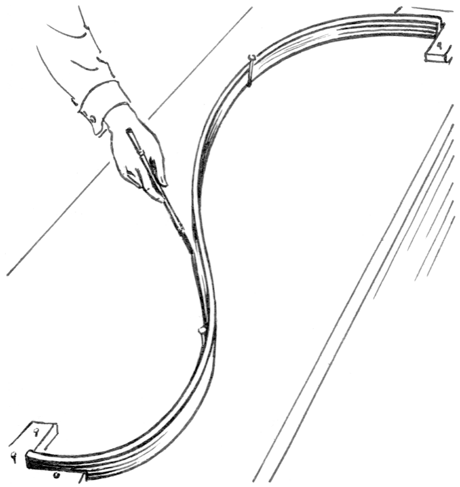
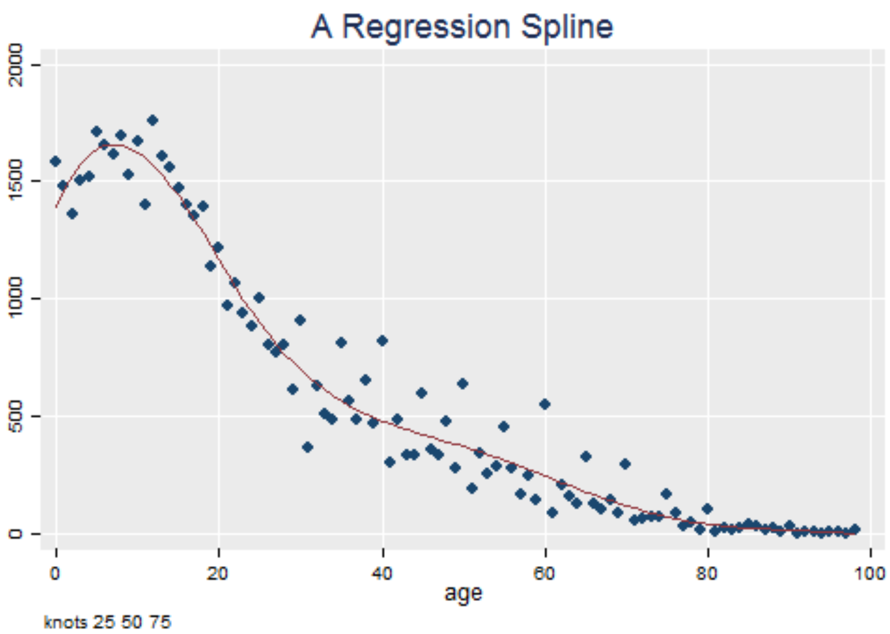
Smoothing Spline and Kernel Method

Ref.: Fasshauer and Ye, Reproducing kernels of Sobolev spaces via a green kernel approach with differential operators and boundary operators, 2013

- Spline regression: smooth curve connecting observed points.





data : $(x_1, y_1), \ldots, (x_n, y_n) \in (0,1) \times \mathbb{R}$.

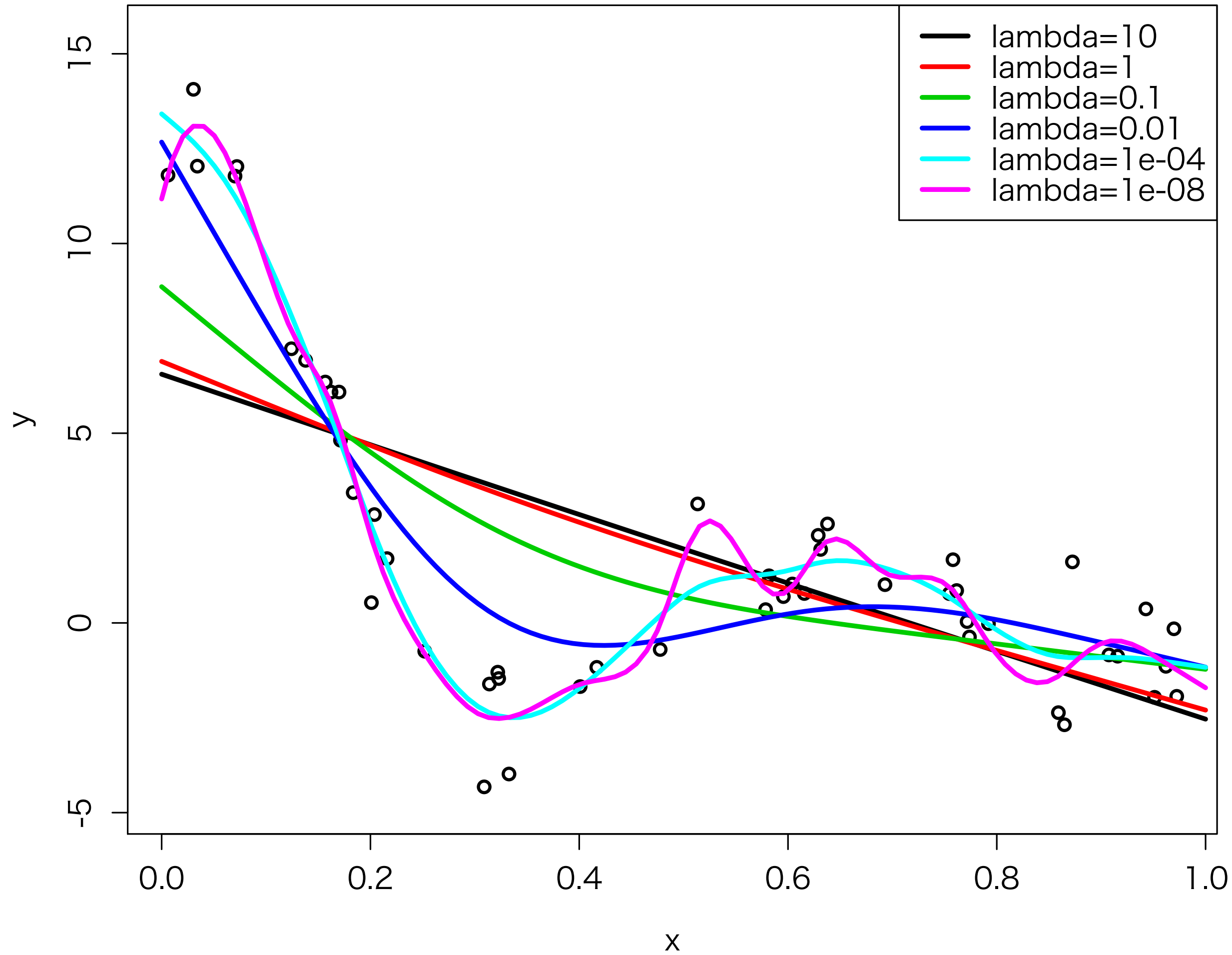
model : $\quad y_i = f(x_i) + \varepsilon_i, \quad i = 1, \ldots, n$

- data fitting: $\displaystyle\sum_{i=1}^{n} (y_i - f(x_i))^2$
- smoothness: $\displaystyle\int_0^1 (f''(x))^2 dx$

Smoothing Spline Regression

Estimator: $\displaystyle\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx \quad \longrightarrow \quad \widehat{f}(x).$

$\lambda \longrightarrow \infty \implies f(x) = ax + b$



$$n = 50,\ y = \frac{\sin(4\pi x)}{x} + N(0,1),\ x \sim U(0,1)$$

Regularization and RKHS

- Our Aim: construct the RKHS $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ such that

$$\langle g, g \rangle = \int_0^1 (g''(x))^2 dx, \quad g \in \mathcal{H}.$$

Then, representor theorem works to compute the estimator.

- Suppose the kernel function of $\mathcal{H}$ is $k(x, x')$. Then,

$$\mathcal{H} = \left\{ g(x) = \sum_i \beta_i k(x_i, x) \,\middle|\, \beta_i \in \mathbb{R},\ x_i \in (0,1) \right\} \text{ \& completion,}$$

$$\text{boundary cond.:} \quad g \in \mathcal{H} \implies g(0) = g'(0) = g(1) = g'(1) = 0 \ \cdots \ (\star)$$

Model of $f(x)$: $f(x) = a + bx + g(x),\ a, b \in \mathbb{R},\ g \in \mathcal{H}$.

- inner product of $\mathcal{H}$: $\langle g_1, g_2 \rangle := \int_0^1 g_1''(x) g_2''(x) dx$.

  $\langle g, g \rangle = \int (g'')^2 dx = 0 \implies g'' = 0 \implies g(x) = a + bx$.

  boundary condition: $g(x) = 0$.

Reproducing property, $\langle g, k_z \rangle = g(z)$, should hold: for $g \in \mathcal{H}$,

$$\begin{aligned}
\langle g, k_z \rangle &= \int_0^1 g''(x) k_z''(x) dx = \Big[g' k_z'' - g k_z''' \Big]_0^1 + \int_0^1 g(x) k_z''''(x) dx \\
&= \int_0^1 g(x) k_z''''(x) dx \\
\Longrightarrow & \int_0^1 g(x) k_z''''(x) dx = g(z) \quad \text{is required.}
\end{aligned}$$

Dirac's delta function:

$$\delta(x) = \begin{cases} 0, & x \neq 0, \\ \infty, & x = 0. \end{cases}, \quad \int_A g(x)\delta(x)dx = \begin{cases} g(0), & 0 \in A, \\ 0, & 0 \notin A. \end{cases}$$

Find $k(x, z) = k_z(x)$ such that

$$\frac{d^4}{dx^4} k_z(x) = \delta(x - z),$$
$$k_z(0) = k'_z(0) = k_z(1) = k'_z(1) = 0 \quad \text{(condition for } k(z, \cdot) \in \mathcal{H})$$

cf. Green function of the differential operator $\dfrac{d^4}{dx^4}$.

Use integral by part. Define $[a]_+ = \max\{a, 0\}$.

$$
\begin{aligned}
k_z(x) &= k(z, x) \\
&= \frac{([x-z]_+)^3}{6} + [\text{cubic poly. of } x] \longleftarrow \text{ such that } (\star) \text{ is satisfied} \\
&= \frac{([x-z]_+)^3}{6} - \frac{x^3}{6}\left(2z^3 - 3z^2 + 1\right) + \frac{x^2}{2}\left(z^3 - 2z^2 + z\right).
\end{aligned}
$$

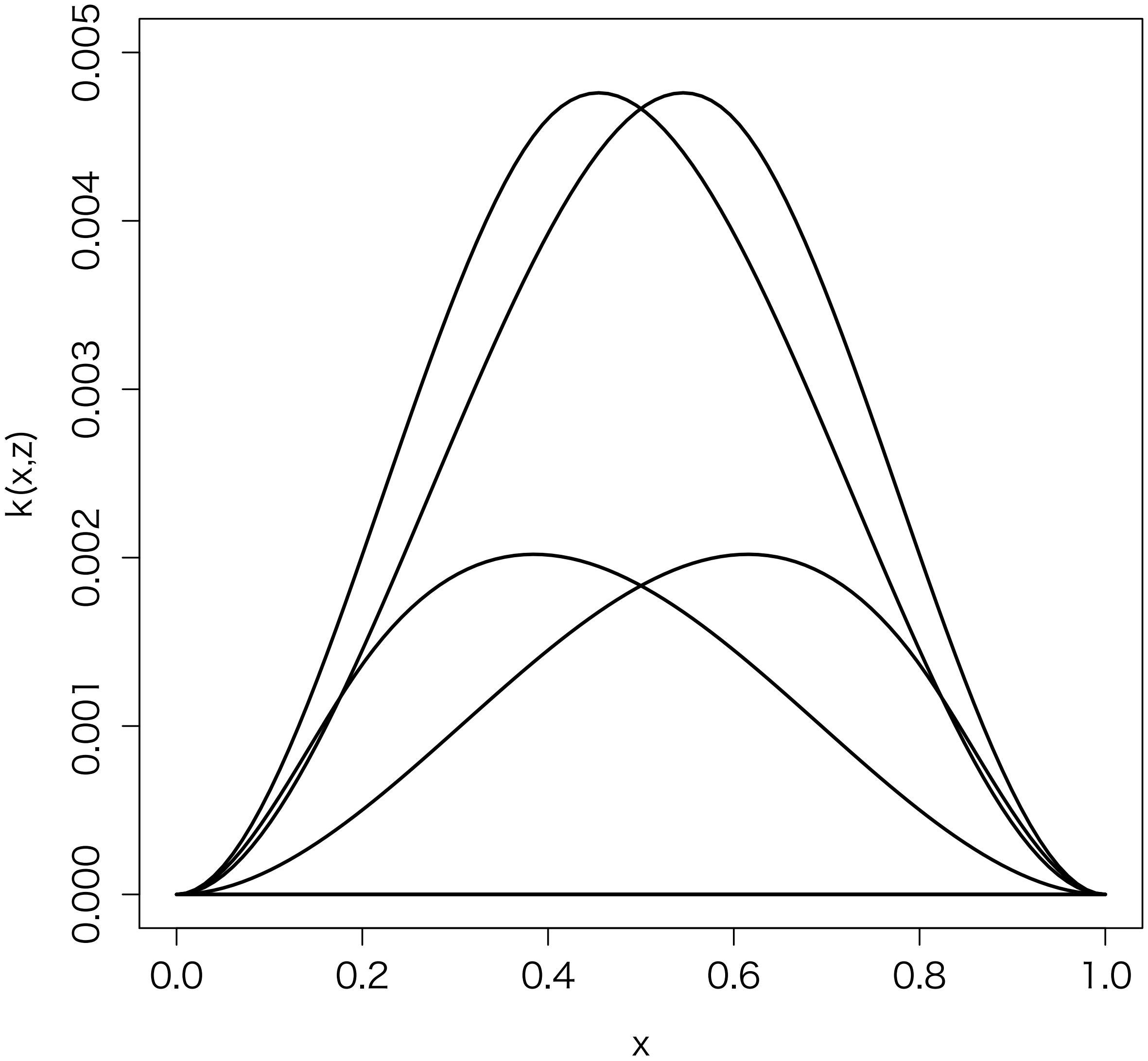
Note that $k(x, z) = k(z, x)$ holds.

Intuitively,

$$
\mathcal{H}_k = \left\{ g : (0,1) \to \mathbb{R} \,\middle|\, g(0) = g'(0) = g(1) = g'(1) = 0,\ g'' \in L_2(0,1) \right\},
$$

$$
\langle g_1, g_2 \rangle = \int_0^1 g_1''(x) g_2''(x) dx.
$$

Plot: $k_z(x), \quad z = 0, 0.2, 0.4, \ldots, 1$



- Estimate by the model $f(x) = a + bx + g(x),\ g \in \mathcal{H}_k$.

  note: $(a+bx)'' = 0$.

representor theorem for smoothing spline

$$\min_f \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_0^1 (f''(x))^2 dx,$$

$$\text{s.t.}\quad f(x) = a + bx + g(x),\ a, b \in \mathbb{R},\ g \in \mathcal{H}_k$$

$$\Longrightarrow\ \text{opt. sol.}:\quad \widehat{f}(x) = a + bx + \sum_{j=1}^{n} \beta_j k(x_j, x)$$

A larger model is available $\rightarrow$ see Assignment-1.

Optimization problem:

$$\min_{a,b,\beta_i} \sum_{i=1}^{n} \left(y_i - a - bx_i - \sum_{j=1}^{n} \beta_j k(x_j, x_i) \right)^2 + \lambda \sum_{i,j} \beta_i \beta_j k(x_i, x_j).$$

Note that for $f(x) = a + bx + g(x),\ g(x) = \sum_j \beta_j k(x_j, x) \in \mathcal{H}_k$,

$$\int_0^1 (f''(x))^2 dx = \int_0^1 (g''(x))^2 dx = \langle g, g \rangle = \sum_{i,j} \beta_i \beta_j k(x_i, x_j).$$

The derivative and integral to calculate $\int_0^1 (f''(x))^2 dx$ are no need.

- d-dimensional expansion: Matérn kernel.

Proof. [representor theorem for smoothing spline]

Define $\mathcal{M} = \mathrm{span}\{k(x_1, \cdot), \dots, k(x_n, \cdot)\}$. Decompose $g \in \mathcal{H}$ into

$$g = g_1 + g_2, \quad g_1 \in \mathcal{M},\ g_2 \in \mathcal{M}^{\perp}.$$

$$\begin{aligned}
\sum_{i=1}^{n} (y_i - f(x_i))^2 &= \sum_{i=1}^{n} (y_i - a - bx_i - \langle k_{x_i},\ g\rangle)^2 \\
&= \sum_{i=1}^{n} (y_i - a - bx_i - \langle k_{x_i},\ g_1\rangle)^2. \\
\int_0^1 (f''(x))^2 dx &= \int_0^1 (g''(x))^2 dx \\
&= \langle g, g\rangle = \langle g_1, g_1\rangle + \langle g_2, g_2\rangle \geq \langle g_1, g_1\rangle.
\end{aligned}$$

∎

Supplementary: completeness of $\mathcal{H}_k$

$$\mathcal{H}_k = \left\{ f : (0,1) \to \mathbb{R} \;\middle|\; \begin{array}{l} \exists f_1, f_2 \text{ s.t. } f(x) = \int_0^x f_1(t)dt, f_1(x) = \int_0^x f_2(t)dt \\ f_2 \in L_2(0,1),\; f(1) = f_1(1) = 0. \end{array} \right\}$$

with the inner product $\langle f, g \rangle = \int_0^1 f_2(x) g_2(x) dx$. In the below, f_1 and f_2 are respectively denoted by f' and f''.

One can prove that $(\mathcal{H}_k, \langle , \rangle)$ is inner product space. Suppose that $\{f_n\} \subset \mathcal{H}_k$ is a Cauchy sequence.

Then, there exists $g \in L_2(0,1)$ s.t. $\|f_n'' - g\|_2 \to 0$. Then, Cauchy-Schwarz inequality leads to $\|f_n'' - g\|_1 \to 0$ (note that the domain $(0,1)$ is bounded). Define $g_1(x) = \int_0^x g(t)dt$ and $g_0(x) = \int_0^x g_1(t)dt$, then $g_0(0) = 0$ and $g_1(0) = 0$ hold. Below, we prove $g_0(1) = g_1(1) = 0$.

Since $\|f_n'' - g\|_1 \to 0$ holds, we have $|g_1(1)| = |f_n'(1) - g_1(1)| \leq \int_0^1 |f_n''(t) - g(t)|dt \to 0$. Hence, $g_1(1) = 0$ holds.

Furthermore, we have $|g_0(1)| = |f_n(1) - g_0(1)| \leq \int_0^1 |f_n'(t) - g_1(t)|dt = \int_0^1 \left| \int_0^t (f_n''(s) - g(s))ds \right| dt \leq \int_0^1 \int_0^t |f_n''(s) - g(s)|dsdt \leq \int_0^1 \int_0^1 |f_n''(s) - g(s)|dsdt = \|f_n'' - g\|_1 \to 0$. Hence, we have $g_0(1) = 0$.