Kernel-Based Estimators

刘佳玲

2022 年 8 月 19 日

Table of Contents

Table of Contents

Kernel Density Estimator

- Let $X_1, , ..., X_n$ be i.i.d random variables that have a probability density p.

Def:kernel density estimator

The function $x \mapsto \hat{p}_n(x)$ is called the kernel density estimator:

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

where $K : \mathbf{R} \to \mathbf{R}$ is an integrable function satisfying $\int K(u)du = 1$.

- We always rewrite $\hat{p}_n(x)$ as

$$\hat{p}_n(x) = \frac{1}{n} \sum_{i=1}^{n} K_h\left(X_i - x\right),$$

where $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$

Kernel Density Estimator

- The choice of kernel function is not crucial in nonparametric estimation.
- A commonly used kernel is the Gaussian kernel:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right)$$

- Epanechnikov kernel: $K(x) = 0.75\left(1 - x^2\right) I(|x| \leq 1)$

Mean Squared Error of Kernel Density Estimators

- At an arbitrary fixed point $x_0 \in \mathbf{R}$:

$$\begin{aligned}
&\mathrm{MSE}(x_0) \\
&\qquad \triangleq \mathbf{E}_p\left[\left(\hat{p}_n(x_0) - p(x_0)\right)^2\right] \\
&\qquad \triangleq \int \cdots \int \left(\hat{p}_n(x_0, x_1, \ldots, x_n) - p(x_0)\right)^2 \prod_{i=1}^{n} \left[p(x_i)\, dx_i\right]
\end{aligned}$$

($\mathbf{E}_p$ denotes the expectation with respect to the distribution of $(X_1, \ldots, X_n)$)

- We have

$$\mathrm{MSE}(x_0) = b^2(x_0) + \sigma^2(x_0)$$

where

$$b(x_0) = \mathbf{E}_p\left[\hat{p}_n(x_0)\right] - p(x_0)$$

and

$$\sigma^2(x_0) = \mathbf{E}_p\left[\left(\hat{p}_n(x_0) - \mathbf{E}_p\left[\hat{p}_n(x_0)\right]\right)^2\right].$$

Variance of the Estimator $\hat{p}_n$

Proposition1

Suppose that the density p satisfies $p(x) \leq p_{\max} < \infty$ for all $x \in \mathbf{R}$. Let $K : \mathbf{R} \to \mathbf{R}$ be a function such that

$$\int K^2(u)du < \infty,$$

Then for any $x_0 \in \mathbf{R}$, $h > 0$, and $n \geq 1$ we have

$$\sigma^2(x_0) \leq \frac{C_1}{nh},$$

where $C_1 = p_{\max} \int K^2(u)du$.

Bias of the Estimator $\hat{p}_n$

Def:Hölder class

Let T be an interval in $\mathbf{R}$ and let β and L be two positive numbers. The Hölder class $\Sigma(\beta, L)$ on T is defined as the set of $\ell = \lfloor \beta \rfloor$ times differentiable functions $f\colon T \to \mathbf{R}$ whose derivative $f^{(\ell)}$ satisfies

$$\left| f^{(\ell)}(x) - f^{(\ell)}\left(x'\right) \right| \leq L \left| x - x' \right|^{\beta-\ell}, \quad \forall x, x' \in T$$

Note: $\lfloor \beta \rfloor$ denote the greatest integer strictly less than the real number β

Def:kernel of order ℓ

Let $\ell \geq 1$ be an integer. We say that $K : \mathbf{R} \to \mathbf{R}$ is a kernel of order ℓ if the functions $u \mapsto u^j K(u), j = 0, 1, \ldots, \ell$, are integrable and satisfy

$$\int K(u) du = 1, \quad \int u^j K(u) du = 0, \quad j = 1, \ldots, \ell$$

Suppose now that p belongs to the class of densities $\mathcal{P} = \mathcal{P}(\beta, L)$ defined as follows:

$$\mathcal{P}(\beta, L) = \left\{ p \mid p \geq 0, \int p(x)dx = 1, \text{ and } p \in \Sigma(\beta, L) \text{ on } \mathbf{R} \right\}$$

Proposition2

Assume that $p \in \mathcal{P}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying

$$\int |u|^\beta |K(u)| du < \infty.$$

Then for all $x_0 \in \mathbf{R}$, $h > 0$ and $n \geq 1$ we have

$$|b(x_0)| \leq C_2 h^\beta,$$

where

$$C_2 = \frac{L}{\ell!} \int |u|^\beta |K(u)| du.$$

Upper Bound on the Mean Squared Risk

If p and K satisfy the assumptions of Propositions 1 and 2, we obtain

$$\mathrm{MSE} \leq C_2^2 h^{2\beta} + \frac{C_1}{nh}.$$

The minimum with respect to h is attained at

$$h_n^* = \left(\frac{C_1}{2\beta C_2^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}.$$

Therefore, the choice $h = h_n^*$ gives

$$\mathrm{MSE}\left(x_0\right) = O\left(n^{-\frac{2\beta}{2\beta+1}}\right), \quad n \to \infty$$

uniformly in x_0.

Upper Bound on the Mean Squared Risk

Theorem1

Assume that $\int K^2(u)du < \infty$ and the assumptions of Proposition 2 are satisfied. Fix $\alpha > 0$ and take $h = \alpha n^{-\frac{1}{2\beta+1}}$. Then for $n \geq 1$ the kernel estimator $\hat{p}_n$ satisfies

$$\sup_{x_0 \in \mathbf{R}} \sup_{p \in \mathcal{P}(\beta, L)} \mathbf{E}_p \left[\left(\hat{p}_n(x_0) - p(x_0) \right)^2 \right] \leq C n^{-\frac{2\beta}{2\beta+1}},$$

where $C > 0$ is a constant depending only on β, L, α and on the kernel K.

- We can see the rate of convergence of the estimator $\hat{p}_n(x_0)$ is $\psi_n = n^{-\frac{\beta}{2\beta+1}}$

Positive Constraint

- We can find that kernels of order $\ell \geq 2$ must take negative values on a set of positive Lebesgue measure.
- The estimators $\hat{p}_n$ based on such kernels can also take negative values.
- However,we can always use the positive part estimator

$$\hat{p}_n^+(x) \triangleq \max\{0, \hat{p}_n(x)\}$$

- $$\mathbf{E}_p\left[\left(\hat{p}_n^+\left(x_0\right)-p\left(x_0\right)\right)^2\right] \leq \mathbf{E}_p\left[\left(\hat{p}_n\left(x_0\right)-p\left(x_0\right)\right)^2\right], \quad \forall x_0 \in \mathbf{R}.$$

Integrated Squared Risk of Kernel Estimators

Another important global criterion is the mean integrated squared error (MISE)

$$\mathrm{MISE} \triangleq \mathbf{E}_p \int \left(\hat{p}_n(x) - p(x)\right)^2 dx.$$

By the Tonelli-Fubini theorem, we have

$$\mathrm{MISE} = \int \mathrm{MSE}(x) dx = \int b^2(x) dx + \int \sigma^2(x) dx.$$

Integrated Squared Risk of Kernel Estimators

Proposition3

Suppose that $K : \mathbf{R} \to \mathbf{R}$ is a function satisfying

$$\int K^2(u)du < \infty.$$

Then for any $h > 0, n \geq 1$ and any probability density p we have

$$\int \sigma^2(x)dx \leq \frac{1}{nh} \int K^2(u)du.$$

Integrated Squared Risk of Kernel Estimators

Def:Nikolski Class

Let $\beta > 0$ and $L > 0$. The Nikol'ski class $\mathcal{H}(\beta, L)$ is defined as the set of functions $f : \mathbf{R} \to \mathbf{R}$ whose derivatives $f^{(\ell)}$ of order $\ell = \lfloor \beta \rfloor$ exist and satisfy

$$\left[\int \left(f^{(\ell)}(x+t) - f^{(\ell)}(x) \right)^2 dx \right]^{1/2} \leq L|t|^{\beta-\ell}, \quad \forall t \in \mathbf{R}.$$

Integrated Squared Risk of Kernel Estimators

Proposition4

Assume that density function $p \in \mathcal{H}(\beta, L)$ and let K be a kernel of order $\ell = \lfloor \beta \rfloor$ satisfying

$$\int |u|^{\beta} |K(u)| du < \infty.$$

Then, for any $h > 0$ and $n \geq 1$,

$$\int b^2(x) dx \leq C_2^2 h^{2\beta},$$

where

$$C_2 = \frac{L}{\ell!} \int |u|^{\beta} |K(u)| du.$$

(Proof of proposition4):

Lemma1:Generalized Minkowski Inequality

For any Borel function g on $\mathbf{R} \times \mathbf{R}$, we have

$$\int \left(\int g(u,x) du \right)^2 dx \leq \left[\int \left(\int g^2(u,x) dx \right)^{1/2} du \right]^2 .$$

Lemma2:Taylor expansion (Integral form of the remainder)

If f is k times differentiable function and $f^{(k)}$ is absolutely continous on the observed interval between a and x,we have

$$f(x) = f(a) + f'(a)(x-a) + \ldots + R_k(x),$$

where $R_k(x) = \int_a^x \frac{f^{(k+1)}(t)}{k!}(x-t)^k dt.$

(Proof of proposition4):

Proof. Take any $x \in \mathbf{R}$, $u \in \mathbf{R}$, $h > 0$ and write the Taylor expansion

$$p(x+uh) = p(x) + p'(x)uh + \cdots + \frac{(uh)^\ell}{(\ell-1)!}\int_0^1 (1-\tau)^{\ell-1} p^{(\ell)}(x+\tau uh)d\tau.$$

Since the kernel K is of order $\ell = \lfloor \beta \rfloor$ we obtain

$$\begin{aligned} b(x) &= \int K(u)\frac{(uh)^\ell}{(\ell-1)!}\left[\int_0^1 (1-\tau)^{\ell-1} p^{(\ell)}(x+\tau uh)d\tau\right] du \\ &= \int K(u)\frac{(uh)^\ell}{(\ell-1)!}\left[\int_0^1 (1-\tau)^{\ell-1} \left(p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x)\right) d\tau\right] du. \end{aligned}$$

Applying twice the generalized Minkowski inequality and using the fact that p belongs to the class $\mathcal{H}(\beta, L)$, we get the following upper bound for the bias term:

(Proof of proposition4):

$$
\begin{aligned}
\int b^2(x)dx \leq & \int \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \times \right.\\
& \left. \int_0^1 (1-\tau)^{\ell-1} \left| p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x) \right| d\tau du \right)^2 dx \\
\leq & \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \times \right.\\
& \left. \left[\int \left(\int_0^1 (1-\tau)^{\ell-1} \left| p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x) \right| d\tau \right)^2 dx \right]^{1/2} du \right)^2 \\
\leq & \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \times \right.\\
& \left. \left[\int_0^1 (1-\tau)^{\ell-1} \left[\int \left(p^{(\ell)}(x+\tau uh) - p^{(\ell)}(x) \right)^2 dx \right]^{1/2} d\tau \right] du \right)^2 \\
\leq & \left(\int |K(u)| \frac{|uh|^\ell}{(\ell-1)!} \left[\int_0^1 (1-\tau)^{\ell-1} L|uh|^{\beta-\ell} d\tau \right] du \right)^2 \\
= & C_2^2 h^{2\beta}
\end{aligned}
$$

- We obtain

$$\mathrm{MISE} \le C_2^2 h^{2\beta} + \frac{1}{nh}\int K^2(u)du,$$

and the minimizer $h = h_n^*$ of the right hand side is

$$h_n^* = \left(\frac{\int K^2(u)du}{2\beta C_2^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$$

Taking $h = h_n^*$ we get

$$\mathrm{MISE} = O\left(n^{-\frac{2\beta}{2\beta+1}}\right), \quad n \to \infty.$$

Selecting the bandwidth

- In particular,when $\beta = 2$ and the kernel K is a density (i.e. nonnegative),

$$\mathrm{MISE}(h) = \frac{\int K^2(x)dx}{nh} + \frac{h^4 \mu_2^2(K) \int \left[p''(x)\right]^2 dx}{4} + o\left(\frac{1}{nh} + h^4\right)$$

- The optimal (asymptotic) bandwidth:

$$h = \left(\frac{\int K^2(x)dx}{n\mu_2^2(K) \int \left[p''(x)\right]^2 dx}\right)^{1/5}$$

where $\mu_2(K) = \int u^2 K(u)du$.

Selecting the bandwidth:rule of thumb

- Suppose that K is a standard normal density and $p(x) \sim N(\mu, \sigma^2)$
- Direct calculation then show that $h = 1.06\sigma n^{-\frac{1}{5}}$
- Using the sample standard deviation $\widehat{\sigma}$,we get $h = 1.06\hat{\sigma} n^{-\frac{1}{5}}$

Selecting the bandwidth:plug-in

- We can estimate $\int (p''(x))^2 dx$ and plug in the optimal (asymptotic) bandwidth.
- A simple way is :

$$\widehat{p''}(x) = \frac{\partial^2}{\partial x^2}\left\{\frac{1}{nh_0}\sum_{i=1}^{n} L\left(\frac{x-x_i}{h_0}\right)\right\}$$
$$= \frac{1}{nh_0^3}\sum_{i=1}^{n} L''\left(\frac{x-x_i}{h_0}\right)$$

  - h_0 is another bandwidth and L is another kernel function.
  - We may use the rule of thumb for choosing h_0.
- We use $\int (\widehat{p''}(x))^2 dx$ as the estimator of $\int (p''(x))^2 dx$.

Selecting the bandwidth:cross-validation

A data-driven approach

- Consider $\mathrm{ISE}(h) = \int (\hat{p}(x) - p(x))^2 dx$;
- The key idea is to minimize $\mathrm{ISE}(h)$;
- Write

$$\mathrm{ISE}(h) = \int \widehat{p}^2(x) dx - 2\mathrm{E}\left\{\widehat{p}(X)\right\} + \int p^2(x) dx$$

- Use $\frac{1}{n}\sum_{i=1}^{n} \widehat{p}_{-i}(x_i)$ to estimate the second term (why?), where

$$\widehat{p}_{-i}(x_i) = \frac{1}{h(n-1)} \sum_{j \neq i} K\left(\frac{x_i - x_j}{h}\right)$$

is the leave-one-out density estimator.

Table of Contents

Nonparametric kernel regression

- $\{y_i, x_i\}_{i=1}^n$. Linear regression: $Y = \beta_0 + \beta_1 X + \epsilon$
- One successful technique to relax the linear assumption is the nonparametric regression model

$$Y = m(X) + \epsilon,$$

Nadaraya-Watson estimator

Def:Nadaraya-Watson estimator

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(x_i - x) y_i}{\sum_{i=1}^{n} K_h(x_i - x)},$$

where $K_h(\cdot) = \frac{1}{h} K(\frac{\cdot}{h})$

Nadaraya-Watson estimator

- One derivation:
  Assume $E(\epsilon_i) = 0$,x_i are i.i.d. random variables.

$$m(x) = E(Y|X = x) = \int yf(y|x)dy = \frac{\int yf(x, y)dy}{f(x)}.$$

$f(x, y)$: the joint density of (X, Y) and $f(x)$: the marginal density

- Idea: use kernel estimators of $f(x)$ and $f(x, y)$ in the above equation

$$\widehat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

$$\widehat{f}(x, y) = \frac{1}{nhh_y}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) K_y\left(\frac{y - y_i}{h_y}\right)$$

Finite-Sample Properties of N-W estimator

Assumptions

1. ε_i is i.i.d. $(0; \sigma^2)$
2. m and f are twice continuously differentiable in a neighborhood of the point x.
3. The kernel K is a symmetric function satisfying (i) $\int K(\psi)d\psi = 1$, (ii) $\int \psi K(\psi)d\psi = 0$, (iii) $\int \psi^2 K(\psi)d\psi = \mu_2 < \infty$.
4. $h = h_n \to 0$ and $nh \to \infty$ as $n \to \infty$.
5. x_i is i.i.d. and independent of the ε_is.
6. The second-order derivatives of the marginal density f of x_i are continuous and bounded in a neighborhood of x, and x is a point in the interior of the support of x_i.

Finite-Sample Properties of N-W estimator

Theorem2

When assumpthions (1)-(6) hold and $f > 0$,

$$\mathrm{Bias}(\hat{m}) \approx \frac{h^2}{2f}\mu_2 \left(m^{(2)} f + 2f^{(1)} m^{(1)} \right)$$

$$V(\hat{m}) \approx \frac{\sigma^2}{nhf(x)} \int K^2(\psi) d\psi.$$

(Proof of Theorem2)
To obtain the bias, up to $O\left(h^2\right)$, we approximate

$$
\begin{aligned}
y_i = m\left(x_i\right) + u_i &\simeq m(x) + \left(x_i - x\right) m^{(1)}(x) + \frac{1}{2}\left(x_i - x\right)^2 m^{(2)}(x) + u_i \\
&= m + h\psi_i m^{(1)} + \frac{h^2}{2}\psi_i^2 m^{(2)} + u_i
\end{aligned}
$$

Then,

$$
\begin{aligned}
\hat{m}(x) - m(x) =& \frac{m^{(1)}}{\hat{f}}\frac{1}{n}\sum_{i=1}^{n} K\left(\psi_i\right)\psi_i + \frac{hm^{(2)}}{2\hat{f}}\frac{1}{n}\sum_{i=1}^{n} K\left(\psi_i\right)\psi_i^2 \\
&+ \frac{1}{\hat{f}}\frac{1}{nh}\sum_{i=1}^{n} K\left(\psi_i\right) u_i.
\end{aligned}
$$

Thus, conditional on x_i, the bias and variance are

$$
E_X(\hat{m}(x) - m(x)) = \frac{m^{(1)}}{\hat{f}}\frac{1}{n}\sum_{i=1}^{n} K\left(\psi_i\right)\psi_i + \frac{hm^{(2)}}{2\hat{f}}\frac{1}{n}\sum_{i=1}^{n} K\left(\psi_i\right)\psi_i^2,
$$

$$
V_X(\hat{m}(x)) = \frac{\sigma^2}{\hat{f}^2}\frac{1}{n^2h^2}\sum_{i=1}^{n} K^2\left(\psi_i\right).
$$

(Proof of Theorem2)
For large n,

$$\hat{f} = f + o_p(1),$$

$$\frac{1}{n}\sum_{i=1}^{n} K(\psi_i)\psi_i = h^2\mu_2 f^{(1)} + o_p\left(h^2\right),$$

$$\frac{h}{n}\sum_{i=1}^{n} K(\psi_i)\psi_i^2 = h^2\mu_2 f + o_p\left(h^2\right),$$

$$\frac{1}{nh}\sum_{i=1}^{n} K^2(\psi_i) = f\int K^2(\psi)d\psi + o_p(1)$$

Then,

$$E_X(\hat{m}(x) - m(x)) = \frac{h^2}{2f}\mu_2\left(m^{(2)}f + 2f^{(1)}m^{(1)}\right),$$

$$V_X(\hat{m}(x)) = \frac{\sigma^2}{nhf}\int K^2(\psi)d\psi.$$

Since the right-hand sides of these expressions are free from x_i, they become the approximate unconditional bias and variance as given in Theorem 2.

The Local Linear Regression Estimators

The Local Linear Regression Estimators

Minimizes

$$\sum_{i=1}^{n} \{y_i - m - (x_i - x)\,\beta\}^2\, K\left(\frac{x_i - x}{h}\right),$$

with respect to m and β.

- By the way,the Nadaraya-Watson estimator of $m(x) = m$ minimizes $\sum_{i=1}^{n} \{y_i - \ m\}^2 K\left(\frac{x_i - x}{h}\right)$ with respect to m.

Finite sample property of local linear regression estimator

Theorem3

The approximate bias and variance of the local linear regression estimator of $m(x)$ are

$$\mathrm{Bias}(\hat{m}(x)) \approx \frac{1}{2}\mu_2 h^2 m^{(2)}(x),$$

$$V(\hat{m}(x)) \approx \sigma^2 \frac{(nh)^{-1}}{f(x)} \int K^2(\psi) d\psi.$$

(Proof of theorem3)

- The local linear regression estimator performs a weighted regression of y_i against $z_i' = (1, (x_i - x))$ using weights $w_i^{1/2} = \left[K\left(\frac{x_i - x}{h}\right)\right]^{1/2}$.
- $$\hat{m}(x) = e_1'\left(\Sigma z_i w_i z_i'\right)^{-1} \Sigma z_i w_i y_i,$$

  where e_1 is a vector with unity in the first place.
- Since

$$\begin{aligned} y_i &= m(x_i) + u_i \\ &\simeq m(x) + (x_i - x)\,\beta(x) + (x_i - x)^2\,\gamma(x^*) + u_i \\ &= z_i'\delta + (x_i - x)^2\,\gamma(x^*) + u_i \end{aligned}$$

where $\beta(x) = m^{(1)}(x), \gamma(x^*) = m^{(2)}(x^*)$, and $\delta(x)' = \delta' = (m(x)\beta(x))$,$x^*$ lines between x and x_i.

(Proof of theorem3)

- one can find

$$\begin{aligned}\hat{m}(x) =& e_1'\delta(x) + e_1' \left(\Sigma z_i w_i z_i'\right)^{-1} \Sigma z_i w_i \left(x_i - x\right)^2 \gamma\left(x^*\right) \\ &+ e_1' \left(\Sigma z_i w_i z_i'\right)^{-1} \Sigma z_i w_i u_i\end{aligned}$$

- Because $e_1'\delta(x) = m(x)$, the conditional bias and variance are

$$E_X(\dot{m}(x) - m(x)) = e_1' \left(\Sigma z_i w_i z_i'\right)^{-1} \Sigma z_i w_i \left(x_i - x\right)^2 \gamma\left(\overline{x^*}\right)$$

and

$$V_X(\hat{m}(x)) = \sigma^2 e_1' \left(\Sigma z_i w_i z_i'\right)^{-1} \left(\sum z_{i,} w_i^2 z_i'\right) \left(\sum z_i w_i z_i'\right)^{-1} e_1$$

(Proof of theorem3)

- For large n, following Ruppert and Wand, we can evaluate this expression by using the asymptotic results in Section 3.3.1 to get

$$\left((nh)^{-1}\Sigma z_i w_i z_i'\right)^{-1} \xrightarrow{p} \begin{bmatrix} f^{-1}(x) & -f^{(1)}(x)f(x)^{-2} \\ -f^{(1)}f(x)^{-2} & \left\{\mu_2 f(x)h^2\right\}^{-1} \end{bmatrix}$$

$$\left((nh)^{-1}\Sigma z_i w_i^2 z_i'\right) \xrightarrow{p} \begin{bmatrix} f(x)\int K^2(\psi)d\psi & hf(x)\int K^2(\psi)\psi d\psi \\ hf(x)\int K^2(\psi)\psi d\psi & h^2 f(x)\int K^2(\psi)\psi^2 d\psi \end{bmatrix},$$

where $\mu_2 = \int \psi^2 K(\psi)d\psi$ and we have used

$$(nh)^{-1}\sum_{i=1}^{n} K^2(\psi_i)\,\psi_i^2 = f\int K^2(\psi)\psi^2 d\psi + o_p(1)$$

Using these results gives the asymptotic bias and variance of $\hat{m}$. Being free of x_i these are also the unconditional quantities.

Local Polynomial Estimators

If $f \in \Sigma(\beta, L), \beta > 1, \ell = \lfloor \beta \rfloor$, then for z sufficiently close to x :

$$f(z) \approx f(x) + f'(x)(z-x) + \cdots + \frac{f^{(\ell)}(x)}{\ell!}(z-x)^{\ell} = \theta^T(x) U\left(\frac{z-x}{h}\right)$$

where

$$U(u) = \left(1, u, u^2/2!, \ldots, u^{\ell}/\ell!\right)^T$$
$$\theta(x) = \left(f(x), f'(x)h, f''(x)h^2, \ldots, f^{(\ell)}(x)h^{\ell}\right)^T$$

Local Polynomial Estimators

Def:local polynomial estimator of order ℓ of $f(x)$

Let $K : \mathbf{R} \to \mathbf{R}$ be a kernel, $h > 0$ be a bandwidth, and $\ell \geq 0$ be an integer. A vector $\hat{\theta}_n(x) \in \mathbf{R}^{\ell+1}$ defined by

$$\hat{\theta}_n(x) = \arg \min_{\theta \in \mathbf{R}^{\ell+1}} \sum_{i=1}^{n} \left[Y_i - \theta^T U\left(\frac{X_i - x}{h}\right)\right]^2 K\left(\frac{X_i - x}{h}\right)$$

is called a local polynomial estimator of order ℓ of $\theta(x)$ or $\mathrm{LP}(\ell)$ estimator of $\theta(x)$ for short.

The statistic

$$\hat{f}_n(x) = U^T(0)\hat{\theta}_n(x)$$

is called a local polynomial estimator of order ℓ of $f(x)$ or $\mathrm{LP}(\ell)$ estimator of $f(x)$ for short.

Local Polynomial Estimators

For a fixed x the estimator $\hat{\theta}_n(x)$is a weighted least squares estimator. Indeed, we can write $\hat{\theta}_n(x)$ as follows:

$$\hat{\theta}_n(x) = \arg \min_{\theta \in \mathbf{R}^{\ell+1}} \left(-2\theta^T \mathbf{a}_{nx} + \theta^T \mathcal{B}_{nx} \theta\right),$$

where the matrix $\mathcal{B}_{nx}$ and the vector $\mathbf{a}_{nx}$ are defined by the formulas

$$\mathcal{B}_{nx} = \frac{1}{nh} \sum_{i=1}^{n} U\left(\frac{X_i - x}{h}\right) U^T \left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)$$

$$\mathbf{a}_{nx} = \frac{1}{nh} \sum_{i=1}^{n} Y_i U\left(\frac{X_i - x}{h}\right) K\left(\frac{X_i - x}{h}\right)$$

Normal equation: $\mathcal{B}_{nx} \hat{\theta}_n(x) = \mathbf{a}_{nx}$

Local Polynomial Estimators

Pointwise and integrated risk of local polynomial estimators
setting and MSE

Assumptions(LP1)

There exist a real number $\lambda_0 > 0$ and a positive integer n_0 such that the smallest eigenvalue $\lambda_{\min}(\mathcal{B}_{nx})$ of $\mathcal{B}_{nx}$ satisfies

$$\lambda_{\min}(\mathcal{B}_{nx}) \geq \lambda_0$$

for all $n \geq n_0$ and any $x \in [0, 1]$.

Local Polynomial Estimators

Assumptions(LP2)

There exists a real number $a_0 > 0$ such that for any interval $A \subseteq [0,1]$ and all $n \geq 1$

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i \in A) \leq a_0 \max(\mathrm{Leb}(A), 1/n)$$

where $\mathrm{Leb}(A)$ denotes the Lebesgue measure of A.

Local Polynomial Estimators

Assumptions(LP3)

The kernel K has compact support belonging to $[-1, 1]$ and there exists a number $K_{\max} < \infty$ such that $|K(u)| \leq K_{\max}, \forall u \in \mathbf{R}$.

Lemma2

Under Assumptions (LP1) - (LP3), for all $n \geq n_0, h \geq 1/(2n)$, and $x \in [0,1]$, the weights W^*_{ni} of the $\mathrm{LP}(\ell)$ estimator are such that:

- (i) $\sup_{i,x} |W^*_{ni}(x)| \leq \frac{C_*}{nh}$;
- (ii) $\sum_{i=1}^{n} |W^*_{ni}(x)| \leq C_*$;
- (iii) $W^*_{ni}(x) = 0$ if $|X_i - x| > h$,

where the constant C_* depends only on λ_0, a_0, and $K_{\max}$.

Proposition3

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 $X_1, \ldots, X_n$ are deterministic;
(ii) Assumptions (LP1)-(LP3) hold;
(iii) the random variables ξ_i are independent and such that for all $i = 1, \ldots, n$,

$$\mathbf{E}\left(\xi_i\right) = 0, \quad \mathbf{E}\left(\xi_i^2\right) \leq \sigma_{\max}^2 < \infty.$$

Then for all $x_0 \in [0,1]$, $n \geq n_0$, and $h \geq 1/(2n)$ the following upper bounds hold:

$$\left|b\left(x_0\right)\right| \leq q_1 h^{\beta}, \quad \sigma^2\left(x_0\right) \leq \frac{q_2}{nh},$$

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

Theorem5

Under the assumptions of Proposition 3 and if the bandwidth is chosen to be $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}, \alpha > 0$, the following upper bound holds:

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathbf{E}_f \left[\psi_n^{-2} \left| \hat{f}_n(x_0) - f(x_0) \right|^2 \right] \leq C < \infty$$

where $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ is the rate of convergence and C is a constant depending only on $\beta, L, \lambda_0, a_0, \sigma^2_{\max}, K_{\max}$, and α.

Corollary

Under the assumptions of Theorem 2 we have

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f \left[\psi_n^{-2} \left\| \hat{f}_n - f \right\|_2^2 \right] \leq C < \infty,$$

where $\|f\|_2^2 = \int_0^1 f^2(x)dx$, $\psi_n = n^{-\frac{\beta}{2\beta+1}}$ and where C is a constant depending only on $\beta, L, \lambda_0, a_0, \sigma_{\max}^2, K_{\max}$, and α.

Local Polynomial Estimators

Theorem6

Assume that f belongs to the Hölder class $\Sigma(\beta, L)$ on $[0,1]$ where $\beta > 0$ and $L > 0$. Let $\hat{f}_n$ be the $\mathrm{LP}(\ell)$ estimator of f with $\ell = \lfloor \beta \rfloor$.
Suppose also that:
(i) $X_i = i/n$ for $i = 1, \ldots, n$;
(ii) the random variables ξ_i are independent and satisfy

$$\mathbf{E}\left(\xi_i\right) = 0, \quad \mathbf{E}\left(\xi_i^2\right) \leq \sigma_{\max}^2 < \infty$$

for all $i = 1, \ldots, n$;
(iii) there exist constants $K_{\min} > 0, \Delta > 0$ and $K_{\max} < \infty$ such that

$$K_{\min} I(|u| \leq \Delta) \leq K(u) \leq K_{\max} I(|u| \leq 1), \quad \forall u \in \mathbf{R}$$

(iv) $h = h_n = \alpha n^{-\frac{1}{2\beta+1}}$ for some $\alpha > 0$.

Local Polynomial Estimators

Theorem6(Cont'd)

Then the estimator $\hat{f}_n$ satisfies

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \sup_{x_0\in[0,1]} \mathbf{E}_f \left[\psi_n^{-2} \left| \hat{f}_n(x_0) - f(x_0) \right|^2 \right] \leq C < \infty$$

and

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f \left[\psi_n^{-2} \left\| \hat{f}_n - f \right\|_2^2 \right] \leq C < \infty$$

Local Polynomial Estimators

Theorem7–Convergence in the sup-norm

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

$$h_n = \alpha \left(\frac{\log n}{n} \right)^{\frac{1}{2\beta+1}}$$

for some $\alpha > 0$. Suppose also that: (i) the design points $X_1, \ldots, X_n$ are deterministic;
(ii) Assumptions (LP1)-(LP3) hold;
(iii) the random variables ξ_i are i.i.d. Gaussian $\mathcal{N}\left(0, \sigma_\xi^2\right)$ with $0 < \sigma_\xi^2 < \infty$;
(iv) K is a Lipschitz kernel: $K \in \Sigma\left(1, L_K\right)$ on $\mathbf{R}$ with $0 < L_K < \infty$.

Local Polynomial Estimators

Theorem7(Cont'd)

Then there exists a constant $C < \infty$ such that

$$\limsup_{n\to\infty} \sup_{f\in\Sigma(\beta,L)} \mathbf{E}_f \left[\psi_n^{-2} \left\| \hat{f}_n - f \right\|_\infty^2 \right] \leq C$$

where

$$\psi_n = \left(\frac{\log n}{n} \right)^{\frac{\beta}{2\beta+1}}$$

Selecting the bandwidth