} - \frac{1}{6}(t - t_i)(t_{i+1} - t)\frac{\delta_{i+1} - \delta_i}{h_i} \\ - \frac{1}{6}\{(t_{i+1} - t) + (t_i - t)\}\left\{\left(1 + \frac{t - t_i}{h_i}\right)\delta_{i+1} + \left(1 + \frac{t_{i+1} - t}{h_i}\right)\delta_i\right\}.$$

The definition of the natural boundary conditions implies $\mu''(t) = 0$ for t outside of the interval (t_1, t_n), that is, $\delta_1 = \delta_n = 0$. Thus

$$\mu'(t_1) = \frac{\mu_2 - \mu_1}{h_1} - \frac{1}{6}h_1\delta_2,$$

and

$$\mu'(t_n) = \frac{\mu_n - \mu_{n-1}}{n_1} + \frac{1}{6}h_{n-1}\delta_{n-1}.$$

It is clear now that for $t_i, i = 2, \ldots, n-1$,

$$\mu'(t_i^+) = \frac{\mu_{i+1} - \mu_i}{h_i} - \frac{1}{6}h_i(\delta_{i+1} + 2\delta_i) \tag{4.3}$$

and

$$\mu'(t_i^-) = \frac{\mu_i - \mu_{i-1}}{h_{i-1}} + \frac{1}{6}h_{i-1}(2\delta_i + \delta_{i-1}), \tag{4.4}$$

where $\mu'(t_i^+) = \lim_{x \searrow t_i^+} \mu'(x)$ and $\mu'(t_i^-) = \lim_{x \nearrow t_i^-} \mu'(x)$. From the definition of a natural spline in (1.6), equations (4.3) and (4.4) must be equal. Therefore

$$\frac{\mu_{i+1} - \mu_i}{h_i} - \frac{\mu_i - \mu_{i-1}}{h_{i-1}} = \frac{1}{6}h_{i-1}\delta_{i-1} + \frac{1}{3}(h_{i-1} + h_i)\delta_i + \frac{1}{6}h_i\delta_{i+1}. \tag{4.5}$$

Thus, we have shown that

$$\boldsymbol{Q}^\top \boldsymbol{\mu} = \boldsymbol{R}\boldsymbol{\delta}.$$

Now, for the penalty term we have

$$\begin{aligned}\int_0^1 (\mu''(t))^2 dt &= \mu''(t)\mu'(t)|_0^1 - \int_0^1 \mu'''(t)\mu'(t)dt \\ &= -\int_{t_1}^{t_n} \mu'''(t)\mu'(t)dt \\ &= -\sum_{i=1}^{n-1} \frac{\delta_{i+1}-\delta_i}{h_i} \int_{t_i}^{t_{i+1}} \mu'(t)dt \\ &= -\sum_{i=1}^{n-1} \frac{\delta_{i+1}-\delta_i}{h_i}(\mu_{i+1}-\mu_i).\end{aligned} \tag{4.6}$$

where the third equality follows from (4.2) and the fact that $\delta_1 = \delta_n = 0$. From (4.6) we obtain

$$\begin{aligned}\int_0^1 (\mu''(t))^2 dt &= \frac{\delta_2}{h_1}(\mu_1-\mu_2) - \frac{\delta_2}{h_2}(\mu_2-\mu_3) - \cdots - \frac{\delta_{n-1}}{h_{n-1}}(\mu_n-\mu_{n-1}) \\ &= \sum_{j=2}^{n-1} \delta_i \left(\frac{\mu_{i+1}-\mu_i}{h_i} - \frac{\mu_i-\mu_{i-1}}{h_{i-1}} \right) \\ &= \boldsymbol{\delta}^\top \boldsymbol{Q}^\top \boldsymbol{\mu} = \boldsymbol{\delta}^\top \boldsymbol{R}\boldsymbol{\delta} = (\boldsymbol{R}^{-1}\boldsymbol{Q}^\top\boldsymbol{\mu})^\top \boldsymbol{R}(\boldsymbol{R}^{-1}\boldsymbol{Q}^\top\boldsymbol{\mu}) \\ &= \boldsymbol{\mu}^\top \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top\boldsymbol{\mu} \\ &= \boldsymbol{\mu}^\top \boldsymbol{K}\boldsymbol{\mu}.\end{aligned}$$

□

Now, by Theorem 3, we can express the PLS in the cubic spline case as

$$\frac{1}{n}(\boldsymbol{Y}-\boldsymbol{\mu})^\top(\boldsymbol{Y}-\boldsymbol{\mu}) + \lambda \boldsymbol{\mu}^\top \boldsymbol{K}\boldsymbol{\mu}. \tag{4.7}$$

Then, taking the derivative w.r.t $\boldsymbol{\mu}$, setting it equal to zero and solve the obtained equation, we get the minimizer of (4.7) in the form

$$\widehat{\boldsymbol{\mu}}_\lambda = (\boldsymbol{I} + n\lambda \boldsymbol{K})^{-1}\boldsymbol{Y}, \tag{4.8}$$

where $\widehat{\boldsymbol{\mu}}_\lambda = (\hat{\mu}_1, \ldots, \hat{\mu}_n)^\top$. Rearranging the terms in (4.8) and substituting $\boldsymbol{K} = \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top$, we get

$$\widehat{\boldsymbol{\mu}}_\lambda = \boldsymbol{Y} - n\lambda \boldsymbol{Q}\boldsymbol{R}^{-1}\boldsymbol{Q}^\top \widehat{\boldsymbol{\mu}}_\lambda.$$

From Theorem 2 in Section 2.1, we know that $\widehat{\boldsymbol{\mu}}_\lambda$ is indeed a natural cubic spline. Put $\hat{\delta}_i = \hat{\mu}''_\lambda(t_i)$ and $\widehat{\boldsymbol{\delta}}_\lambda = (\hat{\delta}_1, \ldots, \hat{\delta}_n)^\top$. Using Theorem 3, we get

$$\widehat{\boldsymbol{\mu}}_\lambda = \boldsymbol{Y} - n\lambda \boldsymbol{Q}\widehat{\boldsymbol{\delta}}_\lambda. \tag{4.9}$$

Then, we multiply both sides of (4.9) by $\boldsymbol{Q}^\top$, and obtain

$$\boldsymbol{Q}^\top \widehat{\boldsymbol{\mu}}_\lambda = \boldsymbol{Q}^\top \boldsymbol{Y} - n\lambda \boldsymbol{Q}^\top \boldsymbol{Q}\widehat{\boldsymbol{\delta}}_\lambda,$$

that is,

$$\boldsymbol{R}\widehat{\boldsymbol{\delta}}_\lambda = \boldsymbol{Q}^\top \boldsymbol{Y} - n\lambda \boldsymbol{Q}^\top \boldsymbol{Q}\widehat{\boldsymbol{\delta}}_\lambda.$$

Thus, we arrive at the following important result:

$$(\boldsymbol{R} + n\lambda \boldsymbol{Q}^\top \boldsymbol{Q})\widehat{\boldsymbol{\delta}}_\lambda = \boldsymbol{Q}^\top \boldsymbol{Y}. \tag{4.10}$$

The matrix $\boldsymbol{R} + n\lambda \boldsymbol{Q}^\top \boldsymbol{Q}$ is a $(n-2) \times (n-2)$ band-limited matrix of bandwith 5 which is symmetric. Therefore, it has a Cholesky decomposition (see Appendix for detail) of the form

$$\boldsymbol{R} + n\lambda \boldsymbol{Q}^\top \boldsymbol{Q} = \boldsymbol{B} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\top,$$

where $\boldsymbol{D}$ is a diagonal matrix with elements D_j, and $\boldsymbol{L}$ is a lower triangular band matrix which has elements L_{ij} with $L_{ii} = 1$ for all i, and $L_{ij} = 0$ for $j < i - 2$ and $j > i$. The following recursive relation holds for the elements of $\boldsymbol{D}$ and $\boldsymbol{L}$:

$$D_1 = B_{11}, \quad D_j = B_{jj} - \sum_{k=1}^{j-1} L_{jk}^2 D_k, \quad \text{for} \quad j = 2, \ldots, n-2,$$

$$L_{ij} = \frac{1}{D_j}\left(B_{ij} - \sum_{k=1}^{j-1} L_{ik}L_{jk}D_k\right) \text{ for } 1 \leq j < i \leq n-2.$$

In the case of bandwith of 5, we have that

$$D_1 = B_{11}, \quad D_2 = B_{22} - L_{21}^2 D_1, \quad D_3 = B_{33} - L_{31}^2 D_1 - L_{32}^2 D_2,$$
$$D_4 = B_{44} - L_{42}^2 D_2 - L_{43}^2 D_3, \ldots,$$

and

$$L_{21} = \frac{B_{21}}{D_1}, \quad L_{31} = \frac{B_{31}}{D_1}, \quad L_{32} = \frac{B_{32} - L_{31}L_{21}D_1}{D_2},$$
$$L_{42} = \frac{B_{42}}{D_2}, \quad L_{43} = \frac{B_{43} - L_{42}L_{32}D_2}{D_3}, \ldots,$$

deducing that for $j = 3, \ldots, n-2$,

$$D_j = B_{jj} - L_{j,j-2}^2 D_{j-2} - L_{j,j-1}^2 D_{j-1},$$
$$L_{j,j-2} = \frac{B_{j,j-2}}{D_{j-2}},$$
$$L_{j,j-1} = \frac{B_{j,j-1} - L_{j,j-2}L_{j-1,j-2}D_{j-2}}{D_{j-1}}.$$

Therefore, the total number of operations needed for the Cholesky decomposition is $O(n)$. To find $\boldsymbol{B}^{-1}$ we have

$$\boldsymbol{B}^{-1} = (\boldsymbol{L}^\top)^{-1}\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}$$

which can be expressed as

$$\boldsymbol{B}^{-1} = \boldsymbol{D}^{-1}\boldsymbol{L}^{-1} - \boldsymbol{L}^\top\boldsymbol{B}^{-1} + \boldsymbol{B}^{-1} = \boldsymbol{D}^{-1}\boldsymbol{L}^{-1} + (\boldsymbol{I} - \boldsymbol{L}^\top)\boldsymbol{B}^{-1}. \tag{4.11}$$

Note that $\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}$ is a lower triangular matrix with diagonal elements D_j^{-1} and $(\boldsymbol{I} - \boldsymbol{L}^\top)$ is a upper triangular matrix with the diagonal element equals

to 0. Let b_{ij} be the elements of $\boldsymbol{B}^{-1}$. Since the inverse of symmetric matrix remains symmetric, we only need to find b_{ij} for all $j > i$, which means that only the diagonal part of $\boldsymbol{D}^{-1}\boldsymbol{L}^{-1}$ is needed. Therefore, from (4.11), we have that for $i = 1, \ldots, n-4$,

$$\begin{aligned} b_{ii} &= D_i^{-1} - L_{i+1,i}b_{i+1,i} - L_{i+2,i}b_{i+2,i}, \\ b_{i+1,i} &= b_{i,i+1} = -L_{i+1,i}b_{i+1,i+1} - L_{i+2,i}b_{i+2,i+1}, \\ b_{i+2,i} &= b_{i,i+2} = -L_{i+1,i}b_{i+1,i+2} - L_{i+2,i}b_{i+2,i+2}, \\ &\vdots \end{aligned}$$

Therefore,

$$b_{i,i+k} = -L_{i+1,i}b_{i+1,i+k} - L_{i+2,i}b_{i+2,i+k} \text{ for } k = 1, \ldots, n-2-i. \tag{4.12}$$

We can see that this algorithm needs to be calculated backwards, starting with

$$\begin{aligned} b_{n-2,n-2} &= D_{n-2}^{-1}, \\ b_{n-3,n-2} &= -L_{n-2,n-3}b_{n-2,n-2}, \\ b_{n-3,n-3} &= D_{n-3}^{-1} - L_{n-2,n-3}b_{n-2,n-3}. \end{aligned}$$

The order of elements of $\boldsymbol{B}^{-1}$ to be obtained from this algorithm is $(n-2, n-2), (n-3, n-2), (n-3, n-3), (n-4, n-2), (n-4, n-3), (n-4, n-4), \ldots, (3,1), (2,1), (1,1)$. By this method, $\boldsymbol{B}^{-1}$ can be found in $O(n)$ operations.

Now, let us write the smoothing matrix, using (4.8) and (4.9), as

$$\widehat{\boldsymbol{S}}_\lambda = \boldsymbol{I} - n\lambda\boldsymbol{Q}(\boldsymbol{R} + n\lambda\boldsymbol{Q}^\top\boldsymbol{Q})^{-1}\boldsymbol{Q}^\top.$$

Given that the matrices $\boldsymbol{Q}$ and $\boldsymbol{R}$ can be found in $O(n)$ operations, the Cholesky decomposition can be obtained in $O(n)$ operations, and the inverse can also be found in $O(n)$ operations. This results in the total operations for finding $\boldsymbol{S}_\lambda$ to be $O(n)$. Therefore, the cubic smoothing spline computation can be very fast in practice.

5 Large-Sample Properties

We shall use the linear smoothing spline case developed in Section 2 to motivate more general developments. Our goal now is to obtain the large-sample approximation to the sequence of risks:

$$R_n(\lambda) = \boldsymbol{E}(L_n(\lambda)) = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{E}(\mu(t_i) - \hat{\mu}_\lambda(t_i))^2,$$

where $t_i = \frac{2i-1}{2n}, i = 1, \ldots, n$. Observe that

$$R_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\left(\hat{\mu}_\lambda(t_i)\right) + \frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{E}\hat{\mu}_\lambda(t_i) - \mu(t_i))^2.$$

Recall from Section 2.2 that the linear smoothing spline can be expressed as

$$\hat{\mu}_\lambda(t) = \frac{1}{n}\sum_{i=1}^{n} K_n(t, t_i; \lambda) Y_i, \quad t \in [0,1] \tag{5.1}$$

where for $t, s \in [0,1]$

$$K_n(t, s; \lambda) = 1 + \sqrt{2}\sum_{j=1}^{n-1} \frac{\cos(j\pi s) x_{j+1}(t)}{1 + \lambda\gamma_j}. \tag{5.2}$$

Here, x_j and γ_j are the Demmler-Reinsch basis functions and eigenvalues given for $j = 1, \ldots, n-1$ by

$$\gamma_j = \gamma_{j,n} = \left(2n \sin\left(\frac{j\pi}{2n}\right)\right)^2,$$

and

$$x_{j+1}(t) = \begin{cases} \sqrt{2}\cos(j\pi t_1), & 0 \le t < t_1, \\ \sqrt{2}\cos(j\pi t_i) & \\ +\sqrt{2}\frac{t-t_i}{t_{i+1}-t_i}\left[\cos(j\pi t_{i+1}) - \cos(j\pi t_i)\right], & t_i \le t < t_{i+1}, \\ & i = 1, \ldots, n-1, \\ \sqrt{2}\cos(j\pi t_n), & t_n \le t \le 1. \end{cases}$$

To obtain a large-sample approximation of $K_n(t, s; \lambda)$, let us first consider the special case of $t = t_i$, which results in

$$K_n(t, s; \lambda) = 1 + 2\sum_{j=1}^{n-1} \frac{\cos(j\pi s)\cos(j\pi t)}{1 + \lambda\gamma_j}, \tag{5.3}$$

where, by L'Hôpital's Rule,

$$\lim_{n\to\infty} \gamma_j = \left(2 \lim_{n\to\infty} \frac{\sin\left(\frac{j\pi}{2n}\right)}{\frac{1}{n}}\right)^2 = \left(2 \lim_{n\to\infty} \frac{\cos\left(\frac{j\pi}{2n}\right)\frac{j\pi}{2n^2}}{\frac{1}{n^2}}\right)^2 = (j\pi)^2.$$

Thus, for $t, s \in [0, 1]$

$$K^+(t, s; \lambda) := \lim_{n\to\infty} K_n(t, s; \lambda) = 1 + 2\sum_{j=1}^{\infty} \frac{\cos(j\pi s)\cos(j\pi t)}{1 + \lambda(j\pi)^2}, \tag{5.4}$$

and, using the identity $\cos(u)\cos(v) = \frac{1}{2}[\cos(u+v) + \cos(u-v)]$, relation

(5.4) can be written as

$$\begin{aligned} K^+(t,s;\lambda) &= 1+\sum_{j=1}^{\infty}\frac{\cos(j\pi(s+t))}{1+\lambda(j\pi)^2}+\sum_{j=1}^{\infty}\frac{\cos(j\pi(s-t))}{1+\lambda(j\pi)^2} \\ &= 1+\sum_{j=1}^{\infty}\frac{\cos(j\pi(s+t))}{\lambda\pi^2(\frac{1}{\lambda\pi^2}+j^2)}+\sum_{j=1}^{\infty}\frac{\cos(j\pi(s-t))}{\lambda\pi^2(\frac{1}{\lambda\pi^2}+j^2)}. \end{aligned} \tag{5.5}$$

Then, using the identity

$$\sum_{k=0}^{\infty}\frac{\cos(kx)}{a^2+k^2}=\frac{\pi}{2a}\frac{e^{-a(\pi-|x|)}+e^{a(\pi-|x|)}}{e^{a\pi}-e^{-a\pi}}+\frac{1}{2a^2},\quad |x|\leq 2\pi,$$

which gives that

$$\sum_{k=1}^{\infty}\frac{\cos(kx)}{a^2+k^2}=\frac{\pi}{2a}\frac{e^{-a(\pi-|x|)}+e^{a(\pi-|x|)}}{e^{a\pi}-e^{-a\pi}}-\frac{1}{2a^2},\quad |x|\leq 2\pi,$$

and setting $a=\frac{1}{\sqrt{\lambda}\pi}$, we obtain from (5.5)

$$\begin{aligned} K^+(t,s;\lambda) &= 1+\frac{1}{\lambda\pi^2}\left(\frac{\pi}{2(\sqrt{\lambda}\pi)^{-1}}\frac{e^{-\frac{\pi-|\pi(s+t)|}{\sqrt{\lambda}\pi}}+e^{\frac{\pi-|\pi(s+t)|}{\sqrt{\lambda}\pi}}}{e^{\frac{1}{\sqrt{\lambda}}}-e^{-\frac{1}{\sqrt{\lambda}}}}-\frac{\lambda\pi^2}{2}\right) \\ &\quad+\frac{1}{\lambda\pi^2}\left(\frac{\pi}{2(\sqrt{\lambda}\pi)^{-1}}\frac{e^{-\frac{\pi-|\pi(s-t)|}{\sqrt{\lambda}\pi}}+e^{\frac{\pi-|\pi(s-t)|}{\sqrt{\lambda}\pi}}}{e^{\frac{1}{\sqrt{\lambda}}}-e^{-\frac{1}{\sqrt{\lambda}}}}-\frac{\lambda\pi^2}{2}\right) \\ &= \frac{1}{2\sqrt{\lambda}}\frac{e^{-\frac{1-(s+t)}{\sqrt{\lambda}}}+e^{\frac{1-(s+t)}{\sqrt{\lambda}}}+e^{-\frac{1-|s-t|}{\sqrt{\lambda}}}+e^{\frac{1-|s-t|}{\sqrt{\lambda}}}}{e^{\frac{1}{\sqrt{\lambda}}}-e^{-\frac{1}{\sqrt{\lambda}}}} \\ &= \frac{e^{\frac{s+t-2}{\sqrt{\lambda}}}+e^{\frac{-(s+t)}{\sqrt{\lambda}}}+e^{\frac{|s-t|-2}{\sqrt{\lambda}}}+e^{\frac{-|s-t|}{\sqrt{\lambda}}}}{2\sqrt{\lambda}(1-e^{-\frac{2}{\sqrt{\lambda}}})}. \end{aligned} \tag{5.6}$$

This approximation holds true not only point-wisely at the design points but also uniformly over $[0,1]$, as seen from the following theorem.

Theorem 4 (Theorem 5.4 of [**2**])**.** *Assume that $\lambda\to 0$ as $n\to\infty$ in such a way that $n\lambda\to\infty$ as $n\to\infty$. Then,*

$$K_n(t,s;\lambda)=K^+(t,s;\lambda)+O\left(\frac{1}{n\lambda}\right),$$

uniformly for $t,s\in[0,1]$.

Proof. We begin by finding the upper bound for

$$\left| x_{j+1}(t) - \sqrt{2}\cos(j\pi t) \right|,$$

where $t \in [0,1]$. From Taylor's formula, we have that for $t_i \leq t < t_{i+1}$ and all large enough n

$$\cos(j\pi t) = \cos(j\pi t_i) - j\pi \sin(j\pi t_i)(t - t_i) + R_1(t), \tag{5.7}$$

where $R_1(t) = o(|t - t_i|) = o(n^{-1})$, and

$$\cos(j\pi t_i) = \cos(j\pi t_{i+1}) + j\pi \sin(j\pi t_{i+1})(t_{i+1} - t_i) + R_1(t_i), \tag{5.8}$$

where $R_1(t_i) = o(|t_{i+1} - t_i|) = o(n^{-1})$, and

$$\sin(j\pi t_i) = \sin(j\pi t_{i+1}) + j\pi \cos(j\pi t_{i+1})(t_i - t_{i+1}) + R_1(t_i), \tag{5.9}$$

where $R_1(t_i) = o(|t_{i+1} - t_i|) = o(n^{-1})$.

Also, from the Maclaurin series of $\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$, we have that

$$x - \frac{x^3}{3!} \leq \sin(x) \leq x, \quad x \in \mathbb{R}, \tag{5.10}$$

and hence, for the cases when $0 \leq t < t_1$ and $t_n \leq t \leq 1$, using (5.7), (5.10), and the fact that $t_{i+1} - t_i = \frac{1}{n}$, we obtain that for all large enough n

$$\left| \sqrt{2}\cos(j\pi t_1) - \sqrt{2}\cos(j\pi t) \right| \leq \sqrt{2}\left| j\pi \sin(j\pi t_1)(t - t_1) \right| \leq \sqrt{2}\left(\frac{j\pi}{n}\right)^2,$$

$$\left| \sqrt{2}\cos(j\pi t_n) - \sqrt{2}\cos(j\pi t) \right| \leq \sqrt{2}\left| j\pi \sin(j\pi t_n)(t - t_n) \right| \leq \sqrt{2}\left(\frac{j\pi}{n}\right)^2.$$

Now, for $t_i \leq t < t_{i+1}, i = 1, \ldots, n-1$, we have that

$$x_{j+1}(t) = \sqrt{2}\cos(j\pi t_i) + \sqrt{2}\frac{t - t_i}{t_{i+1} - t_i}\left[\cos(j\pi t_{i+1}) - \cos(j\pi t_i)\right].$$

Hence, using (5.7), (5.8) and (5.9), we obtain that for all large enough n

$$\begin{aligned}
&\left|x_{j+1}(t) - \sqrt{2}\cos(j\pi t)\right| \\
&= \left|\sqrt{2}\cos(j\pi t_i) - \sqrt{2}\cos(j\pi t) + \sqrt{2}\frac{t-t_i}{t_{i+1}-t_i}[\cos(j\pi t_{i+1}) - \cos(j\pi t_i)]\right| \\
&= \Big|\sqrt{2}j\pi\sin(j\pi t_i)(t-t_i) + o(n^{-1}) \\
&\quad -\sqrt{2}\frac{t-t_i}{t_{i+1}-t_i}[j\pi\sin(j\pi t_{i+1})(t_{i+1}-t_i) + o\left(n^{-1}\right)]\Big| \\
&= \sqrt{2}j\pi(t-t_i)\left|\sin(j\pi t_i) - \sin(j\pi t_{i+1}) + o(n^{-1})\right| \\
&\leq \sqrt{2}j\pi(t-t_i)|j\pi\cos(j\pi t_{i+1})(t_i - t_{i+1}) + o(n^{-1})| \leq \sqrt{2}\left(\frac{j\pi}{n}\right)^2.
\end{aligned}$$

Therefore, we have that for all large enough n

$$\left|x_{j+1}(t) - \sqrt{2}\cos(j\pi t)\right| \leq \sqrt{2}\left(\frac{j\pi}{n}\right)^2 \tag{5.11}$$

for all $t \in [0,1]$.

Using relation (5.11), we can write for all large enough n

$$\begin{aligned}
&\left|K_n(t,s;\lambda) - \left(1 + 2\sum_{j=1}^{n-1}\frac{\cos(j\pi s)\cos(j\pi t))}{1+\lambda\gamma_j}\right)\right| \\
&= \left|\sum_{j=1}^{n-1}\frac{\sqrt{2}\cos(j\pi s)[x_{j+1}(t) - \sqrt{2}\cos(j\pi t)]}{1+\lambda\gamma_j}\right| \leq 2\sum_{j=1}^{n-1}\frac{\cos(j\pi s)(\frac{j\pi}{n})^2}{1+\lambda[2n\sin(\frac{j\pi}{2n})]^2} \\
&\leq 2\sum_{j=1}^{n-1}\frac{\cos(j\pi s)(\frac{j\pi}{n})^2}{1+\lambda(\frac{j\pi}{2})^2} \leq 2\sum_{j=1}^{n-1}\frac{(\frac{j\pi}{n})^2}{1+\lambda(\frac{j\pi}{2})^2} \\
&= \frac{8}{n^2\lambda}\sum_{j=1}^{n-1}\frac{(j\pi)^2}{\frac{1}{4\lambda}+(j\pi)^2} < \frac{8}{n\lambda},
\end{aligned}$$

and

$$\begin{aligned}
&\left|K^+(t,s;\lambda) - \left(1 + 2\sum_{j=1}^{n-1} \frac{\cos(j\pi s)\cos(j\pi t))}{1+\lambda\gamma_j}\right)\right| \\
&= \left|2\sum_{j=1}^{\infty} \frac{\cos(j\pi s)\cos(j\pi t)}{1+\lambda(j\pi)^2} - 2\sum_{j=1}^{n-1} \frac{\cos(j\pi s)\cos(j\pi t)}{1+\lambda\gamma_j}\right| \\
&\le 2\left|\sum_{j=1}^{n-1}\left(\frac{1}{1+\lambda(j\pi)^2} - \frac{1}{1+\lambda\gamma_j}\right) + \sum_{j=n}^{\infty} \frac{2\cos(j\pi s)\cos(j\pi t_n)}{1+\lambda(j\pi)^2}\right| \\
&\le 2\sum_{j=1}^{n-1} \frac{\lambda|\gamma_j - (j\pi)^2|}{(1+\lambda(j\pi)^2)(1+\lambda\gamma_j)} + \sum_{j=n}^{\infty} \frac{2}{1+\lambda(j\pi)^2} \\
&\le 2\lambda\sum_{j=1}^{n-1} \frac{|\sqrt{\gamma_j} - j\pi|(\sqrt{\gamma_j} + j\pi)}{(1+\lambda(j\pi)^2)(1+\lambda(\frac{j\pi}{2})^2)} + \frac{2}{\lambda}\sum_{j=n}^{\infty} \frac{1}{(j\pi)^2}. \qquad (5.12)
\end{aligned}$$

From relation (5.10), it is true that

$$\frac{j\pi}{2} < j\pi\left(1 - \frac{1}{3!}\left(\frac{j\pi}{2n}\right)^2\right) \le 2n\sin\left(\frac{j\pi}{2n}\right) \le j\pi,$$

which gives

$$\left|2n\sin\left(\frac{j\pi}{2n}\right) - j\pi\right| \le \frac{(j\pi)^3}{24n^2}. \qquad (5.13)$$

Using (5.12) and (5.13), we obtain that for all large enough n

$$\begin{aligned}
&\left|K^+(t,s;\lambda) - \left(1 + 2\sum_{j=1}^{n-1} \frac{\cos(j\pi s)\cos(j\pi t))}{1+\lambda\gamma_j}\right)\right| \\
&\le 2\lambda\sum_{j=1}^{n-1} \frac{\left(\frac{(j\pi)^3}{24n^2}\right)(2j\pi)}{(1+\lambda(\frac{j\pi}{2})^2)^2} + O\left(\frac{1}{n\lambda}\right) \le \frac{4\lambda}{24n^2}\sum_{j=1}^{n-1} \frac{(j\pi)^4}{(\frac{\lambda}{4})^2(\frac{4}{\lambda} + (j\pi)^2)^2} + O\left(\frac{1}{n\lambda}\right) \\
&\le \frac{8}{3\lambda n^2}\sum_{j=1}^{n-1} \frac{(j\pi)^4}{\left(\frac{4}{\lambda} + (j\pi)^2\right)^2} + O\left(\frac{1}{n\lambda}\right) < \frac{8}{3\lambda n} + O\left(\frac{1}{n\lambda}\right) = O\left(\frac{1}{n\lambda}\right).
\end{aligned}$$

The proof of Theorem 4 is complete. □

Typically, we are most concerned with the case where λ decays to 0 at the rate of n^{-a} for some $a \in (0,1)$, to satisfy the condition $n\lambda \to \infty$. But looking at (5.6), we have a term $\exp(-\frac{1}{\sqrt{\lambda}})$, which decays to 0 very fast. Recall the following property:

$$\begin{aligned} &\text{For all} \quad a, p > 0, \quad \exp(x^a) \to \infty \\ &\text{faster than any function} \quad x^p \quad \text{as} \quad x \to \infty, \end{aligned} \tag{5.14}$$

which gives that $\exp(-x^a) = o(x^{-p})$ as $x \to \infty$. In our situation, we have $\exp(-\frac{1}{\sqrt{\lambda}}) = \exp(-n^{a/2}) = o(n^{-p})$ as $x \to \infty$. Therefore, Theorem 4 can be simplified to the following statement: if $\lambda = n^{-a}$ for some $a \in (0,1)$, then as $n \to \infty$

$$K_n(t, s; \lambda) = K(t, s; \lambda) + O\left(\frac{1}{n\lambda}\right), \quad t, s \in [0,1], \tag{5.15}$$

where $K(t, s; \lambda) = \frac{1}{2\sqrt{\lambda}} \left(e^{\frac{s+t-2}{\sqrt{\lambda}}} + e^{\frac{-(s+t)}{\sqrt{\lambda}}} + e^{\frac{-|s-t|}{\sqrt{\lambda}}} \right)$. This approximation allows us to analyse the point-wise variance and bias of the linear smoothing spline needed for the risk approximation. From (5.1) and (5.15), we obtain that for all $t \in [0,1]$ and all large enough n

$$\begin{aligned} \boldsymbol{E}(\hat{\mu}_\lambda(t)) &= \frac{1}{n} \sum_{i=1}^{n} K_n(t, t_i; \lambda) \mu(t_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} \left[K(t, t_i; \lambda) + O\left(\frac{1}{n\lambda}\right) \right] \mu(t_i) \\ &= \frac{1}{n} \sum_{i=1}^{n} K(t, t_i; \lambda) \mu(t_i) + O\left(\frac{1}{n\lambda}\right) \\ &= \int_0^1 K(t, s; \lambda) \mu(s) ds + O\left(\frac{1}{n\lambda}\right), \end{aligned} \tag{5.16}$$

where the last equality is obtained from the following arguments. Set $s_0 = 0, s_i = \frac{i}{n}, i = 1, \ldots, n$. Using the mean value theorem with $t_i, \xi_i \in [s_{i-1}, s_i]$,

we have that for all large enough n

$$\begin{aligned}
&\left|\frac{1}{n}\sum_{i=1}^{n} K(t,t_i;\lambda)\mu(t_i) - \int_0^1 K(t,s;\lambda)\mu(s)ds\right| \\
&= \left|\sum_{i=1}^{n} \mu(t_i)\int_{s_{i-1}}^{s_i} K(t,s;\lambda)ds - \int_0^1 K(t,s;\lambda)\mu(s)ds\right| \\
&= \left|\sum_{i=1}^{n} [\mu(t_i)-\mu(\xi_i)]\int_{s_{i-1}}^{s_i} K(t,s;\lambda)ds\right| \\
&\le \max_{t\in[0,1]}|\mu'(t)|\frac{1}{n}\left|\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K(t,s;\lambda)ds\right| \\
&\le \max_{t\in[0,1]}|\mu'(t)|\frac{1}{n}\max_{t\in[0,1]}\max_{s\in[0,1]}|K(t,s;\lambda)| = O\left(\frac{1}{n\sqrt{\lambda}}\right).
\end{aligned}$$

Now, for the variance of $\hat{\mu}_\lambda(t)$, by (5.15) for all large enough n we have that

$$\begin{aligned}
\mathrm{Var}(\hat{\mu}_\lambda(t)) &= \frac{\sigma^2}{n^2}\sum_{i=1}^{n} K_n^2(t,t_i;\lambda) = \frac{\sigma^2}{n^2}\sum_{i=1}^{n}\left[K(t,t_i;\lambda)+O\left(\frac{1}{n\lambda}\right)\right]^2 \\
&= \frac{\sigma^2}{n^2}\left[\sum_{i=1}^{n} K^2(t,t_i;\lambda) + O\left(\frac{1}{n\lambda}\right)\sum_{i=1}^{n} K(t,t_i;\lambda) + O\left(\frac{1}{(n\lambda)^2}\right)\right].
\end{aligned} \tag{5.17}$$

To approximate $\sum_{i=1}^{n} K^2(t,t_i;\lambda)$ by an integral, we again set $s_0 = 0, s_i = \frac{i}{n}, i = 1,\ldots,n$, and observe that

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} K^2(t,t_i;\lambda) - \int_0^1 K^2(t,s;\lambda)ds \\
&= \sum_{i=1}^{n}\left[\frac{1}{n}K^2(t,t_i;\lambda) - \int_{s_{i-1}}^{s_i} K^2(t,s;\lambda)\right]ds.
\end{aligned} \tag{5.18}$$

From the mean value theorem for integrals, we have that $\int_{s_{i-1}}^{s_i} K^2(t,s;\lambda)ds = \frac{1}{n}K^2(t,\xi_i;\lambda)$ with some $\xi_i \in [s_{i-1},s_i]$. Thus, the problem is now to find the

difference of

$$\sum_{i=1}^{n} \frac{1}{n}[K^2(t,t_i;\lambda) - K^2(t,\xi_i;\lambda)].$$

Using the fact that $t_i \in [s_{i-1}, s_i]$ and applying the mean value theorem, we obtain

$$\begin{aligned} |K^2(t,t_i;\lambda) - K^2(t,\xi_i;\lambda)| &\leq \max_t \max_i \left| \frac{d}{ds} K^2(t,s;\lambda)|_{s=\theta_i} \right| |(t_i - \xi_i)| \\ &\leq \max_t \max_i \left| \frac{d}{ds} K^2(t,s;\lambda)|_{s=\theta_i} \right| \frac{1}{n} \end{aligned} \tag{5.19}$$

with some $\theta_i \in [t_i, \xi_i]$. Taking the derivative on the right-hand side of (5.19), we obtain

$$\begin{aligned} \frac{d}{ds} K^2(t,s;\lambda)|_{s=\theta_i} = 2\left(\frac{1}{4\lambda}\right)\left(e^{\frac{\theta_i+t-2}{\sqrt{\lambda}}} + e^{\frac{-(\theta_i+t)}{\sqrt{\lambda}}} + e^{\frac{-|\theta_i-t|}{\sqrt{\lambda}}}\right)\left(\frac{1}{\sqrt{\lambda}}\right) \\ \times \left(e^{\frac{\theta_i+t-2}{\sqrt{\lambda}}} - e^{\frac{-(\theta_i+t)}{\sqrt{\lambda}}} - \frac{(\theta_i - t)}{|\theta_i - t|} e^{\frac{-|\theta_i-t|}{\sqrt{\lambda}}}\right). \end{aligned} \tag{5.20}$$

Hence, using (5.19), (5.20), and noticing that $t, \theta_i \in [0,1]$, we find the following upper bound for all $t \in [0,1]$:

$$|K^2(t,t_i;\lambda) - K^2(t,\xi_i;\lambda)| \leq \frac{3}{2\lambda\sqrt{\lambda}} \frac{1}{n} = O\left(\frac{1}{n\lambda^{3/2}}\right), \quad n \to \infty.$$

This gives the approximation

$$\frac{1}{n}\sum_{i=1}^{n} K^2(t,t_i;\lambda) = \int_0^1 K^2(t,s;\lambda)ds + O\left(\frac{1}{n\lambda^{3/2}}\right), \quad n \to \infty. \tag{5.21}$$

On the other hand, we have that $\max_t \max_i |K(t,t_i;\lambda)| = O\left(\frac{1}{\sqrt{\lambda}}\right)$, which results in

$$O\left(\frac{1}{n\lambda}\right)\frac{1}{n}\sum_{i=1}^{n} K(t,t_i;\lambda) = O\left(\frac{1}{n\lambda}\right) O\left(\frac{1}{\sqrt{\lambda}}\right) = O\left(\frac{1}{n\lambda^{3/2}}\right), \quad n \to \infty. \tag{5.22}$$

Now, combining (5.21) and (5.22) into (5.17), we get for all large enough n

$$\begin{aligned}
&\mathrm{Var}(\hat{\mu}_\lambda(t)) \\
&= \frac{\sigma^2}{n}\left[\int_0^1 K^2(t,s;\lambda)ds + O\left(\frac{1}{n\lambda^{3/2}}\right) + O\left(\frac{1}{n\lambda^{3/2}}\right) + \frac{1}{n}O\left(\frac{1}{(n\lambda)^2}\right)\right] \\
&= \frac{\sigma^2}{n}\left[\int_0^1 K^2(t,s;\lambda)ds + O\left(\frac{1}{n\lambda^{3/2}}\right)\right]. \qquad (5.23)
\end{aligned}$$

Next, we have

$$\begin{aligned}
\int_0^1 K^2(t,s;\lambda)ds &= \frac{1}{4\lambda}\int_0^1 \Big(e^{\frac{2(s+t-2)}{\sqrt{\lambda}}} + e^{\frac{-2(s+t)}{\sqrt{\lambda}}} + e^{\frac{-2|s-t|}{\sqrt{\lambda}}} \\
&\qquad +2e^{\frac{-2}{\sqrt{\lambda}}} + 2e^{\frac{(s+t-2)-|s-t|}{\sqrt{\lambda}}} + 2e^{\frac{-(s+t)-|s-t|}{\sqrt{\lambda}}}\Big)\,ds \\
&= \frac{1}{4\lambda}\int_0^1 \Big(e^{\frac{2(s+t-2)}{\sqrt{\lambda}}} + e^{\frac{-2(s+t)}{\sqrt{\lambda}}} + 2e^{\frac{-2}{\sqrt{\lambda}}}\Big)\,ds \\
&\qquad + \frac{1}{4\lambda}\int_0^t \Big(e^{\frac{-2(t-s)}{\sqrt{\lambda}}} + 2e^{\frac{2s-2}{\sqrt{\lambda}}} + 2e^{\frac{-2t}{\sqrt{\lambda}}}\Big)\,ds \\
&\qquad + \frac{1}{4\lambda}\int_t^1 \Big(e^{\frac{-2(s-t)}{\sqrt{\lambda}}} + 2e^{\frac{2t-2}{\sqrt{\lambda}}} + 2e^{\frac{-2s}{\sqrt{\lambda}}}\Big)\,ds \\
&= \frac{1}{4\lambda}\left[\sqrt{\lambda} + e^{\frac{2(t-1)}{\sqrt{\lambda}}}(\sqrt{\lambda} - 2t + 2) + e^{\frac{-2t}{\sqrt{\lambda}}}(\sqrt{\lambda} + 2t)\right] \\
&\qquad + \frac{1}{4\lambda}\left[e^{\frac{-2}{\sqrt{\lambda}}}(2 - 2\sqrt{\lambda}) - \frac{1}{2}\sqrt{\lambda}e^{\frac{2t-4}{\sqrt{\lambda}}} - \frac{1}{2}\sqrt{\lambda}e^{\frac{-2t-2}{\sqrt{\lambda}}}\right] \\
&= \frac{1}{4\sqrt{\lambda}}\left[1 + e^{\frac{2(t-1)}{\sqrt{\lambda}}}\left(1 - \frac{2(t-1)}{\sqrt{\lambda}}\right) + e^{\frac{-2t}{\sqrt{\lambda}}}\left(1 + \frac{2t}{\sqrt{\lambda}}\right)\right] \\
&\qquad + \frac{1}{4\sqrt{\lambda}}\left[e^{\frac{-2}{\sqrt{\lambda}}}\left(\frac{2}{\sqrt{\lambda}} - 2\right) - \frac{1}{2}e^{\frac{2t-4}{\sqrt{\lambda}}} - \frac{1}{2}e^{\frac{-2t-2}{\sqrt{\lambda}}}\right],
\end{aligned}$$

where for the latter term for all $t \in [0,1]$

$$\lim_{\lambda\to 0} e^{\frac{-2}{\sqrt{\lambda}}}\left(\frac{2}{\sqrt{\lambda}} - 2\right) - \frac{1}{2}e^{\frac{2t-4}{\sqrt{\lambda}}} - \frac{1}{2}e^{\frac{-2t-2}{\sqrt{\lambda}}} = 0.$$

Therefore, it follows from (5.23) that for all $t \in [0,1]$ as $n \to \infty$

$$\begin{aligned}
&\mathrm{Var}(\hat{\mu}_\lambda(t)) \\
&= \frac{\sigma^2}{4n\sqrt{\lambda}} \left[1 + e^{\frac{2(t-1)}{\sqrt{\lambda}}} \left(1 - \frac{2(t-1)}{\sqrt{\lambda}} \right) + e^{\frac{-2t}{\sqrt{\lambda}}} \left(1 + \frac{2t}{\sqrt{\lambda}} \right) + o(1) + O\left(\frac{1}{n\lambda} \right) \right] \\
&= \frac{\sigma^2}{4n\sqrt{\lambda}} \left[1 + e^{\frac{2(t-1)}{\sqrt{\lambda}}} \left(1 - \frac{2(t-1)}{\sqrt{\lambda}} \right) + e^{\frac{-2t}{\sqrt{\lambda}}} \left(1 + \frac{2t}{\sqrt{\lambda}} \right) + o(1) \right]. \qquad (5.24)
\end{aligned}$$

Before we approximate the risk of $\hat{\mu}_\lambda(t)$, let us derive the point-wise approximation for the bias term. Suppose μ'' do not change very rapidly, and satisfies Lipshitz condition of order 2η, that is,

$$|\mu''(t_1) - \mu''(t_2)| \leq M|t_1 - t_2|^{2\eta}. \qquad (5.25)$$

By Taylor's formula, we can express $\mu(s)$ in (5.16) as

$$\mu(s) = \mu(t) + \mu'(t)(s-t) + \frac{\mu''(t)}{2}(s-t)^2 + R_2(s),$$

where $R_2(s) = \int_t^s \frac{\mu'''(x)}{2!}(s-x)^2 dx$. Note that for a small enough a, we have that

$$|\mu'''(x)| = \left| \frac{\mu''(x + a(s-x)) - \mu''(x)}{a(s-x)} \right| \leq M|s-x|^{2\eta-1},$$

and hence,

$$|R_2(s)| \leq \int_t^s \frac{M}{2}|s-x|^{2\eta+1} dx = \frac{M}{2(2\eta+2)}(s-t)^{2\eta+2}.$$

Therefore, we can rewrite (5.16) as follows:

$$\begin{aligned}
\boldsymbol{E}(\hat{\mu}_\lambda(t)) = &\int_0^1 \mu(t) K(t,s;\lambda) ds + \int_0^1 \mu'(t)(s-t) K(t,s;\lambda) ds \\
&+ \int_0^1 \frac{\mu''(t)}{2}(s-t)^2 K(t,s;\lambda) ds \\
&+ \int_0^1 R_2(s) K(t,s;\lambda) ds + O\left(\frac{1}{n\lambda} \right). \qquad (5.26)
\end{aligned}$$

We can compute the integrals in (5.26) by substituting $s = t - \sqrt{\lambda}u$, which gives

$$\begin{aligned}
&\int_0^1 K(t,s;\lambda)ds = -\int_{\frac{t}{\sqrt{\lambda}}}^{\frac{t-1}{\sqrt{\lambda}}} \frac{1}{2\sqrt{\lambda}} \left(e^{\frac{2t-2-\sqrt{\lambda}u}{\sqrt{\lambda}}} + e^{\frac{-2t+\sqrt{\lambda}u}{\sqrt{\lambda}}} + e^{-|u|} \right) \sqrt{\lambda} du \\
&= \frac{1}{2} \left[\int_{\frac{t-1}{\sqrt{\lambda}}}^{\frac{t}{\sqrt{\lambda}}} e^{\frac{2t-2-\sqrt{\lambda}u}{\sqrt{\lambda}}} + e^{\frac{-2t+\sqrt{\lambda}u}{\sqrt{\lambda}}} du + \int_{\frac{t-1}{\sqrt{\lambda}}}^{0} e^u du + \int_0^{\frac{t}{\sqrt{\lambda}}} e^{-u} du \right] \\
&= \frac{1}{2} \left[-e^{\frac{-2+t}{\sqrt{\lambda}}} - e^{\frac{-t-1}{\sqrt{\lambda}}} + 2 \right].
\end{aligned}$$

Observe that for $t \in [0,1]$, $e^{\frac{-2+t}{\sqrt{\lambda}}} \leq e^{\frac{-1}{\sqrt{\lambda}}}$ and $e^{\frac{-t-1}{\sqrt{\lambda}}} \leq e^{\frac{-1}{\sqrt{\lambda}}}$. Hence, we have that as $n \to \infty$

$$\int_0^1 K(t,s;\lambda)ds = 1 + O(e^{\frac{-1}{\sqrt{\lambda}}}), \quad t,s \in [0,1].$$

For the following integrals, using the similar arguments and omitting lengthy computation, we obtain as $n \to \infty$

$$\begin{aligned}
&\int_0^1 (s-t)K(t,s;\lambda)ds = -\lambda \int_{\frac{t-1}{\sqrt{\lambda}}}^{\frac{t}{\sqrt{\lambda}}} uK(t,t-\sqrt{\lambda}u;\lambda)du \\
&= -\lambda \left[\frac{1}{\sqrt{\lambda}} \left(e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{-t}{\sqrt{\lambda}}} + O(e^{\frac{-1}{\sqrt{\lambda}}}) \right) \right] = -\sqrt{\lambda}(e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{-t}{\sqrt{\lambda}}}) + O(e^{\frac{-1}{\sqrt{\lambda}}}),
\end{aligned}$$

and

$$\begin{aligned}
&\int_0^1 \frac{(s-t)^2}{2} K(t,s;\lambda)ds = \lambda^{\frac{3}{2}} \int_{\frac{t-1}{\sqrt{\lambda}}}^{\frac{t}{\sqrt{\lambda}}} u^2 K(t,t-\sqrt{\lambda}u;\lambda)du \\
&= \lambda^{\frac{3}{2}} \left[\frac{1}{\sqrt{\lambda}} \left(1 + \frac{t-1}{\sqrt{\lambda}} e^{\frac{t-1}{\sqrt{\lambda}}} + \frac{t}{\sqrt{\lambda}} e^{\frac{-t}{\sqrt{\lambda}}} + O(e^{\frac{-1}{\sqrt{\lambda}}}) \right) \right] \\
&= \lambda \left(1 + \frac{t-1}{\sqrt{\lambda}} e^{\frac{t-1}{\sqrt{\lambda}}} + \frac{t}{\sqrt{\lambda}} e^{\frac{-t}{\sqrt{\lambda}}} \right) + O(e^{\frac{-1}{\sqrt{\lambda}}}),
\end{aligned}$$

and also, under condition (5.25),

$$\int_0^1 R_2(s)K(t,s;\lambda)ds$$
$$= O\left(\lambda^{\eta+1}\right)\int_{\frac{t-1}{\sqrt{\lambda}}}^{\frac{t}{\sqrt{\lambda}}} u^{2\eta+2}K(t,t-\sqrt{\lambda}u;\lambda)\sqrt{\lambda}du = O\left(\lambda^{\eta+1}\right).$$

Therefore, from (5.26), we have that as $n \to \infty$

$$\begin{aligned}\boldsymbol{E}(\hat{\mu}_\lambda(t)) = \mu(t) - \sqrt{\lambda}\mu'(t)\left(e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{-t}{\sqrt{\lambda}}}\right)\\ + \lambda\mu''(t)\left(1 + \frac{t-1}{\sqrt{\lambda}}e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}}e^{\frac{-t}{\sqrt{\lambda}}}\right) + O\left(\frac{1}{n\lambda} + \lambda^{\eta+1}\right),\end{aligned} \tag{5.27}$$

uniformly in $t \in [0,1]$. Suppose that $n\lambda^{\frac{3}{2}} \to \infty$, that is, $\lambda = n^{-a}$ with $a \in \left(0, \frac{2}{3}\right)$ and consider the following expression for the bias of the estimator $\hat{\mu}_\lambda(t)$: $\text{Bias}\hat{\mu}_\lambda(t) = \boldsymbol{E}(\hat{\mu}_\lambda(t)) - \mu(t)$, which can be used for the pointwise squared bias, denoted by $(\text{Bias}\hat{\mu}_\lambda(t_i))^2, i = 1, \ldots, n$. Now, approximating the overall squared bias by the integral, we have from (5.27) that as $n \to \infty$

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n(\text{Bias}\hat{\mu}_\lambda(t_i))^2 = \int_0^1(\text{Bias}\hat{\mu}_\lambda(t))^2dt + O\left(\frac{1}{n}\right)\\
&= \lambda\int_0^1(\mu'(t))^2\left(e^{\frac{2(t-1)}{\sqrt{\lambda}}} + e^{\frac{-2t}{\sqrt{\lambda}}}\right)dt\\
&\quad + \lambda^2\int_0^1(\mu''(t))^2\left(1 + \frac{t-1}{\sqrt{\lambda}}e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}}e^{\frac{-t}{\sqrt{\lambda}}}\right)^2dt\\
&\quad - 2\lambda^{\frac{3}{2}}\int_0^1\mu'(t)\mu''(t)\left(e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{-t}{\sqrt{\lambda}}}\right)\left(1 + \frac{t-1}{\sqrt{\lambda}}e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}}e^{\frac{-t}{\sqrt{\lambda}}}\right)dt\\
&\quad + O\left(\frac{1}{n\lambda} + \lambda^{\eta+1}\right)\sqrt{\lambda}\int_0^1\mu'(t)\left(e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{-t}{\sqrt{\lambda}}}\right)dt\\
&\quad + O\left(\frac{1}{n\lambda} + \lambda^{\eta+1}\right)\lambda\int_0^1\mu''(t)\left(1 + \frac{t-1}{\sqrt{\lambda}}e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}}e^{\frac{-t}{\sqrt{\lambda}}}\right)dt\\
&\quad + O\left[\left(\frac{1}{n\lambda} + \lambda^{\eta+1}\right)^2\right] + O\left(\frac{1}{n}\right).\end{aligned} \tag{5.28}$$

From (5.28), let $y = \frac{t-1}{\sqrt{\lambda}}$ and using integration by parts, we have as $n\lambda^{\frac{3}{2}} \to \infty$ with one of $\mu'(0)$ or $\mu'(1)$ not being zero

$$\begin{aligned}
&\lambda \int_0^1 (\mu'(t))^2 \left(e^{\frac{2(t-1)}{\sqrt{\lambda}}} + e^{\frac{-2t}{\sqrt{\lambda}}}\right) dt \\
&= \lambda^{\frac{3}{2}} \int_{-\frac{1}{\sqrt{\lambda}}}^0 (\mu'(\sqrt{\lambda}y+1))^2 \left(e^{2y} + e^{-2y-\frac{2}{\sqrt{\lambda}}}\right) dy \\
&= \lambda^{\frac{3}{2}} (\mu'(\sqrt{\lambda}y+1))^2 \left(\frac{1}{2}e^{2y} - \frac{1}{2}e^{-2y-\frac{2}{\sqrt{\lambda}}}\right)\Bigg|_{-\frac{1}{\sqrt{\lambda}}}^0 \\
&\quad - \lambda^{\frac{3}{2}} \int_{-\frac{1}{\sqrt{\lambda}}}^0 2(\mu'(\sqrt{\lambda}y+1))(\mu''(\sqrt{\lambda}y+1))\sqrt{\lambda} \left(\frac{1}{2}e^{2y} - \frac{1}{2}e^{-2y-\frac{2}{\sqrt{\lambda}}}\right) dy \\
&= \frac{\lambda^{\frac{3}{2}}}{2}[(\mu'(1))^2 + \mu'(0))^2] + o(\lambda^{\frac{3}{2}}).
\end{aligned}$$

Similarly, using the change of variable and integration by parts, we obtain

$$\lambda^2 \int_0^1 (\mu''(t))^2 \left(1 + \frac{t-1}{\sqrt{\lambda}} e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{-t}{\sqrt{\lambda}} e^{\frac{-t}{\sqrt{\lambda}}}\right)^2 dt = O(\lambda^2),$$

and

$$\lambda^{\frac{3}{2}} \int_0^1 \mu'(t)\mu''(t) \left(e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{-t}{\sqrt{\lambda}}}\right) \left(1 + \frac{t-1}{\sqrt{\lambda}} e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}} e^{\frac{-t}{\sqrt{\lambda}}}\right) dt = O(\lambda^2).$$

We can therefore express (5.28) as follows:

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n (\text{Bias}\hat{\mu}_\lambda(t_i))^2 &= \frac{\lambda^{\frac{3}{2}}}{2}[(\mu'(1))^2 + \mu'(0))^2] + o(\lambda^{\frac{3}{2}}) + O(\lambda^2) \\
&\quad + O(\lambda^2) + O\left(\left(\frac{1}{n\lambda} + \lambda^{\eta+1}\right)\lambda\right) + O\left(\left(\frac{1}{n\lambda} + \lambda^{\eta+1}\right)\lambda\right) \\
&\quad + O\left[\left(\frac{1}{n\lambda} + \lambda^{\eta+1}\right)^2\right] + O\left(\frac{1}{n}\right). \qquad (5.29)
\end{aligned}$$

Since, by assumption, $\lambda = n^{-a}, a \in \left(0, \frac{2}{3}\right)$, it follows that $n^{-1} = o(\lambda^{\frac{3}{2}})$. It is also easy to see that all the "big oh" terms on the right side of (5.29) are

much smaller than $\lambda^{\frac{3}{2}}$ when $n\lambda^{\frac{3}{2}} \to \infty$. Therefore, we obtain from (5.29) that when $n\lambda^{\frac{3}{2}} \to \infty$

$$\frac{1}{n}\sum_{i=1}^{n}(\text{Bias}\hat{\mu}_\lambda(t_i))^2 = \frac{\lambda^{\frac{3}{2}}}{2}[\mu'(0)^2 + \mu'(1)^2] + o(\lambda^{\frac{3}{2}}). \tag{5.30}$$

In the case that $\mu'(0)^2 = \mu'(1)^2 = 0$ and $n\lambda^2 \to \infty$, we have from (5.28)

$$\lambda\int_0^1 (\mu'(t))^2\left(e^{\frac{2(t-1)}{\sqrt{\lambda}}} + e^{\frac{-2t}{\sqrt{\lambda}}}\right)dt = o(\lambda^2),$$

$$\lambda^2\int_0^1 (\mu''(t))^2\left(1 + \frac{t-1}{\sqrt{\lambda}}e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}}e^{\frac{-t}{\sqrt{\lambda}}}\right)^2 dt = \lambda^2\int_0^1 \mu''(t)^2 dt + O(\lambda^{\frac{5}{2}}),$$

$$\lambda^{\frac{3}{2}}\int_0^1 \mu'(t)\mu''(t)\left(e^{\frac{t-1}{\sqrt{\lambda}}} - e^{\frac{-t}{\sqrt{\lambda}}}\right)\left(1 + \frac{t-1}{\sqrt{\lambda}}e^{\frac{t-1}{\sqrt{\lambda}}} - \frac{t}{\sqrt{\lambda}}e^{\frac{-t}{\sqrt{\lambda}}}\right)dt = O(\lambda^{\frac{5}{2}}).$$

Hence, relation (5.28) can be written as follows:

$$\frac{1}{n}\sum_{i=1}^{n}(\text{Bias}\hat{\mu}_\lambda(t_i))^2 = \lambda^2\int_0^1 (\mu''(t))^2 dt + o(\lambda^2). \tag{5.31}$$

The final step is to approximate the overall variance of $\hat{\mu}_\lambda(t)$, using the averaged point-wise variance. Using (5.24), we have as $n \to \infty$

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\text{Var}(\hat{\mu}_\lambda(t_i)) = \int_0^1 \text{Var}(\hat{\mu}_\lambda(t))dt + O\left(\frac{1}{n}\right) \\
&= \frac{\sigma^2}{4n\sqrt{\lambda}}\left\{1 + \int_0^1\left[e^{\frac{2(t-1)}{\sqrt{\lambda}}}\left(1 - \frac{2(t-1)}{\sqrt{\lambda}}\right) + e^{\frac{-2t}{\sqrt{\lambda}}}\left(1 + \frac{2t}{\sqrt{\lambda}}\right)\right]dt + o(1)\right\} \\
&\quad + O\left(\frac{1}{n}\right) \\
&= \frac{\sigma^2}{4n\sqrt{\lambda}}(1 + 2\sqrt{\lambda} - 2\sqrt{\lambda}e^{\frac{-2}{\sqrt{\lambda}}} - 2e^{\frac{-2}{\sqrt{\lambda}}} + o(1)) + O\left(\frac{1}{n}\right),
\end{aligned}$$

where the terms with $e^{\frac{-2}{\sqrt{\lambda}}}$ are negligible, and $n^{-1} = o((n\sqrt{\lambda})^{-1})$. This gives for all large enough n

$$\frac{1}{n}\sum_{i=1}^{n}\text{Var}(\hat{\mu}_\lambda(t_i)) = \frac{\sigma^2}{4n\sqrt{\lambda}}(1 + O(\sqrt{\lambda}) + o(1)) = \frac{\sigma^2}{4n\sqrt{\lambda}}(1 + o(1)). \tag{5.32}$$

Noting that $R_n(\lambda) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{Var}\,(\hat{\mu}_\lambda(t_i)) + \frac{1}{n}\sum_{i=1}^{n}(\mathrm{Bias}\hat{\mu}_\lambda(t_i))^2$ and combining relation (5.30), (5.31), and (5.32), we obtain the asymptotic expression for the risk of the linear smoothing spline estimator $\hat{\mu}_\lambda$. Namely, if μ' is bounded with $\mu'(0)$ or $\mu'(1)$ not equal to zero, we have

$$R_n(\lambda) \approx \frac{\lambda^{\frac{3}{2}}}{2}[(\mu'(0))^2 + (\mu'(1))^2] + \frac{\sigma^2}{4n\sqrt{\lambda}}, \quad n\lambda^{\frac{3}{2}} \to \infty, \tag{5.33}$$

and the respective asymptotically optimal smoothing parameter, which minimizes the right-hand side of (5.33), is

$$\lambda_{\mathrm{opt}} = n^{-\frac{1}{2}} \left\{ \frac{\sigma^2}{6[(\mu'(0))^2 + (\mu'(1))^2]} \right\}^2.$$

We can see that λ_{opt} decays at the rate of $n^{-\frac{1}{2}}$, which will produce a $n^{-\frac{3}{4}}$ rate of decay for the risk of $\hat{\mu}_\lambda$. Next, if μ'' is bounded with $\mu'(0) = \mu'(1) = 0$, we have

$$R_n(\lambda) \approx \lambda^2 \int_0^1 (\mu''(t))^2 dt + \frac{\sigma^2}{4n\sqrt{\lambda}}, \quad n\lambda^2 \to \infty.$$

The respective optimal smoothing parameter in this case is

$$\lambda_{\mathrm{opt}} = n^{-\frac{2}{5}} \left\{ \frac{\sigma^2}{16\int_0^1 (\mu''(t))^2 dt} \right\}^{\frac{2}{5}}.$$

This optimal smoothing parameter decays at the rate of $n^{-\frac{2}{5}}$ giving the risk decaying at the rate of $n^{-\frac{4}{5}}$. In general, smoothing spline with penalty function $\int_0^1 (\mu^{(m)}(t))^2 dt$ is able to attain the $O\left(n^{\frac{-2m}{2m+1}}\right)$ optimal rate of decay for risk when regression function $\mu \in W_2^m[0,1]$.

Exercise 5.9.12 from [**2**]. Under model (1.1), show that for a general m, we have

$$\frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{E}\hat{\mu}_\lambda(t_i) - \mu(t_i))^2 \le \lambda \int_0^1 (\mu^{(m)}(t))^2 dt.$$

Solution: We begin by noting that (see formula (2.8) in Section 2.1)

$$\boldsymbol{E}(\widehat{\boldsymbol{b}}_\lambda) = (\boldsymbol{X}^\top \boldsymbol{X} + n\lambda \boldsymbol{\Omega})^{-1} \boldsymbol{X}^\top \boldsymbol{\mu}$$

is a minimizer of the following w.r.t $\boldsymbol{d} = (d_1, \ldots, d_n)^\top \in \mathbb{R}^n$:

$$\frac{1}{n} \sum_{i=1}^{n} \left(\mu(t_i) - \sum_{j=1}^{n} d_j x_j(t_i) \right)^2 + \lambda \int_0^1 \left(\sum_{j=1}^{n} d_j x_j^{(m)}(t) \right)^2 dt.$$

In the view of the arguments in Section 2.1 just above relation, this implies that

$$\boldsymbol{E}(\mu_\lambda(t_i)) = \sum_{j=1}^{n} \boldsymbol{E}(\hat{b}_{\lambda j}) x_j(t)$$

is a minimizer of the following w.r.t $g \in W_2^m[0,1]$:

$$\frac{1}{n} \sum_{i=1}^{n} (\mu(t_i) - g(t_i))^2 + \lambda \int_0^1 (g^{(m)}(t))^2 dt.$$

Hence, we have that

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{E}\hat{\mu}_\lambda(t_i) - \mu(t_i))^2 &\leq \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{E}\hat{\mu}_\lambda(t_i) - \mu(t_i))^2 + \lambda \int_0^1 (\boldsymbol{E}\hat{\mu}_\lambda^{(m)}(t))^2 dt \\ &\leq \frac{1}{n} \sum_{i=1}^{n} (g(t_i) - \mu(t_i))^2 + \lambda \int_0^1 (g^{(m)}(t))^2 dt \end{aligned}$$

for all $g \in W_2^m[0,1]$. Then taking $g = \mu$ leads to the desired result.

6 Conclusion

In this project, we considered some topics related to smoothing splines in nonparametric regression analysis and solved selected exercises from Chapter 5 of [**2**]. We started with the discussion of what is a spline. Then, we demonstrated that spline can be used to conduct nonparametric regression: adding a measure of smoothness to the conventional residual sum of squares method gives us the penalized least squares (PLS) criterion. We also showed that the minimizer to the PLS, the smoothing spline, is a natural spline of order $2m$ with knots $t_1, \ldots, t_n$, which can be constructed using basis functions. We gave an example of the basis in the case of $m = 1$, and through the example, we saw the impact of the smoothing parameter selection. We also discussed the prediction risk and the cross-validation criteria that provide ways to determine a reasonable smoothing parameter λ.

From a computational perspective, estimation by means of a smoothing spline and determination of smoothing parameter rely on finding the smoothing matrix. One efficient method is to make the design matrix $\boldsymbol{X}$ band-limited: by applying the Cholesky decomposition, we are able to compute the fitted values in $O(n)$ calculations. Finally, we discussed the large-sample properties of smoothing splines. With the in-depth review of the case $m = 1$, we saw that, generally, the estimation risk for smoothing splines is capable of attaining the $O(n^{-\frac{2m}{2m+1}})$ optimal rate of decay. Thus, the smoothing spline is a very good estimator for the regression curve from the Sobolev space $W_2^m[0, 1]$.

7 Appendix

1. **Proof of Lemma 1.**

Proof. In the proof, we shall use the following identities from [**5**]:

$$\cos x \cos y = \frac{1}{2}[\cos(x-y) + \cos(x+y)], \tag{7.1}$$

$$\cos(x+y) = \cos x \cos y - \sin x \sin y, \tag{7.2}$$

$$\cos^2 x = \frac{1+\cos(2x)}{2}, \tag{7.3}$$

$$\sum_{k=1}^{n} \cos((2k-1)x) = \frac{1}{2}\frac{\sin(2nx)}{\sin(x)}, \quad \text{if} \quad \sin x \neq 0. \tag{7.4}$$

Recall that $t_i = \frac{2i-1}{2n}, i = 1, \ldots, n$. Using (7.1), we have that for $j = 1, \ldots, n-1$,

$$\begin{aligned} d_{jr} &= \frac{2}{n}\sum_{i=1}^{n} \cos(j\pi t_i)\cos(r\pi t_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \cos\left((2i-1)\frac{(r-j)\pi}{2n}\right) + \frac{1}{n}\sum_{i=1}^{n} \cos\left((2i-1)\frac{(r+j)\pi}{2n}\right). \end{aligned}$$

Taking x in (7.4) to be $\frac{(r-j)\pi}{2n}$ and $\frac{(r+j)\pi}{2n}$, we obtain

$$d_{jr} = \frac{1}{2n}\left[\frac{\sin((r-j)\pi)}{\sin\left(\frac{(r-j)\pi}{2n}\right)} + \frac{\sin((r+j)\pi)}{\sin\left(\frac{(r+j)\pi}{2n}\right)}\right] = 0$$

provided $\frac{r-j}{2n} \neq k$ and $\frac{r+j}{2n} \neq k$ for any $k = 1, 2, \ldots$. For the remaining cases, suppose $r - j = 2kn$ and odd k. Then,

$$\begin{aligned} d_{jr} &= \frac{2}{n}\sum_{i=1}^{n} \cos(j\pi t_i)\cos((j+2kn)\pi t_i) \\ &= \frac{2}{n}\sum_{i=1}^{n} \cos(j\pi t_i)\cos((j\pi t_i + k\pi(2i-1)), \end{aligned} \tag{7.5}$$

apply relation (7.2) to equation (7.5), we have that

$$\cos((j\pi t_i + k\pi(2i-1)) = -\cos(j\pi t_i),$$

and using relations (7.3) and (7.4), we obtain

$$\begin{aligned} d_{jr} &= -\frac{2}{n}\sum_{i=1}^{n}\cos^2(j\pi t_i) = -\frac{1}{n}\sum_{i=1}^{n}(1+\cos(2j\pi t_i)) \\ &= -\frac{1}{n}\left(n + \frac{1}{2}\frac{\sin(2j\pi)}{\sin\left(\frac{2j\pi}{2n}\right)}\right) = -1. \end{aligned}$$

Noting that $\cos(-x) = \cos(x)$, we obtain analogously that for $r + j = 2kn$ and odd k

$$\begin{aligned} d_{jr} &= \frac{2}{n}\sum_{i=1}^{n}\cos(j\pi t_i)\cos((-j+2kn)\pi t_i) \\ &= \frac{2}{n}\sum_{i=1}^{n}\cos(j\pi t_i)\cos((-j\pi t_i + k\pi(2i-1)) \\ &= -\frac{2}{n}\sum_{i=1}^{n}\cos^2(j\pi t_i) = -1. \end{aligned}$$

Now, for $r - j = 2kn$ or $r + j = 2kn$ and even k, we have that

$$\cos((j\pi t_i + k\pi(2i-1)) = \cos(j\pi t_i) = \cos((-j\pi t_i + k\pi(2i-1))$$

resulting in

$$d_{jr} = \frac{2}{n}\sum_{i=1}^{n}\cos(j\pi t_i)\cos((j+2kn)\pi t_i) = \frac{2}{n}\sum_{i=1}^{n}\cos^2(j\pi t_i) = 1,$$

and

$$d_{jr} = \frac{2}{n}\sum_{i=1}^{n}\cos(j\pi t_i)\cos((-j+2kn)\pi t_i) = \frac{2}{n}\sum_{i=1}^{n}\cos^2(j\pi t_i) = 1.$$

Finally, for the case of $r = j$, using (7.3) and (7.4) we have that

$$d_{jj} = \frac{2}{n}\sum_{i=1}^{n}\cos^2(j\pi t_i) = 1.$$

□

2. **Proof of Lemma 2.**

Proof. In addition to (7.1)–(7.4), we shall also use the following identities from [**5**]:

$$\sin x \sin y = \frac{1}{2}[\cos(x-y) - \cos(x+y)], \tag{7.6}$$

$$\sin(x+y) = \sin x \cos y + \cos x \sin y, \tag{7.7}$$

$$\sin^2(x) = \frac{1}{2}[1 - \cos(2x)] \tag{7.8}$$

$$\cos x - \cos y = -2\sin\left(\frac{x+y}{2}\right)\sin\left(\frac{x-y}{2}\right), \tag{7.9}$$

$$\sum_{k=0}^{n}\cos(kx) = \frac{\sin\left(\frac{nx}{2}\right)}{\sin\left(\frac{x}{2}\right)}\cos\left(\frac{(n+1)x}{2}\right) + 1, \quad \text{if} \quad \sin\left(\frac{x}{2}\right) \neq 0. \tag{7.10}$$

Recall that $t_i = \frac{i}{n}, i = 1, \ldots, n$. Applying (7.6) we obtain that for $j = 1, \ldots, n-1$,

$$\begin{aligned} d_{jr} &= \frac{2}{n}\sum_{i=1}^{n}\sin(j\pi t_i)\sin(r\pi t_i) \\ &= \frac{1}{n}\sum_{i=1}^{n}\cos\left(i\frac{(r-j)\pi}{n}\right) - \frac{1}{n}\sum_{i=1}^{n}\cos\left(i\frac{(r+j)\pi}{n}\right). \end{aligned}$$

Rearranging (7.10) we obtain

$$\sum_{k=1}^{n}\cos(kx) = \frac{\sin\left(\frac{nx}{2}\right)}{\sin\left(\frac{x}{2}\right)}\cos\left(\frac{(n+1)x}{2}\right), \quad \text{if} \quad \sin\left(\frac{x}{2}\right) \neq 0. \tag{7.11}$$

Taking x in (7.11) to be $\frac{(r-j)\pi}{n}$ and $\frac{(r+j)\pi}{n}$, we obtain

$$d_{jr} = \frac{1}{n}\frac{\sin\left(\frac{(r-j)\pi}{2}\right)}{\sin\left(\frac{(r-j)\pi}{2n}\right)}\cos\left(\frac{(n+1)(r-j)\pi}{2n}\right) - \frac{1}{n}\frac{\sin\left(\frac{(r+j)\pi}{2}\right)}{\sin\left(\frac{(r+j)\pi}{2n}\right)}\cos\left(\frac{(n+1)(r+j)\pi}{2n}\right),$$

and using (7.2), we have that

$$d_{jr} = \frac{1}{n}\left[\frac{\sin\left(\frac{(r-j)\pi}{2}\right)}{\sin\left(\frac{(r-j)\pi}{2n}\right)}\cos\left(\frac{(r-j)\pi}{2}\right)\cos\left(\frac{(r-j)\pi}{2n}\right) - \sin^2\left(\frac{(r-j)\pi}{2}\right)\right] - \frac{1}{n}\left[\frac{\sin\left(\frac{(r+j)\pi}{2}\right)}{\sin\left(\frac{(r+j)\pi}{2n}\right)}\cos\left(\frac{(r+j)\pi}{2}\right)\cos\left(\frac{(r+j)\pi}{2n}\right) - \sin^2\left(\frac{(r+j)\pi}{2}\right)\right].$$

Then, applying (7.9) and (7.11), we have that

$$\begin{aligned} d_{jr} &= -\frac{1}{n}(1-\cos((r-j)\pi)) + \frac{1}{n}(1-\cos(r+j)\pi) \\ &= \frac{1}{n}\left(-2\sin(r\pi)\sin(-j\pi)\right) = 0, \end{aligned}$$

provided $\frac{r-j}{2n} \neq k$ and $\frac{r+j}{2n} \neq k$ for any $k = 1, 2, \ldots$. For the remaining cases, suppose $r - j = 2kn$. Then, applying (7.3), (7.4) and (7.7), we obtain

$$\begin{aligned} d_{jr} &= \frac{2}{n}\sum_{i=1}^{n}\sin(j\pi t_i)\sin((j+2kn)\pi t_i) \\ &= \frac{2}{n}\sum_{i=1}^{n}\sin(j\pi t_i)\sin(j\pi t_i + 2k\pi i) \\ &= \frac{2}{n}\sum_{i=1}^{n}\sin^2(j\pi t_i) = \frac{2}{n}\sum_{i=1}^{n}(1-\cos^2(j\pi t_i)) \\ &= 2 + \frac{2}{n}\sum_{i=1}^{n}\cos^2(j\pi t_i) = 1. \end{aligned}$$

Similarly, noting that $\sin(-x) = -\sin x$, for the case of $r + j = 2kn$ we have that

$$\begin{aligned} d_{jr} &= \frac{2}{n}\sum_{i=1}^{n} \sin(j\pi t_i)\sin((-j+2kn)\pi t_i) \\ &= \frac{2}{n}\sum_{i=1}^{n} \sin(j\pi t_i)\sin(-j\pi t_i + 2k\pi i) \\ &= -\frac{2}{n}\sum_{i=1}^{n} \sin^2(j\pi t_i) = -\frac{2}{n}\sum_{i=1}^{n}(1-\cos^2(j\pi t_i)) \\ &= -2 + \frac{2}{n}\sum_{i=1}^{n} \cos^2(j\pi t_i) = -1. \end{aligned}$$

Finally, for the case of $j = r$ we obtain

$$d_{jj} = \frac{2}{n}\sum_{i=1}^{n} \sin^2(j\pi t_i) = 1.$$

□

3. **Sherman-Morrison formula** (see [**9**]).

   Suppose $\boldsymbol{A}$ is a $n \times n$ invertible square matrix and $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^n$ are column vectors. Then $\boldsymbol{A} - \boldsymbol{u}\boldsymbol{v}^\top$ is invertible if and ony if $1 - \boldsymbol{v}^\top \boldsymbol{A}\boldsymbol{u} \neq 0$. In this case,

   $$(\boldsymbol{A} - \boldsymbol{u}\boldsymbol{v}^\top)^{-1} = \boldsymbol{A}^{-1} + \frac{\boldsymbol{A}^{-1}\boldsymbol{u}\boldsymbol{v}^\top\boldsymbol{A}^{-1}}{1 - \boldsymbol{v}^\top\boldsymbol{A}^{-1}\boldsymbol{u}}.$$

4. **Cholesky decomposition (LDLT decomposition)** (see page 84 in [**10**]).

   If $\boldsymbol{A}$ is a symmetric $n \times n$ matrix, then

   $$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^\top,$$

where

$\boldsymbol{L}$ is a $n \times n$ lower triangular matrix,

$\boldsymbol{D}$ is a $n \times n$ diagonal matrix.

For example, when $n = 3$ we have

$$\begin{aligned}\boldsymbol{A}_{3\times 3} &= \begin{bmatrix} 1 & 0 & 0 \\ L_{21} & 1 & 0 \\ L_{31} & L_{32} & 1 \end{bmatrix} \begin{bmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & D_3 \end{bmatrix} \begin{bmatrix} 1 & L_{21} & L_{31} \\ 0 & 1 & L_{32} \\ 0 & 0 & 1 \end{bmatrix} \\ &= \begin{bmatrix} D_1 & L_{21}D_1 & L_{31}D_1 \\ L_{21}D_1 & L_{21}^2 D_1 + D_2 & L_{31}L_{21}D_1 + L_{32}D_2 \\ L_{31}D_1 & L_{31}L_{21}D_1 + L_{32}D_2 & L_{31}^2 D_1 + L_{32}^2 D_2 + D_3 \end{bmatrix}.\end{aligned}$$

The following recursive relations apply to the elements of $\boldsymbol{D}$ and $\boldsymbol{L}$:

$$D_1 = A_{11}, \quad D_j = A_{jj} - \sum_{k=1}^{j-1} L_{jk}^2 D_k, \quad j = 2, \ldots, n,$$

$$L_{ij} = \frac{1}{D_j}\left(A_{ij} - \sum_{k=1}^{j-1} L_{ik} L_{jk} D_k\right) \text{ for } 1 \le j < i \le n.$$

8 References

[1] A. Demmler and C. Reinsch. Oscillation matrices with spline smoothing. *Numerische Mathematik*, 24: 375–382, 1975.

[2] R. Eubank. *Nonparametric Regression and Spline Smoothing*. Marcel Dekker, New York, 1999.

[3] Th. Gasser, L. Sroka, and C. Jennen-Steinmetz. Residual variance and residual pattern in nonlinear regression. *Biometrika*, 73: 624–633, 1986.

[4] P. J. Green and B. W. Sliverman. *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach*. Chapman and Hall, London, 1995.

[5] I. S. Gradshteyn and I. M. Ryzhik. *Table of Integrals, Series, and Products, Eighth Edition*. Academic Press, London, 2015.

[6] T. Hsing and R. Eubank. *Theoretical Foundations of Functional Data Analysis, with an Introduction to Linear Operators*. John Wiley and Sons, Hoboken, 2015.

[7] D. Pollock. *Handbook of Time Series Analysis, Signal Processing, and Dynamics*, Academic Press, London, 1999.

[8] J. Schoenberg. Spline functions and the problem of graduation. *Proceedings of the National Academy of Science USA*, 52: 947–950, 1964.

[9] J. Sherman and W. J. Morrison. Adjustment of an inverse matrix corresponding to changes in the elements of a given Column or a given Row

of the original matrix. *Annals of Mathematical Statistics*, 21: 124–127, 1950.

[10] D. Watkins. *Fundamentals of Matrix Computations*. Wiley, New York, 1991.

[11] *Amazon Machine Learning: Developer Guide*. Version Latest, Amazon Web Services, Seattle, 2021.

9 R Code

1. Figure 4 (Exercise 5.1 from [**2**]):

```
1
2  n=100
3  t=rep(NA,n)
4  i=1
5  for (i in 1:n) {
6    t[i] = (2*i-1)/(2*n)
7  }
8  I = ifelse(t>=0.7,1,0)
9  obs = 16*(t^2)*(1-t)^2 +
10   32*(t-0.7)^3*I+rnorm(n,0,sd = 0.05)
11 plot(t,obs)
12
13 x = matrix(NA,n,n)
14 x[,1] = 1
15 i=1
16 j=2
17 for (i in 1:n) {
18   for (j in 2:n) {
19     x[i,j]=sqrt(2)*cos((j-1)*pi*t[i])
20   }
21 }
22
23 gamma = rep(NA,n)
24 gamma[1] = 0
25 i=1
26 for (i in 1:n) {
27   gamma[i+1] = (2*n*sin(i*pi/(2*n)))^2
```

```
}

lambda = 0.00006
yhat = rep(0,n)
j=1
obs = as.matrix(obs)
for (j in 1:n) {
  yhat = yhat + as.vector((1)/
  (1+lambda*gamma[j])*x[,j]%*%t(x[,j])%*%obs)/n
}
lines(t,yhat,col="2")

lambda = 0.0006
yhat = rep(0,n)
j=1
obs = as.matrix(obs)
for (j in 1:n) {
  yhat = yhat + as.vector((1)/
  (1+lambda*gamma[j])*x[,j]%*%t(x[,j])%*%obs)/n
}
lines(t,yhat,col="3")

lambda = 0.006
yhat = rep(0,n)
j=1
obs = as.matrix(obs)
for (j in 1:n) {
  yhat = yhat + as.vector((1)/
  (1+lambda*gamma[j])*x[,j]%*%t(x[,j])%*%obs)/n
}
```

```
lines(t,yhat,col="4")

#Kernel weights
lambda = 0.00006
kernel = rep(0,n)
j=1
s=seq(0,1,1/(n-1))
temp=0
i=1
for (i in 1:n) {
  for (j in 1:n) {
    temp = temp + (cos(j*pi*s[i]))/
      (1+lambda*gamma[j])*(sqrt(2)*cos(j*pi*0.495)
      +sqrt(2)*(0.5-0.496)/(0.505-0.495)*
      (cos(j*pi*0.505)-cos(j*pi*0.495)))
  }
  kernel[i] = 1 + sqrt(2)*temp
  i=i+1
  temp=0
}
plot(s,kernel,type = "l",col="2")

lambda = 0.0006
kernel = rep(0,n)
j=1
s=seq(0,1,1/(n-1))
temp=0
i=1
for (i in 1:n) {
  for (j in 1:n) {
```

```
    temp = temp + (cos(j*pi*s[i]))/
      (1+lambda*gamma[j])*(sqrt(2)*cos(j*pi*0.495)
      +sqrt(2)*(0.5-0.496)/(0.505-0.495)*
      (cos(j*pi*0.505)-cos(j*pi*0.495)))
  }
  kernel[i] = 1 + sqrt(2)*temp
  i=i+1
  temp=0
}
lines(s,kernel,col="3")

lambda = 0.006
kernel = rep(0,n)
j=1
s=seq(0,1,1/(n-1))
temp=0
i=1
for (i in 1:n) {
  for (j in 1:n) {
    temp = temp + (cos(j*pi*s[i]))/
      (1+lambda*gamma[j])*(sqrt(2)*cos(j*pi*0.495)
      +sqrt(2)*(0.5-0.496)/(0.505-0.495)*
      (cos(j*pi*0.505)-cos(j*pi*0.495)))
  }
  kernel[i] = 1 + sqrt(2)*temp
  i=i+1
  temp=0
}
lines(s,kernel,col="4")
```