SDS8102: Advanced Mathematical Statistics

Week 13 – Lecture 2
November 26, 2026
Lecturer: Alexandre Tsybakov

---

11 Mean squared error of local polynomial estimator (contd.)

Proposition 5. *Suppose that f belongs to a Hölder class $\Sigma(\beta, L)$ on $[0,1]$, with $\beta > 0$ and $L > 0$. Let $\hat{f}_n$ be the $\mathrm{LP}(\ell)$ estimator of f with $\ell = \lfloor \beta \rfloor$. Assume also that:*

(i) the design points $X_1, \dots, X_n$ are deterministic;

(ii) Assumptions (LP1)–(LP3) hold;

(iii) the random variables ξ_i are independent and such that for all $i = 1, \dots, n$,

$$\mathbb{E}(\xi_i) = 0, \qquad \mathbb{E}(\xi_i^2) \le \sigma_{\max}^2 < \infty.$$

Then for all $n \ge n_0$ and $h \ge 1/(2n)$ the following upper bounds hold:

$$|b(x)| \le q_1 \, h^\beta, \qquad \sigma^2(x) \le \frac{q_2}{nh},$$

where $q_1 = C_* L/\ell!$ and $q_2 = \sigma_{\max}^2 C_*^2$.

Proof. Set for brevity $W_{ni}(x) = W_{ni}^{LP}(x)$. Using Corollary 3 and the Taylor expansion of f we obtain that, for $f \in \Sigma(\beta, L)$,

$$\begin{aligned} b(x) &= \mathbb{E}\big[\hat{f}_n(x)\big] - f(x) = \sum_{i=1}^n f(X_i) \, W_{ni}(x) - f(x) \\ &= \sum_{i=1}^n (f(X_i) - f(x)) W_{ni}(x) \\ &= \sum_{i=1}^n \frac{f^{(\ell)}(x + \tau_i(X_i - x)) - f^{(\ell)}(x)}{\ell!} (X_i - x)^\ell \, W_{ni}(x), \end{aligned}$$

where $0 \le \tau_i \le 1$. This representation and statements (ii) and (iii) of Lemma 4 imply that

$$\begin{aligned} |b(x)| &\le \sum_{i=1}^n \frac{L|X_i - x|^\beta}{\ell!} |W_{ni}(x)| \\ &= L \sum_{i=1}^n \frac{|X_i - x|^\beta}{\ell!} |W_{ni}(x)| I(|X_i - x| \le h) \\ &\le L \sum_{i=1}^n \frac{h^\beta}{\ell!} |W_{ni}(x)| \le \frac{LC_*}{\ell!} \, h^\beta = q_1 \, h^\beta. \end{aligned}$$

The variance satisfies

$$\begin{aligned}\sigma^2(x) &= \mathbb{E}\left[\left(\sum_{i=1}^n \xi_i W_{ni}(x)\right)^2\right] = \sum_{i=1}^n (W_{ni}(x))^2 \mathbb{E}(\xi_i^2) \\ &\leq \sigma^2_{\max} \max_i |W_{ni}(x)| \sum_{i=1}^n |W_{ni}(x)| \leq \frac{\sigma^2_{\max} C_*^2}{nh} = \frac{q_2}{nh}.\end{aligned}$$

Proposition 5 implies that

$$\mathrm{MSE} \leq q_1^2 h^{2\beta} + \frac{q_2}{nh}$$

and that the minimizer h_n^* with respect to h of this upper bound on the risk is given by

$$h_n^* = \left(\frac{q_2}{2\beta q_1^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

Therefore we obtain the following result.

Theorem 6. *Under the assumptions of Proposition 5 and if the bandwidth is $h = \alpha n^{-\frac{1}{2\beta+1}}$ for some $\alpha > 0$, the following upper bound holds:*

$$\sup_{f \in \Sigma(\beta, L)} \mathbb{E}\left[|\hat{f}_n(x) - f(x)|^2\right] \leq C n^{-\frac{2\beta}{2\beta+1}}, \tag{1}$$

where C is a constant depending only on the parameters $\beta, L, \lambda_0, a_0, \sigma^2_{\max}, K_{\max}$, and α.

We now discuss Assumption (LP1) in more detail. If the design is regular and n is large enough, $\mathcal{B}_x$ is close to the matrix $\mathcal{B} = \int U(u)U^T(u)K(u)du$, which is independent of n and x. Therefore, for Assumption (LP1) to hold it is crucial to assure that $\mathcal{B}$ is positive definite. This is indeed true, except for pathological cases, as the following lemma states.

Lemma 5. *Let $K : \mathbb{R} \to [0, \infty)$ be a function such that the Lebesgue measure* $\mathrm{Leb}(u : K(u) > 0) > 0$. *Then the matrix*

$$\mathcal{B} = \int U(u)U^T(u)K(u)du$$

is positive definite.

Proof. It is sufficient to prove that for all $v \in \mathbb{R}^{\ell+1}$ satisfying $v \neq 0$ we have

$$v^T \mathcal{B} v > 0.$$

Clearly,

$$v^T \mathcal{B} v = \int (v^T U(u))^2 K(u) du \geq 0.$$

If there exists $v \neq 0$ such that $\int [v^T U(u)]^2 K(u)\, du = 0$, then $v^T U(u) = 0$ for almost all u on the set $\{u : K(u) > 0\}$, which has a positive Lebesgue measure by assumption of the lemma. But the function $u \mapsto v^T U(u)$ is a polynomial of degree $\leq \ell$ which cannot be equal to zero except for a finite number of points. Thus, we come to a contradiction showing that $\int [v^T U(u)]^2 K(u)\, du = 0$ is impossible for $v \neq 0$.

12 Spline estimators

There is no guarantee of smoothness for the local polynomial estimator. In this section, we introduce another estimator that is a smooth function by construction. We will consider the regression functions defined on the interval $[0,1]$.

One of the classical problems in numerical analysis is the following. Given points $(X_i, Y_i), i = 1, \ldots, n$ in a 2-dimensional space, construct a function that passes trough all the points and has a given smoothness. Schoenberg (1946)[1] was the first to refer to spline in this context (spline is a flexible wood or metal strip that was used by draftsmen to construct smooth functions). The spline takes a form that minimizes the potential energy, which is proportional to the second derivative of the function. Also, it passes through the points $(X_i, Y_i), i = 1, \ldots, n$. So, formally, it solves the following minimization problem

$$\min \int (g''(x))^2 dx,$$
$$s.t. Y_i = g(X_i), i = 1, ..., n.$$

The solution of this problem is called the *interpolating spline*. In a similar spirit, the *approximating spline* is defined as a solution of the problem

$$\min \int (g''(x))^2 dx,$$
$$s.t. RSS(g) \leq \epsilon,$$

where $RSS(g) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))^2$, and $\epsilon > 0$ is a given constant. The Lagrange functional of this problem is

$$\tilde{\mathcal{L}}(g) = \int (g''(x))^2 dx + \tilde{\lambda}(\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))^2 - \epsilon)$$

where $\tilde{\lambda} > 0$ is the Lagrange multiplier. The goal is to minimize $\tilde{\mathcal{L}}(g)$. After a re-writing, this leads to the *smoothing spline* defined as a solution of

$$\min_g \mathcal{L}(g),$$

where

$$\mathcal{L}(g) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))^2 + \lambda \int (g''(x))^2 dx$$

for some $\lambda > 0$. It is a different type of estimator, and it is similar to the penalized versions of MLE type estimators.

Remarks.

- The smoothing spline realizes a trade-off between the goodness-of-fit represented by the $RSS(g)$ term and smoothness represented by the penalty $\int (g''(x))^2 dx$. The role of the term $\int (g''(x))^2 dx$ is to penalize the functions that are not smooth.

[1] Schoenberg, I.J. (1946) Contributions to the problem of approximation of equidistant data by analytic functions. *Quart. Appl. Math.*, 4 , 45–99, 112–141.

For instance, if $g(t) = \sin(wt)$, then $\int (g''(x))^2 dx \sim w^2$. If the frequency w is large, then $g(t)$ is not smooth, and the penalizing term is large. For the minimizer of $\mathcal{L}(g)$, the integral $\int (g''(x))^2 dx$ cannot be too large, so the solution to the optimization problem should not be too oscillating.

- The choice of λ is important. It plays the role of smoothing parameter, and it is analogous to the bandwidth h in kernel density estimation.

  If λ goes to infinity, then the problem becomes

  $$\min_g \int (g''(x))^2 dx.$$

  The solution is that g is a linear function of x. In this case the bias is large.

  If λ goes to 0, then the solution satisfies

  $$Y_i = g(X_i), i = 1, \ldots, n.$$

  In this case, the smoothness of g is not guaranteed. Both extremes are not interesting. We encounter here again a kind of bias-variance trade-off phenomenon that appears in all nonparametric estimation problems.

More generally, for integer l, consider the functional

$$\mathcal{L}_l(g) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - g(X_i))^2 + \lambda \int (g^{(l)}(x))^2 dx.$$

The function that minimizes $\mathcal{L}_l(g)$ is called the spline estimator.

Now we give the assumption that will be used throughout the discussion of spline estimators.

Assumption (SP): The points X_i are distinct, $0 < X_1 < X_2 < ... < X_n < 1$, and $n \geq l$.

Definition 6. *Let Assumption (SP) be satisfied. The space of splines of degree $2l-1$ with knots at $X_1, X_2, ..., X_n$ (denoted as S_n^l) is the space of all functions $s : [0,1] \to \mathbb{R}$, such that the following holds.*

(i) $s^{(2l-1)}$ *is piecewise constant on the intervals* $[0, X_1], (X_1, X_2], (X_2, X_3], ..., (X_n, 1]$.

(ii) $s^{(j)}(0) = s^{(j)}(1) = 0,\ j = l, ..., 2l-1$.

Item (ii) of Definition 6 allows us to represent $s^{(2l-2)}$ in the following form:

$$s^{(2l-2)}(x) = a_1 + a_0 x + b_1(x - X_1)_+ + ... + b_n(x - X_n)_+.$$

Thus, s is determined only by a finite number of coefficients. Next, using the constraint at the left endpoint, $s^{(j)}(0) = 0, j = l, ..., 2l-1$, we reduce s to the form

$$s(x) = \sum_{j=0}^{l-1} c_j x^j + \sum_{i=1}^{n} b_i (x - X_i)_+^{2l-1}$$

(check it as an exercise). This representation has $l+n$ coefficients. There are l more linear equations coming from the condition $s^{(j)}(1) = 0,\ j = l, ..., 2l-1$. Therefore, the dimension of S_n^l equals to n under Assumption (SP).

Theorem 7. *Let Assumption (SP) be satisfied. Then the problem* $\min_g \mathcal{L}_l(g)$ *has a unique solution, and this solution belongs to* S_n^l.

Note that this theorem gives us the right to define the spline estimator of degree $2l-1$ as

$$\hat{f}_n^{SP} = \operatorname*{argmin}_g \mathcal{L}_l(g).$$

It is a polynomial of degree $2l-1$ on each interval.

Remarks.

- If $l = 2$, then $2l - 1 = 3$. Between the knots X_i, $\hat{f}_n^{SP}$ is a polynomial of degree 3 (called the cubic spline).
- Reproduction of polynomials.

  Assume that there is no noise, $Y_i = Q(X_i), i = 1, \dots, n$, and Q is a polynomial of degree less than or equal to $l-1$. Then $\mathcal{L}_l(Q) = 0$. On the other hand, for all g we have $\mathcal{L}_l(g) \geq 0$, so the minimum is achieved at the polynomials. Since we also know that the minimizer is unique under assumption (SP), so it is exactly this polynomial Q.
- The estimator $\hat{f}_n^{SP}$ is a linear estimator under Assumption (SP). Indeed, under this assumption we can reduce the minimization to the n-dimensional space S_n^l:

$$\hat{f}_n^{SP} = \operatorname*{argmin}_{s \in S_n^l} \Big(\frac{1}{n} \sum_{i=1}^n (Y_i - s(X_i))^2 + \lambda \int (s^{(l)}(x))^2 dx \Big).$$

  If $\{\varphi_j\}_{j=1}^n$ is a basis in S_n^l we can write

$$s(x) = \sum_{j=1}^n \theta_j \varphi_j(x) \stackrel{\text{def}}{=} < \theta, \varphi > (x),$$

$$s^{(l)}(x) = \sum_{j=1}^n \theta_j \varphi_j^{(l)}(x),$$

  provided that functions φ_j are l times differentiable. Then, the problem reduces to minimization of a quadratic form:

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \mathbb{R}^n} \left\{ \theta^T \Psi \theta - \frac{2}{n} \sum_{i=1}^n Y_i < \theta, \varphi > (X_i) \right\}, \qquad \hat{f}_n^{SP}(x) = < \hat{\theta}, \varphi > (x),$$

  where Ψ is a matrix depending only $X_1, \dots, X_n$. Taking the derivative in θ we can explicitly obtain the minimizer, and deduce that $\hat{f}_n^{SP}$ is a linear estimator.

  Moreover, one can prove that there exists a basis $\{\varphi_j\}_{j=1}^n$ in S_n^l such that

$$\hat{f}_n^{SP}(x) = \sum_{i=1}^n W_{ni}^{SP}(x) Y_i = \sum_{j=1}^n \hat{\theta}_j^{SP} \varphi_j(x),$$

  where the coefficients $\hat{\theta}_j^{SP}$ have the following explicit form

$$\hat{\theta}_j^{SP} = \frac{1}{n} \sum_{i=1}^n Y_i \varphi_j(X_i).$$

  These functions φ_j are obtained from the so-called Demmler-Reinsch basis.

- Some heuristics: From discrete to continuous observations.

  Assume that instead of $Y_1, \ldots, Y_n$, we observe a continuous function y. Consider the minimization problem

$$\min_g \Big[\int (y(t) - g(t))^2 dt + \lambda \int (g''(x))^2 dx \Big].$$

  Note that the first term here is a continuous analog of $RSS(g)$. Passing to the Fourier domain via Plancherel's formula we obtain an equivalent minimization problem

$$\min_{\hat{g}} \Big[\int |\hat{y}(w) - \hat{g}(w)|^2 dw + \lambda \int w^4 |\hat{g}(w)|^2 dw \Big].$$

  To solve this problem, it is enough to find $\hat{g}$ such that, for any fixed w, it minimizes

$$|\hat{y}(w) - \hat{g}(w)|^2 + \lambda w^4 |\hat{g}(w)|^2 \tag{2}$$

  over $\hat{g}(w)$. Notice that for any $c > 0$ and any $\hat{g}(w)$ we have

$$|\hat{y}(w) - \hat{g}(w)|^2 + c|\hat{g}(w)|^2 \geq |\hat{y}(w) - \hat{g}^*(w)|^2 + c|\hat{g}^*(w)|^2,$$

  where $\hat{g}^*(w) = \frac{\hat{y}(w)}{1+c}$. It follows that the minimizer of the expression in (2) over $\hat{g}(w)$ is

$$\hat{g}^*(w) = \frac{\hat{y}(w)}{1 + \lambda w^4} = \hat{y}(w)\hat{K}(hw),$$

  where $h = \lambda^{1/4}$ and $\hat{K}$ is the Fourier transform of the spline kernel K (for $\beta = 4$):

$$\hat{K}(w) = \frac{1}{1 + w^4}. \tag{3}$$

  Note that the product $\hat{y}(w)\hat{K}(hw)$ in the Fourier domain corresponds to the convolution of $y(\cdot)$ and $(1/h)K(\cdot/h)$ in the original domain. Thus, the "spline estimator"' for the continuous seting is given by

$$\hat{f}_n^{SP}(x) = \int y(t) \frac{1}{h} K\left(\frac{t-x}{h}\right) dt.$$

  This can be viewed as an analog of the following estimator based on discrete observations:

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right).$$

  In conclusion, the above heuristics suggests that the cubic spline is "equivalent" to the kernel estimator where the kernel has Fourier transform (3). It is called the Silverman kernel.