Lecture 7

Nonparametric regression: local polynomials

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be i.i.d. as (Y, X) random variables, $Y \in \mathbb{R}$ and $X \in \mathbb{R}^d$. Consider the nonparametric regression model

$$Y_i = f(X_i) + \epsilon_i, \;\; \mathrm{E}(\epsilon_i | X_i) = 0, \;\; i = 1, \ldots, n$$

If f were a constant, then $\widehat{f}_n = n^{-1} \sum_{i=1}^n Y_i \to f$ a.s. (LLN).

If f is sufficiently smooth, then consider a finite (or countably infinite) partition $\{A_1, A_2, \ldots\}$ of $\mathbb{R}^d$, for Borel sets $A_j \subset \mathbb{R}^d$ and for all $x \in A_j$ estimate

$$\widehat{f}_n(x) = \frac{\sum_{i=1}^n \mathbb{I}\{X_i \in A_j\} Y_i}{\sum_{i=1}^n \mathbb{I}\{X_i \in A_j\}}, \; x \in A_j$$

(here and subsequently the convention $0/0 = 0$ is used).
This estimator is called **partitioning estimator** and in $d = 1$ is just a piecewise constant.

If instead of taking all $x \in A_j$, one estimates at each $x \in \mathbb{R}^d$ and generalizes the weight to some suitable $K : \mathbb{R}^d \to \mathbb{R}_+$, then

$$\widehat{f}_n(x; h) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)} =: \sum_{i=1}^n W_i(x; h) Y_i, \; \forall x \in \mathbb{R}^d$$

for some $h > 0$. The function $W_i(x; h) = W_i(x; h, X_1, \ldots, X_n)$ is a weight function. A naive kernel would be $K(x) = \mathbb{I}\{\|x\| \leq 1\}$. This estimator is called **Nadaraya-Watson kernel estimator**.

It is easy to see that the Nadaraya-Watson kernel estimator can also be obtained as

$$\widehat{f}_n(x; h) = \arg\min_{c \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)(Y_i - c)^2, \; \forall x \in \mathbb{R}^d.$$

This can be generalized as follows.

Let $g(\cdot; a) : \mathbb{R}^d \to \mathbb{R}$ be a parametric function of unknown parameters $a \in \mathbb{R}^{\ell+1}$, then define the estimator

$$\begin{aligned} \widehat{f}_n(x; h) &= g(x; \widehat{a}) \\ \widehat{a} &= \arg\min_{a \in \mathbb{R}^{\ell+1}} \frac{1}{n} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \{Y_i - g(X_i; a)\}^2 . \end{aligned}$$

For $d = 1$ and $g(x; a) = \sum_{i=1}^{\ell+1} a_i x^{i-1}$, this estimator is referred to as a **local polynomial kernel estimator** and is motivated by the Taylor expansion for some x_0 that is close to x

$$f(x) \approx f(x_0) + f^{'}(x_0)(x - x_0) + \cdots + \frac{f^{(\ell)}(x_0)}{\ell!}(x - x_0)^l =: \sum_{i=1}^{\ell+1} a_i x^{i-1}.$$

Consider now local polynomial estimators in more detail. Consider a random design nonparametric regression model

$$Y_i = f(X_i) + \epsilon_i, \;\; i = 1, \ldots, n$$
$$\mathrm{E}(\epsilon_i|X_i) = 0, \; \mathrm{E}(\epsilon_i^2|X_i) = \sigma^2.$$

For the regression function $f(x) = \mathrm{E}(Y|X = x)$ we assume that $f \in \Sigma(\beta, L)$ (a Hölder class with parameters β and L, $\lfloor\beta\rfloor = \ell$). If $f \in \Sigma(\beta, L)$ then for x_0 sufficiently close to some fixed $x \in [0, 1]$ we may write

$$f(x_0) \approx f(x) + f^{'}(x)(x_0 - x) + \ldots + \frac{f^{(\ell)}(x)}{\ell!}(x_0 - x)^\ell = A(x)^t P(x_0 - x) \in \mathcal{P}_{\ell+1},$$

where $A(x) = \left\{f(x), f^{'}(x), \ldots, f^{(\ell)}(x)/\ell!\right\}^t$ and $P(x_0 - x) = \left\{1, (x_0 - x), \ldots, (x_0 - x)^\ell\right\}^t$.

With this,

$$\widehat{A}_n(x) = \arg\min_{A \in \mathbb{R}^{\ell+1}} \sum_{i=1}^{n} \left\{Y_i - A(x)^t P(X_i - x)\right\}^2 K\left(\frac{X_i - x}{h}\right)$$

is the local polynomial estimator of order $\ell + 1$ (degree ℓ) of $A(x)$.

Denote $e_k = (0, \ldots, 0, 1, 0, \ldots, 0) \in \mathbb{R}^{\ell+1}$ a unit vector with 1 at k-th position, $k = 1, \ldots, \ell + 1$. Then,

$$\widehat{f}^{(k-1)}(x) = (k - 1)!\, e_k^t\, \widehat{A}_n(x)$$

is the local polynomial estimator of $f^{(k-1)}(x)$, $k = 1, \ldots, \ell + 1$.

In matrix notation

$$\begin{aligned} X &= \begin{pmatrix} 1 & (X_1 - x) & \ldots & (X_1 - x)^\ell \\ \vdots & \vdots & \ldots & \vdots \\ 1 & (X_n - x) & \ldots & (X_n - x)^\ell \end{pmatrix}, \quad Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \\ V &= \mathrm{diag}\left\{K\left(\frac{X_1 - x}{h}\right), \ldots, K\left(\frac{X_n - x}{h}\right)\right\} \end{aligned}$$

we can write

$$\widehat{A}_n(x) = \arg\min_{A\in\mathbb{R}^{\ell+1}} \{Y - XA(x)\}^t V \{Y - XA(x)\} = (X^tVX)^{-1}X^tVY,$$

which is unique, if X^tVX is a positive definite matrix.

This representation makes obvious, that a local polynomial estimator of $f^{(k-1)}(x)$ is a linear estimator

$$\widehat{f}^{(k-1)}(x) = (k-1)!\, e_k^t\, (X^tVX)^{-1}X^tVY = \sum_{i=1}^n W_{k,i}(x)\, Y_i,$$

with the weight function

$$W_{k,i}(x) = \frac{(k-1)!}{nh}\, e_k^t \left(\frac{1}{nh}X^tVX\right)^{-1} P(X_i - x)K\left(\frac{X_i - x}{h}\right).$$

Theorem

Let $\widehat{f}^{(k-1)}$, $k = 1, \dots, \ell+1$ be the degree $\ell \geq 0$ local polynomial estimator of $f^{(k-1)}$, where f is the regression function in a random design nonparametric regression model

$$Y_i = f(X_i) + \epsilon_i, \quad E(\epsilon_i|X_i) = 0, \; E(\epsilon_i^2|X_i) = \sigma^2, \quad i = 1, \dots, n,$$

with unknown $\sigma^2 > 0$ and $f^{(\ell+1)}$ is bounded and continuous in a neighbourhood of x. Assume that

(i) kernel $K : [-1, 1] \to [0, \infty)$ is a symmetric first order kernel with finite moments $\mu_j = \int_{-1}^1 x^j K(x)dx < \infty$, $j = 1, 2, \dots$ and $\int_{-1}^1 \{K(x)\}^2 dx < \infty$;

(ii) the bandwidth h is such that $h = h(n) \to 0$ and $nh \to \infty$;

(iii) the marginal Lebesgue density of X_i, denoted by q, is assumed to be differentiable, bounded and bounded away from zero with q' being Lipschitz continuous.

Then, at $x \in [h, 1-h]$

$$\begin{aligned}
\text{var}\left\{ \widehat{f}^{(k-1)}(x) \middle| \mathbf{X} \right\} &= \frac{\sigma^2\{(k-1)!\}^2}{nh^{2k-1}q(x)} \int_{-1}^1 \{\mathcal{W}_k(u)\}^2 du\{1 + o_p(1)\} \\
\textit{Bias}\left\{ \widehat{f}^{(k-1)}(x) \middle| \mathbf{X} \right\} &= \begin{cases} \frac{h^{\ell+1-(k-1)}(k-1)!\, f^{(\ell+1)}(x)\kappa_{\ell+1}}{(\ell+1)!} \{1 + o_p(1)\}, & (\ell + k - 1) \textit{ odd} \\ \frac{h^{\ell+2-(k-1)}(k-1)!\, f^{(\ell+1)}(x)\, q'(x)\, \kappa_{\ell+2}}{q(x)(\ell+1)!} \{1 + o_p(1)\}, & (\ell + k - 1) \textit{ even} \end{cases}
\end{aligned}$$

where $\kappa_\ell = \int_{-1}^1 u^\ell \mathcal{W}_k(u)du$.

Remarks

1. For $(\ell + k - 1)$ odd, the asymptotic conditional bias is independent of $q(x)$ and is therefore **design-adaptive**.
   For $(\ell + k - 1)$ even, the asymptotic conditional bias depends on $q^{'}(x)/q(x)$.

2. For $(\ell + k - 1)$ even, the asymptotic conditional bias has the same asymptotic order $\mathcal{O}(h^{\ell+2-(k-1)})$ for $(\ell + k - 1)$ and $(\ell + k)$. However, the constants are different.

3. Similar to kernel density estimation, we observe the bias-variance trade-off: increasing h increases the bias, while reducing the variance (oversmoothing) and decreasing h decreases the bias, while increasing the variance (undersmoothing).

The following theorem gives the asymptotic conditional bias and variance at a left boundary point, that is $x \in [0, h)$. For the right boundary point the result is completely analogous.

Theorem

Under assumptions of previous Theorem, a local polynomial estimator $\widehat{f}^{(k-1)}$ of f^{k-1} has the following asymptotic variance and bias at some $x \in [0, h)$:

$$\begin{aligned}
var\left\{\widehat{f}^{(k-1)}(x)|\mathbf{X}\right\} &= \frac{\sigma^2(0)\{(k-1)!\}^2}{nh^{2k-1}q(0)}\int_{-x/h}^{1}\{\mathcal{W}_k(u)\}^2du\{1+o_p(1)\} \\
Bias\left\{\widehat{f}^{(k-1)}(x)|\mathbf{X}\right\} &= \frac{h^{\ell+1-(k-1)}(k-1)!f^{(\ell+1)}(0)}{(\ell+1)!}\int_{-x/h}^{1}u^{\ell+1}\mathcal{W}_k(u)du\ \{1+o_p(1)\}.
\end{aligned}$$

Remarks

1. For $(\ell + k - 1)$ odd the rate of the bias $\mathcal{O}(h^{\ell+1-(k-1)})$ is the same for all $x \in [0, 1]$, however, at the boundaries the constants are different and depend on x/h.

2. For $(\ell + k - 1)$ even, the rate of the bias at the boundary is larger, than in the interior (=boundary effect).

Under certain assumptions one can show that estimator $\widehat{f}^{(k-1)}$ of $f^{(k-1)}$, $k = 1, \ldots, \ell + 1$, $f \in \Sigma(\beta, L)$ satisfies

$$\limsup_{n\to\infty}\sup_{f\in\Sigma(\beta,L)} \mathrm{E}\left(n^{\frac{2\beta-2(k-1)}{2\beta+1}}\|\widehat{f}^{(k-1)} - f^{(k-1)}\|_2^2\right) \leq C < \infty,$$

if $h = c\,n^{-1/(2\beta+1)}$, $c > 0$ is taken.

Remarks

1. The convergence rate for derivatives is slower.

2. The optimal bandwidth is independent on k.

Let us now discuss the choice of the bandwidth. Similar to kernel density estimation we are looking for an unbiased estimator of the L_2 risk of $\widehat{f}$ $(=MISE\{\widehat{f}(h)\})$. However, in regression models one can obtain, in general, only approximately unbiased estimators of $MISE\{\widehat{f}(h)\}$. In particular, we are able to find an unbiased estimator of a discretised version of the L_2 risk, that is of

$$\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-\widehat{f}_n(X_i)\right\}^2 .$$

Consider the empirical L_2 risk $n^{-1}\sum_{i=1}^{n}\left\{Y_i-\widehat{f}_n(X_i;h)\right\}^2$. Obviously, minimizing this expression w.r.t. the smoothing parameter will result in an estimator $\widehat{f}_n$ which is closest to Y_i.

Let $\widehat{f}_n$ be an estimator, that can be written as

$$\widehat{f}_n(X_i;h)=\sum_{j=1}^{n}W_j(X_i;h)Y_j,$$

where $W_j(x;h)=W_j(x;h,X_1,\dots,X_n)$ are some weight functions.
Assume $\mathrm{E}(\epsilon_i|X_1,\dots,X_n)=0$ and $\mathrm{E}(\epsilon_i\epsilon_j|X_1,\dots,X_n)=\sigma^2\delta_{ij}$, $\sigma\in(0,\infty)$.
Consider

$$\begin{aligned}
&\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i-\widehat{f}_n(X_i;h)\right\}^2\right]=\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i^2-2Y_i\widehat{f}_n(X_i;h)+\widehat{f}(X_i;h)^2\right\}\right]\\
= &\ \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-\widehat{f}_n(X_i;h)\right\}^2\right]+\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{Y_i^2-f(X_i)^2\right\}\right]\\
- &\ \mathrm{E}\left(\frac{2}{n}\mathrm{E}\left[\sum_{i=1}^{n}\{Y_i-f(X_i)\}\widehat{f}(X_i;h)\middle| X_1,\dots,X_n\right]\right)\\
= &\ \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-\widehat{f}(X_i;h)\right\}^2\right]+\mathrm{E}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i^2\right)\\
- &\ \mathrm{E}\left[\frac{2}{n}\mathrm{E}\left\{\sum_{i=1}^{n}\epsilon_i\sum_{j=1}^{n}\epsilon_jW_j(X_i;h)|X_1,\dots,X_n\right\}\right]\\
= &\ \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i)-\widehat{f}_n(X_i;h)\right\}^2\right]+\sigma^2-\mathrm{E}\left\{2\sigma^2\frac{1}{n}\sum_{i=1}^{n}W_i(X_i;h)\right\},
\end{aligned}$$

so that the last term is “disturbing”.

Mallows' C_p criterion is a simple way to correct for this term

$$C_p(h) = \frac{1}{n}\sum_{i=1}^{n}\left\{Y_i - \widehat{f}_n(X_i;h)\right\}^2 + 2\sigma^2\frac{1}{n}\sum_{i=1}^{n} W_i(X_i;h).$$

Apparently,

$$\mathrm{E}\left\{C_p(h)\right\} = \mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{f(X_i) - \widehat{f}_n(X_i;h)\right\}^2\right] + \sigma^2$$

and h can be chosen as

$$\widehat{h} = \arg\min_{h>0} C_p(h)$$

Note that $C_p(h)$ criterion depends on an unknown σ^2, which needs to be estimated.

Other methods for smoothing parameter selection that (asymptotically) correct for the "disturbing" term include

$$\begin{aligned} AIC(h) &= \log\left[\sum_{i=1}^{n}\left\{Y_i - \widehat{f}_n(X_i;h)\right\}^2\right] + \frac{2}{n}\sum_{i=1}^{n} W_i(X_i;h) \\ GCV(h) &= \frac{\sum_{i=1}^{n}\left\{Y_i - \widehat{f}_n(X_i;h)\right\}^2}{\left\{1 - n^{-1}\sum_{i=1}^{n} W_i(X_i;h)\right\}^2}, \end{aligned}$$