Local Polynomial Methods

Zhang Jin

University of Science and Technology of China

October, 2022

Table of Contents

Kernel density estimation

Let $X_1, X_2, ..., X_n$ be a sample from a distribution F with density f, a common nonparametric density estimation is KDE:

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

We refer to K as a kernel function and to h as a bandwidth.

If K satisfies **(i)**. $\int K(v)dv = 1$, **(ii)**. $K(v) = K(-v)$, and **(iii)**. $\int v^2 K(v)dv = \kappa_2 > 0$, then $\hat{f}(x)$ is a consistent estimator of $f(x)$.

Nonparametric regression model

In the nonparametric approach, no prior restrictions on $m(\cdot)$

$$Y_i = m(X_i) + u_i, \quad i = 1, \ldots, n$$

We assume that (X_i, Y_i) are i.i.d. distributions, and (X, Y) denote a generic member of the sample, whose conditional mean and variance are denoted by $m(x) = E(Y \mid X = x)$ and $\sigma^2(x) = \operatorname{Var}(Y \mid X = x)$.

Theorem 1

Let $\mathcal{G}$ denote the class of Borel measurable functions having finite second moment. Assume that $g(x) \equiv \mathrm{E}(Y \mid X = x)$ belongs to $\mathcal{G}$, and that $\mathrm{E}\left(Y^2\right)$ is finite. Then $\mathrm{E}(Y \mid X)$ is the optimal predictor of Y given X, in the following MSE sense:

$$\mathrm{E}\left\{[Y - r(X)]^2\right\} \geq \mathrm{E}\left\{[Y - \mathrm{E}(Y \mid X)]^2\right\} \text{ for all } r(\cdot) \in \mathcal{G}$$

Nadaraya-Watson estimator

Using marginal PDF of X and joint PDF of (X, Y):

$$\mathrm{E}(Y \mid X = x) = \int y f_{y|x}(y \mid x) dy = \frac{\int y f_{y,x}(x, y) dy}{f(x)}$$

We can simply use KDE to give estimate of f and $f_{y,x}$ as

$$\hat{m}(x) = \frac{\int y \hat{f}_{y,x}(x, y) dy}{\hat{f}(x)} = \frac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}$$

This is Nadaraya-Watson kernel regression estimator.

Intuition: Taylor's expansion locally

Suppose that locally the regression function m can be approximated by

$$m(z) \approx \sum_{j=0}^{p} \frac{m^{(j)}(x)}{j!}(z-x)^j \equiv \sum_{j=0}^{p} \beta_j (z-x)^j$$

It models $m(z)$ locally by a simple polynomial model. This suggests using a locally weighted polynomial regression

$$\sum_{i=1}^{n} \left\{ Y_i - \sum_{j=0}^{p} \beta_j \left(X_i - x_0\right)^j \right\}^2 K_h\left(X_i - x_0\right)$$

Denote by $\widehat{\beta}_j (j = 0, \cdots, p)$ the minimizer. The above exposition suggests that an estimator for $m^{(\nu)}(x_0)$ is

$$\widehat{m}_\nu(x_0) = \nu! \widehat{\beta}_\nu$$

The weighted least squares problem

It is more convenient to work with matrix notation. Denote by $\mathbf{X}$ the design matrix of problem :

$$\mathbf{X} = \begin{pmatrix} 1 & (X_1 - x_0) & \cdots & (X_1 - x_0)^p \\ \vdots & \vdots & & \vdots \\ 1 & (X_n - x_0) & \cdots & (X_n - x_0)^p \end{pmatrix}$$

and put

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \quad \text{and} \quad \widehat{\beta} = \begin{pmatrix} \widehat{\beta}_0 \\ \vdots \\ \widehat{\beta}_p \end{pmatrix}.$$

Further, let $\mathbf{W}$ be the $n \times n$ diagonal matrix of weights:

$$\mathbf{W} = \operatorname{diag}\left\{K_h\left(X_i - x_0\right)\right\}$$

The weighted least squares problem

Then the weighted least squares problem can be written as

$$\min_{\beta}(\mathbf{y}-\mathbf{X}\beta)^T\mathbf{W}(\mathbf{y}-\mathbf{X}\beta),$$

with $\beta = \left(\beta_0, \cdots, \beta_p\right)^T$. The solution vector is provided by weighted least squares theory and is given by

$$\widehat{\beta} = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{y}$$

When $p = 0$, the minimizer is Nadaraya-Watson estimator with kernel K. Therefore, N-W estimator is also called **local constant estimator**.

Bias and variance

Let $\mathcal{Z}_n = \{X_i\}_{i=1}^n$, the conditional bias and variance of the estimator $\widehat{\beta}$ are derived immediately from its definition:

$$
\begin{aligned}
E(\widehat{\beta} \mid \mathcal{Z}_n) &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{m} \\
&= \beta + \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r} \\
\operatorname{Var}(\widehat{\beta} \mid \mathcal{Z}_n) &= \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\Sigma\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}
\end{aligned}
$$

where $\mathbf{m} = \{m(X_1), \cdots, m(X_n)\}^T, \beta = \left\{m(x_0), \cdots, m^{(p)}(x_0)/p!\right\}^T$, $\mathbf{r} = \mathbf{m} - \mathbf{X}\beta$, the vector of residuals of the local polynomial approximation, and $\Sigma = \operatorname{diag}\left\{K_h^2(X_i - x_0)\sigma^2(X_i)\right\}$

Bias and variance

These exact bias and variance expressions are not directly usable!

We will give the first order asymptotic expansions for the bias and the variance of the estimator $\widehat{m}_\nu(x_0) = \nu!\widehat{\beta}_\nu$. The following notation will be used. The moments of K and K^2 are denoted respectively by

$$\mu_j = \int u^j K(u) du \quad \text{and} \quad \nu_j = \int u^j K^2(u) du.$$

Some matrices and vectors of moments appear in the asymptotic expressions. Let

$$\begin{aligned} S &= \left(\mu_{j+\ell}\right)_{0\le j,\ell\le p} & c_p &= \left(\mu_{p+1}, \cdots, \mu_{2p+1}\right)^T \\ \tilde{S} &= \left(\mu_{j+\ell+1}\right)_{0\le j,\ell\le p} & \tilde{c}_p &= \left(\mu_{p+2}, \cdots, \mu_{2p+2}\right)^T \\ S^* &= \left(\nu_{j+\ell}\right)_{0\le j,\ell\le p}. \end{aligned}$$

Asymptotic bias and variance

Theorem 2

Assume that $f(x_0) > 0$ and that $f(\cdot), m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are continuous in a neighborhood of x_0. Further, assume that $h \to 0$ and $nh \to \infty$. Then the asymptotic conditional variance of $\widehat{m}_\nu(x_0)$ is given by

$$\begin{aligned} \operatorname{Var}\{\widehat{m}_\nu(x_0) \mid \mathcal{Z}_n\} =& e_{\nu+1}^T S^{-1} S^* S^{-1} e_{\nu+1} \frac{\nu!^2 \sigma^2(x_0)}{f(x_0)\, nh^{1+2\nu}} \\ &+ o_P\left(\frac{1}{nh^{1+2\nu}}\right) \end{aligned}$$

Asymptotic bias and variance (Cont'd)

Theorem 3

The asymptotic conditional bias for $p - \nu$ odd is given by

$$\begin{aligned}\operatorname{Bias}\left\{\widehat{m}_{\nu}\left(x_0\right) \mid \mathcal{Z}_n\right\} = & e_{\nu+1}^T S^{-1} c_p \frac{\nu!}{(p+1)!} m^{(p+1)}\left(x_0\right) h^{p+1-\nu} \\ & + o_P\left(h^{p+1-\nu}\right).\end{aligned}$$

Further, for $p - \nu$ even the asymptotic conditional bias is

$$\begin{aligned}\operatorname{Bias}\left\{\widehat{m}_{\nu}\left(x_0\right) \mid \mathcal{Z}_n\right\} = & e_{\nu+1}^T S^{-1} \tilde{c}_p \frac{\nu!}{(p+2)!}\left\{m^{(p+2)}\left(x_0\right)\right. \\ & \left.+(p+2) m^{(p+1)}\left(x_0\right) \frac{f'\left(x_0\right)}{f\left(x_0\right)}\right\} h^{p+2-\nu} + o_P\left(h^{p+2-\nu}\right)\end{aligned}$$

provided that $f'(\cdot)$ and $m^{(p+2)}(\cdot)$ are continuous in a neighborhood of x_0 and $nh^3 \to \infty$

Calculation of asymptotic variance

Denote by $S_n \equiv \mathbf{X}^T\mathbf{W}\mathbf{X}$ the $(p+1)\times(p+1)$ matrix $\left(S_{n,j+\ell}\right)_{0\le j,\ell\le p}$ with

$$S_{n,j} = \sum_{i=1}^{n} K_h\left(X_i - x_0\right)\left(X_i - x_0\right)^j$$

Denote by $S_n^* \equiv \mathbf{X}^T\Sigma\mathbf{X}$ the $(p+1)\times(p+1)$ matrix $\left(S^*_{n,j+\ell}\right)_{0\le j,\ell\le p}$ with

$$S^*_{n,j} = \sum_{i=1}^{n}\left(X_i - x_0\right)^j K_h^2\left(X_i - x_0\right)\sigma^2\left(X_i\right)$$

Then, the exact conditional variance

$$\operatorname{Var}(\widehat{\beta} \mid \mathcal{Z}_n) = \left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\Sigma\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{W}\mathbf{X}\right)^{-1}$$

can be re-expressed as $S_n^{-1}S_n^*S_n^{-1}$

The task is now to find approximations for the two matrices S_n and S_n^*. Note that

$$
\begin{aligned}
S_{n,j} =& ES_{n,j} + O_P\left\{\sqrt{\operatorname{Var}\left(S_{n,j}\right)}\right\} \\
=& nh^j \int u^j K(u) f\left(x_0 + hu\right) du \\
& + O_P\left(\sqrt{nE\left\{\left(X_1 - x_0\right)^{2j} K_h^2\left(X_1 - x_0\right)\right\}}\right) \\
=& nh^j\left\{f\left(x_0\right)\mu_j + o(1) + O_P(1/\sqrt{nh})\right\} \\
=& nh^j f\left(x_0\right)\mu_j\left\{1 + o_P(1)\right\},
\end{aligned}
$$

provided that $h \to 0$ and $nh \to \infty$.

Calculation of asymptotic variance

From this we obtain immediately that

$$S_n = nf(x_0) HSH \{1 + o_P(1)\},$$

where $H = \operatorname{diag}(1, h, \cdots, h^p)$. Using similar arguments, we find that

$$S_{n,j}^* = nh^{j-1} f(x_0) \sigma^2(x_0) \nu_j \{1 + o_P(1)\}$$

and therefore,

$$S_n^* = nh^{-1} f(x_0) \sigma^2(x_0) HS^*H \{1 + o_P(1)\}$$

Finally, we have

$$\operatorname{Var}(\widehat{\beta} \mid \mathcal{Z}_n) = \frac{\sigma^2(x_0)}{f(x_0) nh} H^{-1} S^{-1} S^* S^{-1} H^{-1} \{1 + o_P(1)\}$$

and since $\widehat{m}_\nu(x_0) = \nu! e_{\nu+1}^T \widehat{\beta}$ this leads directly to the asymptotic expression for the conditional variance.

Calculation of asymptotic bias ($p-\nu$ odd)

For the bias we have to distinguish between the case that $p-\nu$ is odd and $p-\nu$ is even. By using the Taylor expansion the conditional bias $S_n^{-1}\mathbf{X}^T\mathbf{W}\mathbf{r}$ of $\widehat{\beta}$ can be written as

$$\begin{aligned}
&S_n^{-1}\mathbf{X}^T\mathbf{W}\left[\beta_{p+1}\left(X_i-x_0\right)^{p+1}+o_P\left\{\left(X_i-x_0\right)^{p+1}\right\}\right]_{1\leq i\leq n}\\
=&S_n^{-1}\left\{\beta_{p+1}c_n+o_P\left(nh^{p+1}\right)\right\}
\end{aligned}$$

where $c_n=\left(S_{n,p+1},\cdots,S_{n,2p+1}\right)^T$ and $\beta_{p+1}=m^{(p+1)}\left(x_0\right)/(p+1)!$. Applying the expression of $S_{n,j}$ and S_n, we obtain

$$\mathrm{Bias}(\widehat{\beta}\mid\mathcal{Z}_n)=H^{-1}S^{-1}c_p\beta_{p+1}h^{p+1}\left\{1+o_P(1)\right\}$$

where we recall that $c_p=\left(\mu_{p+1},\cdots,\mu_{2p+1}\right)^T$. This immediately leads to the asymptotic expression for the conditional bias as given.

Calculation of asymptotic bias ($p - \nu$ even)

The above derivation of course holds for any value of $p - \nu$, but the problem is that for $p - \nu$ even the $(\nu + 1)^{th}$ element of the vector $S^{-1}c_p$ is zero. (recalling that the odd order moments of a symmetric kernel are zero.)

Hence, the $(\nu + 1)^{th}$ element of the main term in expansion (3.60) is zero, and one clearly has to proceed to higher order expansions.

The justifications of these expansions are similar to those given above, but now relying on the more stringent conditions on $f(\cdot)$ and $m^{(p+2)}(\cdot)$.

Calculation of asymptotic bias ($p - \nu$ even)

The expansion of $S_{n,j}$ can be extended to

$$S_{n,j} = nh^j \left\{ f\left(x_0\right) \mu_j + hf'\left(x_0\right) \mu_{j+1} + O_P\left(a_n\right) \right\}$$

where $a_n = h^2 + 1/\sqrt{nh}$. Hence, we get

$$S_n = nH \left\{ f\left(x_0\right) S + hf'\left(x_0\right) \tilde{S} + O_P\left(a_n\right) \right\} H,$$

with $\tilde{S} = \left(\mu_{j+\ell+1}\right)_{0 \le j,\ell \le p}$. Using a higher order Taylor expansion we can write the conditional bias as

$$\begin{aligned} S_n^{-1} \mathbf{X}^T \mathbf{W} \Big[\beta_{p+1} \left(X_i - x_0\right)^{p+1} + \beta_{p+2} \left(X_i - x_0\right)^{p+2} \\ + o_P \left\{ \left(X_i - x_0\right)^{p+2} \right\} \Big]_{1 \le i \le n} \\ = S_n^{-1} \left\{ \beta_{p+1} c_n + \beta_{p+2} \tilde{c}_n + o_P\left(nh^{p+2}\right) \right\} \end{aligned}$$

where $\tilde{c}_n = \left(S_{n,p+2}, \cdots, S_{n,2p+2}\right)^T$.

Calculation of asymptotic bias ($p - \nu$ even)

Then, substituting $S_{n,j}$ and S_n into the conditional bias, we obtain

$$\begin{aligned}\operatorname{Bias}(\widehat{\beta} \mid \mathcal{Z}_n) =& H^{-1}\left\{f\left(x_0\right) S+h f^{\prime}\left(x_0\right) \tilde{S}+O_P\left(a_n\right)\right\}^{-1} \\ & \times h^{p+1}\left[\beta_{p+1} f\left(x_0\right) c_p+h\left\{f^{\prime}\left(x_0\right) \beta_{p+1}\right.\right. \\ & \left.\left.+\beta_{p+2} f\left(x_0\right)\right\} \tilde{c}_p+O_P\left(a_n\right)\right]\end{aligned}$$

where $\tilde{c}_p=\left(\mu_{p+2}, \cdots, \mu_{2 p+2}\right)^T$. Finally, we find the following asymptotic expansion for the bias term

$$\operatorname{Bias}(\widehat{\beta} \mid \mathcal{Z}_n)=h^{p+1} H^{-1}\left\{\beta_{p+1} S^{-1} c_p+h b^*\left(x_0\right)+O_P\left(a_n\right)\right\}$$

where

$$b^*\left(x_0\right)=\frac{f^{\prime}\left(x_0\right) \beta_{p+1}+\beta_{p+2} f\left(x_0\right)}{f\left(x_0\right)} S^{-1} \tilde{c}_p-\frac{f^{\prime}\left(x_0\right)}{f\left(x_0\right)} \beta_{p+1} S^{-1} \tilde{S} S^{-1} c_p$$

Taking the $(\nu+1)^{th}$ element of this bias vector, we obtain the result.

Equivalent kernels

Now we show how the local polynomial approximation method assigns weights to each datum point. Note first of all that the estimator $\widehat{\beta}_\nu$ can be written as

$$\begin{aligned}\widehat{\beta}_\nu = e_{\nu+1}^T \widehat{\beta} &= e_{\nu+1}^T S_n^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y} \\ &= \sum_{i=1}^{n} W_\nu^n \left(\frac{X_i - x}{h} \right) Y_i,\end{aligned}$$

where $W_\nu^n(t) = e_{\nu+1}^T S_n^{-1} \{1, th, \cdots, (th)^p\}^T K(t)/h$.

The above expression reveals that the estimator $\widehat{\beta}_\nu$ is very much like a conventional kernel estimator except that the 'kernel' W_ν^n depends on the design points and locations.

Equivalent kernels

The weights W_ν^n satisfy the following discrete moment conditions:

$$\sum_{i=1}^{n} (X_i - x)^q \, W_\nu^n \left(\frac{X_i - x}{h} \right) = \delta_{\nu,q} \quad 0 \leq \nu, q \leq p.$$

A direct consequence of this relation is that the finite sample bias when estimating polynomials up to order p is zero.

Equivalent kernels

Substituting $S_n = nf(x)\,HSH\,\{1 + o_P(1)\}$ into the definition of W_n^ν,

$$W_\nu^n(t) = \frac{1}{nh^{\nu+1}f(x)} e_{\nu+1}^T S^{-1} (1, t, \cdots, t^p)^T K(t)\,\{1 + o_P(1)\}$$

and therefore

$$\widehat{\beta}_\nu = \frac{1}{nh^{\nu+1}f(x)} \sum_{i=1}^n K_\nu^* \left(\frac{X_i - x}{h} \right) Y_i \,\{1 + o_P(1)\}$$

where

$$K_\nu^*(t) = e_{\nu+1}^T S^{-1} (1, t, \cdots, t^p)^T K(t) = \left(\sum_{\ell=0}^p S^{\nu\ell} t^\ell \right) K(t)$$

with $S^{-1} = \left(S^{j\ell}\right)_{0 \le j, \ell \le p}$.

Equivalent kernels

We refer to K_ν^* as the equivalent kernel. This kernel satisfies the following moment conditions:

$$\int u^q K_\nu^*(u) du = \delta_{\nu,q} \quad 0 \le \nu, q \le p$$

which are an asymptotic version of the discrete moment conditions.

Table 3.1. *The equivalent kernel functions $K_{\nu,p}^*$.*

ν	p	Equivalent kernel function $K_{\nu,p}^*(t)$
0	1	$K(t)$
0	3	$(\mu_4 - \mu_2^2)^{-1}(\mu_4 - \mu_2 t^2)K(t)$
1	2	$\mu_2^{-1} t K(t)$
2	3	$(\mu_4 - \mu_2^2)^{-1}(t^2 - \mu_2)K(t)$

Re-express asymptotic conditional bias and variance

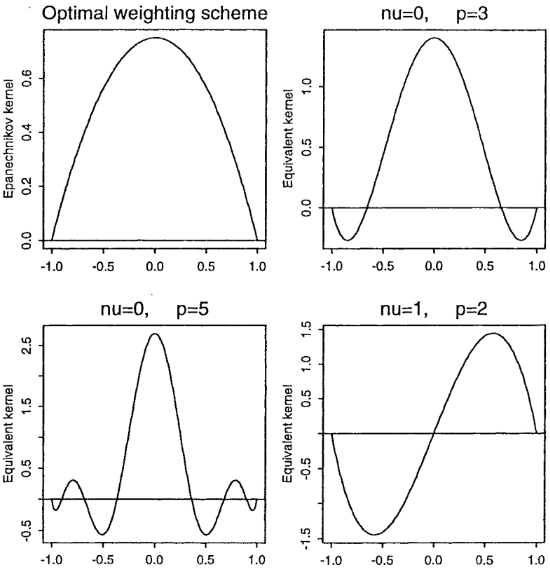
The conditional bias and variance of the estimator $\widehat{m}_\nu(x_0)$ can equally well be re-expressed in terms of the equivalent kernel K_ν^*, leading to the asymptotic expression

$$\operatorname{Bias}\left\{\widehat{m}_\nu(x_0) \mid \mathcal{Z}_n\right\} = \left\{\int t^{p+1} K_\nu^*(t) dt\right\} \frac{\nu!}{(p+1)!} m^{(p+1)}(x_0) \times h^{p+1-\nu} + o_P\left(h^{p+1-\nu}\right)$$

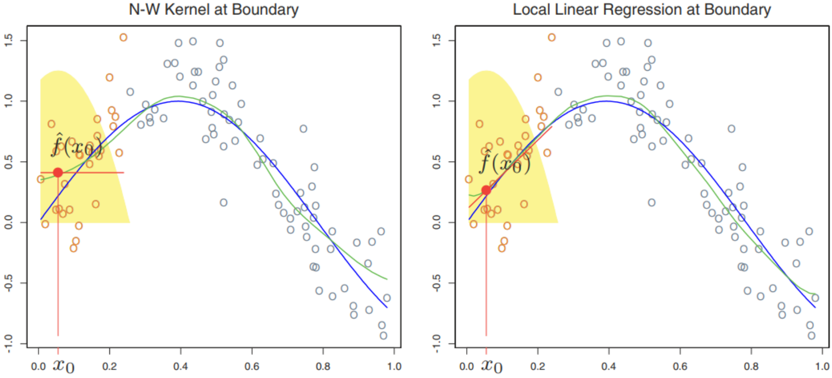
and its asymptotic variance equals

$$\operatorname{Var}\left\{\widehat{m}_\nu(x_0) \mid \mathcal{Z}_n\right\} = \int K_\nu^{*2}(t) dt \frac{\nu!^2 \sigma^2(x_0)}{f(x_0) n h^{1+2\nu}} + o_P\left(\frac{1}{n h^{1+2\nu}}\right)$$

Figure of Equivalent kernels

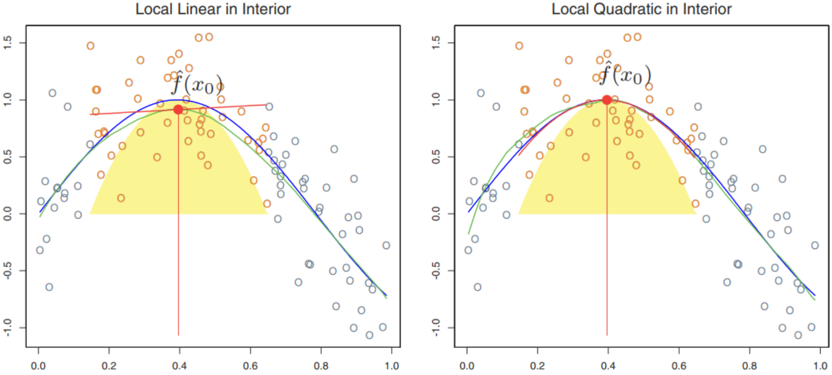


Boundary behavior of N-W estimators



N-W estimators has bias problems at or near the boundaries of the domain. By fitting a locally weighted linear regression, this bias is removed to first order.

Bias of local linear fits



Local linear fits exhibit bias in regions of high curvature of the true function, as "trimming the hills and filling in the valleys".

Asymptotic MSE

Conditional on $\mathcal{Z}_n$, When $p-\nu$ is odd

$$\hat{m}_\nu(x_0) - m^{(\nu)}(x_0) = O_p\left(h^{p-\nu+1} + \left(nh^{2\nu+1}\right)^{-1/2}\right)$$

When $p-\nu$ is even

$$\hat{m}_\nu(x_0) - m^{(\nu)}(x_0) = O_p\left(h^{p-\nu+2} + \left(nh^{2\nu+1}\right)^{-1/2}\right)$$

Assuming that $p-\nu$ is odd, we have

$$\text{MSE}\left(\hat{m}_\nu(x_0)\right) = O\left(m^{2(p-\nu+1)} + \left(nh^{1+2\nu}\right)^{-1}\right)$$

Using Liapunov's CLT, we can also establish the asymptotic normality of $\hat{m}(x)$ and $\hat{m}_\nu(x)$.

"It is an odd world"

Polynomial fits with $p - \nu$ odd outperform those with $p - \nu$ even.

- The asymptotic variance only increases whenever $p - \nu$ goes from odd to even. The gain in bias appears to be "free". (Fan and Gijbels, 1995, Ruppert and Wand, 1994)
- For $p - \nu$ odd the asymptotic bias has a simpler structure and does not involve $f'(x)$.
- When $p - \nu$ is odd, the conditional MSE is a continuous function in c, which implies that the risk changes continuously from a boundary point to an interior point. However, when $p - \nu$ is even, the bias at the boundary is of a larger order.

"It is an odd world"

Table 3.3. *Increase of the variability with the order of the polynomial approximation p.*

p	Gaussian	Uniform	Epanechnikov	Biweight	Triweight
1	1	1	1	1	1
2	1.6876	2.2500	2.0833	1.9703	1.9059
3	1.6876	2.2500	2.0833	1.9703	1.9059
4	2.2152	3.5156	3.1550	2.8997	2.7499
5	2.2152	3.5156	3.1550	2.8997	2.7499
6	2.6762	4.7852	4.2222	3.8133	3.5689
7	2.6762	4.7852	4.2222	3.8133	3.5689
8	3.1224	6.0562	5.2872	4.7193	4.3753
9	3.1224	6.0562	5.2872	4.7193	4.3753
10	3.5704	7.3281	6.3509	5.6210	5.1744

Boundary effects

Boundary bias: If a bandwidth is chosen to be 25% of the data range, then for about 50% of the data range the local neighborhood will lie partly outside the design region. Hence the boundary region is about 50% of the whole data range.

What if in higher dimensions?

WLOG we assume that the design density has a bounded support $[0, 1]$. A left boundary point is thought of as being of the form $x = ch$, with $c \geq 0$, whereas a right boundary point is of the form $x = 1 - ch$.

Automatic boundary carpentry

Consider a point at the left boundary $x = ch$. The finite sample moments behave as

$$S_{n,j} = nh^j f(0+)\mu_{j,c}\left\{1 + o_P(1)\right\}$$

where $\mu_{j,c} = \int_{-c}^{\infty} u^j K(u)du$. This leads to the following equivalent kernel at the boundary

$$K^*_{\nu,c}(t) = e^T_{\nu+1} S_c^{-1}\left(1, t, \ldots, t^p\right)^T K(t) \quad \text{with} \quad S_c = \left(\mu_{j+\ell,c}\right)_{0\le j,\ell\le p}$$

This equivalent kernel differs from K^*_ν only in the matrix S, and satisfies the boundary moment conditions. This reflects the automatic adaptation to the boundary.

Automatic boundary carpentry

Theorem 4

Assume that $f(0+) > 0$ and that $f(\cdot), m^{(p+1)}(\cdot)$ and $\sigma^2(\cdot)$ are right continuous at the point 0 . Then, the conditional MSE of the estimator $\widehat{m}_\nu(x)$ at the left boundary point $x = ch$ is given by

$$\left[\left\{\int_{-c}^{\infty} t^{p+1} K_{\nu,c}^{*}(t)dt\right\}^2 \left\{\nu! \frac{m^{(p+1)}(0+)}{(p+1)!}\right\}^2 h^{2(p+1-\nu)} \right.$$
$$\left. + \int_{-c}^{\infty} K_{\nu,c}^{*2}(t)dt \frac{\nu!^2\sigma^2(0+)}{f(0+)nh^{1+2\nu}}\right] \{1 + o_P(1)\}$$

Automatic boundary carpentry

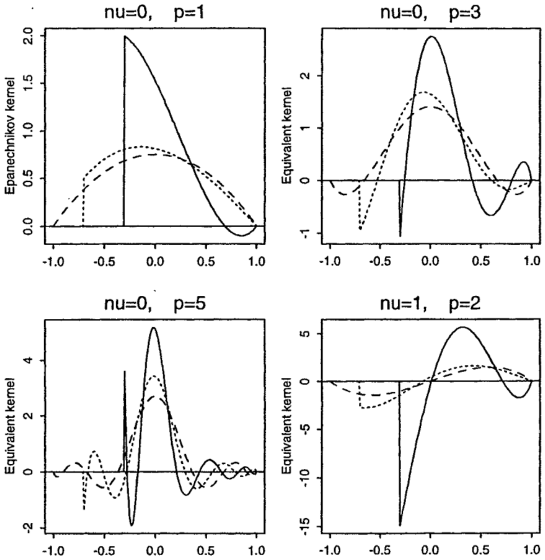


Table of Contents

Ideal choice of bandwidth

Re-expressed the asymptotic conditional bias of the estimator, in terms of the equivalent kernel:

$$\operatorname{Bias}\left\{\widehat{m}_{\nu}\left(x_{0}\right) \mid \mathcal{Z}_{n}\right\}=\left\{\int t^{p+1} K_{\nu}^{*}(t) d t\right\} \frac{\nu !}{(p+1) !} m^{(p+1)}\left(x_{0}\right) \\ \times h^{p+1-\nu}+o_{P}\left(h^{p+1-\nu}\right)$$

and its asymptotic variance equals

$$\operatorname{Var}\left\{\widehat{m}_{\nu}\left(x_{0}\right) \mid \mathcal{Z}_{n}\right\}=\int K_{\nu}^{* 2}(t) d t \frac{\nu !^{2} \sigma^{2}\left(x_{0}\right)}{f\left(x_{0}\right) n h^{1+2 \nu}}+o_{P}\left(\frac{1}{n h^{1+2 \nu}}\right)$$

Ideal choice of bandwidth

A theoretical optimal local bandwidth for estimating $m^{(\nu)}(x)$ is obtained by minimizing the conditional Mean Squared Error (MSE) given by

$$\left[\text{Bias}\left\{\widehat{m}_\nu(x_0) \mid \mathcal{Z}_n\right\}\right]^2 + \text{Var}\left\{\widehat{m}_\nu(x_0) \mid \mathcal{Z}_n\right\}$$

Minimization of it leads to

$$h_{\text{opt}}(x_0) = C_{\nu,p}(K)\left[\frac{\sigma^2(x_0)}{\left\{m^{(p+1)}(x_0)\right\}^2 f(x_0)}\right]^{1/(2p+3)} n^{-1/(2p+3)}$$

where

$$C_{\nu,p}(K) = \left[\frac{(p+1)!^2(2\nu+1)\int K_\nu^{*2}(t)dt}{2(p+1-\nu)\left\{\int t^{p+1}K_\nu^*(t)dt\right\}^2}\right]^{1/(2p+3)}$$

Ideal choice of bandwidth

A commonly used, simple measure of global loss is the weighted Mean Integrated Squared Error (MISE). Minimization of the conditional weighted MISE

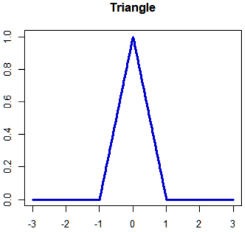
$$\int \left(\{\text{Bias}\,\{\widehat{m}_\nu(x) \mid \mathcal{Z}_n\}]^2 + \text{Var}\,\{\widehat{m}_\nu(x) \mid \mathcal{Z}_n\} \right) w(x)dx$$

with $w \geq 0$ some weight function, leads to a theoretical optimal constant bandwidth. We find an asymptotically optimal constant bandwidth given by

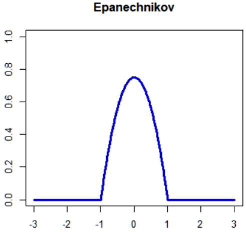
$$h_{\text{opt}} = C_{\nu,p}(K) \left[\frac{\int \sigma^2(x)w(x)/f(x)dx}{\int \left\{m^{(p+1)}(x)\right\}^2 w(x)dx} \right]^{1/(2p+3)} n^{-1/(2p+3)}$$

It depends on unknown quantities!

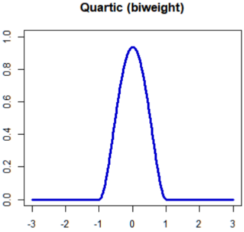
Choose the kernel K



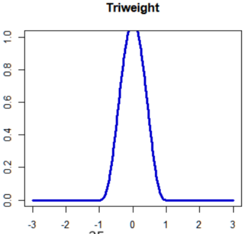
$K(u) = (1 - |u|)I(|u| \leq 1)$



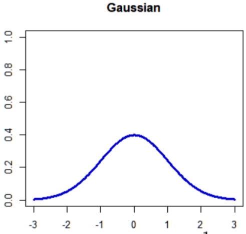
$K(u) = \frac{3}{4}(1 - u^2)I(|u| \leq 1)$



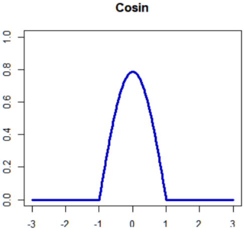
$K(u) = \frac{15}{16}(1 - u^2)^2 I(|u| \leq 1)$



$K(u) = \frac{35}{32}(1 - u^2)^3 I(|u| \leq 1)$



$K(u) = (2\pi)^{-1/2} exp(-\frac{1}{2}u^2)$



$K(u) = \frac{\pi}{4} cos(\frac{\pi u}{2}) I(|u| \leq 1)$

Choose the kernel K

The optimal bandwidths depend on the kernel function K through

$$T_{\nu,p}(K) \equiv \left| \int t^{p+1} K_\nu^*(t) dt \right|^{2\nu+1} \left\{ \int K_\nu^{*2}(t) dt \right\}^{p+1-\nu}$$

Theorem 5

The Epanechnikov weight function $K(z) = 3/4(1 - z^2)_+$ is the optimal kernel in the sense that it minimizes $T_{\nu,p}(K)$ over all nonnegative, symmetric, and Lipschitz continuous functions.

Choose the kernel K

When $p = 1, \nu = 0$,

Kernel	$T(K)$	$T(K)/T(K_{Epa})$
Uniform	0.3701	1.0602
Triangle	0.3531	1.0114
Epanechnikov	0.3491	1.0000
Quartic	0.3507	1.0049
Triweight	0.3699	1.0595
Gaussian	0.3633	1.0408
Cosine	0.3494	1.0004

- The choice of the kernel function K is **not very important** for the performance of the resulting estimators.
- However, since the Epanechnikov kernel is optimal in minimizing MSE and MISE at an interior point, it's recommended.

Choose the order p

- For many applications the choice $p = \nu + 1$ suffices. Such an order selection procedure is mainly proposed for recovering spatially inhomogeneous curves.
- Intuitively it is clear that in a flat non-sloped region a local constant or linear fit is recommendable, whereas at peaks and valleys local quadratic and cubic fits are preferable.

Choose the order p

Suppose that a bandwidth h is given, the estimated curve is usually evaluated at grid points of the form

$$x_j = x_L + j\Delta; \quad j = 0, \cdots, n_{\text{grid}}$$

The algorithm of adaptive order approximation:

1. For each order $p(\nu < p \leq R)$ and for each grid point obtain $\widehat{\mathrm{MSE}}_{\nu,p}\left(x_j; h\right)$
2. For each order p, and for each grid point calculate the smoothed estimated MSE by taking the weighted local average of the estimated MSE in the neighboring $2[h/\Delta] + 1$ grid points
3. For each grid point x_j choose the order p_j which has the smallest smoothed estimated MSE, and use a p_j order polynomial approximation to estimate $m^{(\nu)}\left(x_j\right)$.

Rule of thumb for bandwidth selection

Asymptotically optimal constant bandwidth:

$$h_{\text{opt}} = C_{v,p}(K) \left[\frac{\int \sigma^2(x) w(x) / f(x) dx}{\int \left\{ m^{(p+1)}(x) \right\}^2 w(x) dx} \right]^{1/(2p+3)} n^{-1/(2p+3)}$$

A simple way to estimate the unknown quantities is by fitting a polynomial of order $p+3$ **globally** to $m(x)$:

$$\check{m}(x) = \check{\alpha}_0 + \cdots + \check{\alpha}_{p+3} x^{p+3}$$

The standardized residual sum of squares from this parametric fit is denoted by $\check{\sigma}^2$. Suppose that we are interested in estimating $m^{(\nu)}(\cdot)$ and we take $w(x) = f(x) w_0(x)$ for some specific function w_0.

Rule of thumb for bandwidth selection

Regarding the conditional variance $\sigma^2(x)$ as a constant, we obtain the following expression:

$$C_{v,p}(K)\left[\frac{\check{\sigma}^2\int w_0(x)dx}{n\int\left\{\check{m}^{(p+1)}(x)\right\}^2 w_0(x)f(x)dx}\right]^{1/(2p+3)}$$

The denominator in the above expression can be estimated by

$$\sum_{i=1}^{n}\left\{\check{m}^{(p+1)}\left(X_i\right)\right\}^2 w_0\left(X_i\right)$$

which leads to the rule of thumb bandwidth selector

$$\check{h}_{\mathrm{ROT}}=C_{v,p}(K)\left[\frac{\check{\sigma}^2\int w_0(x)dx}{\sum_{i=1}^{n}\left\{\check{m}^{(p+1)}\left(X_i\right)\right\}^2 w_0\left(X_i\right)}\right]^{1/(2p+3)}$$

Plug-in ideas

An asymptotic expression for the MISE (mean integrated squared error) of the KDE is given by

$$\frac{1}{4}\left\{\int u^2 K(u)du\right\}^2 R\left(f''\right)h^4 + R(K)\frac{1}{nh}$$

with $R(g) = \int g^2(x)dx$. Minimization led to

$$h_{\text{opt}} = \alpha(K)\left[\int \left\{f''(x)\right\}^2 dx\right]^{-1/5} n^{-1/5}$$

Starting with an initial bandwidth h_0 find subsequent values $h_1, h_2, \ldots$ satisfying

$$h_i = \alpha(K)\left\{R\left(\hat{f}''_{h_{i-1}}\right)\right\}^{-1/5} n^{-1/5}$$

until convergence.

Plug-in ideas

Similar plug-in techniques are applied in the regression estimation setup. However, some limitations including:

1. Nice initial nonparametric estimates are required
2. Strong conditions
3. Very computationally intensive

Cross-validation

For each given i, we use data $\left\{\left(X_j, Y_j\right), j \neq i\right\}$ to build a regression function $\widehat{m}_{h,-i}(\cdot)$ and then validate the model by examining the prediction error $Y_i - \widehat{m}_{h,-i}\left(X_i\right)$. The least squares cross-validation technique uses the weighted average of squared errors

$$n^{-1} \sum_{i=1}^{n} \left\{Y_i - \widehat{m}_{h,-i}\left(X_i\right)\right\}^2 w\left(X_i\right),$$

as an overall measure of the effectiveness of the estimation $\widehat{m}_h(\cdot)$.

The least squares cross-validation bandwidth selector is the one that minimizes above formula.

Let $\hat{g}_{-i,L}(X_i)$ denote the leave-one-out local linear estimator. That is, let $\left(\hat{a}_i, \hat{b}_i\right)$ be the solution of (a, b) in the following minimization problem:

$$\min_{\{a,b\}} \sum_{j \neq i, j=1}^{n} \left[Y_j - a - \left(X_j - X_i\right)' b \right]^2 K\left(\frac{X_i - X_j}{h}\right)$$

where $K\left(\frac{X_i - X_j}{h}\right) = \prod_{s=1}^{q} k\left(\frac{X_{is} - X_{js}}{h_s}\right)$. Then $\hat{a}_i \equiv \hat{g}_{-i,L}(X_i)$ is the leave-one-out local linear kernel estimator of $g(X_i)$.

Cross-validation in local linear bandwidth selection

The local linear cross-validation approach to bandwidth selection chooses those h_s 's which minimize

$$CV_{ll}\left(h_1, \ldots, h_q\right) = \min_h \frac{1}{n} \sum_i \left[Y_i - \hat{g}_{-i,L}\left(X_i\right)\right]^2 w\left(X_i\right),$$

where $w(\cdot)$ is a weight function.

Cross-validation in local linear bandwidth selection

From the asymptotic bias and variance term of local linear estimator, we have

$$\mathrm{E}[\hat{g}(x) - g(x)]^2 = \left[\frac{\kappa_2}{2}\sum_{s=1}^{q} g_{ss}(x)h_s^2\right]^2 + \frac{\kappa^q}{nh_1 \dots h_q}\frac{\sigma^2(x)}{f(x)} + o\left(\eta_2^2 + \eta_1\right)$$

where
$\kappa = \int k(v)^2 dv, \kappa_2 = \int k(v)v^2 dv, \eta_1 = \left(nh_1 \dots h_q\right)^{-1}, \eta_2 = \sum_{s=1}^{q} h_s^2$

The leading term of CV_{ll} is

$$
\begin{aligned}
CV_{ll,0} &\sim \int \mathrm{E}[\hat{g}(x) - g(x)]^2 f(x) w(x) dx \\
&= \int \left[\frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) h_s^2 \right]^2 f(x) w(x) dx + \frac{\kappa^q \int \sigma^2(x) w(x) dx}{n h_1 \ldots h_q} \\
&\quad + o\left(\eta_2^2 + \eta_1\right)
\end{aligned}
$$

The leading term can be expressed as $n^{-4/(q+4)} \chi_{ll}\left(a_1, \ldots, a_q\right)$, where the a_s 's are defined by $h_s = a_s n^{-1/(q+4)} (s = 1, \ldots, q)$, and

$$
\begin{aligned}
\chi_{ll}\left(a_1, \ldots, a_q\right) = \int \left[\frac{\kappa_2}{2} \sum_{s=1}^{q} g_{ss}(x) a_s^2 \right]^2 f(x) M(x) dx \\
+ \frac{\kappa^q}{a_1 \ldots a_q} \int \sigma^2(x) M(x) dx
\end{aligned}
$$

Cross-validation in local linear bandwidth selection

Let $a_{1,l}^0, \ldots, a_{q,l}^0$ denote those values of $a_1, \ldots, a_q$ that minimize χ_{ll}, and assume that

Each $a_{s,l}^0$ is uniquely defined and is positive and finite.

Letting the $\hat{h}_s$'s denote those values of the h_s 's that minimize $CV_{ll}\left(h_1, \ldots, h_q\right)$, Li and Racine (2004) showed that

$$n^{1/(q+4)} \hat{h}_s \rightarrow a_{s,l}^0$$

in probability for $s = 1, \ldots, q$

Table of Contents

Derivative estimation

In this section, we present a fully automated framework to estimate derivatives nonparametrically without estimating the regression function.

Difference quotient

Using the first order difference quotient

$$Y_i^{(1)} = \frac{Y_i - Y_{i-1}}{x_i - x_{i-1}}$$

as a noise corrupted version of $m'\left(x_i\right)$. Such an approach produces a very noisy estimate of the derivative which is of the order $O\left(n^2\right)$ and as a result it will be difficult to estimate the derivative function. For equispaced design yields

$$\operatorname{Var}\left(Y_i^{(1)}\right) = \frac{2\sigma^2}{\left(x_i - x_{i-1}\right)^2} = \frac{2\sigma^2(n-1)^2}{d(\mathcal{X})^2}$$

where $d(\mathcal{X}) := \sup \mathcal{X} - \inf \mathcal{X}$.

Linear combination of difference quotients

In order to reduce the variance we use a variance-reducing linear combination of symmetric (about i) difference quotients

$$Y_i^{(1)} = Y^{(1)}(x_i) = \sum_{j=1}^{k} w_j \cdot \left(\frac{Y_{i+j} - Y_{i-j}}{x_{i+j} - x_{i-j}} \right)$$

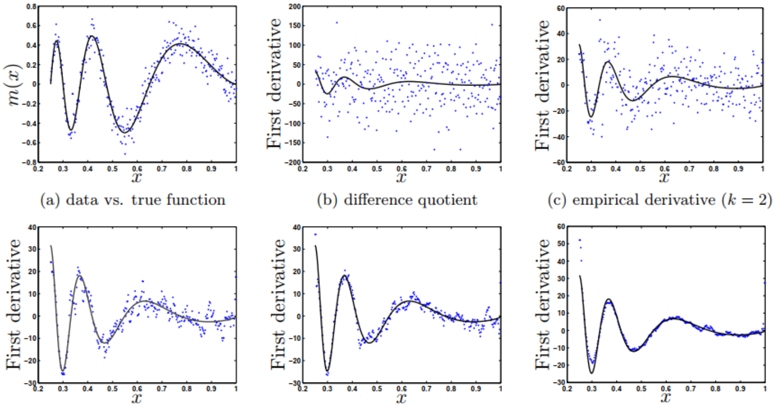
Theorem 6

Assume equispaced design and let $\sum_{j=1}^{k} w_j = 1$. Then, for $k+1 \leq i \leq n-k$, the weights

$$w_j = \frac{6j^2}{k(k+1)(2k+1)}, \quad j = 1, \ldots, k$$

minimize the variance of $Y_i^{(1)}$.

Figure of empirical first derivatives



(a) data vs. true function (b) difference quotient (c) empirical derivative ($k = 2$)

(d) empirical derivative ($k = 5$) (e) empirical derivative ($k = 7$) (f) empirical derivative ($k = 12$)

Figure 1: (a) Simulated data set of size $n = 300$ equispaced points from model (1) with $m(x) = \sqrt{x(1-x)}\sin((2.1\pi)/(x + 0.05))$ and $e \sim \mathcal{N}(0, 0.1^2)$; (b) first order difference quotients which are barely distinguishable from noise. As a reference, the true derivative is also displayed (full line); (c)-(f) empirical first derivatives for $k \in \{2, 5, 7, 12\}$.

Asymptotic bias and variance

Theorem 7

Assume the model holds with equispaced design and m is twice continuously differentiable on $\mathcal{X} \subseteq \mathbb{R}$. Further, assume that the second order derivative $m^{(2)}$ is finite on $\mathcal{X}$. Then the bias and variance of the empirical first order derivative, with weights assigned by Theorem 6, satisfy

$$\operatorname{bias}\left(Y_i^{(1)}\right) = O\left(n^{-1}k\right) \quad \textit{and} \quad \operatorname{Var}\left(Y_i^{(1)}\right) = O\left(n^2 k^{-3}\right)$$

uniformly for $k+1 \leq i \leq n-k$.

Pointwise consistency

Theorem 8

Assume $k \to \infty$ as $n \to \infty$ such that $nk^{-3/2} \to 0$ and $n^{-1}k \to 0$. Further assume that m is twice continuously differentiable on $\mathcal{X} \subseteq \mathbb{R}$. Then, for the minimum variance weights given in Theorem 6, we have for any $\epsilon > 0$

$$\mathbf{P}\left(\left|Y_i^{(1)} - m'\left(x_i\right)\right| \geq \epsilon\right) \to 0.$$

According to Theorem 7 and Theorem 8, the bias and variance of the empirical first order derivative tends to zero and $k \to \infty$ faster than $O\left(n^{2/3}\right)$ but slower than $O(n)$.

L_1 and L_2 rates

- The optimal rate at which $k \to \infty$ such that the MSE of the empirical first order derivatives will tend to zero at the fastest possible rate is a direct consequence of Theorem 7. This optimal L_2 rate is achieved for $k = O\left(n^{4/5}\right)$ and consequently

$$\text{MSE}\left(Y_i^{(1)}\right) = O\left(n^{-2/5} + n^{-1/5}\right)$$

- Similar, one can also establish the rate of the MAD or L_1 rate of the estimator i.e. $\mathbf{E}\left|Y_i^{(1)} - m'\left(x_i\right)\right|$. By Jensen's inequality

$$\begin{aligned}\mathbf{E}\left|Y_i^{(1)} - m'\left(x_i\right)\right| &\leq \mathbf{E}\left|Y_i^{(1)} - \mathbf{E}\left(Y_i^{(1)}\right)\right| + \left|\mathbf{E}\left(Y_i^{(1)}\right) - m'\left(x_i\right)\right| \\ &\leq \sqrt{\text{Var}\left(Y_i^{(1)}\right)} + \text{bias}\left(Y_i^{(1)}\right) = O\left(n^{-1/5}\right)\end{aligned}$$

for the optimal L_1 rate of $k = O\left(n^{4/5}\right)$.

Choose k

Even though we know the optimal asymptotic order, how to choose k in practice?

An upperbound for the MSE is given by

$$\begin{aligned}\mathrm{MSE}\left(Y_i^{(1)}\right) &= \mathrm{bias}^2\left(Y_i^{(1)}\right) + \mathrm{Var}\left(Y_i^{(1)}\right)\\ &\leq \frac{9k^2(k+1)^2\mathcal{B}^2 d(\mathcal{X})^2}{16(n-1)^2(2k+1)^2} + \frac{3\sigma^2(n-1)^2}{k(k+1)(2k+1)d(\mathcal{X})^2},\end{aligned}$$

where $\mathcal{B} = \sup_{x\in\mathcal{X}}\left|m^{(2)}(x)\right|$. Setting the derivative w.r.t. k to zero yields

$$3\mathcal{B}^2 d(\mathcal{X})^4 k^3(1+k)^3\left(1+2k+2k^2\right) = 8\left(1+8k+18k^2+12k^3\right)(n-1)^4\sigma^2$$

Solving this with the constraint that $k>0$ will result in the value k for which the MSE is lowest.

Choose k (rule of thumb)

A much simpler rule of thumb is obtained by only considering the highest order terms yielding

$$k = \left(\frac{16\sigma^2}{\mathcal{B}^2 d(\mathcal{X})^4} \right)^{1/5} n^{4/5}$$

The error variance σ^2 can be estimated by means of Hall's $\sqrt{n}$-consistent estimator

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^{n-2} (0.809Y_i - 0.5Y_{i+1} - 0.309Y_{i+2})^2$$

Choose k (rule of thumb)

For the second unknown quantity $\mathcal{B}$ one can use the local polynomial regression estimate of order $p = 3$ leading to the following (rough) estimate of the second derivative $\hat{m}^{(2)}(x_0) = 2\hat{\beta}_2$. Consequently, a rule of thumb selector for k is given by

$$\hat{k} = \left(\frac{16\hat{\sigma}^2}{\left(\sup_{x_0 \in \mathcal{X}} \left|\hat{m}^{(2)}(x_0)\right|\right)^2 d(\mathcal{X})^4} \right)^{1/5} n^{4/5}.$$

In practice we round the obtained k value closest to the next integer value. As an alternative, one could also consider cross-validation to find an optimal k.

Higher order empirical derivatives

We generalize the idea of first order empirical derivatives to higher order derivatives. Let q denote the order of the derivative and assume further that $q \geq 2$, then higher order empirical derivatives can be defined inductively as

$$Y_i^{(l)} = \sum_{j=1}^{k_l} w_{j,l} \cdot \left(\frac{Y_{i+j}^{(l-1)} - Y_{i-j}^{(l-1)}}{x_{i+j} - x_{i-j}} \right) \quad \text{with} \quad l \in \{2, \ldots, q\}$$

where $k_1, k_2, \ldots, k_q$ are positive integers (not necessary equal), the weights at each level l sum up to one and $Y_i^{(0)} = Y_i$ by definition.

Higher order empirical derivatives

Theorem 9

Assume equispaced design and let $\sum_{j=1}^{k_l} w_{j,l} = 1$. *Further assume that the first* $(q+1)$ *derivatives of* m *are continuous on the interval* $\mathcal{X}$. *Assume that there exist* $\lambda \in (0,1)$ *and* $c_l \in (0,\infty)$ *such that* $k_l n^{-\lambda} \to c_l$ *for* $n \to \infty$ *and* $l \in \{1,2,\ldots,q\}$. *Further, assume that* $w_{j,1}$ *given in Theorem 6 and*

$$w_{j,l} = \frac{2j}{k_l\,(k_l+1)} \quad \textit{for} \quad j = 1,\ldots,k_l \quad \textit{and} \quad l \in \{2,\ldots,q\}.$$

Then the asymptotic bias and variance of the empirical qth order derivative are given by

$$\text{bias}\left(Y_i^{(q)}\right) = O\left(n^{\lambda-1}\right) \quad \textit{and} \quad \text{Var}\left(Y_i^{(q)}\right) = O\left(n^{2q-2\lambda(q+1/2)}\right)$$

uniformly for $\sum_{l=1}^{q} k_q + 1 < i < n - \sum_{l=1}^{q} k_q$.

Rates and pointwise consistency

Theorem 10

Under assumptions, the asymptotic MSE and asymptotic MAD are given by

$$\mathbf{E}\left(Y_i^{(q)} - m^{(q)}\left(x_i\right)\right)^2 = O\left(n^{2(\lambda-1)} + n^{2q-2\lambda(q+1/2)}\right)$$
$$\mathbf{E}\left|Y_i^{(q)} - m^{(q)}\left(x_i\right)\right| = O\left(n^{\lambda-1} + n^{q-\lambda(q+1/2)}\right)$$

Theorem 11

Under the assumptions and $\lambda \in \left(\frac{2q}{2q+1}, 1\right)$, *it follows that*

$$\mathbf{E}\left(Y_i^{(q)} - m^{(q)}\left(x_i\right)\right)^2 \to 0 \quad \textit{and} \quad \mathbf{E}\left|Y_i^{(q)} - m^{(q)}\left(x_i\right)\right| \to 0, \quad n \to \infty$$

Simulation

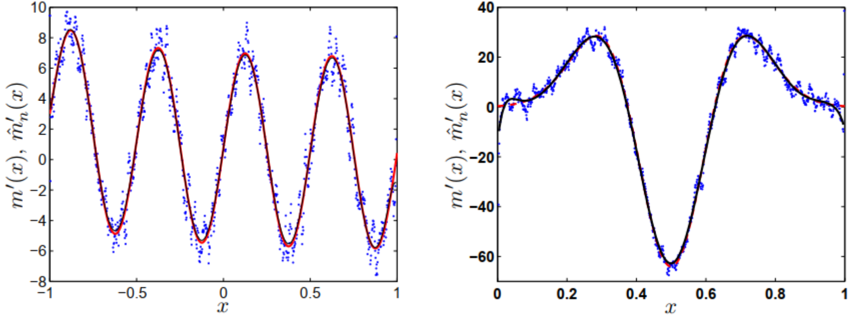


Figure 2: Illustration of the noisy empirical first order derivative (data points), smoothed empirical first order derivative based on a local polynomial regression estimate of order $p = 3$ (bold line) and true derivative (bold dashed line). (a) First order derivative of regression function (11) with $k_1 = 7$; (b) First order derivative of regression function (12) with $k_1 = 12$.

Table of Contents

References

- J.Fan & I.Gijbels (1996), **Local Polynomial Modelling and Its Applications**, Chapman & Hall (Sections 1-4)
- Q.Li & J.S.Racine (2006), **Nonparametric Econometrics: Theory and Practice**, Princeton University Press (Sections 1-2)
- T.Hastie, R.Tibshirani & J.Friedman (2017), **The Elements of Statistical Learning**, Springer (Section 6)
- K.De Brabanter, J.De Brabanter, I.Gijbels, & B.De Moor (2013), **Derivative estimation with local polynomial fitting**. Journal of Machine Learning Research, 14(1), 281-301.