Chapter 8: Smoothing[a]

For a set $A \subseteq \mathbb{R}^p$ we denote by $\mathcal{C}^2(A)$ the set of functions $f : A \to \mathbb{R}$ which are twice continuously differentiable.

In this chapter we let $p = 1$, consider observations $\{(y_i^0, x_i^0)\}_{i=1}^n$ and assume the following non-parametric regression model

$$Y_i^0 = f(x_i^0) + \epsilon_i, \qquad i = 1, \dots n, \quad f \in \mathcal{C}^2(\mathbb{R}) \tag{8.1}$$

where, for all $i \neq l$, $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{E}[\epsilon_i \epsilon_l] = \sigma^2 \delta_{il}$.

Then, for a given $\lambda \in [0, \infty]$, we estimate f in (8.1) using

$$\hat{f}_\lambda \in \underset{f \in \mathcal{C}^2(\mathbb{R})}{\operatorname{argmin}} \sum_{i=1}^n \big(y_i^0 - f(x_i^0)\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x. \tag{8.2}$$

Remark that,

- for $\lambda = 0$ the function $\hat{f}_\lambda$ is any function in $\mathcal{C}^2(\mathbb{R})$ that interpolates the data.

- for $\lambda = \infty$ the function $\hat{f}_\lambda$ is the least squares line fit (i.e. we have $\hat{f}_\infty(x) = x\hat{\beta}$ for all x and with $\hat{\beta}$ the OLS estimator of β in the model $Y_i^0 = \beta x_i^0 + \epsilon_i$).

The function $\hat{f}_\lambda$ is therefore very wiggly for $\lambda = 0$ and very smooth for $\lambda = \infty$, and the hope is that as λ increases from 0 to ∞ the smoothness of $\hat{f}_\lambda$ 'gradually' evolves between these two extreme cases.

Surprisingly, and as we will see below, for $\lambda > 0$ the infinite dimensional optimization problem (8.2) admits an explicit, unique and finite dimensional solution.

[a]The main reference for this chapter is [13, Chapter 5].

Preliminaries: Spline functions

Definition 8.3 *Let $\xi_1 < \xi_2 < \cdots < \xi_K$ be $K \geq 2$ real numbers, called knots. Then, a function $B : [\xi_1, \xi_K] \to \mathbb{R}$ is called a spline of degree $M \in \mathbb{N}$ if*

1. *B is a polynomial of degree M on the interval (ξ_k, ξ_{k+1}), for all $k \in \{1, \dots K-1\}$,*
2. *$B \in \mathcal{C}^{M-1}\big((\xi_1, \xi_K)\big)$ if $M \geq 2$.*

Remark: If B is as Definition 8.3 then there exist polynomials $\{p_k\}_{k=1}^{K-1}$ of degree M such that

$$B(x) = \sum_{k=1}^{K-1} p_k(x) \mathbb{I}_{(\xi_k, \xi_{k+1})}(x), \quad \forall x \in (\xi_1, \xi_K). \tag{8.3}$$

For $M \geq 1$ and $K \geq 2$ we let $\mathcal{S}_{M,K}(\{\xi_k\}_{k=1}^K)$ denote the set of splines of degree M with knots $\{\xi_k\}_{k=1}^K$. The following proposition gives an important property of this set.

Proposition 8.1 *For $M \geq 2$ and $K \geq 3$ the set $\mathcal{S}_{M,K}(\{\xi_k\}_{k=1}^K)$ is a vector space of dimension $M + K - 1$.*

Proof: The fact that $\mathcal{S}_{M,K}(\{\xi_k\}_{k=1}^K)$ is a vector space is trivial. To compute the dimension of this space remark that, in (8.3), each polynomial p_k can be written as $p_k(x) = \sum_{m=0}^M a_m^{(k)} x^m$ for some real numbers $\{a_m^{(k)}\}_{m=0}^M$, and thus the function B has $(K-1)(M+1)$ 'parameters' $\{(a_0^{(k)}, \dots, a_M^{(k)})\}_{k=1}^{K-1}$. However, the condition $B \in \mathcal{C}^{M-1}((\xi_1, \xi_K)))$ implies that not all these parameters can be freely chosen. Indeed, these parameters must be such that the function B and its first $M-1$ derivatives are continuous at each point $x \in \{\xi_k\}_{k=2}^{K-1}$, which imposes $M(K-2)$ constraints of the parameters $\{(a_0^{(k)}, \dots, a_M^{(k)})\}_{k=1}^{K-1}$. Therefore, the set $\{(a_0^{(k)}, \dots, a_M^{(k)})\}_{k=1}^{K-1}$ contains only $(K-1)(M+1) - (K-2)M = K + M - 1$ free parameters. The proof is complete. $\square$

Preliminaries: Natural cubic splines

Definition 8.4 *A spline $B \in \mathcal{S}_{M,K}(\{\xi_k\}_{k=1}^K)$ of degree $M = 3$ is called a natural cubic spline if $B''(\xi_1) = B''(\xi_K) = 0$.*

Remark: The curvature of a natural cubic spline at the first and last knot is therefore zero, so that if we want to extrapolate the value of B outside the interval $[\xi_1, \xi_K]$ we would do it linearly.

For $K \geq 3$ we denote by $\mathcal{S}_K^*(\{\xi_k\}_{k=1}^K)$ the set of natural cubic splines having knots $\{\xi_k\}_{k=1}^K$. The following proposition gives an important property of this set.

Proposition 8.2 *For $K \geq 3$ the set $\mathcal{S}_K^*(\{\xi_k\}_{k=1}^K)$ is a vector space of dimension K.*

Proof: Natural cubic splines impose two additional constraints compared to a "regular" splines. The result then follows by Proposition 8.1. $\square$

The importance of natural cubic splines comes from the following key result.

Theorem 8.1 *Let $K \geq 3$, $\xi_1 < \cdots < \xi_K$ be K knots and let $z \in \mathbb{R}^K$. Then, there exists a unique natural cubic spline $B \in \mathcal{S}_K^*(\{\xi_k\}_{k=1}^K)$ such that $B(\xi_k) = z_k$ for all $k \in \{1, \ldots, K\}$. In addition, for every function $h \in \mathcal{C}^2\big([\xi_1, \xi_K]\big)$ such that $h \neq B$ and such that $h(\xi_k) = z_k$ for all $k \in \{1, \ldots, K\}$, we have*

$$\int_{[\xi_1,\xi_K]} (B''(x))^2 \mathrm{d}x < \int_{[\xi_1,\xi_K]} (h''(x))^2 \mathrm{d}x. \tag{8.4}$$

Remark: In (8.4) the inequality is strict.

Proof of Theorem 8.1

By Proposition 8.2 the set $\mathcal{S}_K^*(\{\xi_k\}_{k=1}^K)$ is a vector space of dimension K so that $\mathcal{S}_K^*(\{\xi_k\}_{k=1}^K) = \mathrm{span}\big(B_1^*, \dots, B_K^*\big)$ for some linearly independent natural cubic splines $\{B_k^*\}_{k=1}^K$ with knots $\{\xi_k\}_{k=1}^K$. Let $\boldsymbol{M} \in \mathbb{R}^{K\times K}$ be the matrix having $B_l^*(\xi_k)$ as element (l,k) and note that the square matrix $\boldsymbol{M}$ is full rank, and thus invertible, since the functions $\{B_k^*\}_{k=1}^K$ are linearly independent and, by assumption, $\xi_k \neq \xi_l$ for all $k \neq l$. Therefore, $B := \sum_{k=1}^k a_k B_k^* \in \mathcal{S}_K^*(\{\xi_k\}_{k=1}^K)$ is such that $B(\xi_k) = z_k$ for all $k \in \{1,\dots,K\}$ if and only if $\boldsymbol{M}a = z \Leftrightarrow a = \boldsymbol{M}^{-1}z$, showing the first part of the theorem.

Next let h be as in the second part of the theorem and let $g = h - B$. As preliminary computations remark that, using integration by parts,

$$
\begin{aligned}
\int_{[\xi_1,\xi_K]} B''(x)g''(x)\mathrm{d}x &= B''(x)g'(x)\Big|_{\xi_1}^{\xi_K} - \int_{[\xi_1,\xi_K]} B'''(x)g'(x)\mathrm{d}x \\
&= -\int_{[\xi_1,\xi_K]} B'''(x)g'(x)\mathrm{d}x \\
&= -\sum_{k=1}^{K-1}\int_{[\xi_k,\xi_{k+1}]} B'''(x)g'(x)\mathrm{d}x \\
&= -\sum_{k=1}^{K-1} B'''\Big(\frac{\xi_{k+1}-\xi_k}{2}\Big)\int_{[\xi_k,\xi_{k+1}]} g'(x)\mathrm{d}x \\
&= -\sum_{k=1}^{K-1} B'''\Big(\frac{\xi_{k+1}-\xi_k}{2}\Big) g(x)\Big|_{\xi_k}^{\xi_{k+1}} \\
&= -\sum_{k=1}^{K-1} B'''\Big(\frac{\xi_{k+1}-\xi_k}{2}\Big)\Big(g(\xi_{k+1})-g(\xi_k)\Big) \\
&= 0
\end{aligned}
\tag{8.5}
$$

where the 2nd equality uses the fact that $B''(\xi_1) = B''(\xi_K) = 0$, the 4th equality the fact that between the knots ξ_k and ξ_{k+1} the third derivative of B is constant, and the last equality holds since $g(\xi_k) = 0$ for all k.

Proof of Theorem 8.1 (end)

Then, using (8.5), we have

$$\begin{aligned}
\int_{[\xi_1,\xi_K]} \big(h''(x)\big)^2 \mathrm{d}x &= \int_{[\xi_1,\xi_K]} \big(h''(x) - B''(x) + B''(x)\big)^2 \mathrm{d}x \\
&= \int_{[\xi_1,\xi_K]} \big(g''(x)\big)^2 \mathrm{d}x + \int_{[\xi_1,\xi_K]} \big(B''(x)\big)^2 \mathrm{d}x \\
&+ 2\int_{[\xi_1,\xi_K]} B''(x) g''(x) \mathrm{d}x \\
&= \int_{[\xi_1,\xi_K]} \big(g''(x)\big)^2 \mathrm{d}x + \int_{[\xi_1,\xi_K]} \big(B''(x)\big)^2 \mathrm{d}x.
\end{aligned} \tag{8.6}$$

To complete the proof remark that since $g(\xi_k) = 0$ for all $k \in \{1, \ldots, K\}$ and $g \in \mathcal{C}^2([\xi_1, \xi_K])$, it follows that because $h \neq B$ there must exist an interval $[a, b] \subset (\xi_1, \xi_K)$ such that $g''(x) \neq 0$ for all $x \in [a, b]$. Hence,

$$\int_{[\xi_1,\xi_K]} \big(g''(x)\big)^2 \mathrm{d}x \geq \int_{[a,b]} \big(g''(x)\big)^2 \mathrm{d}x > 0$$

which, together with (8.6), shows that

$$\int_{[\xi_1,\xi_K]} \big(h''(x)\big)^2 \mathrm{d}x > \int_{[\xi_1,\xi_K]} \big(B''(x)\big)^2 \mathrm{d}x.$$

The proof is complete $\square$

Solution to the optimization problem (8.2)

We are now in position to state and prove the main result of this chapter.

Theorem 8.2 *Let $\lambda > 0$, $x^0_{\min} = \min\{x^0_i\}_{i=1}^n$, $x^0_{\max} = \max\{x^0_i\}_{i=1}^n$ and assume that the set $\{x^0_i\}_{i=1}^n$ contains at least three distinct values. Then, (8.2) has a unique solution and*

$$\hat{f}_\lambda = \underset{f\in\mathcal{C}^2(\mathbb{R})}{\operatorname{argmin}} \sum_{i=1}^n \big(y^0_i - f(x^0_i)\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x \tag{8.7}$$

is such that (i) the restriction of $\hat{f}_\lambda$ on $[x^0_{\min}, x^0_{\max}]$ is a natural cubic spline with knots at the unique values of $x^0_1, \dots, x^0_n$ and (ii) $\hat{f}''_\lambda(x) = 0$ for all $x \notin [x^0_{\min}, x^0_{\max}]$.

Remark: If $\hat{B}_\lambda : [x^0_{\min}, x^0_{\max}] \to \mathbb{R}$ is the natural cubic spline defined by $\hat{B}_\lambda(x) = \hat{f}_\lambda(x)$, $x \in [x^0_{\min}, x^0_{\max}]$ then, for $x \notin [x^0_{\min}, x^0_{\max}]$, the value of $\hat{f}_\lambda(x)$ is obtained by linearly extrapolating $\hat{B}_\lambda$.

Proof of Theorem 8.2: Let $I \subseteq \{1, \dots, n\}$ be such that $\{x^0_i\}_{i\in I}$ is the set of distinct values of $x^0_1, \dots, x^0_n$ and assume that there exists a function $h \in \mathcal{C}^2(\mathbb{R}) \setminus \mathcal{S}^*_{|I|}(\{x^0_i\}_{i\in I})$ such that

$$h \in \underset{f\in\mathcal{C}^2(\mathbb{R})}{\operatorname{argmin}} \sum_{i=1}^n \big(y^0_i - f(x^0_i)\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x. \tag{8.8}$$

Let $z_i = h(x^0_i)$ for all $i \in I$, $B \in \mathcal{S}^*_{|I|}(\{x^0_i\}_{i\in I})$ be as in Theorem 8.1 (for $\{\xi_j\}_{j=1}^K = \{x^0_i\}_{i\in I}$) and let $f \in \mathcal{C}^2(\mathbb{R})$ be such that $f(x) = B(x)$ for all $x \in [x^0_{\min}, x^0_{\max}]$ and such that $f''(x) = 0$ for all $x \notin [x^0_{\min}, x^0_{\max}]$. Then,

$$\sum_{i=1}^n \big(y^0_i - f(x^0_i)\big)^2 = \sum_{i=1}^n \big(y^0_i - B(x^0_i)\big)^2 = \sum_{i=1}^n \big(y^0_i - h(x^0_i)\big)^2 \tag{8.9}$$

while

$$\int_{\mathbb{R}} (h''(x))^2 \mathrm{d}x \geq \int_{[x^0_{\min}, x^0_{\max}]} (h''(x))^2 \mathrm{d}x > \int_{[x^0_{\min}, x^0_{\max}]} (B''(x))^2 \mathrm{d}x = \int_{\mathbb{R}} (f''(x))^2 \mathrm{d}x. \tag{8.10}$$

Therefore, by (8.9)-(8.10), we have

$$\sum_{i=1}^n \big(y^0_i - h(x^0_i)\big)^2 + \lambda \int_{\mathbb{R}} (h''(x))^2 \mathrm{d}x > \sum_{i=1}^n \big(y^0_i - f(x^0_i)\big)^2 + \lambda \int_{\mathbb{R}} (f''(x))^2 \mathrm{d}x$$

which contradicts (8.8). The fact that (8.2) has a unique solution follows from the fact that the spline

$B \in \mathcal{S}^*_{|I|}(\{x^0_i\}_{i\in I})$ defined in theorem Theorem 8.1 is unique. The proof is complete. □

Computation of $\hat{f}_\lambda$

Let $\tilde{x}_0$ be the vector containing the $m \leq n$ distinct values of $x_1^0, \dots, x_n^0$ and let $\{b_j\}_{j=1}^m$ be a basis for the set $\mathcal{S}_m^*(\tilde{x}_0)$.

Let $\boldsymbol{Z}$ be the $n \times m$ matrix having $b_j(x_i^0)$ as entry (i,j) and let $\boldsymbol{S}_{\text{pen}}$ be the $m \times m$ matrix having $\int_{[x_{\min}^0, x_{\max}^0]} b_j''(x) b_l''(x) \mathrm{d}x$ as entry (j,l). Remark that the matrix $\boldsymbol{Z}^\top \boldsymbol{Z}$ is full rank, since $\{b_j\}_{j=1}^m$ are m basis functions and since the set $\{x_i^0\}_{i=1}^n$ contains m distinct values[a].

Corollary 8.1 *Consider the set-up of Theorem 8.2. Then, for any* $\lambda > 0$

$$\hat{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{C}^2(\mathbb{R})} \sum_{i=1}^n \big(y_i^0 - f(x_i^0)\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x$$

if and only if (i) $\hat{f}_\lambda(x) = \sum_{j=1}^m \beta_{\lambda,j} b_j(x)$ *for all* $x \in [x_{\min}^0, x_{\max}^0]$, *with*

$$\beta_\lambda = \big(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{S}_{\text{pen}}\big)^{-1} \boldsymbol{Z}^\top y^0,$$

and (ii) $\hat{f}_\lambda''(x) = 0$ *for all* $x \notin [x_{\min}^0, x_{\max}^0]$.

Proof: Let $\hat{f}_\lambda \in \operatorname{argmin}_{f \in \mathcal{C}^2(\mathbb{R})} \sum_{i=1}^n (y_i^0 - f(x_i^0))^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x$. Then, by Theorem 8.2, there exists a $\beta_\lambda \in \mathbb{R}^m$ such that $\hat{f}_\lambda = \sum_{j=1}^m \beta_{\lambda,j} b_j$. More precisely, β_λ must be such that

$$\begin{aligned}
\beta_\lambda &\in \operatorname*{argmin}_{\beta \in \mathbb{R}^m} \|y^0 - \boldsymbol{Z}\beta\|^2 + \lambda \int_{[x_{\min}^0, x_{\max}^0]} \Big(\sum_{j=1}^m \beta_j b_j''(x)\Big)^2 \\
&= \operatorname*{argmin}_{\beta \in \mathbb{R}^m} \|y^0 - \boldsymbol{Z}\beta\|^2 + \lambda \beta^\top \boldsymbol{S}_{\text{pen}} \beta \\
&= \big(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{S}_{\text{pen}}\big)^{-1} \boldsymbol{Z}^\top y^0. \qquad (8.11)
\end{aligned}$$

The proof is complete. □.

[a] The matrix $\boldsymbol{S}_{\text{pen}}$ is however not full rank.

Choosing the penalty parameter λ

Given the expression (8.11) of β_λ, it follows that, as for ridge regression, the leave-one-out cross validation procedure for choosing the penalty parameter λ can be efficiently implemented for smoothing.

More precisely, recall that using leave-one-out cross validation procedure to choose λ amounts to letting $\lambda = \hat{\lambda}$ where

$$\hat{\lambda} \in \underset{\lambda \in [0,\infty)}{\operatorname{argmin}} \mathrm{OCV}_{\text{smooth}}(\lambda), \quad \mathrm{OCV}_{\text{smooth}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i^0 - \beta_{-i,\lambda}^\top z_i)^2$$

where $\beta_{-i,\lambda}$ is computed as in (8.11) after having removed observation (y_i^0, x_i^0) from the sample.

Letting $\tilde{\mathbf{A}}^{(\lambda)} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{S}_{\text{pen}})^{-1}\boldsymbol{Z}^\top$, it follows from Theorem 6.1[a] that

$$\mathrm{OCV}_{\text{smooth}}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\frac{(y_i^0 - \beta_\lambda^\top z_i)^2}{(1-\tilde{a}_{ii}^{(\lambda)})^2}, \quad \forall \lambda > 0$$

and therefore that computing $\mathrm{OCV}_{\text{smooth}}(\lambda)$ requires only to compute β_λ and $\{\tilde{a}_{ii}^{(\lambda)}\}_{i=1}^n$.

Alternatively, one can choose λ using the generalized cross-validation criterion

$$\mathrm{GCV}_{\text{smooth}}(\lambda) := \frac{n\|y^0 - \boldsymbol{Z}\beta_\lambda\|^2}{\big\{n - \mathrm{tr}(\tilde{\mathbf{A}}^{(\lambda)})\big\}^2}, \quad \lambda > 0 \qquad (8.12)$$

Remark: It can be shown that for $\lambda > 0$ we have $\mathrm{tr}(\tilde{\mathbf{A}}^{(\lambda)}) \in (0, n)$ [13, page 212], so that $\mathrm{GCV}_{\text{smooth}}(\lambda)$ is well-defined for all $\lambda > 0$. By contrast, we can only guarantee that $\tilde{a}_{ii}^{(\lambda)} \in [0,1]$ and therefore the quantity $\mathrm{OCV}_{\text{smooth}}(\lambda)$ may not be well-defined.

[a] Actually, to apply Theorem 6.1 we need (i) that all the x_i^0's are distinct and (ii) to use only $n-1$ out of the n basis functions.

Choice of the basis functions

We start with two important remarks:

- In theory the choice of the basis functions $\{b_j\}_{j=1}^m$ of the set $\mathcal{S}_m^*(\tilde{x}_0)$ does not matter. However, from a computational point of view this choice is important. Indeed, for some basis functions $\{b_j\}_{j=1}^m$ (such as for the truncated power basis) the columns of $\boldsymbol{Z}$ may be highly correlated, which could lead to numerical instabilities.

- In general, we have $m = \mathcal{O}(n)$ so that inverting the matrix $\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{S}_{\text{pen}}$ (that appears in the definition of β_λ) requires $\mathcal{O}(n^3)$ operations.

To avoid the two aforementioned problem, the B-spline basis is often used in practice.

One of the main advantage of this basis is to be such that, for all $x \in [x^0_{\text{min}}, x^0_{\text{max}}]$, we have $b_j(x) \neq 0$ for at most 4 values of $j \in \{1, \dots, m\}$, making the matrix $\boldsymbol{Z}$ sparse. Because $\boldsymbol{Z}$ is sparse the computation of β_λ is usually numerically stable. In addition, the particular sparsity structure of the matrix $\boldsymbol{Z}$ obtained with B-spline makes possible to compute β_λ in $\mathcal{O}(n \log(n))$ operations, even when $m = \mathcal{O}(n)$ (see [3], Appendix of Chapter 5).

Remark: The B-spline functions form a basis of $\mathcal{S}_{3,m}(\tilde{x}^0)$ and not of $\mathcal{S}_m^*(\tilde{x}^0)$, and thus contains $m+2$ functions $\{\tilde{b}_j\}_{j=1}^{m+2}$ (by Proposition 8.1). However, since $\mathcal{S}_m^*(\tilde{x}^0) \subset \mathcal{S}_{3,m}(\tilde{x}^0)$ it follows that $\hat{f}_\lambda$ can be expressed as linear combination of the B-splines basis[a], and thus the resulting value of $\beta_\lambda \in \mathbb{R}^{m+2}$ is guaranteed to be such that $\hat{f}_\lambda = \sum_{j=1}^{m+2} \beta_{\lambda,j} \tilde{b}_j$.

[a] We abuse notation/language here by referring to $\hat{f}_\lambda$ as a spline while, in fact, it is the restriction of $\hat{f}_\lambda$ to $[x^0_{\text{min}}, x^0_{\text{max}}]$ which is a spline.

Thinning

If the use of B-splines allows to compute β_λ in $\mathcal{O}(n)$ operations the cost of computing the cross-validation criterion $\mathrm{OCV}_{\mathrm{smooth}}(\lambda)$ or $\mathrm{GCV}_{\mathrm{smooth}}(\lambda)$ is still of size $\mathcal{O}(\max(m^3, n))$[a].

For this reason, in practice, when m in large not all the m basis functions $\{b_j\}_{j=1}^m$ are used. Luckily, any reasonable thinning strategy will have little impact on the fit.

To understand this latter claim assume that the regression model (8.1) is well-specified, that is that there exists a function $f^0 \in \mathcal{C}^2(\mathbb{R})$ such that, with $\{\epsilon_i\}_{i=1}^n$ as above,

$$Y_i^0 = f^0(x_i^0) + \epsilon_i, \quad i = 1, \dots, n.$$

Assume that $m = n$, that for some $a < \infty$ we have $|x_i^0| \leq a$ for all i and that all the x_i^0's are distinct.

For $f : [-a, a] \to \mathbb{R}$ let

$$\|f\| = \left(\int_{[-a,a]} f(x)^2 \mathrm{d}x \right)^{1/2}$$

be the L_2 norm of f and let

$$f_n^0 = \operatorname*{argmin}_{f \in \mathcal{S}_n^*(x^0)} \|f - f^0\|.$$

[a] One reason why β_λ can be computed in $\mathcal{O}(n)$ is that we can compute β_λ without inverting the matrix $\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{S}_{\mathrm{pen}}$. However,m this matrix needs to be inverted to compute the OCV and GCV criteria

Thinning (end)

Let $\lambda > 0$. Then, the estimation error $\|f^0 - \hat{f}_\lambda\|$ depends on

1. The approximation error $\|f^0 - f_n^0\|$, which is due to the fact that we approximate $f^0 \in \mathcal{C}^2([-a, a])$ by a function in the set $\mathcal{S}_n^*(x^0)$.

   Under some additional conditions on f^0 it can be shown that [**?**]

   $$\|f^0 - f_n^0\| = \mathcal{O}(h_n^4), \quad h_n = \max_{i\in\{1,\dots,n\}} \min_{i\neq l} \|x_i^0 - x_l^0\|.$$

   Typically, $h_n = \mathcal{O}(1/n)$, in which case $\|f^0 - f_n^0\| = \mathcal{O}(n^{-4})$.

2. The statistical error $\|\hat{f}_\lambda - f_n^0\|$, which is at least of size $\mathcal{O}(n^{-1/2})$[a].

When all the n basis functions are used the approximation error is therefore much smaller than the statistical error.

In particular, if for some $\alpha \in (0, 1]$ we only use $m = \mathcal{O}(n^\alpha)$ basis functions associated to m distinct elements of $\{x_i^0\}_{i=1}^n$ which are approximatively equally spaced then the rate at which $\|f^0 - \hat{f}_\lambda\|$ converges to zero is the same for all $\alpha \in [1/8, 1]$.

Remark: When λ is selected from the data (e.g. using cross-validation) then we need a slightly larger value of α if we want the penalization to be effective when n in large [13, Section 5.2, page 199].

[a] Recall that $n^{-1/2}$ is the standard parametric convergence rate.

Illustrative example: The fossil dataset[a]

This dataset contains the ratio of strontium isotopes found in $n = 106$ fossil shells. The fossils shells were formed in the mid-Cretaceous period and are between 91 to 123 million years old. For this example y_i^0 is ratio of strontium isotopes in the ith fossil and x_i^0 is its age (measured in million of years). In this dataset all the x_i^0's are distinct.

Figure 8.1 below shows the function $\hat{f}_\lambda$ obtained when λ has been selected using the GCV criterion (8.12).

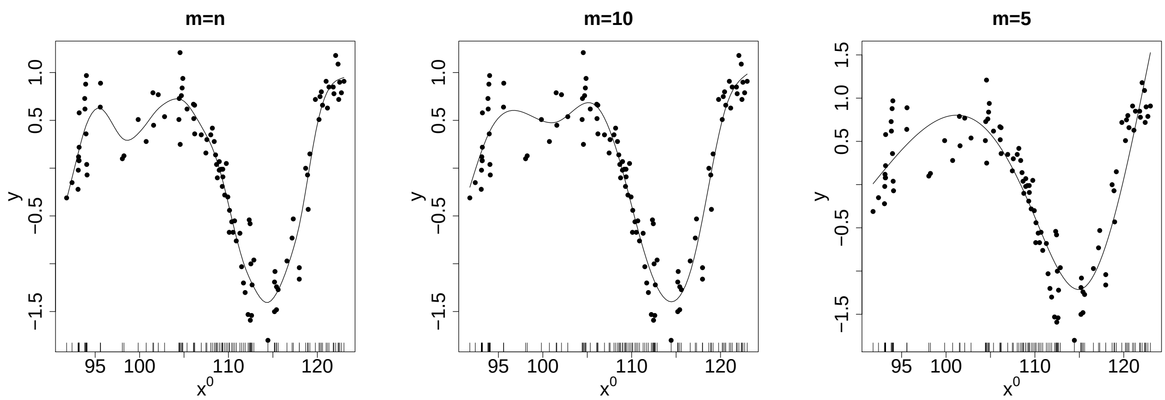

Figure 8.1: Smoothing regression for the fossil dataset with m basis functions, for all $m \in \{n, 10, 5\}$. The dots represents the observations $\{(y_i, x_i^0)\}_{i=1}^n$

We observe that when all the $m = 106$ basis functions of $\mathcal{S}_n^*(\{x_i^0\}_{i=1}^n)$ are used the function $\hat{f}_\lambda$ represents well the relationship between the age and ratio of strontium isotopes of a fossil. The second plot of Figure 8.1 shows that taking only 10 out of the 106 basis functions has little impact on the estimated function. However, decreasing further m to $m = 5$ significantly deteriorates the fit.

[a]This dataset is Available in the `R` package `brinla`.

Inclusion of a smooth function in a larger model

Let $p > 1$ and, using the shorthand $w_i^0 = (x_{i2}^0, \dots, x_{ip}^0)$ and with $\{\epsilon_i\}_{i=1}^n$ as in (8.1), assume the following model for the observations $\{y_i^0\}_{i=1}^n$

$$Y_i^0 = \alpha + f(x_{i1}^0) + \gamma^\top w_i^0 + \epsilon_i, \quad i = 1, \dots, n \tag{8.13}$$

where $f \in \mathcal{C}^2(\mathbb{R})$, $\alpha \in \mathbb{R}$ and where $\gamma \in \mathbb{R}^{p-1}$.

Then, a natural estimator of (f, α, γ) is

$$(\hat{f}_\lambda, \hat{\alpha}_\lambda, \hat{\gamma}_\lambda) \in \underset{f \in \mathcal{C}^2(\mathbb{R}), (\alpha,\gamma) \in \mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n \big(y_i^0 - \alpha - f(x_{i1}^0) - \gamma^\top w_i^0\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x.$$

The model (8.13) is however non-identifiable since, for all $c \in \mathbb{R}$,

$$Y_i^0 = \alpha + f(x_{i1}^0) + \gamma^\top w_i^0 + \epsilon_i = (\alpha - c) + \big(f(x_{i1}^0) + c\big) + \gamma^\top w_i^0 + \epsilon_i$$

where $f + c \in \mathcal{C}^2(\mathbb{R})$ if $f \in \mathcal{C}^2(\mathbb{R})$. Consequently, the solution $(\hat{f}_\lambda, \hat{\alpha}_\lambda, \hat{\gamma}_\lambda)$ to the above optimization problem is not unique.

A first solution to this identifiability issue is to remove the intercept from the model, in which case there exists in general a unique solution $(\hat{f}_\lambda, \hat{\gamma}_\lambda)$ to the optimization problem

$$\min_{f \in \mathcal{C}^2(\mathbb{R}), \gamma \in \mathbb{R}^{p-1}} \sum_{i=1}^n \big(y_i^0 - f(x_{i1}^0) - \gamma^\top w_i^0\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x. \tag{8.14}$$

Remark: If $\{b_j\}_{j=1}^m$ is the B-spline basis of $\mathcal{S}_m^*(\tilde{x}^0)$, with $m = |\tilde{x}^0|$, then $\sum_{j=1}^m b_j(x) = 1$ of all $x \in [x_{\text{min}}^0, x_{\text{max}}^0]$. In this case, we have $\alpha + \sum_{j=1}^m \beta_j b_j(x_i^0) = \sum_{j=1}^m (\beta_j + \alpha) b_j(x_i^0)$ so that omitting α in (8.13) is equivalent to shifting all the β_j's parameter by α (and thus the shape of the estimated function will be unchanged).

Inclusion of a smooth function in a larger model (end)

A second, and more popular, solution to the aforementioned identifiability issue is to impose that $\hat{f}_\lambda$ is such that $\sum_{i=1}^n \hat{f}_\lambda(x_{i1}^0) = 0$, that is to estimate (α, γ, f) using

$$(\tilde{f}_\lambda, \tilde{\alpha}_\lambda, \tilde{\gamma}_\lambda) \in \underset{f \in \tilde{\mathcal{C}}^2(\mathbb{R}), (\alpha,\gamma)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n \big(y_i^0 - \alpha - f(x_{i1}^0) - \gamma^\top w_i^0\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x.$$

where $\tilde{\mathcal{C}}^2(\mathbb{R}) = \{f \in \mathcal{C}^2(\mathbb{R}) : \sum_{i=1}^n f(x_{i1}^0) = 0\}$.

As shown in the next proposition, $(\tilde{f}_\lambda, \tilde{\alpha}_\lambda, \tilde{\gamma}_\lambda)$ is uniquely defined.

Proposition 8.3 *Let* $\boldsymbol{M}_\lambda = \boldsymbol{I}_n - \boldsymbol{Z}\big(\boldsymbol{Z}^\top \boldsymbol{Z} + \lambda \boldsymbol{S}_{\text{pen}}\big)^{-1}\boldsymbol{Z}^\top$ *and assume that the matrix* $\boldsymbol{W}^\top \boldsymbol{M}_\lambda^2 \boldsymbol{W}$ *is invertible. Let*

$$\tilde{\gamma}_\lambda = \big(\boldsymbol{W}^\top \boldsymbol{M}_\lambda^2 \boldsymbol{W}\big)^{-1} \boldsymbol{W}^\top \boldsymbol{M}_\lambda^2 y, \quad \tilde{\alpha}_\lambda = \bar{y}^0 - \tilde{\gamma}_\lambda^\top \bar{w}^0$$

and $\tilde{f}_\lambda = \operatorname{argmin}_{f \in \mathcal{C}^2(\mathbb{R})} \sum_{i=1}^n \big((y_i - \tilde{\gamma}_\lambda^\top w_i) - f(x_{i1}^0)\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x$. *Then,*

$$(\tilde{\alpha}_\lambda, \tilde{\gamma}_\lambda, \tilde{f}_\lambda) = \underset{f \in \tilde{\mathcal{C}}^2(\mathbb{R}), (\alpha,\gamma)\in\mathbb{R}^p}{\operatorname{argmin}} \sum_{i=1}^n \big(y_i^0 - \alpha - f(x_{i1}^0) - \gamma^\top w_i^0\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x.$$

Remark: By Theorem 8.2, $\tilde{f}_\lambda : [x^0_{(1),\min}, x^0_{(1),\max}] \to \mathbb{R}$ is a natural cubic spline with knots at the unique values of $\{x_{i1}^0\}_{i=1}^n$ and $\tilde{f}''_\lambda(x) = 0$ for all $x \notin [x^0_{(1),\min}, x^0_{(1),\max}]$, where $x^0_{(1),\min} = \min\{x_{i1}\}_{i=1}^n$ and $x^0_{(1),\max} = \max\{x_{i1}\}_{i=1}^n$.

Remark: $\tilde{\gamma}_\lambda$ is a generalized least squares estimate of γ in the model $Y = \boldsymbol{W}\gamma + \epsilon$.

Proof of Proposition 8.3

Let $F(\alpha,\gamma,f)=\sum_{i=1}^n\big(y_i^0-\alpha-f(x_{i1}^0)-\gamma^\top w_i^0\big)^2+\lambda\int_{\mathbb{R}}f''(x)^2\mathrm{d}x$ and for $f\in\tilde{\mathcal{C}}^2(\mathbb{R})$ and $\gamma\in\mathbb{R}^p$ let

$$\alpha_{f,\gamma}=\operatorname*{argmin}_{\alpha\in\mathbb{R}}F(\alpha,\gamma,f)=\bar{y}^0-\frac{1}{n}\sum_{i=1}^n f(x_{i1})-\gamma^\top\bar{w}^0=\bar{y}^0-\gamma^\top\bar{w}^0.$$

Next, for $\gamma\in\mathbb{R}^{p-1}$, let $y_{\gamma,i}=y_i-\gamma^\top w_i$ and $f_\gamma\in\tilde{\mathcal{C}}^2(\mathbb{R})$ be such that

$$\begin{aligned}f_\gamma&\in\operatorname*{argmin}_{f\in\tilde{\mathcal{C}}^2(\mathbb{R})}F(\alpha_{f,\gamma},\gamma,f)\\&=\operatorname*{argmin}_{f\in\tilde{\mathcal{C}}^2(\mathbb{R})}\sum_{i=1}^n\big(y_{\gamma,i}-f(x_{i1}^0)\big)^2+\lambda\int_{\mathbb{R}}f''(x)^2\mathrm{d}x\\&=\operatorname*{argmin}_{f\in\mathcal{C}^2(\mathbb{R})}\sum_{i=1}^n\big(y_{\gamma,i}-f(x_{i1}^0)\big)^2+\lambda\int_{\mathbb{R}}g''(x)^2\mathrm{d}x.\end{aligned}\tag{8.15}$$

To show the latter equality let $\tilde{f}_\gamma\in\operatorname{argmin}_{f\in\mathcal{C}^2(\mathbb{R})}F(\alpha_{f,\gamma},\gamma,f)$ and, for every $c\in\mathbb{R}$, let $g_c=\tilde{f}_\gamma-c$. Then, $g_c''(x)\equiv\tilde{f}_\gamma''(x)$ for all $x\in\mathbb{R}$ and

$$\begin{aligned}c^*:=\operatorname*{argmin}_{c\in\mathbb{R}}\sum_{i=1}^n\big(y_{\gamma,i}-(\tilde{f}_\gamma(x_{i1}^0)+c)\big)^2&=\frac{1}{n}\sum_{i=1}^n\tilde{f}_\gamma(x_{i1}^0)-\frac{1}{n}\sum_{i=1}^n y_{\gamma,i}\\&=\frac{1}{n}\sum_{i=1}^n\tilde{f}_\gamma(x_{i1}^0).\end{aligned}$$

Hence, if $\sum_{i=1}^n\tilde{f}_\gamma(x_{i1}^0)\neq0$ we have $F(\alpha_{g_{c^*},\gamma},\gamma,g_{c^*})<F(\alpha_{\tilde{f}_\gamma,\gamma},\gamma,\tilde{f}_\gamma)$, which contradicts the fact that $\tilde{f}_\gamma\in\operatorname{argmin}_{f\in\mathcal{C}^2(\mathbb{R})}F(\alpha_{f,\gamma},\gamma,f)$.

Finally, using the fact that $\alpha_{f_\gamma,\gamma}=\bar{y}^0-\gamma^\top\bar{w}^0$ together with (8.15) and Corollary 8.1, we obtain

$$\begin{aligned}\operatorname*{argmin}_{\gamma\in\mathbb{R}^{p-1}}F(\alpha_{f_\gamma},\gamma,f_\gamma)&=\operatorname*{argmin}_{\gamma\in\mathbb{R}^{p-1}}\|y-\boldsymbol{W}\gamma-\boldsymbol{Z}\big(\boldsymbol{Z}^\top\boldsymbol{Z}+\lambda\boldsymbol{S}_{\text{pen}}\big)^{-1}\boldsymbol{Z}^\top(y-\boldsymbol{W}\gamma)\|^2\\&=\operatorname*{argmin}_{\gamma\in\mathbb{R}^{p-1}}\|\boldsymbol{M}_\lambda(y-\boldsymbol{W}\gamma)\|^2\\&=(\boldsymbol{W}^\top\boldsymbol{M}_\lambda^2\boldsymbol{W})^{-1}\boldsymbol{W}^\top\boldsymbol{M}_\lambda^2 y.\end{aligned}$$

The proof is complete. $\square$

A convenient representation of $\hat{f}_\lambda$

Going back to the case where $p = 1$, Proposition 8.3 shows that the smoothing estimate $\hat{f}_\lambda$ of f, defined in (8.2), can be written as

$$\hat{f}_\lambda = \bar{y}_0 + \tilde{f}_\lambda \tag{8.16}$$

where the function $\tilde{f}_\lambda$ is defined by

$$\tilde{f}_\lambda = \operatorname*{argmin}_{f \in \mathcal{C}^2(\mathbb{R})} \sum_{i=1}^{n} \big(y_i - f(x_i^0)\big)^2 + \lambda \int_{\mathbb{R}} f''(x)^2 \mathrm{d}x.$$

and is such that $\sum_{i=1}^n \tilde{f}_\lambda(x_i^0) = 0$.

Remark that, unlike $\hat{f}_\lambda$, the function $\tilde{f}_\lambda$ remains unchanged if we replace $\{y_i^0\}_{i=1}^n$ by $\{y_i^0 + c\}_{i=1}^n$ for some $c \in \mathbb{R}$.

For this reason, in practice, the function $\tilde{f}_\lambda$ is often the main object of interest in smoothing, and we therefore compute $\hat{f}_\lambda$ by first computing $\tilde{f}_\lambda$ and $\bar{y}_0$ and then using (8.16).

Multi-dimensional smoothing

The smoothing approach introduced in this chapter can be extended to $p > 1$ dimensional input variables $\{x_i^0\}_{i=1}^n$. In this case, the model we consider for $\{y_i^0\}_{i=1}^n$ is

$$Y_i^0 = f(x_i^0) + \epsilon_i, \qquad i = 1, \dots n, \quad f \in \mathcal{C}^2(\mathbb{R}^p) \tag{8.17}$$

where, as per above, we have $\mathbb{E}[\epsilon_i] = 0$, $\mathbb{E}[\epsilon_i \epsilon_l] = \sigma^2 \delta_{il}$ for all i and l. Then, for a given $\lambda \in [0, \infty]$, the smoothing estimate of the function f is given by

$$\hat{f}_{p,\lambda} \in \operatorname*{argmin}_{f \in \mathcal{C}^2(\mathbb{R}^p)} \sum_{i=1}^n \big(y_i^0 - f(x_i^0)\big)^2 + \lambda J_p(f) \tag{8.18}$$

for some penalty functional $J_p : \mathcal{C}^2(\mathbb{R}^p) \to \mathbb{R}$. For instance, for $p = 2$ we have

$$J_2(f) = \int \Big[\Big(\frac{\partial^2 f(x)}{\partial x_1^2}\Big)^2 + 2\Big(\frac{\partial^2 f(x)}{\partial x_1 \partial x_2}\Big)^2 + \Big(\frac{\partial^2 f(x)}{\partial x_2^2}\Big)^2\Big] \mathrm{d}x, \quad \forall f \in \mathcal{C}^2(\mathbb{R}^2).$$

Remark: When $p = 1$ the function $\hat{f}_{p,\lambda}$ defined in (8.18) reduces to the function $\hat{f}_\lambda$ defined in (8.2).

It can be shown that, for some $m_p \in \mathbb{N}$, the function $\hat{f}_{p,\lambda}$ defined in (8.18) can be written as $\hat{f}_{p,\lambda} = \sum_{j=1}^{m_p} \beta_j b_{p,j}$ where $\beta \in \mathbb{R}^{m_p}$ and where $\{b_{p,j}\}_{j=1}^{m_d}$ are known basis functions. Hence, as for the case $p = 1$, the problem of estimating $f \in \mathcal{C}^2(\mathbb{R}^p)$ reduces to the problem of estimating a finite dimensional vector β of parameters.

Key problem: The cost of estimating β is $\mathcal{O}(m_p^3)$ with $m_p = n + c_p$ where (i) c_p increases exponentially fast with p[a] and (ii) unlike in the case $p = 1$, thinning cannot be used to reduce the computational cost without losing too much in term of estimation error[b].

[a] This is because, assuming p is odd, $c_p \geq \binom{(p+1)/2+p-1}{p} \geq (3/2 - 1/p)^p$ [13, page 216].
[b] This is because for $p > 1$ all the x_i^0's are far apart.

Example: Two dimensional smoothing

We let $p = 2$, $f^0 \in \mathcal{C}^2(\mathbb{R}^2)$ be as represented in Figure 8.2 and simulate $n = 200$ independent observations $\{(y_i^0, x_i^0)\}_{i=1}^n$ using

$$Y_i^0 = f^0(x_i^0) + \sigma\epsilon_i, \quad X_i^0 \sim \mathcal{U}((-2,2)^2), \quad i = 1, \dots, n.$$

The function $\hat{f}_{p,\lambda}$ defined in (8.18) is represented in Figure 8.2 for $\sigma = 0.01$ and for $\sigma = 0.1$, and when is chosen using GCV and when all the m_p basis functions $\{b_{p,j}\}_{j=1}^{m_p}$ are used.

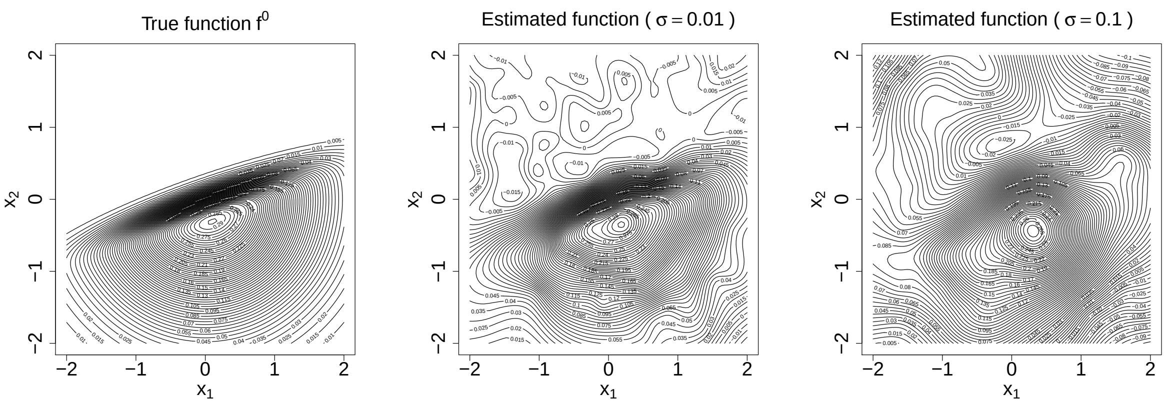

Figure 8.2: True function f^0 and estimated function $\hat{f}_{p,\lambda}$ for $\sigma = 0.01$ and for $\sigma = 0.1$. The value of λ is chosen using GCV.

From Figure 8.2 we observe that we obtain a reasonable estimate of f^0 when the size σ of the noise is very small. We also remark that even for the small value $\sigma = 0.1$ the function $\hat{f}_{p,\lambda}$ only provides a rough estimate of f^0. Improving the estimate would require to increase the sample size n but, as mentioned above, multivariate smoothing is computationally expensive when n is large.

References

[1] Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

[2] Comon, P. (1994). Independent component analysis, a new concept? *Signal processing*, 36(3):287–314.

[3] Friedman, J., Hastie, T., Tibshirani, R., et al. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

[4] Hastie, T., Tibshirani, R., and Wainwright, M. (2019). *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC.

[5] Hyvärinen, A. and Oja, E. (2000). Independent component analysis: algorithms and applications. *Neural networks*, 13(4-5):411–430.

[6] Inaba, M., Katoh, N., and Imai, H. (1994). Applications of weighted voronoi diagrams and randomization to variance-based k-clustering. In *Proceedings of the tenth annual symposium on Computational geometry*, pages 332–339.

[7] Mairal, J. and Yu, B. (2012). Complexity analysis of the lasso regularization path. *arXiv preprint arXiv:1205.0079*.

[8] Mardia, K., Kent, J., and Bibby, J. (1979). *Multivariate analysis*. Probability and mathematical statistics. Academic Press Inc.

[9] Ng, A., Jordan, M., and Weiss, Y. (2001). On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14.

[10] Paulsen, V. I. and Raghupathi, M. (2016). *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press.

[11] van Wieringen, W. N. (2015). Lecture notes on ridge regression. *arXiv preprint arXiv:1509.09169*.

[12] Von Luxburg, U. (2007). A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416.

[13] Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.