SL: SPLINES & SMOOTHING SPLINES

Linear models for regression are popular since easy to fit and highly interpretable, although most of the time not very accurate. In this chapter we discuss techniques for moving beyond linearity by augmenting the feature space with variables that are transformations of the original input points. The main advantage is that once the new features are introduced, the model remains linear in this enlarged feature space, and standard techniques apply. We have already encountered such models when discussing polynomial regression with univariate features in SL: FOUNDATIONS:

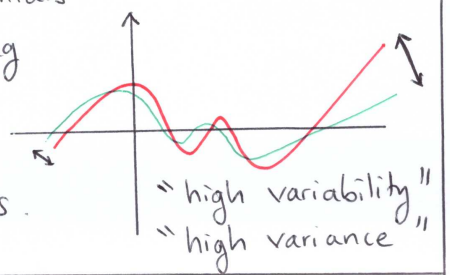
$$\mathcal{F}_d := \{ f : \mathbb{R} \to \mathbb{R} \mid f(x) = \beta_0 + \beta_1 x + \cdots + \beta_d x^d, \; x \in \mathbb{R} \}$$

Introducing $x_1 := x$, $x_2 := x^2$, $\vdots$, $x_d := x^d$,

functions $f \in \mathcal{F}_d$ are linear in $x_1, \cdots, x_d$:

$$f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_d x_d .$$

The "global" nature of polynomials make them unattractive: changing slightly the value of the coefficients may have a dramatic effects on remote regions.



In this chapter, we consider (polynomial) local regression to address this problem, we discuss splines, cubic splines, and natural cubic spliner & we focus on univariate predictors $X \in \mathbb{R}$ for simplicity. In the second part of this chapter, we consider non-parametric regression, and show that the solution to

$$\min_{f \in \mathcal{F}} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \int |f''(u)|^2 du$$

$\mathcal{F}$: space of functions with two continuous derivatives (NO parametric form)

ℓ: square loss

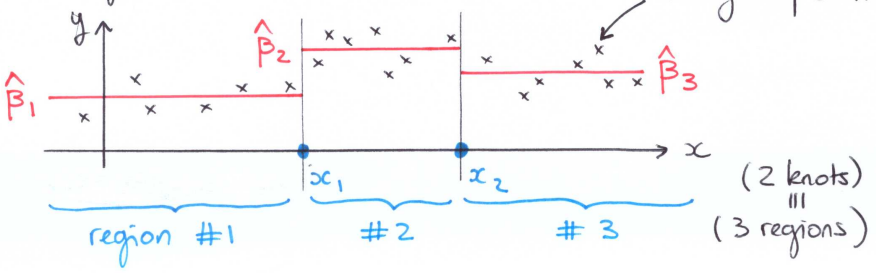
λ: tuning parameter ($\lambda > 0$)

is a natural cubic spline.

I - LOCAL REGRESSION

I.1 Piecewise polynomials

→ Divide the input space $\mathcal{X} = \mathbb{R}$ into non-overlapping regions, say $(-\infty, x_1)$, $[x_1, x_2)$, and $[x_2, +\infty)$, for $x_1 < x_2$ (called KNOTS), and fit a constant in each region.



The resulting fit can be written

$$f_n(x) = \hat{\beta}_1 \mathbb{1}(x < x_1) + \hat{\beta}_2 \mathbb{1}(x_1 \leq x < x_2) + \hat{\beta}_3 \mathbb{1}(x \geq x_2)$$

We have 3 "basis" functions. Denote them by $f_1(x)$, $f_2(x)$, and $f_3(x)$ respectively, so that

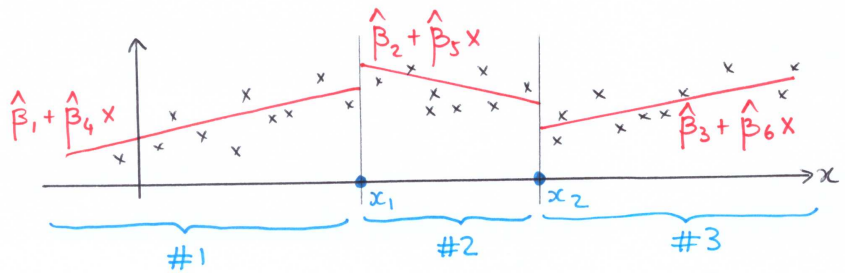
$$f_n(x) = \sum_{j=1}^{3} \hat{\beta}_j f_j(x) .$$

(Coefficients $\hat{\beta}_j$ corresponding to the mean value of the response variables in each region, if a square loss is used).

⊕ Right-most fit does not influence the left-most fit.

⊖ Discontinuities of the regression function at knots x_1, x_2

→ Consider a piecewise linear fit in each region.



To express f_n, we need an additional 3 basis functions $f_{j+3}(x) = x f_j(x)$; $j = 1, 2, 3$, so that

$$f_n(x) = \sum_{j=1}^{6} \hat{\beta}_j f_j(x) .$$

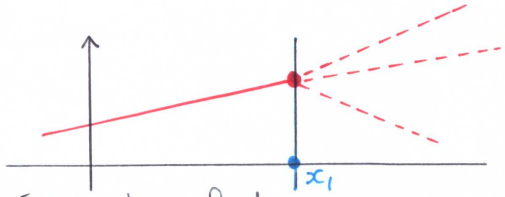
[We have 6 regions, for a total of 6 basis functions]

To remove discontinuities at the knots x_1, x_2, we enforce that

$$\begin{cases} \cdot \underline{\text{at } x_1} : & \beta_1 + \beta_4 x_1 = \beta_2 + \beta_5 x_1 \\ \cdot \underline{\text{at } x_2} : & \beta_2 + \beta_5 x_2 = \beta_3 + \beta_6 x_2 . \end{cases}$$

↳ We are loosing 2 degrees of freedom; i.e. 2 parameters ⇒ we only need 4 basis functions, instead of 6.

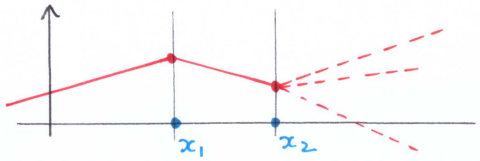
Instead of imposing constraints on the parameters, we work with basis functions that ensure automatically continuity at the knots. This can be achieved as follows:



(a) Consider basis functions $f_1(x) := 1$ & $f_2(x) := x$, so that the fit in region #1 can be written $\hat{\beta}_1 f_1(x) + \hat{\beta}_2 f_2(x)$

(b) To express $f_n(x)$ in region 2, we only have one degree of freedom: the value of the slope (we need continuity at x_1). Introducing $f_3(x) := (x - x_1)_+$, the fit in region 1+2 is $f_n(x) = \sum_{j=1}^{3} \hat{\beta}_j f_j(x)$

(c) Similarly, introducing $f_4(x) := (x - x_2)_+$, we have expressed $f_n(x)$ as a linear combination of 4 basis functions.



(+) . Right-most fit does not influence the left-most fit
. Continuity at the knots.

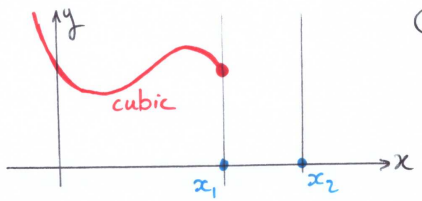
(−) . Not very smooth: fit is not differentiable at the knots.

↳ We increase the smoothness of the fit by increasing the order of the local polynomials.

I.2. Splines & Cubic Splines.

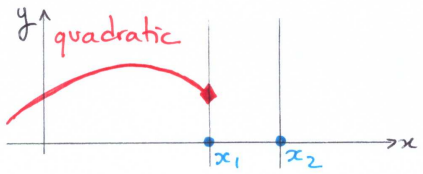
We consider polynomials of degree 3 in each region, enforcing smoothness for the first and second derivatives at the knots (note that enforcing smoothness for the third derivative at the knots would lead to a global cubic polynomial).

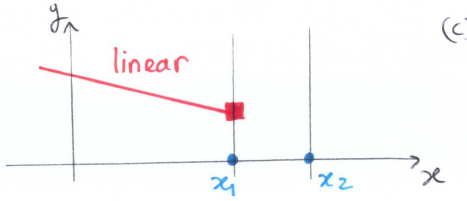
→ How many basis functions do we need ?



(a) Fit a cubic polynomial in the first region ⇒ 4 basis functions
$f_j(x) = x^j$, $j = 0, 1, 2, 3$.
(The location at x_1 is fixed)

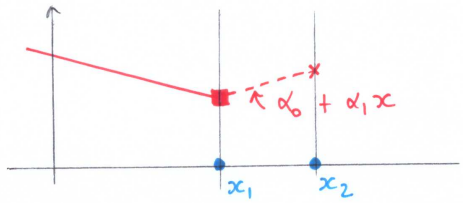
(b) The location of the first derivative at x_1 is fixed as well.





(c) And so is the location of the second derivative at x_1.

The location of the regression function and its first two derivatives are fixed at x_1. To enforce smoothness of the second derivative at x_1, the only degree of freedom we have is in the slope of the linear fit in the second region:



(d) Fit in the second region is

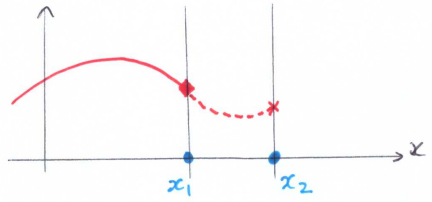
$\alpha_0 + \alpha_1 x$

↑ Fixed ↑ Free.

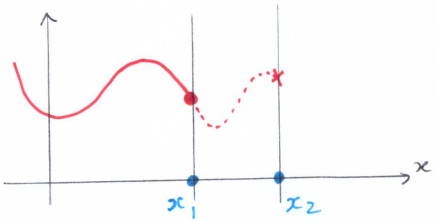
(e) Going back to the first derivative, we obtain the quadratic expression

$\frac{\alpha_1}{2} x^2 + \alpha_0 x + \alpha_{-1}$

↑ Free ↖ ↗ Fixed



(f) And going back to the regression function, we see that the only degree of freedom is in the coefficient in front of the cubic term.
⇒ Introduce $f_5(x) = (x - x_1)_+^3$



(g) Similarly for the right-most region, we need to introduce $f_6(x) = (x - x_2)_+^3$.

⇒ A total of 6 basis functions. (⚠ not unique)

↳ We obtain the representation

$$f_n(x) = \sum_{j=1}^{6} \hat{\beta}_j f_j(x),$$

with

$f_1(x) = 1$ $\quad$ $f_3(x) = x^2$ $\quad$ $f_5(x) = (x - x_1)_+^3$

$f_2(x) = x$ $\quad$ $f_4(x) = x^3$ $\quad$ $f_6(x) = (x - x_2)_+^3$.

A piecewise-cubic polynomial with continuous first & second derivatives at the knots is known as a CUBIC SPLINE.

Space of order p-splines is a vector space. Denote it $S_{p,\ell,x}$ (↑ sequence of ℓ knots)

→ Generalizing: An ORDER p-SPLINE is

(i) a piecewise polynomial of order $(p-1)$

(ii) with continuous derivatives up to order $(p-2)$.

[A cubic spline is an order 4- spline]

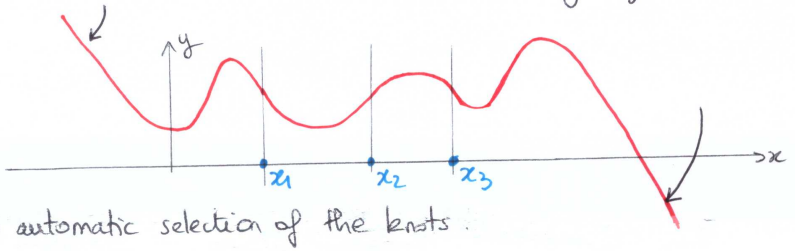
Basis functions for an order p- spline with ℓ knots is

$$\begin{cases} f_j(x) = x^{j-1}, & j = 1, \dots, p \\ f_{p+k}(x) = (x - x_k)_+^{p-1}, & k = 1, \dots, \ell, \end{cases}$$

for a total of $(p+\ell)$ basis functions. And so, can show that given the f_j / f_{p+k}, the decomposition of f is unique.

⊕ . Smoothness

⊖ . Potential wild behaviour at the boundary regions.



. No automatic selection of the knots.

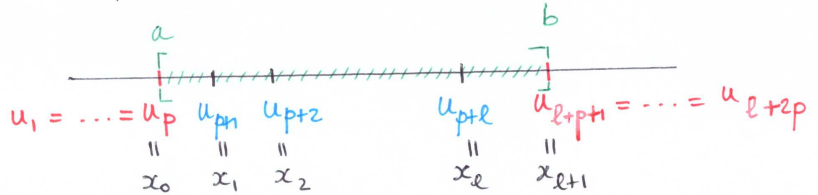
Remark = The basis functions on page 7 do not provide an efficient spline representation for numerical calculations: because the supports of the basis functions are unbanded, we require the evaluation of almost all basis functions at a new input point x, which can be computationally demanding as the number of knots increases (and as we move away from univariate problems). We introduce next a basis where the support of the basis functions is bounded: the B-SPLINE basis.

Consider ℓ knots $x_1, \dots, x_\ell$, $x_1 < \dots < x_\ell$, and two boundary knots $[x_0, x_{\ell+1}] = [a, b]$ (↑ We are looking for a decomposition of f_n on the interval $[a, b]$.)

We augment the knot sequence by introducing a sequence $\{u\}$:

$$\begin{cases} u_1 \le u_2 \le \dots \le u_p \le x_0 \\ u_{j+p} = x_j, & j = 1, \dots, \ell \\ x_{\ell+1} \le u_{\ell+p+1} \le \dots \le u_{\ell+2p} \end{cases}$$

← It is customary to take $u_1, \dots, u_p$ all equal to x_0, and all $u_{\ell+p+1}, \dots, u_{\ell+2p}$ equal to $x_{\ell+1}$.



We define recursively the sequence $B_{i,m}(x)$ = i-th B-spline basis function of order m for the knot family $\{u\}$, with $m \le p$.

Specifically:

$$B_{i,1}(x) = \begin{cases} 1 & \text{if } u_i \le x < u_{i+1} \\ 0 & \text{otherwise} \end{cases}$$

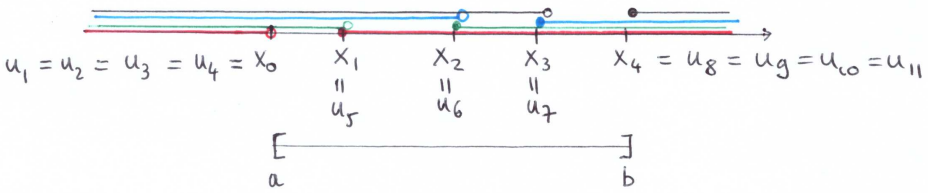
$i = 1, \dots, \ell + 2p - 1$

$$B_{i,m}(x) = \frac{x - u_i}{u_{i+m-1} - u_i} B_{i,m-1}(x) + \frac{u_{i+m} - x}{u_{i+m} - u_{i+1}} B_{i+1,m-1}(x)$$

$i = 1, \dots, \ell + 2p - m$

We illustrate the B-spline basis for $p = 4$, $\ell = 3$. (3 knots x_1, x_2, x_3)

- B-splines of order 1 (discontinuous at interior knots)



- B-splines of order 2 (continuous at interior knots)

$$B_{i,2}(x) = \frac{x - u_i}{u_{i+1} - u_i} B_{i,1}(x) + \frac{u_{i+2} - x}{u_{i+2} - u_{i+1}} B_{i+1,1}(x)$$

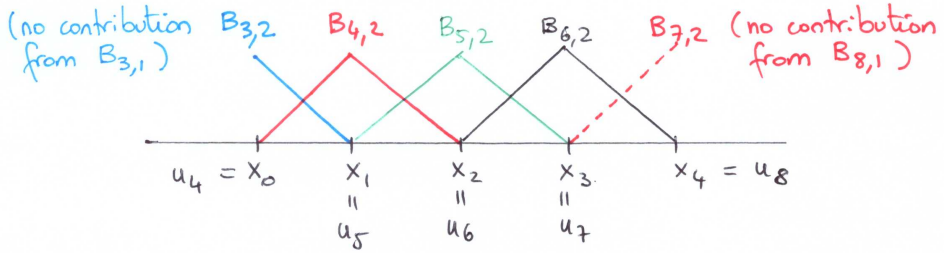
$i = 1, \dots, \ell + 2p - 2$

non-zero, constant on $[u_i, u_{i+1})$

non-zero, constant on $[u_{i+1}, u_{i+2})$

$\Rightarrow B_{i,2}(x)$ is non-zero on $[u_i, u_{i+2})$?
(aka the "support" of $B_{i,2}$)

Putting the contributions from $B_{i,1}$ and $B_{i+1,1}$ together:



$B_{i,2}(x)$ is continuous at interior knots
non-differentiable at the knots. $\Rightarrow B_{i,2} \in C^0$

- B-splines of order 3. (differentiable at interior knots)

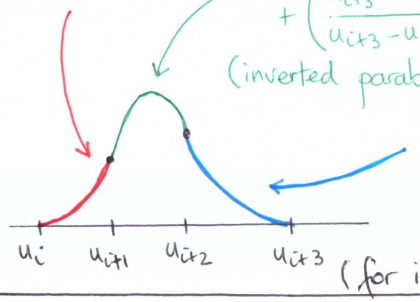
$$B_{i,3}(x) = \frac{x - u_i}{u_{i+2} - u_i} B_{i,2}(x) + \frac{u_{i+3} - x}{u_{i+3} - u_{i+1}} B_{i+1,2}(x)$$

$$= \left(\frac{x - u_i}{u_{i+2} - u_i}\right)\left(\frac{x - u_i}{u_{i+1} - u_i}\right) B_{i,1}(x) \quad \text{(parabola)}$$

$$+ \left(\frac{x - u_i}{u_{i+2} - u_i}\right)\left(\frac{u_{i+2} - x}{u_{i+2} - u_{i+1}}\right) B_{i+1,1}(x)$$

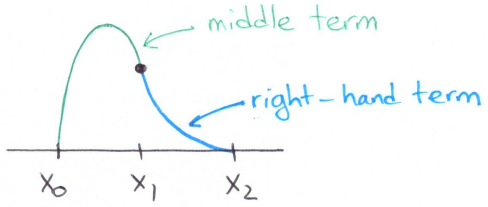
$$+ \left(\frac{u_{i+3} - x}{u_{i+3} - u_{i+1}}\right)\left(\frac{x - u_{i+1}}{u_{i+2} - u_{i+1}}\right) B_{i+1,1}(x) \quad \text{(inverted parabola)}$$

$$+ \left(\frac{u_{i+3} - x}{u_{i+3} - u_{i+1}}\right)\left(\frac{u_{i+3} - x}{u_{i+3} - u_{i+2}}\right) B_{i+2,1}(x) \quad \text{(parabola)}$$

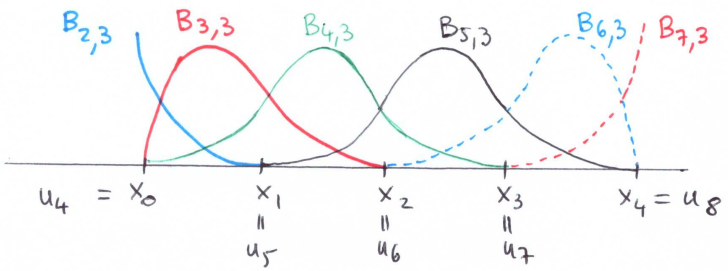


(for interior knots)

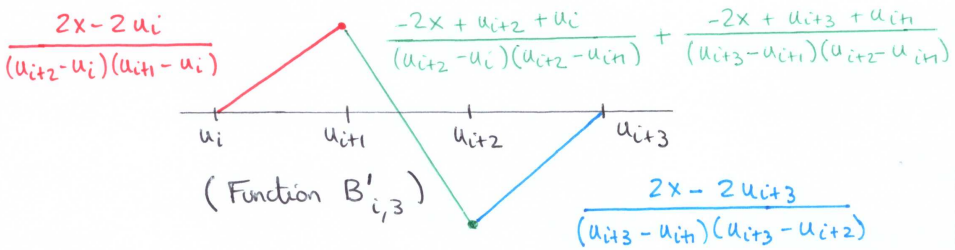
At be boundaries:



Summarizing:



The functions $B_{i,3}$ are differentiable at the interior knots only. At interior knots, we have the following picture:

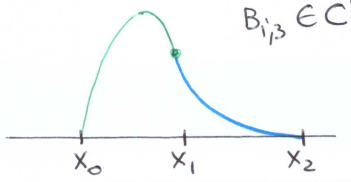


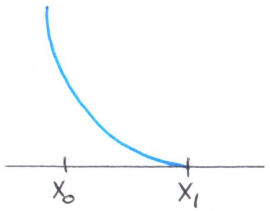
$\Rightarrow$ $B_{i,3}$ is differentiable
$B'_{i,3}$ is continuous and not differentiable at the knots.

$\Rightarrow$ $B_{i,3} \in C^1$ at interior knots.

At boundary knots,

$B_{i,3}$ is continuous,
& not differentiable at x_0

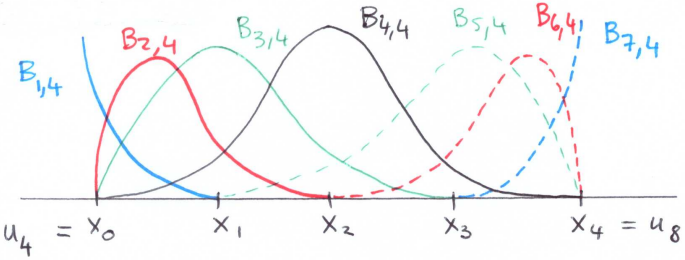




$B_{i,3}$ is discontinuous at x_0
$\Rightarrow$ $B_{i,3} \in C^{-1}$

- B-splines of order 4. (C^2 at interior knots).

We sketch the basis functions, without going through details.



A total of $p+\ell = 7$ functions

↳ We see that $B_{4,4} \in C^2$
$B_{3,4}$ / $B_{5,4} \in C^1$
$B_{2,4}$ / $B_{6,4} \in C^0$
$B_{1,4}$ / $B_{7,4} \in C^{-1}$

In fact, it is possible to show that the family $\{B_{i,4}\}$ provides a basis function for cubic-splines [i.e. an order-4 spline].

More generally, $\{B_{i,p}\}$ = basis function for order-p splines. [Theorem 14.1 in Györfi et al (2002)]: we have a unique representation:

$$f(x) = \sum_{i=1}^{p+\ell} a_i B_{i,p}(x), \quad f \in S_{p,\ell,x}$$

↳ defined on page 7

I.3. Consistency of the spline estimates.

In this section, we address the problem of knot + spline order selection in order to get universally consistent LS estimates. The results presented here are not completely rigorous, and we refer the reader to Section 14.2 in Györfi et al (2002) for precise statements.

For $n \geq 1$, put $p_{max}(n) \geq 1$ = max order of the p-spline
$l_{max}(n) \geq 1$ = max number of knots.

Depending on the learning sample $\mathcal{L}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$, select $p \leq p_{max}(n)$
$l \leq l_{max}(n)$
& a sequence of l knots $\xi_1, \dots, \xi_l$,
and compute the LS estimate over the space $\mathcal{F}_n = S_{p,l,\xi}$ of p-order splines with this sequence of knots:

$$\tilde{f}_n = \underset{f \in S_{p,l,\xi}}{\operatorname{argmin}} \sum_{i=1}^{n} (Y_i - f(X_i))^2$$

For technical reasons, one needs to consider the truncation of the LS estimate to obtain consistency results:

$$f_n(x) = \begin{cases} \tilde{f}_n(x) & \text{if } |\tilde{f}_n(x)| < \beta_n \\ \beta_n \operatorname{sign}(\tilde{f}_n(x)) & \text{otherwise} \end{cases}$$

β_n = a sequence of positive numbers diverging to $+\infty$.

Assume that, as $n \to +\infty$,

(i) $\beta_n \to +\infty$

(ii) $\beta_n^4 / n^{1-\delta} \to 0$, for some $\delta > 0$

(iii) $\dfrac{(p_{max}(n)\, l_{max}(n) + l_{max}^2(n))\, \beta_n^4 \log n}{n} \longrightarrow 0$

In addition, we assume that the distribution $\mathbb{P}_X$ of X satisfies:

[C] $$\mathbb{P}_X\left[\left\{(-\infty, u_p) \cup \bigcup_{\substack{i=1,\dots,l \\ (u_{pti} - u_i) > \delta}} [u_{pti-1}, u_{pti}) \cup [u_{ptl}, \infty)\right\} \cap [-L, L]\right] \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty$$

for each $\delta, L > 0$.

augmented set of knots to construct the B-spline basis, see page 8.

Then, for $\mathbb{E}\, Y^2 < \infty$, [Theorem 14.2 in Györfi et al].

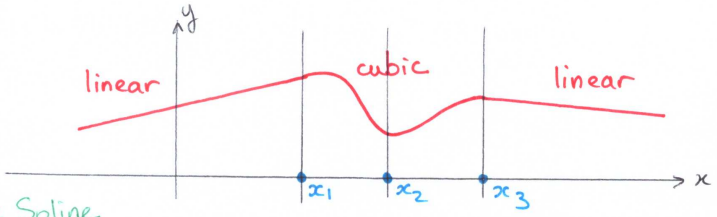
$$\varepsilon(f_n) = \int (f_n(x) - r(x))^2\, \mathbb{P}_X(dx) \xrightarrow{a.s.} 0 \quad \text{as } n \to \infty$$

And so, if one can choose the degree p and knot sequence in such a way that condition [C] holds for every distribution $\mathbb{P}_X$ of X, we obtain strong universal consistency.

Remark: Conditions for weak universal consistency, together with examples under which [C] holds $\forall\, \mathbb{P}_X$ are discussed in Györfi et al (2002), p. 268-270.

I.4 Natural Cubic Splines.

A natural cubic spline extrapolates linearly in the terminal regions :



Cubic Spline

- $f(x) = \sum_{j=0}^{3} \beta_j x^j + \sum_{i=1}^{\ell} \gamma_i (x - x_i)_+^3$ (ℓ knots $x_1, \dots, x_\ell$)

$= \beta_0 + \beta_1 x + \sum_{i=1}^{\ell} \gamma_i (x - x_i)_+^3$ ← linear for $x \leq x_1$

+ additional constraints on the γ_i since for $x \geq x_\ell$, $f''(x) = 0$ and $f'''(x) = 0$.

⇒ We are loosing a total of 4 degrees of freedom ; ie 4 parameters ⇒ we need ℓ basis functions to represent a Natural Cubic Spline (NCS) ≡ number of knots.

- Let f be a NCS interpolating the points (knots) $x_1, \dots, x_\ell \in [a, b]$, for $a < x_1 \leq x_2 \leq \dots \leq x_\ell < b$.

The value of f at the knots is denoted $y_i := f(x_i)$.

In addition, put $\gamma_i := f''(x_i)$, $i = 1, \dots, \ell$. Note that necessarily $\gamma_1 = \gamma_\ell = 0$.

Let $y := (y_1, \dots, y_\ell)^t$

$\gamma := (\gamma_1, \dots, \gamma_\ell)^t$.

We show that it is enough to know the value of the NCS at the ℓ knots to know its value at any x.

To do so, we define (Green & Silverman)

- If $x_i \leq x \leq x_{i+1}$, $i = 1, \dots, \ell - 1$

$$f(x) := \frac{(x - x_i) y_{i+1} + (x_{i+1} - x) y_i}{\Delta_i} - \frac{1}{6}(x - x_i)(x_{i+1} - x)\left\{\left(1 + \frac{x - x_i}{\Delta_i}\right)\gamma_{i+1} + \left(1 + \frac{x_{i+1} - x}{\Delta_i}\right)\gamma_i\right\}$$

← cubic

$\Delta_i := x_{i+1} - x_i$

- If $x \leq x_1$

$$f(x) := y_1 - (x_1 - x) f'(x_1)$$

- If $x \geq x_\ell$

$$f(x) := y_\ell + (x - x_\ell) f'(x_\ell)$$

← linear

↑ We show that under some additional constraints on y and γ, the function f is a NCS

→ linear in the terminal regions, cubic elsewhere

→ $\forall i = 1, \dots, \ell$, $f(x_i) = y_i$

Moreover, for $x \in [x_i, x_{i+1}]$,

$$f'(x) = \frac{y_{i+1} - y_i}{\Delta_i} - \frac{1}{6}\left\{(x - x_i)(x_{i+1} - x)\left(\frac{\gamma_{i+1} - \gamma_i}{\Delta_i}\right) + \left[\left(1 + \frac{x - x_i}{\Delta_i}\right)\gamma_{i+1} + \left(1 + \frac{x_{i+1} - x}{\Delta_i}\right)\gamma_i\right](x_i + x_{i+1} - 2x)\right\}$$

&

$$f''(x) = \frac{\gamma_i (x_{i+1} - x) + \gamma_{i+1}(x - x_i)}{\Delta_i}$$

⇒ $\forall i = 1, \dots, \ell$, $f''(x_i) = \gamma_i$.

→ It remains to establish continuity of f' at the knots:

$$f'(x_i^+) = \frac{y_{i+1}-y_i}{\Delta_i} - \frac{1}{6}\Delta_i(\gamma_{i+1}+2\gamma_i) \quad \text{——— (1)}$$

$$f'(x_{i+1}^-) = \frac{y_{i+1}-y_i}{\Delta_i} + \frac{1}{6}\Delta_i(2\gamma_{i+1}+\gamma_i) \quad \text{——— (2)}$$

We want $f'(x_i^-) = f'(x_i^+)$, that is:

$$\Rightarrow \underbrace{\frac{y_{i+1}-y_i}{\Delta_i} - \frac{1}{6}\Delta_i(\gamma_{i+1}+2\gamma_i)}_{f'(x_i^+) \text{ from (1)}} = \underbrace{\frac{y_i-y_{i-1}}{\Delta_{i-1}} + \frac{1}{6}\Delta_{i-1}(\gamma_{i-1}+2\gamma_i)}_{\text{Replace } i \text{ by } (i-1) \text{ in (2)}}$$

$$\downarrow$$

$$\frac{y_{i+1}-y_i}{\Delta_i} - \frac{y_i-y_{i-1}}{\Delta_{i-1}} = \frac{1}{6}\Delta_{i-1}\gamma_{i-1} + \frac{1}{3}\gamma_i(\Delta_{i-1}+\Delta_i) + \frac{1}{6}\Delta_i\gamma_{i+1}$$

↑ Collect all y terms together ↑ Collect all γ terms together

⇒ In matrix form:

$$\begin{pmatrix} \frac{1}{\Delta_{i-1}} & -\left(\frac{1}{\Delta_{i-1}}+\frac{1}{\Delta_i}\right) & \frac{1}{\Delta_i} \end{pmatrix} \begin{pmatrix} y_{i-1} \\ y_i \\ y_{i+1} \end{pmatrix} = \begin{pmatrix} \frac{\Delta_{i-1}}{6} & \frac{\Delta_{i-1}+\Delta_i}{3} & \frac{\Delta_i}{6} \end{pmatrix} \begin{pmatrix} \gamma_{i-1} \\ \gamma_i \\ \gamma_{i+1} \end{pmatrix}$$

This relation holds for $i = 2, \dots, \ell-1$.

We end up with a system

$$\boxed{\underline{\underline{Q}}^t \underline{y} = \underline{\underline{R}}\,\underline{\gamma}},$$

Relation that y and γ must satisfy to ensure continuity of f' at the ℓ knots

where • $\underline{\underline{Q}}$ is an $\ell \times (\ell-2)$ matrix

• $\underline{\underline{R}}$ is an $(\ell-2)\times(\ell-2)$ matrix.
(not $\ell\times\ell$ since $\gamma_1 = \gamma_\ell = 0$)

Q and R are given on the next page.

$$\underset{\ell \times (\ell-2)}{\underline{\underline{Q}}} = \begin{pmatrix} q_{12} & & \\ q_{22} & \ddots & \\ q_{32} & & q_{\ell-2,\ell-1} \\ & \ddots & q_{\ell-1,\ell-1} \\ & & q_{\ell,\ell-1} \end{pmatrix}$$

→ $q_{j-1,j} = \frac{1}{\Delta_{j-1}}$

→ $q_{j,j} = -\left(\frac{1}{\Delta_{j-1}} + \frac{1}{\Delta_j}\right)$

↳ $q_{j+1,j} = \frac{1}{\Delta_j}$

$$\underset{(\ell-2)\times(\ell-2)}{\underline{\underline{R}}} = \begin{pmatrix} r_{22} & r_{23} & \\ r_{32} & \ddots & \\ & \ddots & r_{\ell-2,\ell-1} \\ & r_{\ell-1,\ell-2} & r_{\ell-1,\ell-1} \end{pmatrix}$$

→ $r_{i,i+1} = \frac{\Delta_i}{6}$

→ $r_{i,i} = \frac{1}{3}(\Delta_{i-1}+\Delta_i)$

↳ $r_{i+1,i} = \frac{\Delta_i}{6}$ $\quad (i = 2, \dots, \ell-1)$

↳ R is symmetric & strictly diagonally dominant (since $|r_{ii}| > \sum_{j\neq i} |r_{ij}| \ \forall i$) ⇒ One can show that R is strictly positive definite; we can define its inverse R^{-1}.

Theorem: (i) y and γ specify a NCS $\Leftrightarrow$ $Q^t y = R\gamma$

(ii) If this is satisfied, then

$$\int_a^b |f''(x)|^2 dx = \gamma^t R \gamma = y^t K y \ ; \quad K := QR^{-1}Q^t$$

proof (i) Follows from the previous derivation.

(ii) $\int_a^b |f''(x)|^2 dx = \underbrace{\left[f''(x) f'(x)\right]_a^b}_{=0 \text{ since } f''(a) = f''(b) = 0} - \int_a^b f'(x) f'''(x)\, dx$

$= -\int_a^b f'(x) f'''(x)\, dx$

since the 3rd order derivative vanishes on the two regions $[a, x_1]$ and $[x_\ell, b]$.

$= -\sum_{j=1}^{\ell-1} \int_{x_j}^{x_{j+1}} f'(x) f'''(x)\, dx$

f''' constant on $[x_j, x_{j+1}]$, and equal to $f'''(x_j^+) = \frac{\gamma_{j+1} - \gamma_j}{\Delta_j}$.

$= -\sum_{j=1}^{\ell-1} f'''(x_j^+) \int_{x_j}^{x_{j+1}} f'(x)\, dx$

$= \sum_{j=1}^{\ell-1} \frac{\gamma_{j+1} - \gamma_j}{\Delta_j} (y_j - y_{j+1})$

since $\gamma_1 = 0$

$= \sum_{j=2}^{\ell-1} \gamma_j \left(\frac{y_{j+1} - y_j}{\Delta_j} - \frac{y_j - y_{j-1}}{\Delta_{j-1}} \right)$

$= \gamma^t \underbrace{Q^t y}_{= R\gamma}$

$= \gamma^t R \gamma$

$\gamma = R^{-1} Q^t y$, $\gamma^t = y^t Q R^{-1}$

$= y^t (Q R^{-1} Q^t) y$

$= y^t K y$. ∎

Remark: Once y is known, γ is as well since R is invertible and $\gamma = R^{-1} Q^t y$

- K is a $(\ell \times \ell)$ symmetric matrix, pos. semi definite

Theorem: The NCS has the minimum value of $\int |f''(x)|^2 dx$ among all smooth curves interpolating the data $\{x_i, y_i\}$, $i = 1, \dots, \ell$. Specifically, for $\ell \geq 2$,

- f = NCS interpolating values $y_1, \dots, y_\ell$ at $x_1, \dots, x_\ell$.
- $\tilde{f}$ = any function on $[a, b]$ s.t. $\tilde{f}(x_i) = y_i$ (twice continuously differentiable)

$\int_a^b |f''(x)|^2 dx \leq \int_a^b |\tilde{f}''(x)|^2 dx$.

proof: Put $h := \tilde{f} - f$.

Then $\rightarrow h(x_i) = 0$ $(i = 1, \dots, \ell)$

$\rightarrow f''(a) = f''(b) = 0$

- $\int_a^b f''(x) h''(x)\, dx = \left[f''(x) h'(x)\right]_a^b - \int_a^b h'(x) f'''(x)\, dx$

$= -\int_a^{x_\ell} h'(x) f'''(x)\, dx$

$= -\sum_{j=1}^{\ell-1} f'''(x_j^+) \int_{x_j}^{x_{j+1}} h'(x)\, dx$

$= -\sum_{j=1}^{\ell-1} f'''(x_j^+) \left(h(x_{j+1}) - h(x_j) \right) = 0$ ($h(x_{j+1}) = 0$, $h(x_j) = 0$)

- $\int_a^b |\tilde{f}''(x)|^2 dx = \int_a^b \left(f''(x) + h''(x) \right)^2 dx$

$= \int_a^b |f''(x)|^2 dx + \int_a^b |h''(x)|^2 dx + 0$

$\geq \int_a^b |f''(x)|^2 dx$.

Equality holds iff $\int |h''(x)|^2 dx = 0$; i.e. if h is linear on $[a, b]$. Since h is zero at $x_1, \dots, x_\ell$, $\ell \geq 2$, this can only happen if $h \equiv 0$; i.e. if $f = \tilde{f}$. ∎

II- SMOOTHING SPLINES.

Consider the following non-parametric regression problem:
Among all functions f twice continuously differentiable, find the one that minimizes the penalized sum of squares:

$$RSS(f, \lambda) = \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b |f''(x)|^2 dx .$$

goodness of fit term — roughness penalty

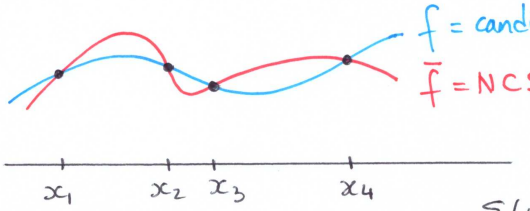
- Penalty term $\int |f''(x)|^2 dx$: the second order derivative picks all all the wiggles of f ; the square gets rid of the sign, and the integral sums it all up. All non-linearities are added into the penalty term ; so that functions with small second order derivatives are preferred. The tuning parameter λ regulates the relative importance of the goodness of fit term, and the penalty.
- $\lambda = 0$: solution is any function interpolating all data points.
- $\lambda = \infty$: solution is linear, as introducing non-linearities result in an infinite cost.

Put $\hat{f} \in \underset{f \in \mathcal{F}}{\operatorname{argmin}} \; RSS(f, \lambda)$

$\mathcal{F}$: space of twice continuously differentiable functions. (∞-dim)

We are looking for a solution in an ∞-dim space.

Let
- f be our current candidate solution. (We are looking for another function reducing further the penalized RSS)
- $\bar{f}$ be a NCS interpolating the values $f(x_i)$ of the candidate solution.



$$\Rightarrow \sum (y_i - f(x_i))^2 = \sum (y_i - \bar{f}(x_i))^2$$

The optimal property of NCSs ensure that (with equality only if f is a NCS itself)

$$\int |\bar{f}(x)|^2 dx < \int |f(x)|^2 dx .$$

Since $\lambda > 0$, we conclude that $RSS(\bar{f}, \lambda) < RSS(f, \lambda)$.

$\Rightarrow$ Unless f is itself a NCS, we can find a NCS which attains a smaller value of the penalized sum of squares.

$\Rightarrow$ The minimizer fn of $RSS(f, \lambda)$ must be a NCS.

finite dimensional : it is a NCS with knots at $x_1, \dots, x_n$ ($\not\Rightarrow$ n degrees of freedom, due to the presence of the penalty term).

For $\mathcal{L}_n = \{(x_1, y_1), \dots, (x_n, y_n)\}$, put $x = (x_1, \dots, x_n)^t$, $y = (y_1, \dots, y_n)^t$,

and $f = (f(x_1), \dots, f(x_n))^t \in \mathbb{R}^n$

The RSS criterion can be rewritten see page 14

$$RSS(f, \lambda) = (y - f)^t (y - f) + \lambda f^t K f$$

$$RSS(f,\lambda) = f^t(I+\lambda K)f - 2f^t y + y^t y$$

$$\frac{\partial RSS(f,\lambda)}{\partial f} = 2\hat{f}(I+\lambda K) - 2y = 0$$

($\hat{f}(I+\lambda K)$: strictly positive definite)

aka a SMOOTHING SPLINE

$$\hat{f} = (I+\lambda K)^{-1} y =: S_\lambda y$$

= solution is a linear estimator

Gives the value of the NCS at the n knots, which uniquely defines the NCS for all x, see p. 9.

- Analysis of the solution.

The matrix $S_\lambda := (I+\lambda K)^{-1}$ is referred to as the SMOOTHER MATRIX. It is symmetric (since $K = QR^{-1}Q^t$ is symmetric), and admits the spectral decomposition

$$S_\lambda = U \Lambda U^t = \sum_{i=1}^{n} \lambda_i u_i u_i^t, \text{ where}$$

(nxn)

$$U = \begin{pmatrix} | & & | \\ u_1 & \cdots & u_n \\ | & & | \end{pmatrix} \qquad \Lambda = \begin{pmatrix} \lambda_1 & & 0 \\ & \ddots & \\ 0 & & \lambda_n \end{pmatrix}$$

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n > 0$$

orthonormal basis of $\mathbb{R}^n$.

We show that both S_λ and K have the same eigenvectors, and thus that the eigenvectors of S_λ do not depend on λ.

↳ An eigenvalue μ of K is such that $\det(K-\mu I) = 0$.

Since

$$\det(K-\mu I) = \det\left(\frac{1}{\lambda}\left[(I+K\lambda) - (1+\lambda\mu)I\right]\right),$$

$1+\lambda\mu$ must be an eigenvalue of $(I+K\lambda)$.

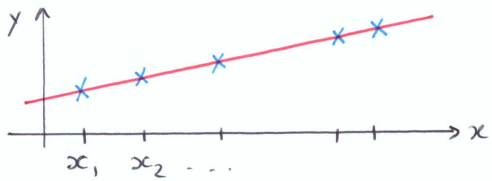
$\Rightarrow (1+\lambda\mu)^{-1}$ must be an eigenvalue of $S_\lambda = (I+K\lambda)^{-1}$.

In the representation

$$\hat{f} = \sum_{j=1}^{n} \frac{1}{1+\lambda\mu_{n-j+1}} \langle u_j, y\rangle u_j = \sum_{j=1}^{n} \lambda_j \langle u_j, y\rangle u_j,$$

we have $\lambda_1 = \lambda_2 = 1$.

Indeed, once the values $x_1, \ldots, x_n$ are observed, the matrices Q, R and thus K are fixed, so that the eigenvalues $\lambda_1, \ldots, \lambda_n$ do not depend on the response variables $y_1, \ldots, y_n$. Suppose now that the relationship between x_i and y_i is perfectly linear.



The solution to the smoothing spline problem is linear, since a linear fit interpolates perfectly the data, with a 0 penalty term $\Rightarrow$ the associated penalized RSS is 0. We conclude that $\hat{f} = y = S_\lambda y$, so that $\lambda_1 = 1$ indeed. In addition, since y is the associated eigenvector, we deduce that the eigenvector associated with $\lambda_1 = 1$ varies linearly with x. Moreover, to represent y as a linear combination of x, we have two degrees of freedom (slope + intercept), so that it can be expressed as a linear combination of two linearly independent vectors. Since the eigenvectors of S_λ are orthogonal, the eigenvalue 1 must have multiplicity 2.

$\Rightarrow \lambda_1 = \lambda_2 = 1$

u_1 & u_2 = linear functions of x.

For some index j, we must have $\lambda_j = (1+\lambda\mu)^{-1}$; a decreasing function of μ.

Ordering the eigenvalues of K: $\mu_1 \geq \dots \geq \mu_n \geq 0$, we conclude that $\lambda_j = \dfrac{1}{1+\lambda\mu_{n-j+1}}$, with associated eigenvector u_j: (necessarily less (or equal) to 1.)

$$S_\lambda u_j = \lambda_j u_j$$
$$\Leftrightarrow \quad u_j = \lambda_j S_\lambda^{-1} u_j$$
$$= \lambda_j (I + \lambda K) u_j$$
$$= \lambda_j u_j + \lambda \lambda_j K u_j$$

$$\Rightarrow K u_j = \frac{1-\lambda_j}{\lambda\lambda_j} u_j = \mu_{n-j+1} u_j$$

$\Rightarrow$ u_j is an eigenvector of K corresponding to the $(n-j+1)$ largest eigenvalue. For all $\lambda > 0$, the eigenvectors of S_λ are eigenvectors of K, and thus do not depend on λ.

We can re-express the NCS solution as:

$$\hat{f} = S_\lambda y = \sum_{j=1}^{n} \lambda_j u_j u_j^t y = \sum_{j=1}^{n} \lambda_j \underbrace{\langle u_j, y\rangle}_{\text{independent of } \lambda} u_j$$

$$\hat{f} = \sum_{j=1}^{n} \frac{1}{1+\lambda\mu_{n-j+1}} \langle u_j, y\rangle u_j$$

- The larger λ, the more severe the shrinkage
- shrinkage factor in the u_j direction.
- The larger j, the smaller λ_j, and the more shrinkage in the u_j direction.

(compare $\hat{f}$ with the ridge solution in SL: RR AND LASSO)

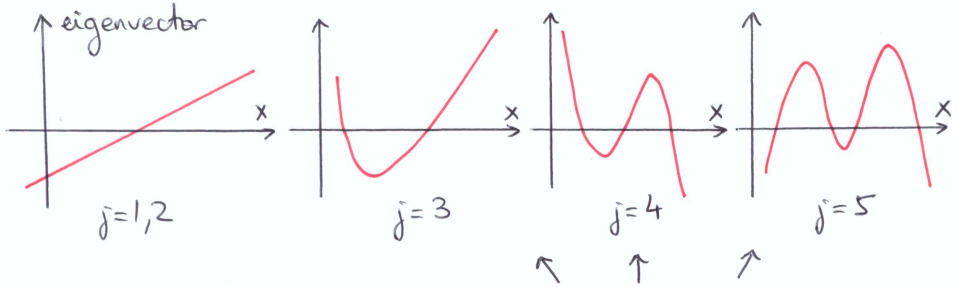
We have the decomposition

$$\hat{f} = \langle u_1, y\rangle u_1 + \langle u_2, y\rangle u_2 + \sum_{j=3}^{n} \lambda_j \langle u_j, y\rangle u_j.$$

the linear component of the solution are not shrunk (as you would expect)

Only the components in the u_j $(j \geq 3)$ directions are shrinked.

Demler & Reinsh (1975) proved that for $j \geq 3$, the number of sign changes in the j-th eigenvector of a cubic spline smoother is $(j-1)$. Eigenvectors thus behave as:



The more wiggly the eigenvectors, the more they are penalized. (coefficient λ_j is smaller).

Remarks:

(i) As in the case of Ridge Regression & Lasso, we may define a "degree of freedom" coefficient which accounts for the correlation of the response variable y with the fitted values $\hat{f}$. For linear predictors, this amounts to computing the trace of the matrix H, where $\hat{y} = Hy$.

$$\Rightarrow df(\lambda) := \mathrm{Tr}(S_\lambda) = \sum_{j=1}^{n} \lambda_j = \sum_{j=1}^{n} (1+\lambda\mu_j)^{-1} \quad (\leq n)\ (\geq 2)$$