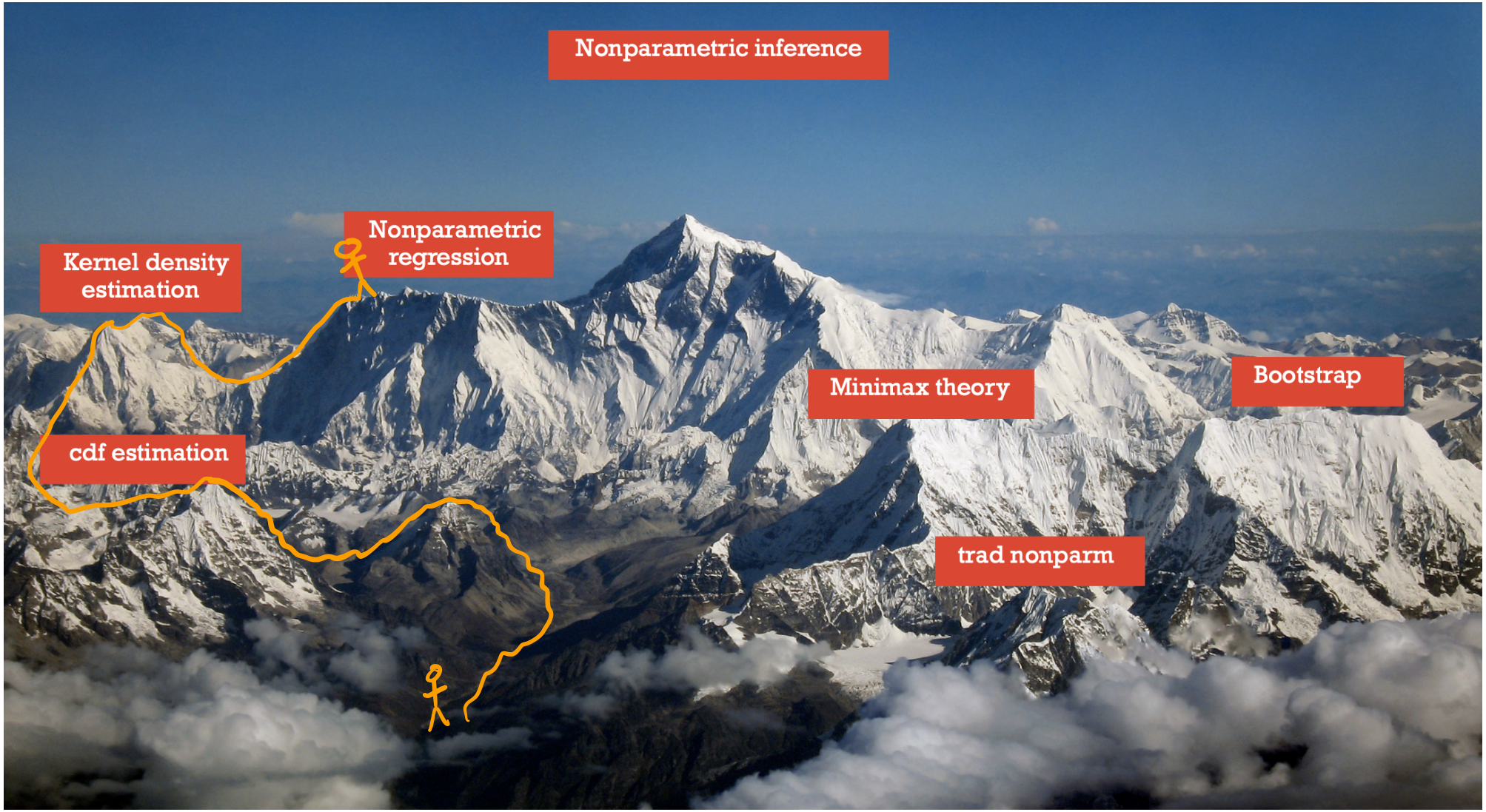
STAT 824 sp 2025 Lec 04 slides

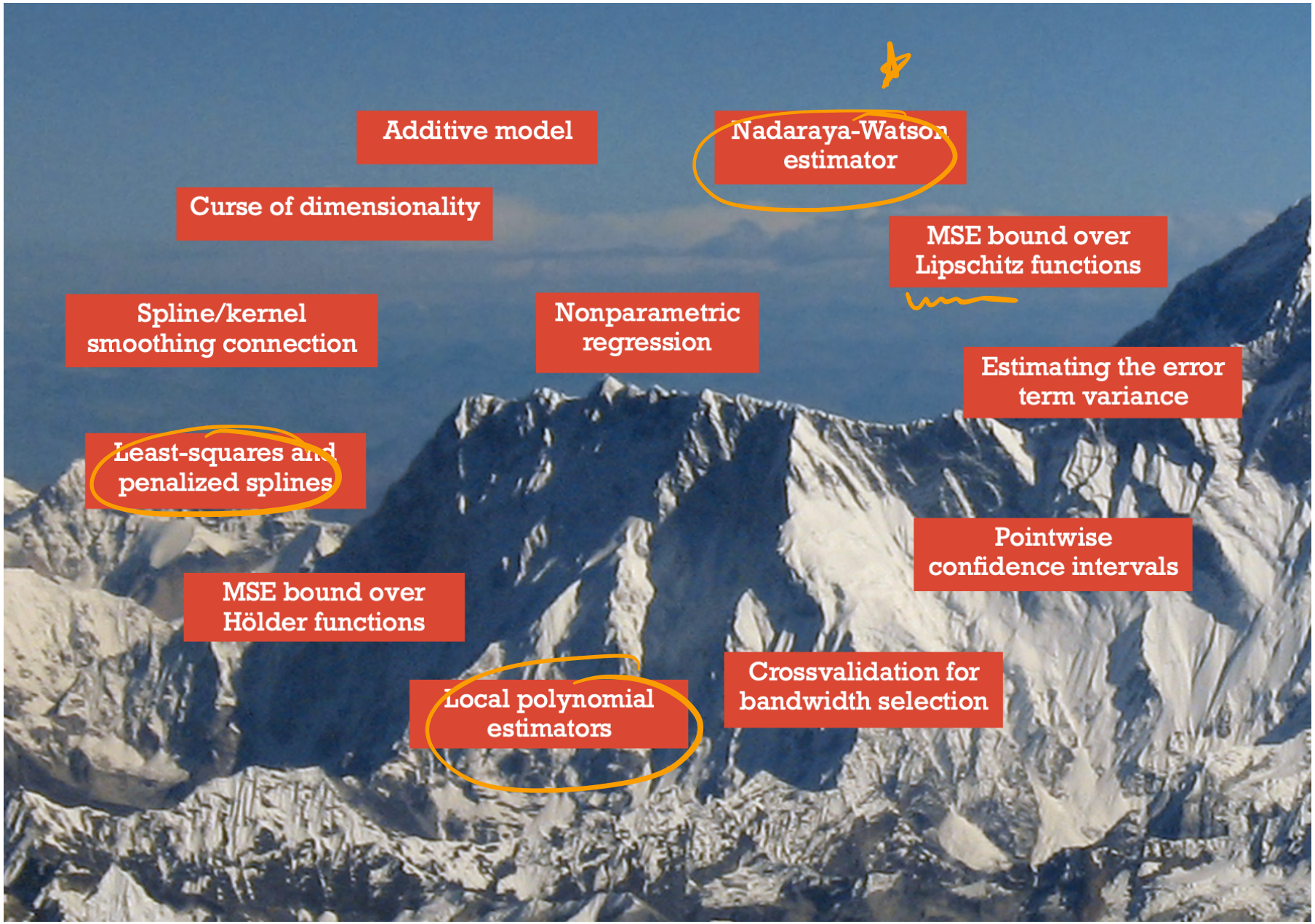
Nonparametric regression: N-W and LP smoothers

Karl Gregory

University of South Carolina

These slides are an instructional aid; their sole purpose is to display, during the lecture, definitions, plots, results, etc. which take too much time to write by hand on the blackboard. They are not intended to explain or expound on any material.


Nonparametric inference
Nonparametric regression
Kernel density estimation
cdf estimation
Minimax theory
Bootstrap
trad nonparm

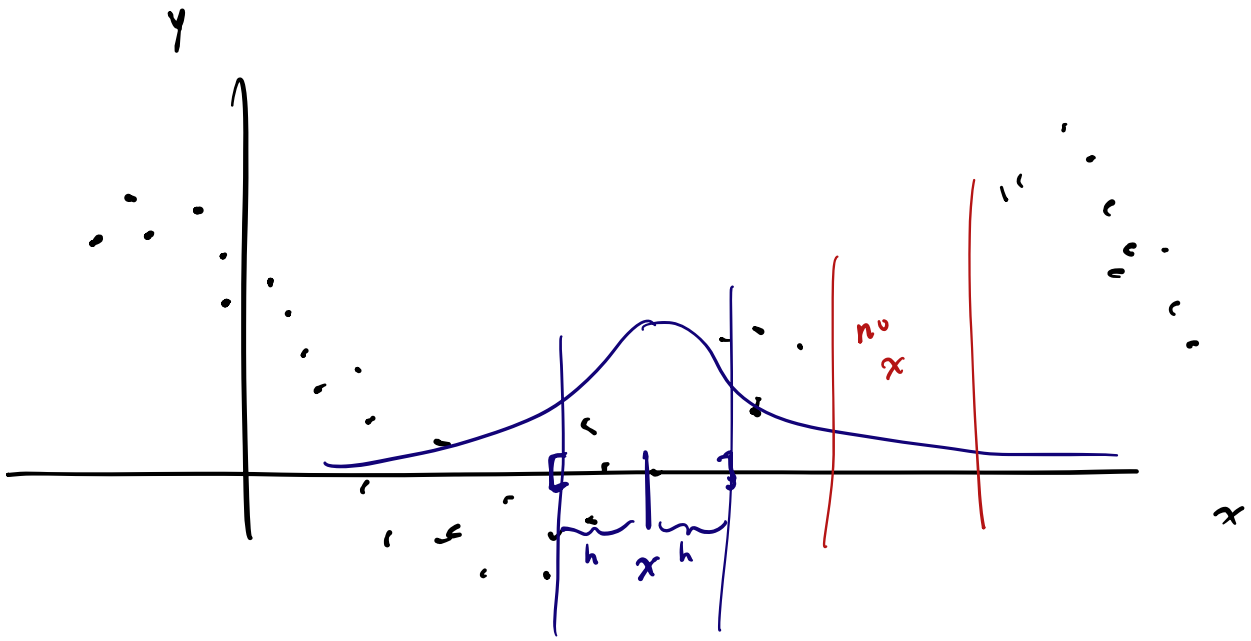

Additive model
Nadaraya-Watson estimator
Curse of dimensionality
MSE bound over Lipschitz functions
Spline/kernel smoothing connection
Nonparametric regression
Estimating the error term variance
Least-squares and penalized splines
Pointwise confidence intervals
MSE bound over Hölder functions
Crossvalidation for bandwidth selection
Local polynomial estimators

Observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ iid copies of (X, Y), where

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

$$\varepsilon \perp\!\!\!\perp X \quad (\text{independent}), \qquad \mathbb{E}\,\varepsilon = 0, \qquad \mathbb{E}\,\varepsilon^2 = \sigma^2.$$

Goal: Estimate function m, σ^2.



"Local average estimator"

$$\hat{m}(x) = \frac{\sum_{i=1}^n Y_i \, \mathbb{1}(x-h \le X_i \le x+h)}{\sum_{i=1}^n \mathbb{1}(x-h \le X_i \le x+h)}$$

$$= \frac{\sum_{i=1}^n Y_i \, \mathbb{1}\left(-1 \le \frac{X_i - x}{h} \le 1\right)}{\sum_{i=1}^n \mathbb{1}\left(-1 \le \frac{X_i - x}{h} \le 1\right)}$$

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

$$k(z) = \mathbb{1}(-1 \le z \le 1)$$

Nonparametric regression model

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be independent realizations of $(X, Y) \in T \times \mathbb{R}$, where

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

where ε is independent of X with $\mathbb{E}\varepsilon = 0$ and $\mathbb{E}\varepsilon^2 = \sigma^2$.

- Note that $m(x) = \mathbb{E}[Y|X = x]$.
- Will assume that m belongs to some class of smooth functions.
- Linear regression assumes $m \in \text{Lin}(T)$, where

$$\text{Lin}(T) = \{m : T \to \mathbb{R} : m(x) = \beta_0 + \beta_1 x, \text{ for some } \beta_0, \beta_1 \in \mathbb{R}\}.$$

- We will proceed treating $X_1, \ldots, X_n$ as deterministic (fixed) in $[0, 1]$.

Consider the "local-average" estimator of m given by

$$\bar{m}_n(x) = \frac{\sum_{i=1}^n Y_i \cdot \mathbf{1}(x - h \leq X_i \leq x + h)}{\sum_{j=1}^n \mathbf{1}(x - h \leq X_j \leq x + h)}$$

for all x, for some $h > 0$.

Discuss: Potential problems? Draw pictures.

Nadaraya-Watson estimator

The *Nadaraya-Watson (N-W)* estimator of m is given by

$$\hat{m}_n^{\mathrm{NW}}(x) = \frac{\sum_{i=1}^n Y_i \cdot K\left(h^{-1}(X_i - x)\right)}{\sum_{j=1}^n K\left(h^{-1}(X_j - x)\right)}$$

for all x, for some $h > 0$ and kernel function K.

Exercise: Show that $\hat{m}_n^{\mathrm{NW}}(x)$ is an estimate of $\mathbb{E}[Y|X = x]$ based on

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) \text{ and } \hat{f}_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right).$$

$$\mathbb{E}_{\hat{f}_n(y|x)}\left[Y \,\middle|\, X=x\right] = \int_{\mathbb{R}} y \;\hat{f}_n(y|x)\, dy$$

$$= \int_{\mathbb{R}} y \;\frac{\hat{f}_n(x,y)}{\hat{f}_n(x)}\, dy$$

$$= \frac{1}{\hat{f}_n(x)} \int_{\mathbb{R}} y \;\hat{f}_n(x,y)\, dy$$

$$= \frac{1}{\frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \int_{\mathbb{R}} y \;\frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) K\left(\frac{y_i - y}{h}\right) dy$$

$$= \frac{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \int_{\mathbb{R}} y \;K\left(\frac{y_i - y}{h}\right) dy}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

$u = \frac{y_i - y}{h}$

$y = y_i - hu$

$dy = -h\,du$

$$= \frac{\frac{1}{h}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \int_{\infty}^{-\infty} (y_i - hu)\, K(u)\, (-h)\, du}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

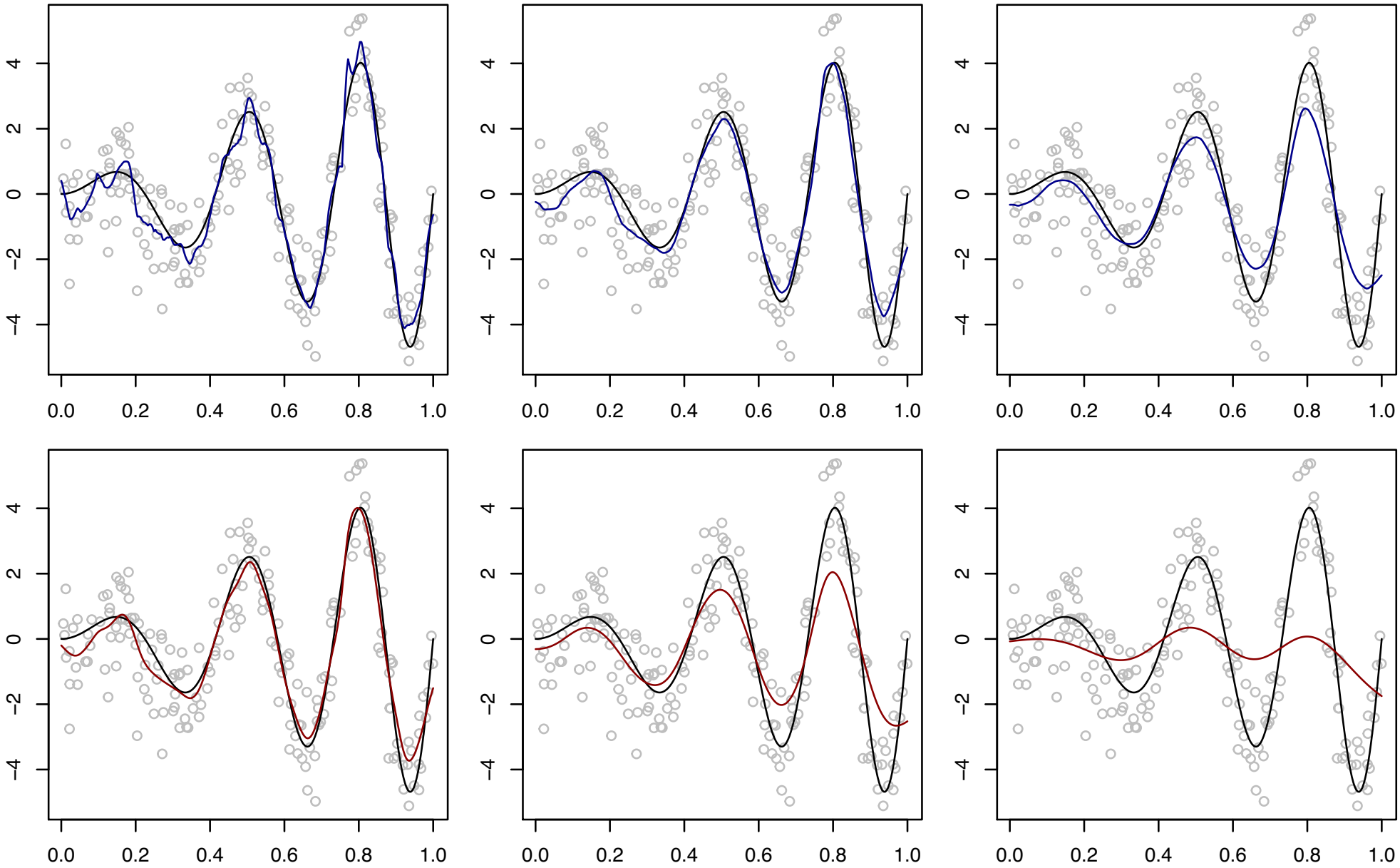
$$= \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \int_{-\infty}^{\infty} (y_i - hu)\, K(u)\, du}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

$$= \frac{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right) \left(\int_{-\infty}^{\infty} y_i\, K(u)\, du - h \int_{-\infty}^{\infty} u\, K(u)\, du \right)}{\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)}$$

$= y_i$

$= 0$ if K symm. around 0.

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

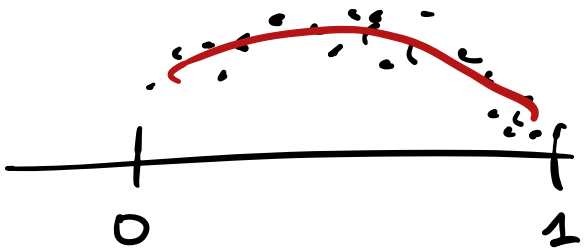

4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0

Exercise: Make plots of the NW estimator as on the previous slide, where

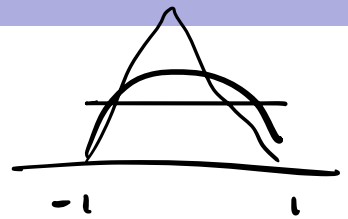
$$m(x) = 5x \sin(2\pi(1+x)^2), \quad \varepsilon \sim \text{Normal}(0, \sigma^2),$$

under the

1. Epinechnikov kernel $K(z) = (3/4)(1-z^2)\mathbf{1}(|z| \leq 1)$.
2. Gaussian kernel $K(z) = 1/\sqrt{2\pi} e^{-z^2/2}$.



We now find a bound for MSE $\hat{m}_n^{\mathrm{NW}}(x)$ under simple conditions:



(K1) Let K have support on $[-1,1]$ and let $0 \le K(u) \le K_{\max} < \infty \ \forall\ u \in \mathbb{R}$.

(D1) Let $X_1, \ldots, X_n$ be deterministic such that for some $n_0 > 0$

$$\hat{f}_n(x) = (nh)^{-1} \sum_{i=1}^{n} K(h^{-1}(X_i - x)) \ge f_{\min} > 0 \quad \text{for all } x \in [0,1]$$

for all $n \ge n_0$.

Pointwise MSE bound for N-W under Lipschitz smoothness

Under (K1) and (D1), if $m \in \mathrm{Lipschitz}(L)$ on $[0,1]$, we have

$$\mathrm{MSE}\, \hat{m}_n^{\mathrm{NW}}(x_0) \le h^2 \cdot L^2 + \frac{\sigma^2}{nh} \cdot \frac{K_{\max}}{f_{\min}} \quad \text{for all } x_0 \in [0,1]$$

for all $n \ge n_0$.

$\Rightarrow h_{opt} = \text{const.}\ n^{-1/3} \quad \Rightarrow \mathrm{MSE}_{opt} \le \text{const.}\ n^{-2/3}$

Exercise:

1. Find the weights $W_{ni}(x)$, $i = 1, \ldots, n$, such that we can write

$$\hat{m}_n^{\mathrm{NW}}(x) = \sum_{i=1}^{n} W_{ni}(x)\, Y_i.$$

2. Prove the result on the previous slide.
3. Find the optimal bandwidth (in terms of minimizing the MSE bound).

(1) $\hat{m}_n^{NW}(x) = \dfrac{\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)} = \sum_{i=1}^{n} W_{ni}(x)\, Y_i$,

where

$$W_{ni}(x) = \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x_j - x}{h}\right)}$$

(2) Upper bound for $\mathrm{Var}\ \hat{m}_n^{NW}(x)$:

$$\mathrm{Var}\ \hat{m}_n^{NW}(x) = \mathrm{Var}\left(\sum_{i=1}^{n} W_{ni}(x)\, Y_i \right)$$

($W_{ni}(x) \geq 0$ by $K \geq 0$)

$$= \sigma^2 \sum_{i=1}^{n} W_{ni}^2(x)$$

$$\leq \sigma^2 \left(\max_{1 \leq i \leq n} W_{ni}(x) \right) \sum_{i=1}^{n} W_{ni}(x)$$

($\sum_{i=1}^{n} W_{ni}(x) = 1$)

$$= \frac{\sigma^2}{nh} \left(\max_{1 \leq i \leq n} \frac{K\left(\frac{x_i - x}{h}\right)}{\frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right)} \right)$$

($K\left(\frac{x_i - x}{h}\right) \leq K_{max}$; $\hat{f}_n(x) \geq f_{min} > 0$)

$$\leq \frac{\sigma^2}{nh} \cdot \frac{K_{max}}{f_{min}}$$

Upper bound of $\left| \text{Bias } \hat{m}_n^{NW}(x) \right|$ $\qquad y_i = m(x_i) + \varepsilon_i$

$$\text{Bias } \hat{m}_n^{NW}(x) = \mathbb{E}\, \hat{m}_n^{NW}(x) - m(x)$$

$$= \mathbb{E} \sum_{i=1}^{n} W_{ni}(x)\, Y_i - m(x)$$

$$= \sum_{i=1}^{n} W_{ni}(x)\, m(x_i) - m(x)$$

$$= \sum_{i=1}^{n} W_{ni}(x) \left[m(x_i) - m(x) \right]$$

$$|m(x) - m(x')| \leq L\, |x - x'|$$

$$\left| \text{Bias } \hat{m}_n^{NW}(x) \right| \leq \sum_{i=1}^{n} W_{ni}(x) \left| m(x_i) - m(x) \right|$$

$$\leq \sum_{i=1}^{n} W_{ni}(x)\, L\, |x_i - x|$$

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

$$= \sum_{i : W_{ni}(x) \neq 0} W_{ni}(x)\, L\, |x_i - x|$$

$= 0 \iff \frac{x_i - x}{h} \notin [-1, 1]$

$\iff |x_i - x| > h$

$$= \sum_{i : W_{ni}(x) \neq 0} W_{ni}(x)\, L\, h$$

$$= L \cdot h \underbrace{\sum_{i=1}^{n} W_{ni}(x)}_{=1}$$

$$= L \cdot h$$

Plugging in the optimal bandwidth choice gives the following result:

Theorem 1: Uniform MSE bound of N-W over Lipschitz functions

Under (K1) and (D1) with $h = \alpha n^{-1/3}$ for some $\alpha > 0$, we have

$$\sup_{m \in \text{Lipschitz}(L) \text{ on } [0,1]} \sup_{x \in [0,1]} \mathbb{E}[\hat{m}_n^{\text{NW}}(x) - m(x)]^2 \leq n^{-2/3} \cdot C,$$

for all $n \geq n_0$, where $C > 0$ depends on $\alpha, L, K_{\text{max}}$, and f_{min}.

$$Y_i = m(x_i) + \varepsilon_i$$

$$\text{Lin}([0,1]) = \left\{ m : [0,1] \to \mathbb{R} \;:\; m(x) = \beta_0 + \beta_1 x, \quad \beta_0, \beta_1 \in \mathbb{R} \right\}$$

For $m \in \text{Lin}([0,1])$, use estimator

$$\hat{m}_n^{LS}(x) = \hat{\beta}_0 + \hat{\beta}_1 x \quad , \quad \hat{\beta}_0, \hat{\beta}_1 \text{ are LS estimators.}$$

Then

$$\text{MSE } \hat{m}_n^{LS}(x) = \text{Var}\left(\hat{\beta}_0 + \hat{\beta}_1 x \right)$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x}_n)^2}{\sum_{i=1}^n (x_i - \bar{x}_n)^2} \right]$$

$$= \frac{\sigma^2}{n} \left[1 + \frac{(x - \bar{x}_n)^2}{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2} \right]$$

$\geq S_{min} > 0$

$$\leq \frac{\sigma^2}{n} \left[1 + \frac{(x - \bar{x}_n)^2}{S_{min}} \right]$$

Consider the MSE bound for $\hat{m}_n^{\text{Lin}}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$, where

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^n [Y_i - (\beta_0 + \beta_1 X_i)]^2.$$

We need one assumption:

(D1lin) Let $X_1, \ldots, X_n \in [0,1]$ be deterministic such that for some $n_0 > 0$

$$n^{-1} \sum_{i=1}^n (X_i - \bar{X}_n)^2 \geq S_{\text{min}} > 0 \text{ for all } n \geq n_0.$$

Uniform MSE bound for simple linear regression

Under (D1lin), for $n \geq n_0$, we have

$$\sup_{m \in \text{Lin}([0,1])} \sup_{x \in [0,1]} \mathbb{E}\left[\hat{m}_n^{\text{Lin}}(x) - m(x)\right]^2 \leq \frac{\sigma^2}{n} \cdot \left[1 + \frac{1}{S_{\text{min}}}\right].$$

Exercise: Prove the above result and discuss.

Table of Contents

1 Nonparametric regression

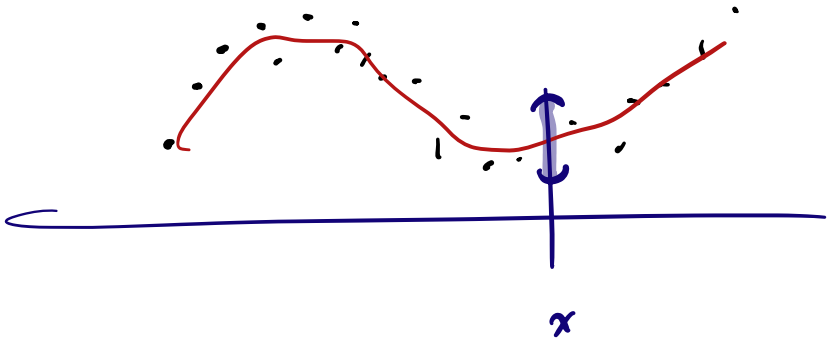
2 The Nadaraya-Watson estimator

3 Pointwise confidence intervals

4 Variance estimation

5 Crossvalidation for bandwidth selection

6 Local polynomial estimators



Want a pivotal quantity ...

If $$\frac{\hat{m}_n^{NW}(x) - m(x)}{\hat{se}\{\hat{m}_n^{NW}(x)\}} \overset{?}{\overset{approx}{\sim}} N(0,1)$$

Then an approximate $(1-\alpha)*100\%$ C.I. for $m(x)$ is

$$\hat{m}_n^{NW}(x) \pm z_{\alpha/2}\, \hat{se}\{\hat{m}_n^{NW}(x)\}$$

First, $$Var\left(\hat{m}_n^{NW}(x)\right) = Var\left(\sum_{i=1}^n W_{ni}(x)\, Y_i\right)$$
$$= \sigma^2 \sum_{i=1}^n W_{ni}^2(x) .$$

So consider this:

$$\frac{\hat{m}_n^{NW}(x) - m(x)}{\sigma\sqrt{\sum_{i=1}^n W_{ni}^2(x)}} = \frac{\hat{m}_n^{NW}(x) - \mathbb{E}\,\hat{m}_n^{NW}(x)}{\sigma\sqrt{\sum_{i=1}^n W_{ni}^2(x)}} + \frac{\mathbb{E}\,\hat{m}_n^{NW}(x) - m(x)}{\sigma\sqrt{\sum_{i=1}^n W_{ni}^2(x)}}$$

Variance term

bias term

$\xrightarrow{D} N(0,1)$

$\to 0$

as $n \to \infty$ under the right choice of h_n.

We have $\hat{m}_n^{NW}(x) - \mathbb{E}\,\hat{m}_n^{NW}(x) = \sum_{i=1}^n W_{ni}(x)\, Y_i - \mathbb{E} \sum_{i=1}^n W_{ni}(x)\, Y_i$

$= \sum_{i=1}^n W_{ni}(x)\, \varepsilon_i$

$$\frac{\hat{m}_n^{NW}(x) - \mathbb{E}\,\hat{m}_n^{NW}(x)}{\sigma \sqrt{\sum_{i=1}^n W_{ni}^2(x)}} = \frac{\sum_{i=1}^n W_{ni}(x) (\varepsilon_i/\sigma)}{\sqrt{\sum_{i=1}^n W_{ni}^2(x)}} \to N(0,1)$$

(a_i: $W_{ni}(x)$; $\xi_i \sim N(0,1)$: ε_i/σ)

provided

$$\frac{\max_{1\le i\le n} |a_i|}{\left[\sum_{j=1}^n a_j^2\right]^{1/2}} \to 0 \quad \text{as } n \to \infty,$$

$$\frac{\max_{1\le i\le n} |W_{ni}(x)|}{\sqrt{\sum_{i=1}^n W_{ni}^2(x)}} \to 0 \quad \text{as } n \to \infty.$$

We have

$$\frac{\max_{1\le i\le n} |W_{ni}(x)|}{\sqrt{\sum_{i=1}^n W_{ni}^2(x)}} = \frac{\max_{1\le i\le n} \left| \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right)} \right|}{\sqrt{\sum_{i=1}^n \frac{K^2\left(\frac{x_i - x}{h}\right)}{\left(\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right)\right)^2}}}$$

$$W_{ni}(x) : \frac{K\left(\frac{x_i - x}{h}\right)}{\sum_{j=1}^n K\left(\frac{x_j - x}{h}\right)}$$

$$= \frac{\max_{1 \le i \le n} \left| K\left(\frac{x_i - x}{h}\right) \right|}{\sqrt{\sum_{i=1}^{n} K^2\left(\frac{x_i - x}{h}\right)}} \qquad \left(\le K_{max} \right)$$

$$\le \frac{K_{max}}{\sqrt{nh}\sqrt{\frac{1}{nh}\sum_{i=1}^{n} K^2\left(\frac{x_i - x}{h}\right)}} \qquad \left(\ge M > 0 \right)$$

$$\le \frac{1}{\sqrt{nh}} \frac{K_{max}}{\sqrt{M}}$$

$$\longrightarrow 0 \quad \text{if} \quad nh \to \infty .$$

We can choose $\underline{h_{opt} = c \cdot n^{-1/3}}$

$$\left| \frac{\mathbb{E}\,\hat{m}_n^{NW}(x) - m(x)}{\sigma \sqrt{\sum_{i=1}^n W_{ni}^2(x)}} \right| = \frac{\left| \text{Bias}\, \hat{m}_n^{NW}(x) \right|}{\sigma \sqrt{\sum_{i=1}^n W_{ni}^2(x)}}$$

$$\leq \frac{h \cdot L}{\sigma \sqrt{\sum_{i=1}^n W_{ni}^2(x)}}$$

$$= \frac{h \cdot L}{\sigma \sqrt{\dfrac{\sum_{i=1}^n K^2\left(\frac{X_i - x}{h}\right)}{\left(\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)\right)^2}}}$$

$$= \frac{nh\, h \cdot L \; \frac{1}{nh} \sum_{i=1}^n K\left(\frac{X_i - x}{h}\right)}{\sigma \sqrt{nh} \sqrt{\frac{1}{nh} \sum_{i=1}^n K^2\left(\frac{X_i - x}{h}\right)}}$$

$\hat{f}_n(x) \leq f_{max} < \infty$

$\geq M > 0$

$$\leq n^{1/2} h^{3/2} \frac{L}{\sigma} \frac{f_{max}}{\sqrt{M}}$$

$$\longrightarrow 0 \quad \text{if} \quad n^{1/2} h^{3/2} \longrightarrow 0$$

$$h_{opt} = C \cdot n^{-1/3} \Rightarrow n^{1/2} h_{opt}^{3/2} = n^{1/2} \left(n^{-1/3}\right)^{3/2} = 1$$

The bias term DOES NOT vanish under the MSE-optimal choice of h!!!!

Can we make a confidence interval for $m(x_0)$ for some $x_0 \in [0,1]$?

If for some sequence $\vartheta_n(x_0)$ we have

$$\left(\hat{m}_n^{\text{NW}}(x_0) - m(x_0)\right) / \vartheta_n(x_0) \to \text{Normal}(0,1) \text{ in dist. as } n \to \infty,$$

then the interval

$$\hat{m}_n^{\text{NW}}(x_0) \pm z_{\alpha/2}\vartheta_n(x_0)$$

would contain $m(x_0)$ with probability approaching $1-\alpha$ as $n \to \infty$.

Since we are talking about sending $n \to \infty$...

Discuss: What must happen to h as $n \to \infty$ for $\hat{m}_n^{\text{NW}}(x_0)$ to be consistent?

To get the desired result sketched on the previous slide, we will use the following:

Corollary to the Lindeberg Central Limit Theorem

For a seq. of iid rvs $\xi_1, \xi_2, \ldots$ with zero mean and unit variance and a seq. of numbers $a_1, a_2, \ldots$ satisfying

$$\frac{\max_{1 \leq i \leq n} |a_i|}{\left[\sum_{j=1}^n a_j^2\right]^{1/2}} \to 0 \quad \text{as } n \to \infty,$$

we have

$$\frac{\sum_{i=1}^n a_i \cdot \xi_i}{\left[\sum_{j=1}^n a_j^2\right]^{1/2}} \to \text{Normal}(0,1) \text{ in dist. as } n \to \infty.$$

$$\text{Var}\left(\sum_{i=1}^n a_i \xi_i\right) = \sum_{i=1}^n a_i^2$$

This CLT is highly useful in regression contexts.

Theorem 1: Asymptotic behavior of N-W under Lipschitz smoothness

Assume (K1) and let $X_1, \ldots, X_n \in [0,1]$ be deterministic such that for any sequence $h_n \to 0$ and $nh_n \to \infty$ there exists $n_0 > 0$ such that

$$\frac{1}{nh_n} \sum_{i=1}^{n} K^2\left(\frac{X_i - x}{h_n}\right) \geq M > 0 \quad \text{and} \quad \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{X_i - x}{h_n}\right) \leq f_{\max} < \infty$$

for all $x_0 \in [0,1]$ for all $n \geq n_0$. Then for $m \in \text{Lipschitz}([0,1])$, we have

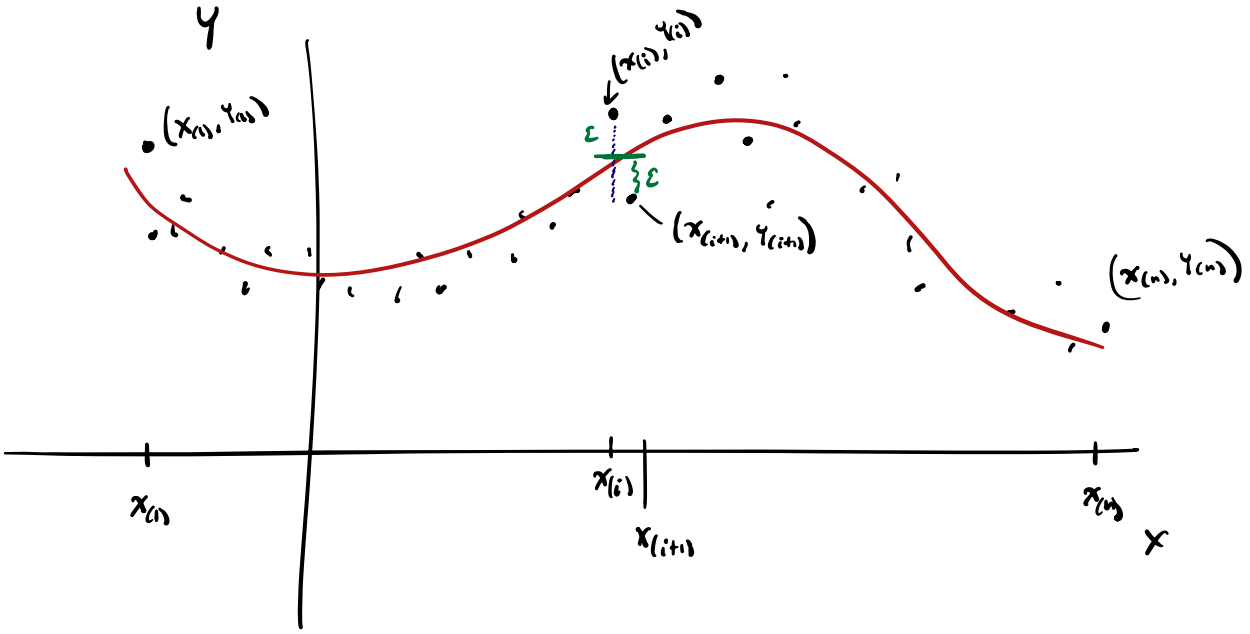
1. $$\frac{\hat{m}_n^{\text{NW}}(x_0) - \mathbb{E}\hat{m}_n^{\text{NW}}(x_0)}{\sigma \left[\sum_{i=1}^{n} W_{ni}^2(x_0)\right]^{1/2}} \to \text{Normal}(0,1) \text{ as } n \to \infty.$$

2. $$\frac{|\mathbb{E}\hat{m}_n^{\text{NW}}(x_0) - m(x_0)|}{\sigma \left[\sum_{i=1}^{n} W_{ni}^2(x_0)\right]^{1/2}} \leq n^{1/2} h_n^{3/2} \cdot \frac{f_{\max} \cdot L}{\sigma \cdot M} \text{ for all } n \geq n_0.$$

Choose on purpose h smaller than h_{opt} (Make your estimator too wiggly on purpose.)

Exercise: Prove result, but discuss: Can we choose h_{opt}? Do *undersmoothing*...

How to estimate σ^2 ?



$$\mathbb{E}\left(y_{(i+1)} - y_{(i)}\right)^2 = \mathbb{E}\left(\left(m(x_{(i+1)}) + \varepsilon_{(i+1)}\right) - \left(m(x_{(i)}) + \varepsilon_{(i)}\right)\right)^2$$

$$= \mathbb{E}\left(\left(m(x_{(i+1)}) - m(x_{(i)})\right) + \left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)\right)^2$$

$$= \left(m(x_{(i+1)}) - m(x_{(i)})\right)^2$$

$$+ \mathbb{E}\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)^2$$

$$+ 2\,\mathbb{E}\left(m(x_{(i+1)}) - m(x_{(i)})\right)\left(\varepsilon_{(i+1)} - \varepsilon_{(i)}\right)$$

$$= \underbrace{\left(m(x_{(i+1)}) - m(x_{(i)})\right)^2}_{\leq L^2 \left| x_{(i+1)} - x_{(i)} \right|^2} + 2\sigma^2$$

Try

$$\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(Y_{(i+1)} - Y_{(i)} \right)^2$$

Then

$$\mathbb{E}\,\hat{\sigma}^2 = \sigma^2 + \underbrace{\frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(m(X_{(i+1)}) - m(X_{(i)}) \right)^2}$$

$$\leq \frac{1}{2(n-1)} \sum_{i=1}^{n} L^2 \left| X_{(i+1)} - X_{(i)} \right|^2$$

$X_1, \ldots, X_n \in [0,1]$

$$\leq \frac{1}{2(n-1)} \sum_{i=1}^{n} L^2 \left| X_{(i+1)} - X_{(i)} \right|$$

$$\leq \frac{1}{2(n-1)} L^2 \longrightarrow 0 \quad \text{as} \quad n \to \infty.$$

Table of Contents

1 Nonparametric regression

2 The Nadaraya-Watson estimator

3 Pointwise confidence intervals

4 Variance estimation

5 Crossvalidation for bandwidth selection

6 Local polynomial estimators

We now consider estimating σ^2.

Estimator of σ^2

Let $(X_{(1)}, Y_{(1)}), \ldots, (X_{(n)}, Y_{(n)})$ be the data reordered st $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$. Rice (1984) [1] suggested the estimator

$$\hat{\sigma}_n^2 = \frac{1}{2(n-1)} \sum_{i=1}^{n-1} \left(Y_{(i+1)} - Y_{(i)} \right)^2 .$$

Consistency of differencing estimator under Lipschitz smoothness

If $X_1, \ldots, X_n \in [0,1]$, $\mathbb{E}\varepsilon_1^4 < \infty$, and $m \in \text{Lipschitz}(L)$ on $[0,1]$,

$$\hat{\sigma}_n^2 \to \sigma^2 \text{ in prob. as } n \to \infty.$$

Exercise: Show that $\mathbb{E}\hat{\sigma}_n^2 \to \sigma^2$ as $n \to \infty$. Discuss how to prove consistency.

Finally, we can construct a confidence interval for $m(x_0)$:

Approximate $(1-\alpha)100\%$ confidence interval for $m(x_0)$

Under the conditions of Theorem 1 and if $\mathbb{E}\varepsilon_1^4 < \infty$, we have

$$P\left(m(x_0) \in \left(\hat{m}_n^{\text{NW}}(x_0) \pm z_{\alpha/2}\hat{\sigma}_n \left[\sum_{i=1}^n W_{ni}^2(x_0)\right]^{1/2}\right)\right) \to 1-\alpha$$

as $n \to \infty$, provided $n^{1/2}h_n^{3/2} \to 0$ and $nh_n \to \infty$, i.e. $n^{-1} \ll h_n \ll n^{-1/3}$.

I have used $a_n \ll b_n$ to mean $a_n/b_n \to 0$ as $n \to \infty$.

Discuss: How do we "undersmooth" to build confidence intervals in practice?

- Maybe we don't.

What about confidence bands?? We'll do this later with the bootstrap.

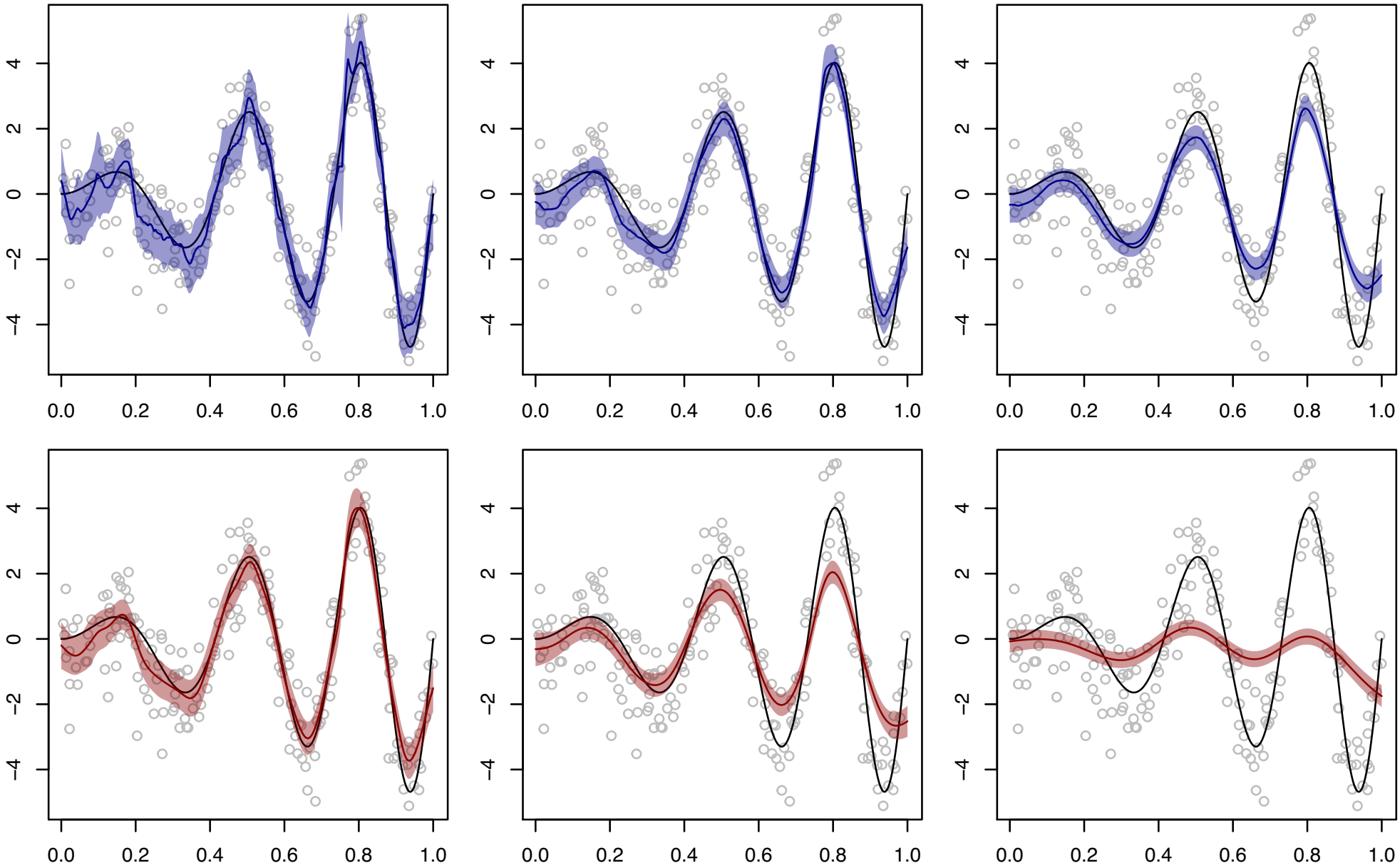

4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0

Table of Contents

1 Nonparametric regression

2 The Nadaraya-Watson estimator

3 Pointwise confidence intervals

4 Variance estimation

5 Crossvalidation for bandwidth selection

6 Local polynomial estimators

Crossvalidation to select h: Define the *crossvalidation prediction risk* as

$$\mathrm{CV}_n(h) = \frac{1}{n}\sum_{i=1}^{n}[Y_i - \hat{m}_{n,-i}(X_i)]^2,$$

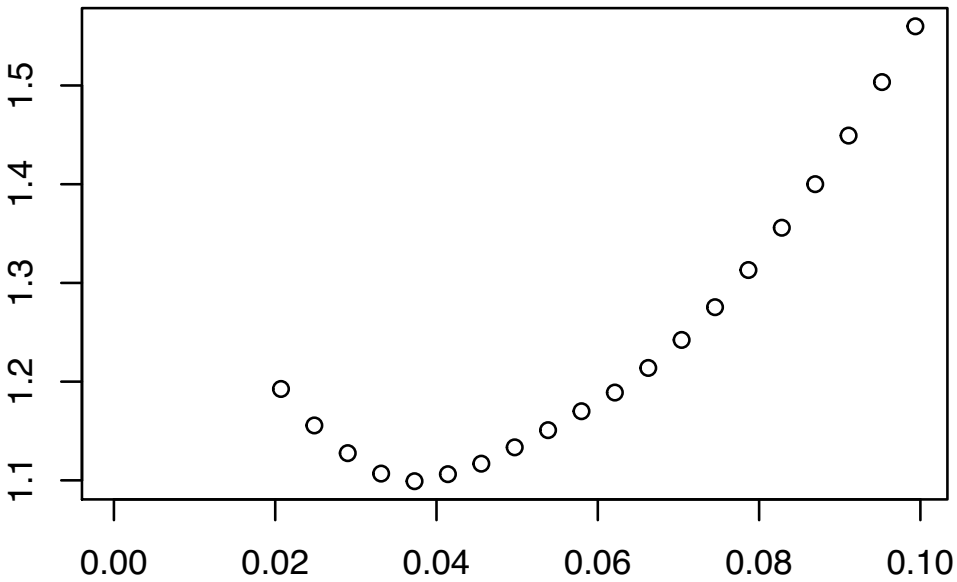
where $\hat{m}_{n,-i}$ is an estimator of m computed after removing the pair (X_i, Y_i).

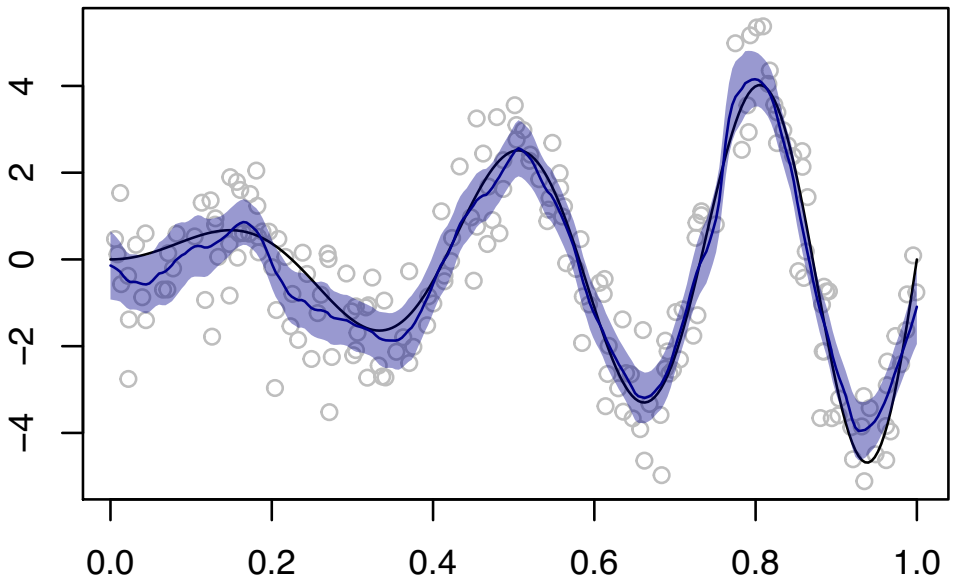
Cheaper computation of crossvalidation prediction risk

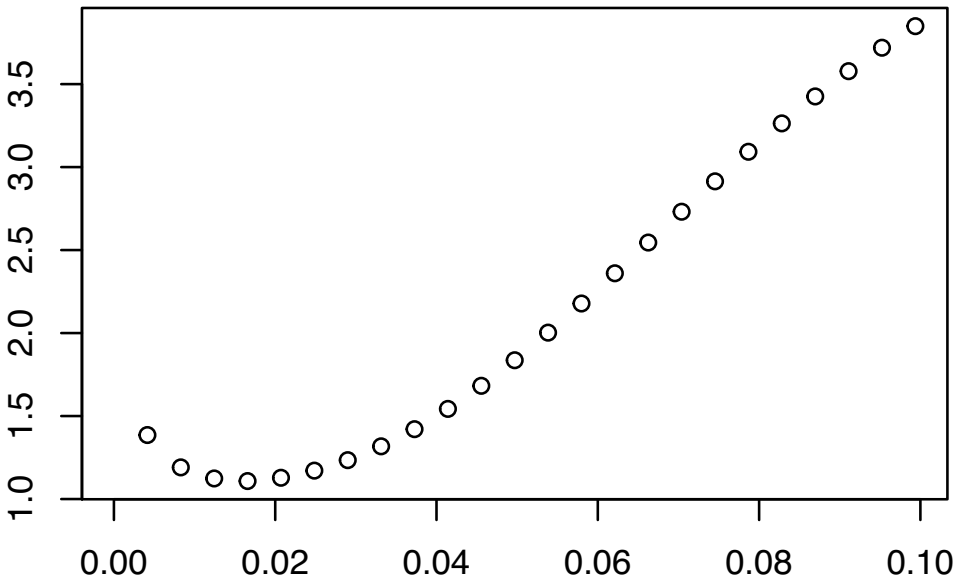
For any estimator $\hat{m}$ of m with the form $\hat{m}_n(x) = \sum_{i=1}^{n} W_{ni}(x)Y_i$, we have

$$\mathrm{CV}_n(h) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{Y_i - \hat{m}_n(X_i)}{1 - W_{ni}(X_i)}\right]^2.$$

The above will work for many estimators—not just the N-W.


1.1
1.2
1.3
1.4
1.5
0.00
0.02
0.04
0.06
0.08
0.10


−4
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0


1.0
1.5
2.0
2.5
3.0
3.5
0.00
0.02
0.04
0.06
0.08
0.10

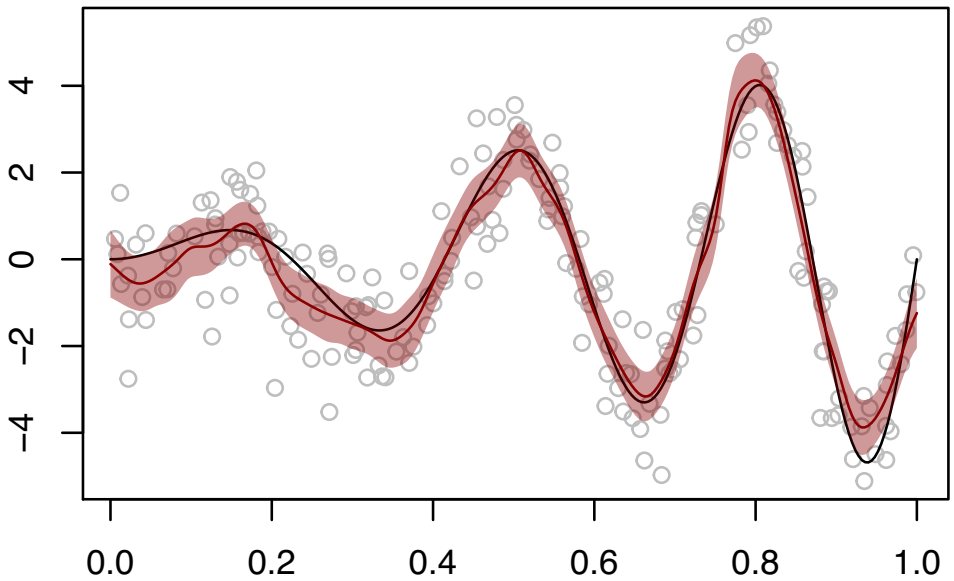

−4
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0

Table of Contents

1 Nonparametric regression

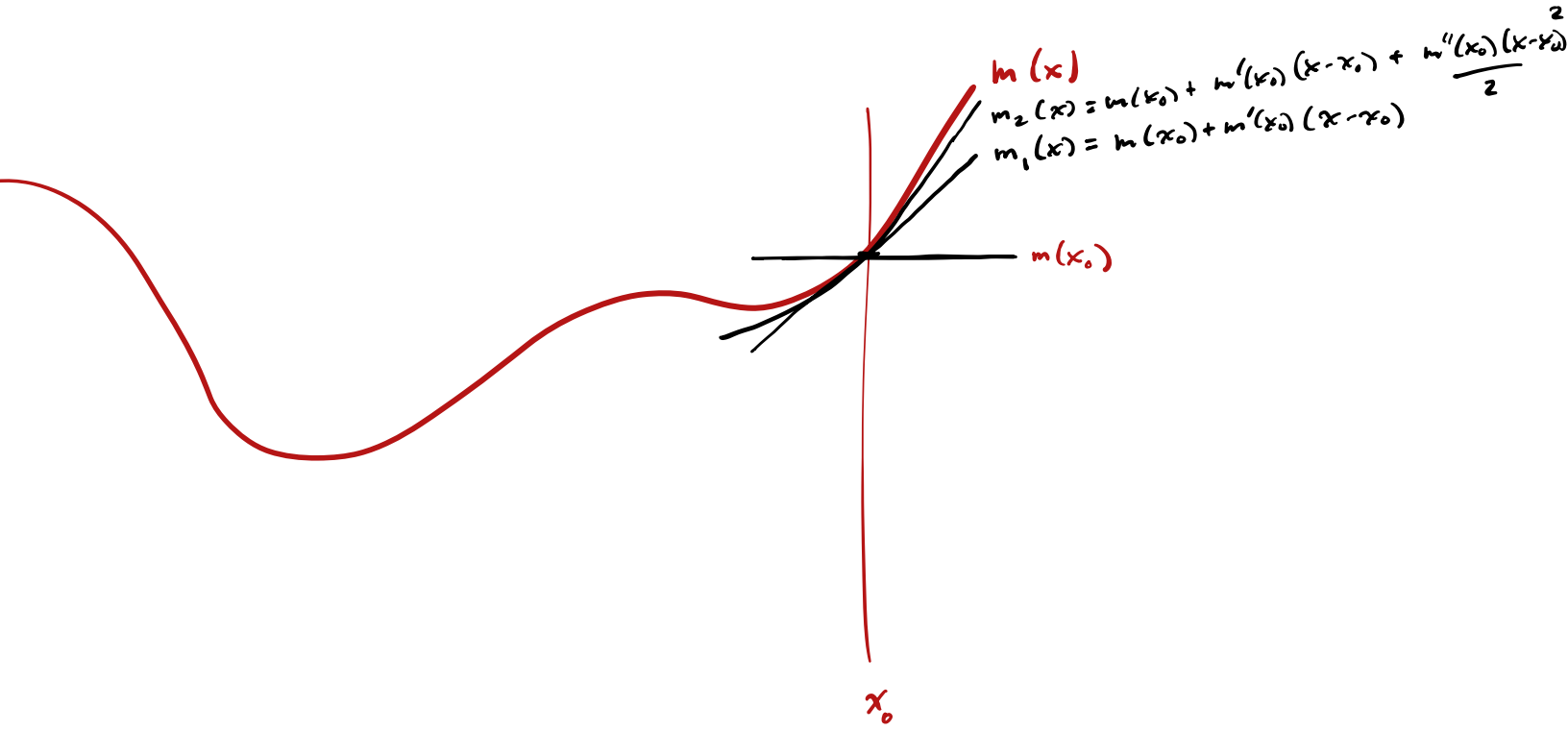
2 The Nadaraya-Watson estimator

3 Pointwise confidence intervals

4 Variance estimation

5 Crossvalidation for bandwidth selection

6 Local polynomial estimators



$$m_2(x) = \underbrace{m(x_0)}_{a_0} + \underbrace{m'(x_0)}_{a_1}(x - x_0) + \underbrace{m''(x_0)}_{a_2}\frac{(x-x_0)^2}{2}$$

Consider, for x' close to x, an order-ℓ polynomial approximation

$$m(x') \approx a_0 + a_1(x' - x) + \cdots + a_\ell \frac{(x' - x)^\ell}{\ell!}.$$

Idea: estimate $a_0, a_1, \ldots, a_\ell$ "locally" for each x.

Local polynomial (LP) estimator

The *LP estimator of order-ℓ* of m is given by $\hat{m}^{\text{LP}}_{n,\ell}(x) = \hat{a}_0(x)$, where this is the first element of the vector

$$(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_\ell)(x) = \underset{a_0, a_1, \ldots, a_\ell \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{\ell} a_j \frac{(X_i - x)^j}{j!} \right]^2 \cdot K\left(\frac{X_i - x}{h} \right).$$

for some kernel function K and bandwidth $h > 0$.

Discuss: How is the N-W a special case of this?

NW: $\ell = 0$

$$\hat{a}_0 = \underset{a_0 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \left[y_i - a_0 \right]^2 K\left(\frac{x_i - x}{h} \right)$$

$$\vdots$$

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

Write $\hat{m}^{LP}_{n,\ell}(x) = \sum_{i=1}^{n} W_{ni}(x)\, y_i$

$$(\underset{\sim}{y} - U_x \underset{\sim}{a})^T K_x (\underset{\sim}{y} - U_x \underset{\sim}{a})$$

$$(\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_\ell)(x) = \underset{a_0, a_1, \ldots, a_\ell \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \left[Y_i - \sum_{j=0}^{\ell} a_j \frac{(X_i - x)^j}{j!} \right]^2 \cdot K\left(\frac{X_i - x}{h} \right).$$

$$\underset{\sim}{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} \qquad U_x = \begin{bmatrix} 1 & X_1 - x & \cdots & \frac{(X_1 - x)^\ell}{\ell!} \\ & & & \\ 1 & X_n - x & & \frac{(X_n - x)^\ell}{\ell!} \end{bmatrix}$$

$$\underset{\sim}{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_\ell \end{bmatrix} \qquad K_x = \operatorname{diag}\left(K\left(\frac{X_1 - x}{h} \right), \ldots, K\left(\frac{X_n - x}{h} \right) \right)$$

Then we can show

$$\hat{\underset{\sim}{a}}_x = \left(U_x^T K_x U_x\right)^{-1} U_x^T K_x \underset{\sim}{Y}$$

$$\hat{m}_n^{LP}(x) = \underset{\sim}{e}_1^T \hat{\underset{\sim}{a}}_x \qquad \underset{\sim}{e}_1^T = (1, 0 \cdots 0)$$

$$= \underset{\sim}{e}_1^T \left(U_x^T K_x U_x\right)^{-1} U_x^T K_x \underset{\sim}{Y}$$

Find $W_{ni}(x)$, $i = 1, \ldots, n$ s.t. that $\hat{m}_n^{LP}(x) = \sum_{i=1}^{n} W_{ni}(x)\, Y_i$.

$$\hat{m}_n^{LP}(x) = \underset{\sim}{e}_1^T \left(U_x^T K_x U_x\right)^{-1} U_x^T K_x \underset{\sim}{Y}$$

$$= \underset{\sim}{e}_1^T \left(U_x^T K_x U_x\right)^{-1} \begin{bmatrix} 1 & \cdots & 1 \\ X_1 - x & \cdots & X_n - x \\ \vdots & \ddots & \vdots \\ \frac{(X_1 - x)^q}{q!} & \cdots & \frac{(X_n - x)^q}{q!} \end{bmatrix} \begin{bmatrix} k\left(\frac{X_1 - x}{h}\right) & & \\ & \ddots & \\ & & k\left(\frac{X_n - x}{h}\right) \end{bmatrix} \begin{bmatrix} Y_1 \\ \\ Y_n \end{bmatrix}$$

$$= \sum_{i=1}^{n} \underset{\sim}{e}_1^T \left(U_x^T K_x U_x\right)^{-1} \begin{bmatrix} 1 \\ X_i - x \\ \vdots \\ \frac{(X_i - x)^q}{q!} \end{bmatrix} k\left(\frac{X_i - x}{h}\right) Y_i$$

$$\Rightarrow \qquad W_{ni}(x) = \underset{\sim}{e}_1^T \left(U_x^T K_x U_x\right)^{-1} \begin{bmatrix} 1 \\ X_i - x \\ \vdots \\ \frac{(X_i - x)^q}{q!} \end{bmatrix} k\left(\frac{X_i - x}{h}\right)$$

OR:

$$\hat{\underset{\sim}{a}}_x = \left(U_x^T K_x U_x\right)^{-1} U_x^T K_x \underset{\sim}{Y}$$

$$= L_x^T \underset{\sim}{Y} \qquad L_x = K_x U_x \left(U_x^T K_x U_x\right)^{-1}$$

$$\hat{a}_0(x) = \underset{\sim}{e}_1^T L_x^T \underset{\sim}{Y}$$

$$= \left(L_x \underset{\sim}{e}_1\right)^T \underset{\sim}{Y},$$

So $$\underset{\sim}{W}_x = \begin{bmatrix} W_{n1}(x) \\ \vdots \\ W_{nn}(x) \end{bmatrix} = L_x \underset{\sim}{e}_1 .$$

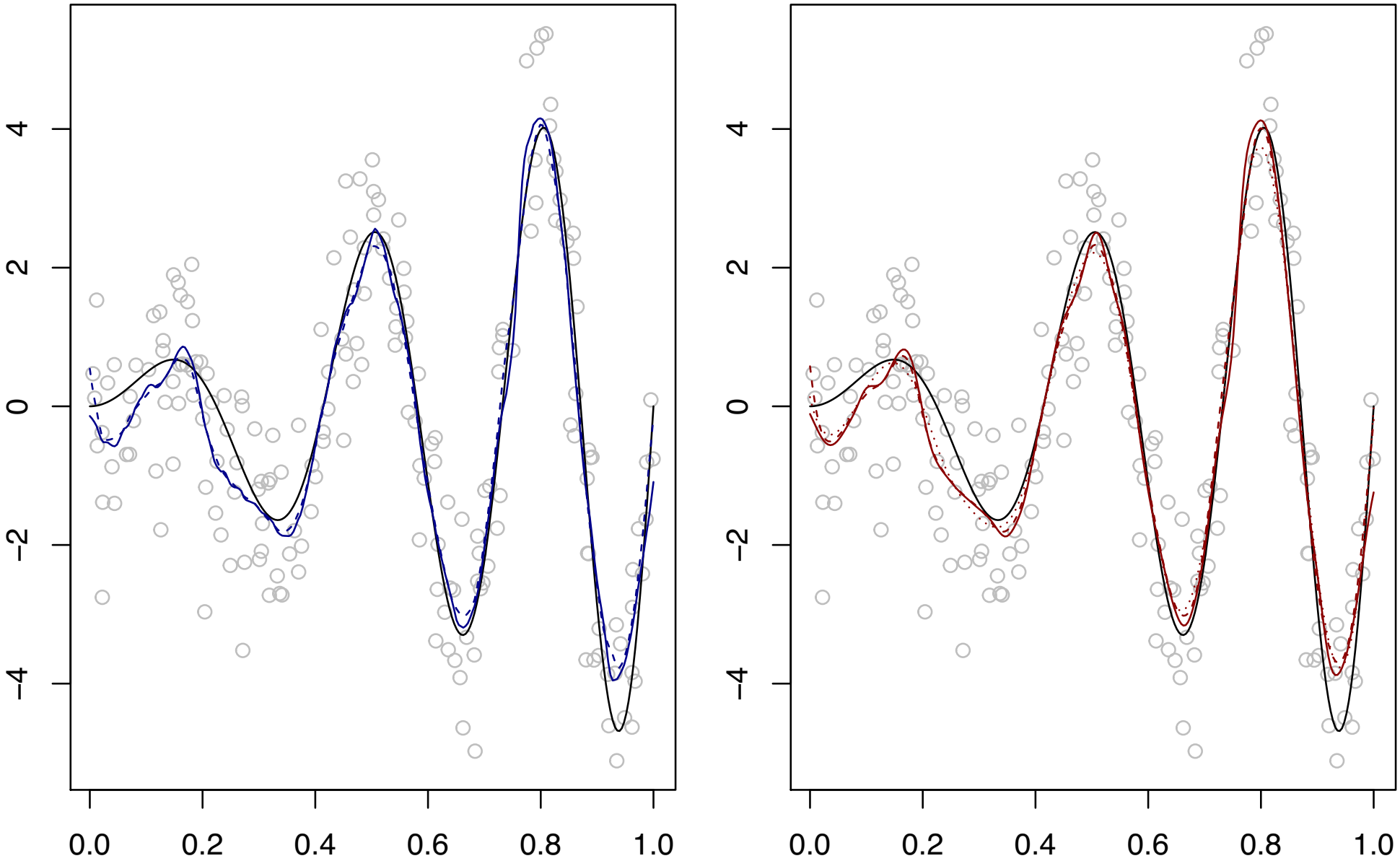
Exercise:

1. Find matrices $\mathbf{U}_x$ and $\mathbf{K}_x$ and the vector $\mathbf{Y}$ such that

$$\hat{\mathbf{a}}(x) = (\hat{a}_0, \hat{a}_1, \ldots, \hat{a}_\ell)(x)^T = \left(\mathbf{U}_x^T \mathbf{K}_x \mathbf{U}_x\right)^{-1} \mathbf{U}_x^T \mathbf{K}_x \mathbf{Y}.$$

2. Find an expression for weights $W_{ni}^*(x)$ such that

$$\hat{m}_{n,\ell}^{\mathrm{LP}}(x) = \sum_{i=1}^{n} W_{ni}^*(x) Y_i.$$


4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0
4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0

Tsybakov [2] gives the following result (we omit the details of the assumptions)

Theorem 2: Uniform MSE bound for LP over Hölder functions

Under conditions in Tsybakov, if $m \in \mathcal{H}(\beta, L)$, the order-$\ell$ LP estimator satisfies

$$\sup_{m \in \mathcal{H}(\beta,L)} \sup_{x \in [0,1]} \underbrace{\mathbb{E}[\hat{m}^{\text{LP}}_{n,\ell}(x) - m(x)]^2}_{\text{MSE } \hat{m}^{LP}_{n,\ell}(x)} \leq n^{-\frac{2\beta}{2\beta+1}} \cdot C$$

for large enough n, where $\ell = \beta - 1$.

$\Rightarrow$ $\beta = 2$, use local linear, LP($\ell = 1$), get $n^{-4/5}$

Discuss:

1. Compare this to our N-W results under Lipschitz smoothness.
2. What does a larger β mean for the rate of convergence?
3. What is the interpretation of larger β?

Much nonparametric regression lit. is written under the settings of this theorem:

Theorem 3: MSE of N-W, LP-1 under bounded 2nd derivative

Let $X_1, \ldots, X_n$ have continuous and differentiable density f, and x st $f(x) > 0$; set $\kappa^2 = \int_{\mathbb{R}} K^2(u)du$, $\mu_2(K) = \int_{\mathbb{R}} u^2 K(u)du$; suppose m'' is continuous and bounded. Then for $h \to 0$ and $nh \to \infty$ as $n \to \infty$, we have

$$\text{Bias}\, \hat{m}_n^{\text{NW}}(x) = \frac{h^2}{2}\left[m''(x) + 2\frac{m'(x)f'(x)}{f(x)}\right]\mu_2(K) + o_p(h^2)$$

$$\text{Bias}\, \hat{m}_{n,1}^{\text{LP}}(x) = \frac{h^2}{2} m''(x)\mu_2(K) + o_p(h^2),$$

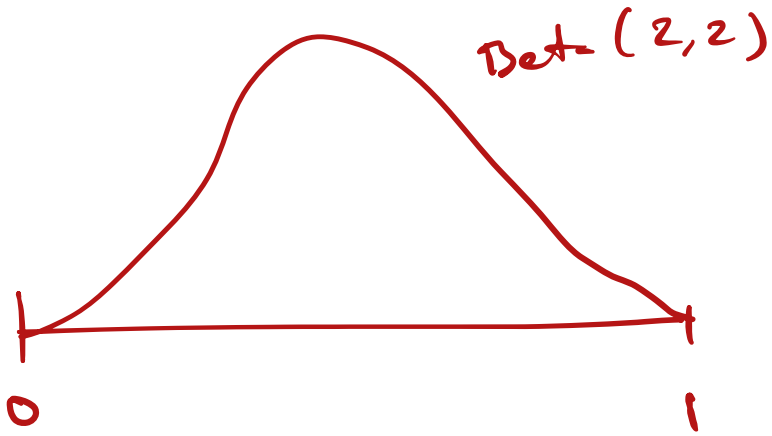
local-linear ($q = 1$)

while $\text{Var}\, \hat{m}_n^{\text{NW}}(x)$ and $\text{Var}\, \hat{m}_{n,1}^{\text{LP}}(x)$ are both given by $\frac{1}{nh}\frac{\sigma^2}{f(x)}\kappa^2 + o_p((nh)^{-1})$.

See Wasserman book [3].

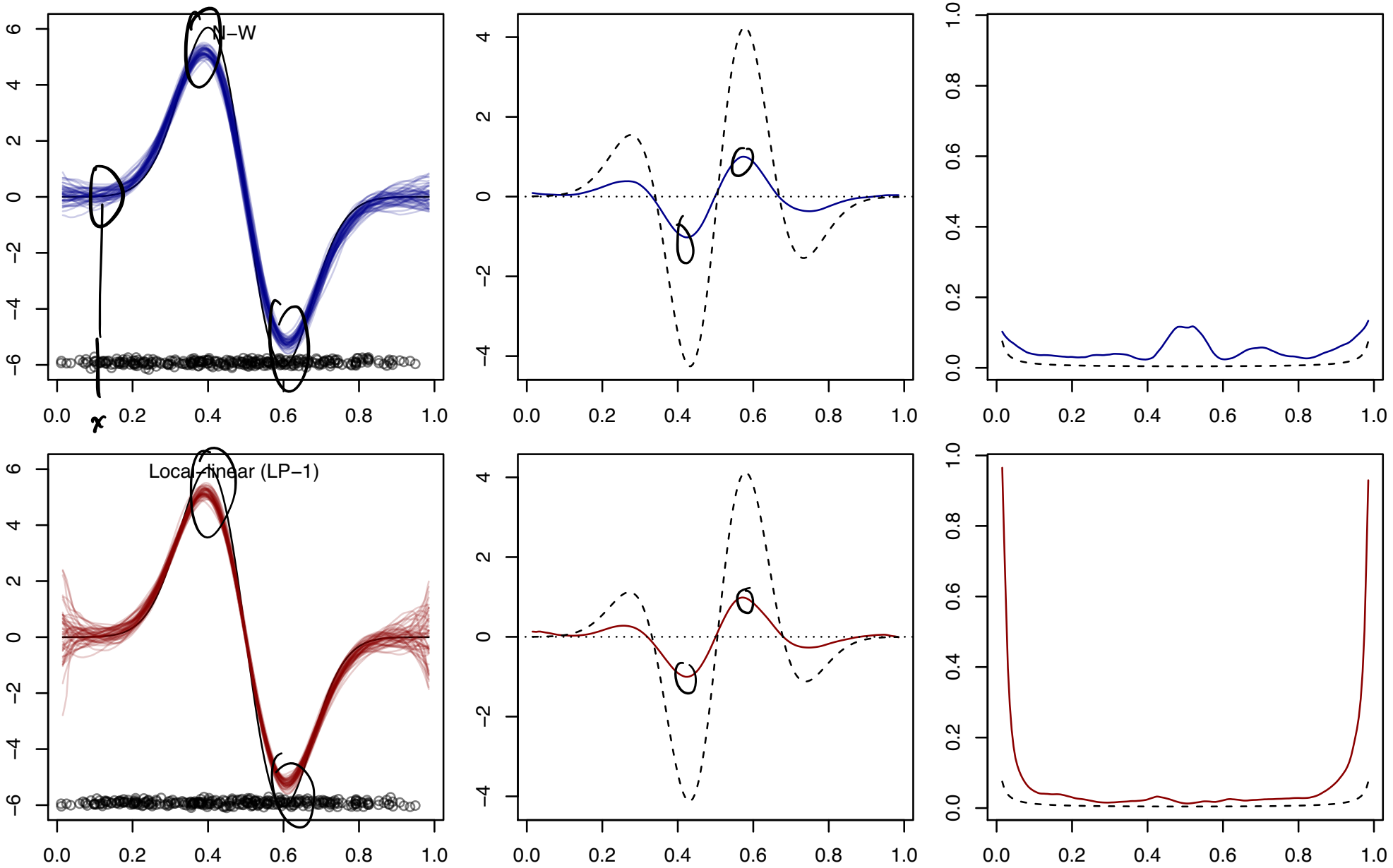
Discuss: Insights from above. Curvature, design bias, optimal rate, etc.

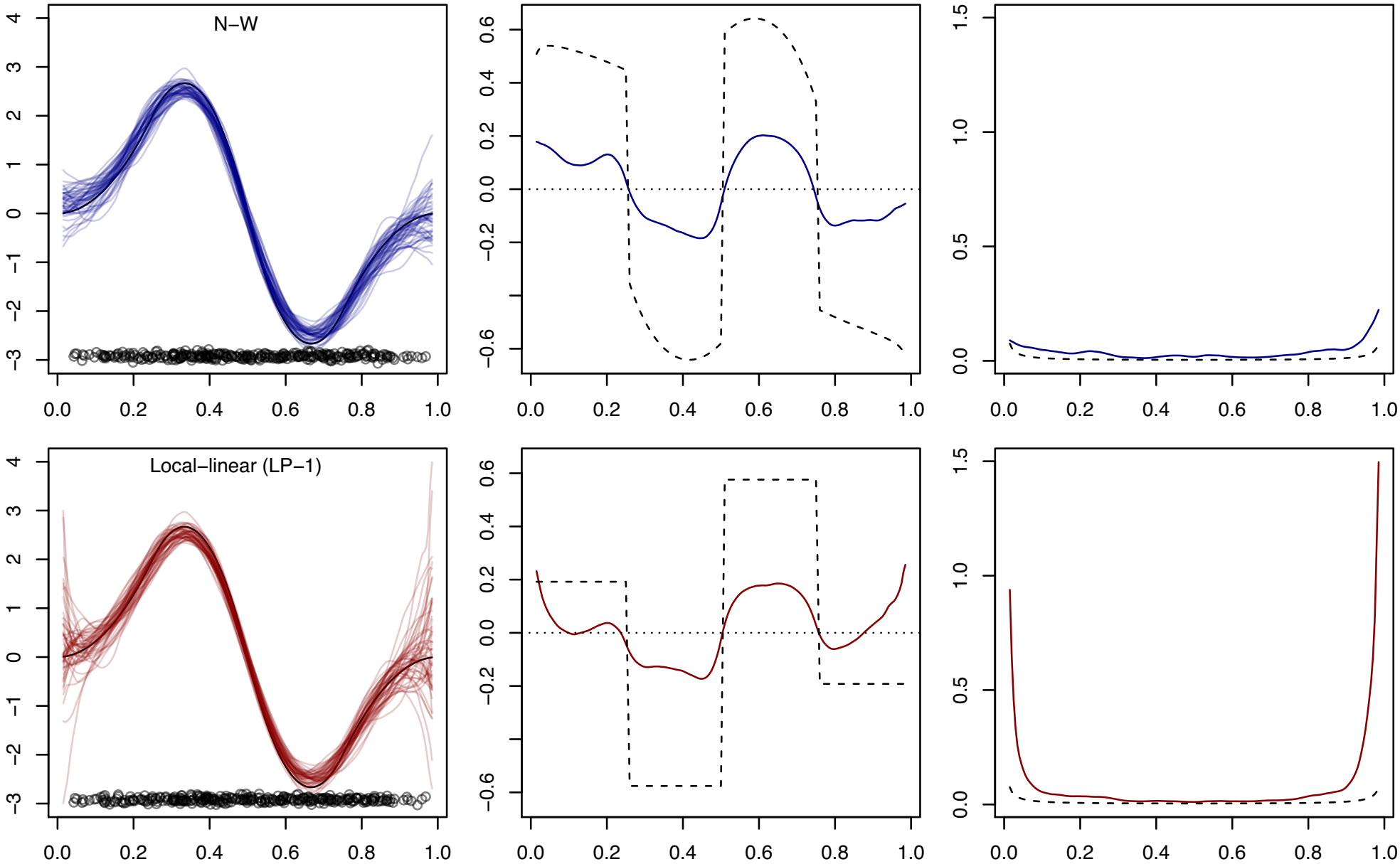
Illustration: In the next four slides, we have

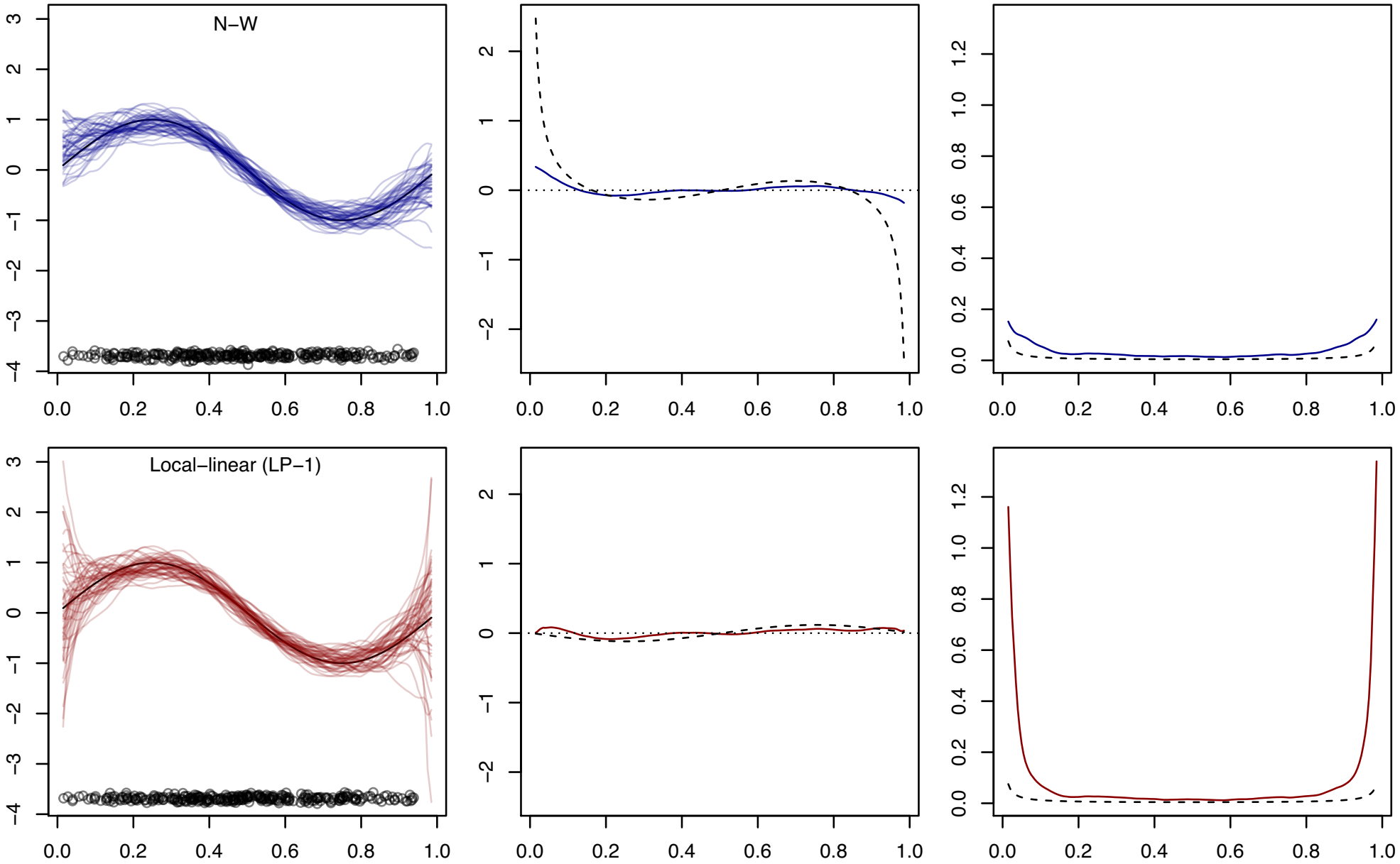


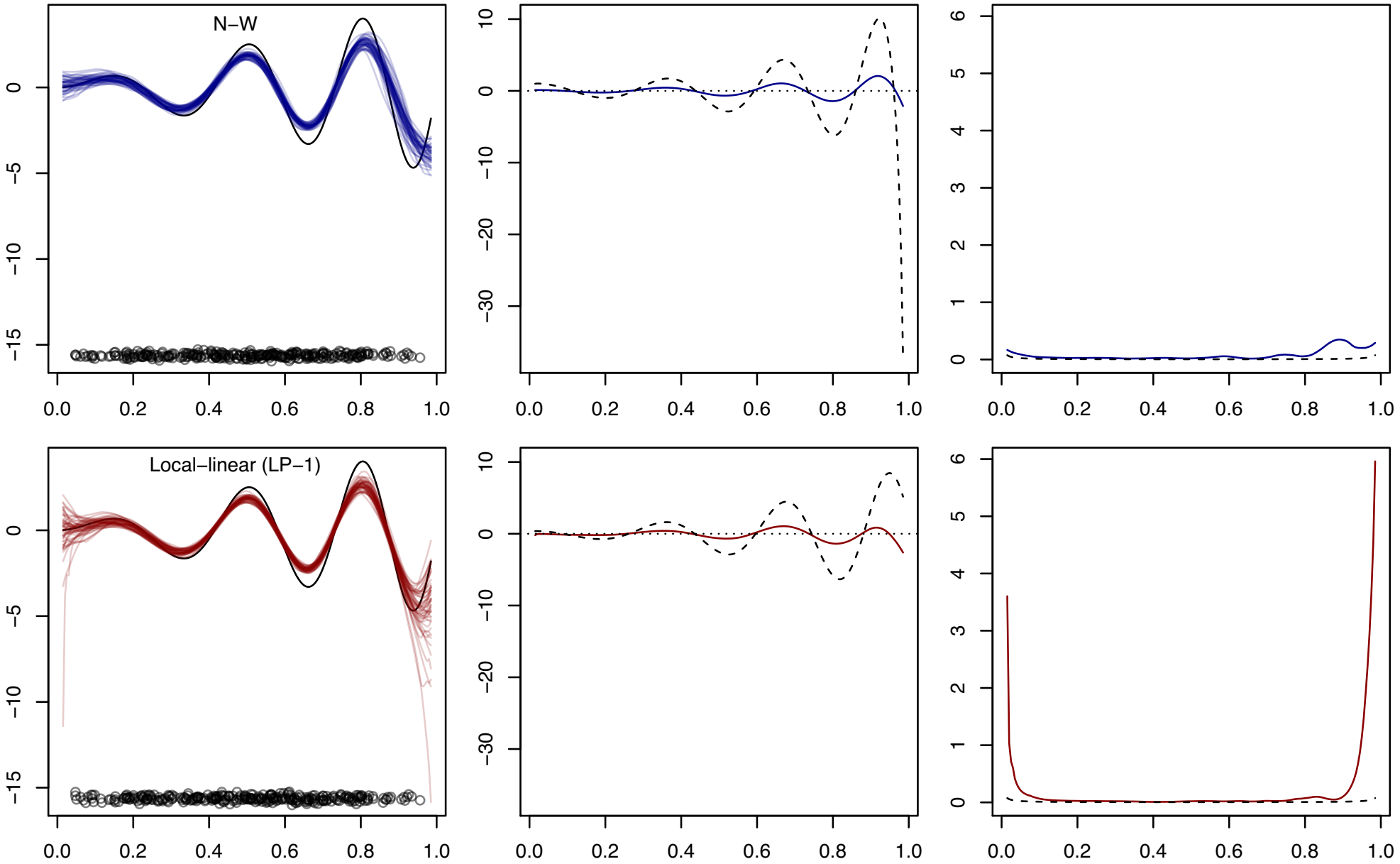
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Beta}(2, 2)$.
- $n = 300$, $h = 0.1$.
- N-W and local linear (LP-1) estimators fit on 50 simulated data sets.
- Four different choices of the function m.

Goal: Compare bias and variance of against the result in the previous slide


N–W
Local–linear (LP–1)
x


N–W
Local–linear (LP–1)


N–W
Local–linear (LP–1)


N–W
Local–linear (LP–1)

John Rice et al.
Bandwidth choice for nonparametric regression.
The Annals of Statistics, 12(4):1215–1230, 1984.

Alexandre B Tsybakov.
Introduction to nonparametric estimation.
Springer Science & Business Media, 2008.

Larry Wasserman.
All of nonparametric statistics.
Springer Science & Business Media, 2006.