STAT 824 sp 2023 Lec 04 slides

Nonparametric regression: N-W and LP smoothers

Karl B. 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.

Table of Contents

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

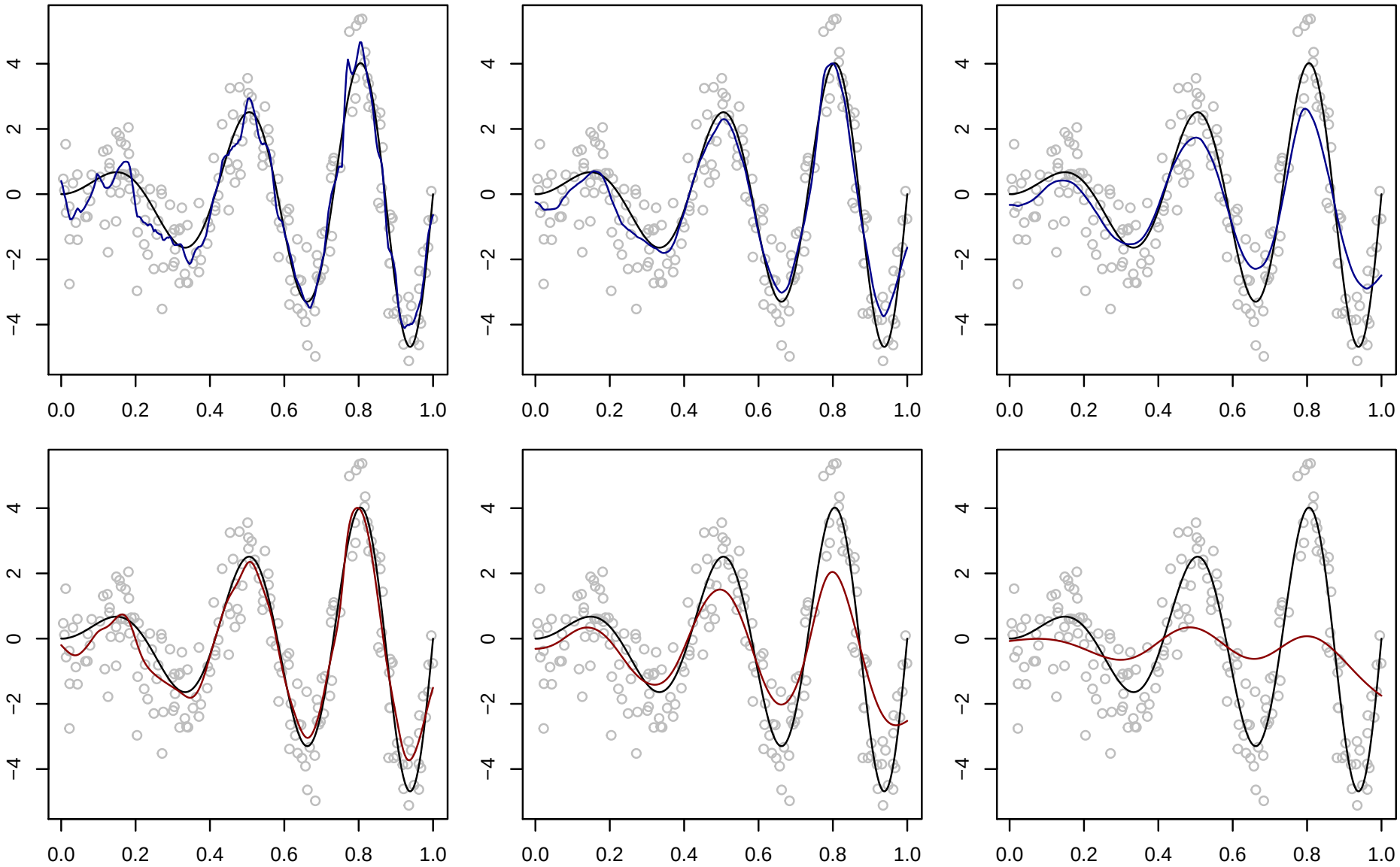
The *Nadaraya-Watson (N-W)* estimator of m is given by

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

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

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

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


4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0

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

(K1) Let K have support on $[-1, 1]$ and let $0 \leq K(u) \leq K_{\max} < \infty \ \forall \ u \in \mathbb{R}$.
(D1) Let $X_1, \ldots, X_n$ be deterministic such that for some $n_0 > 0$

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

for all $n \geq n_0$.

Pointwise MSE bound for N-W under Lipschitz smoothness

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

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

for all $n \geq n_0$.

Exercise:

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

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

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

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}.

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

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

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

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

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

This CLT is highly useful in regression contexts.

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

Let K have support on $[-1,1]$ and $0 \leq K(u) \leq K_{\max} < \infty \; \forall \; u \in \mathbb{R}$ 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. $\dfrac{\hat{m}_n^{\text{NW}}(x_0) - \mathbb{E}\hat{m}_n^{\text{NW}}(x_0)}{\sigma \left[\sum_{i=1}^{n} W_{ni}^2(x_0) \right]^{1/2}} \to \text{Normal}(0,1)$ as $n \to \infty$.

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

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

Table of Contents

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.

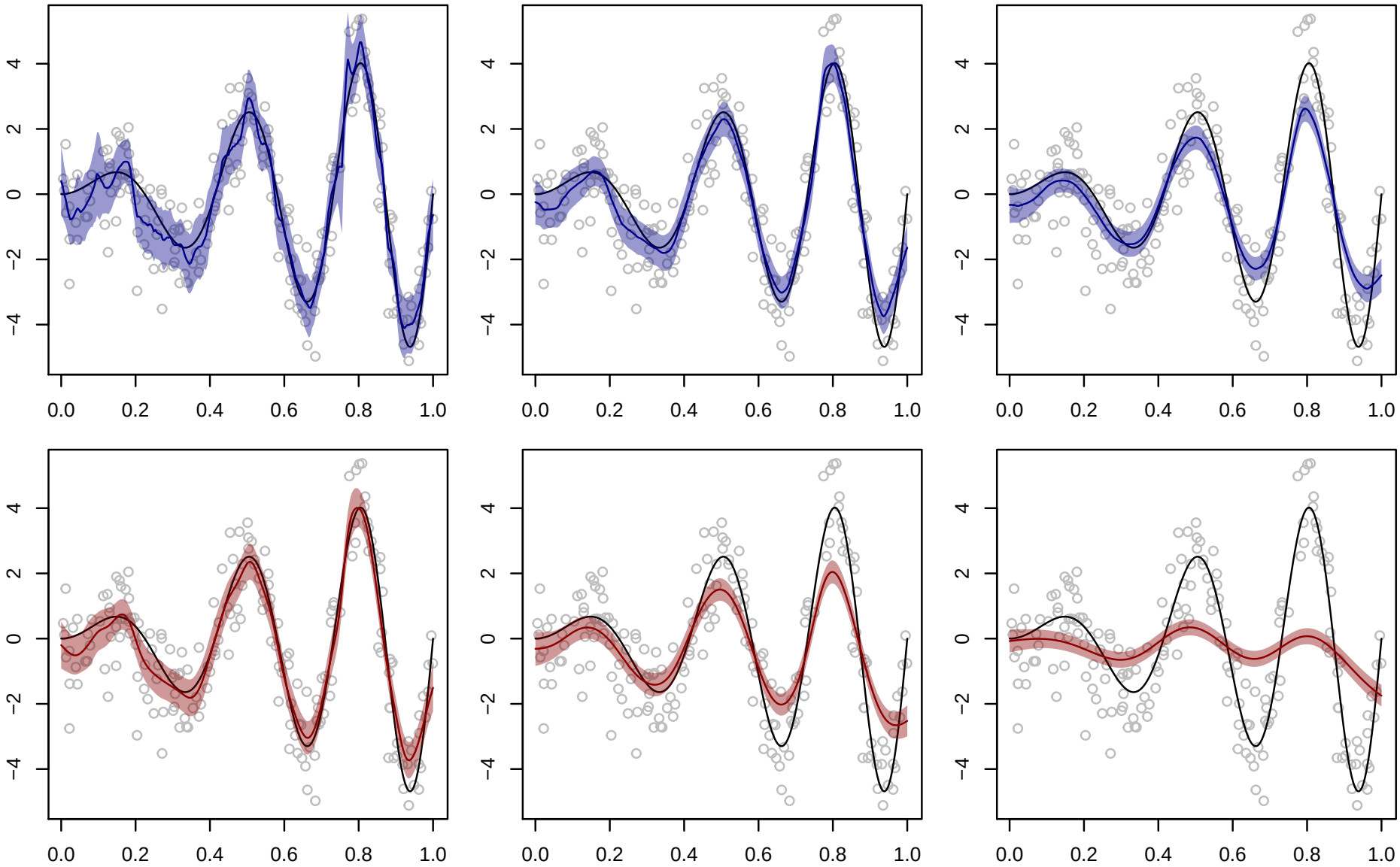

4
2
0
−2
−4
0.0
0.2
0.4
0.6
0.8
1.0

Table of Contents

How can we select the bandwidth h?

We again turn to crossvalidation: Define the *crossvalidation prediction risk* as

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

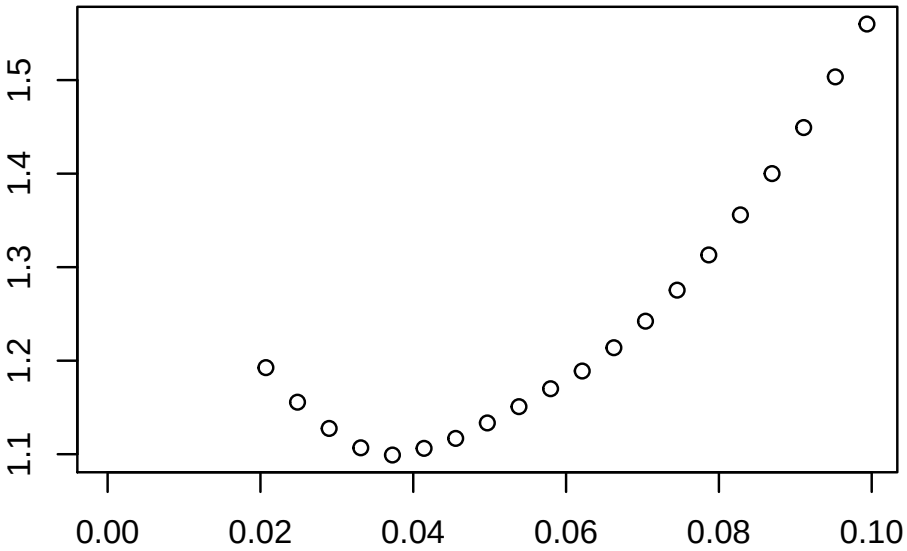
where $\hat{m}_{n,-i}$ is an estimator of m computed after removing the pair (X_i, Y_i).

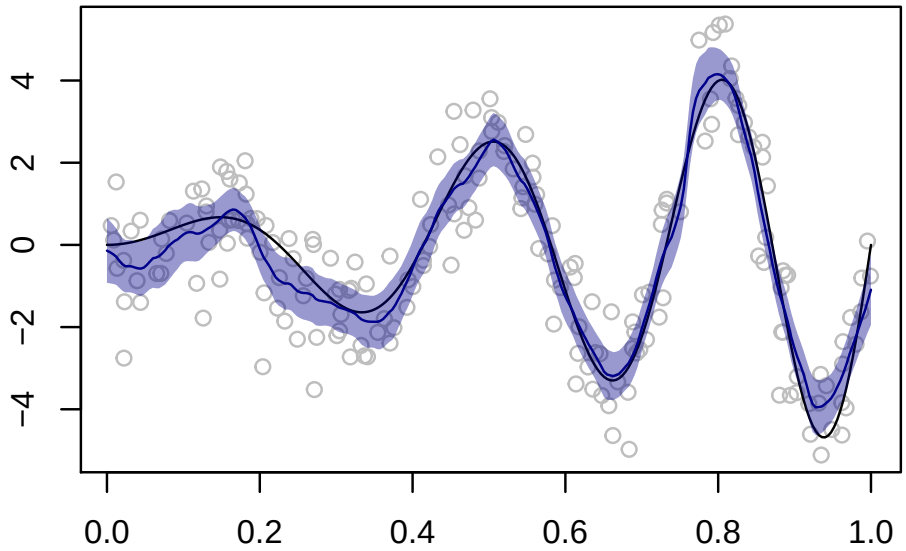
Cheaper computation of crossvalidation prediction risk

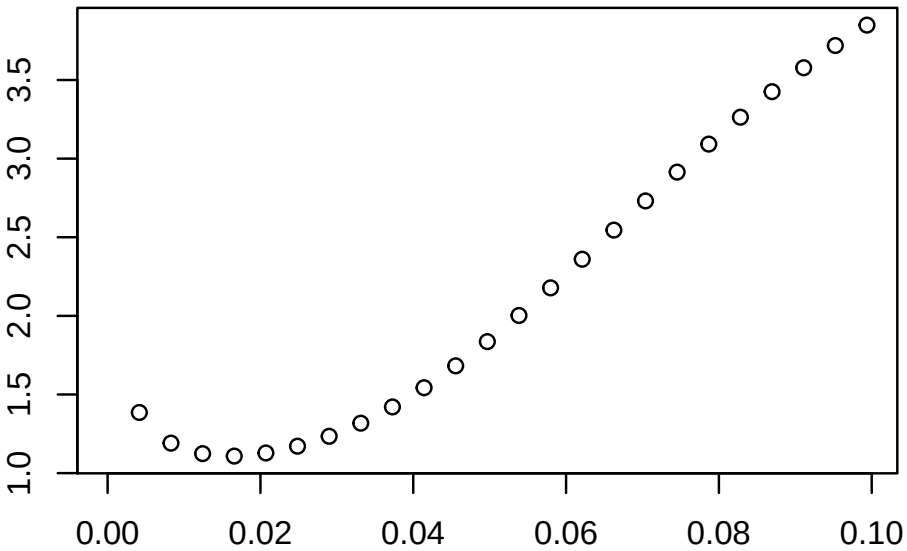
For any estimator $\hat{m}$ of m with the form $\hat{m}_n(x) = \sum_{i=1}^{n} W_{ni}(x)Y_i$, we have

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

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

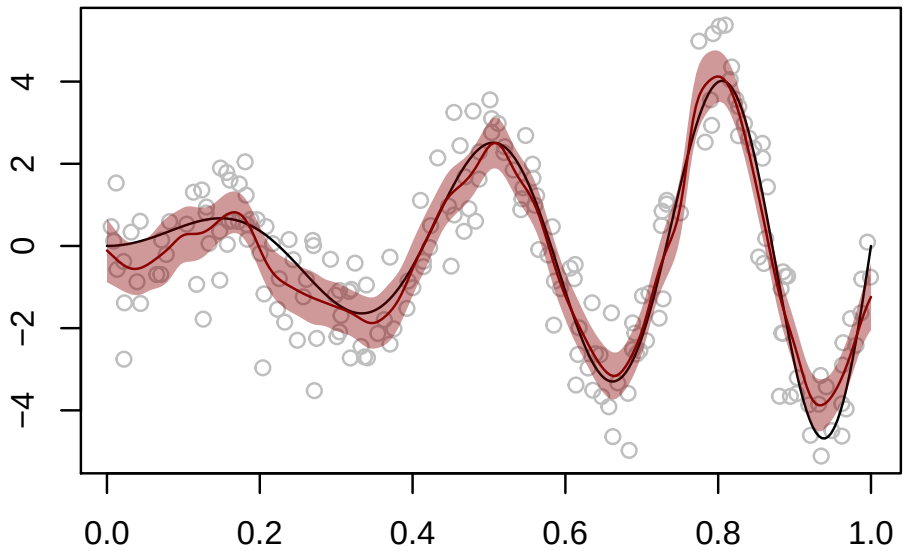

−4
−2
0
2
4
0.0
0.2
0.4
0.6
0.8
1.0

Table of Contents

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

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

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

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

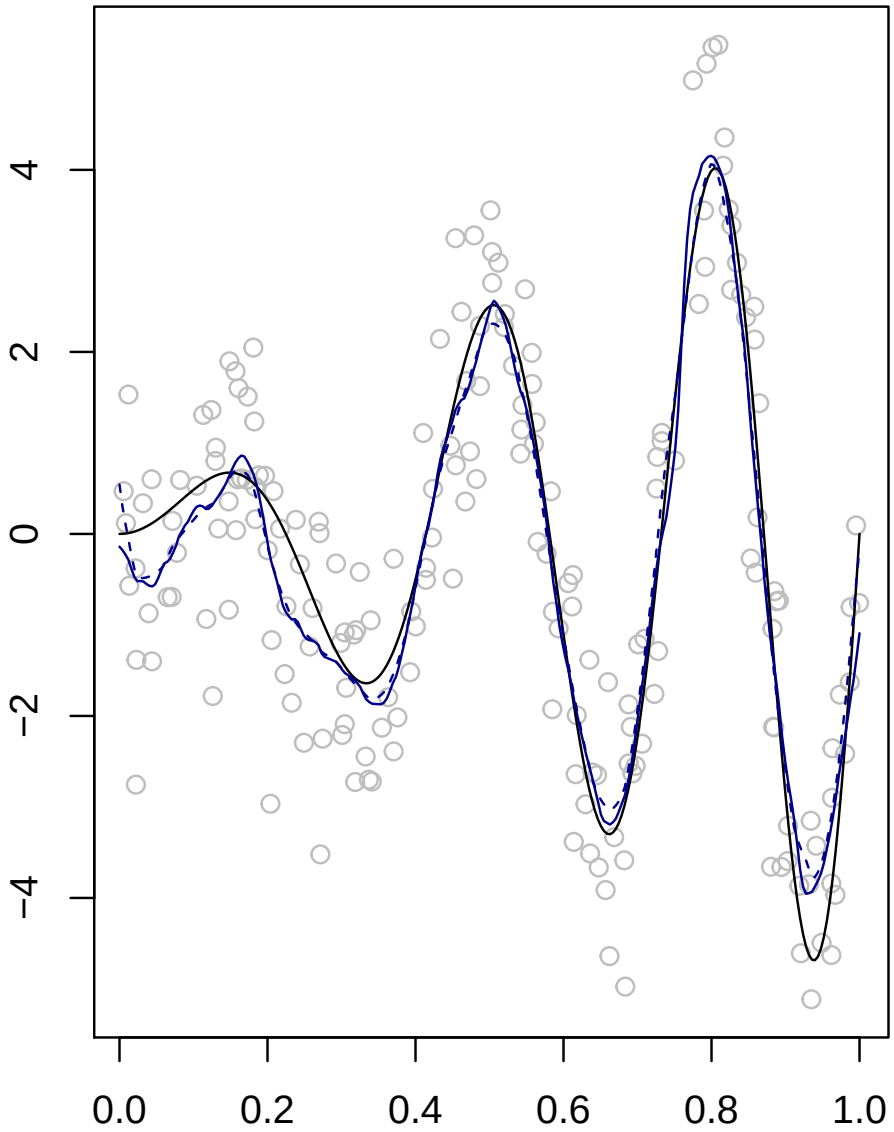
Exercise:

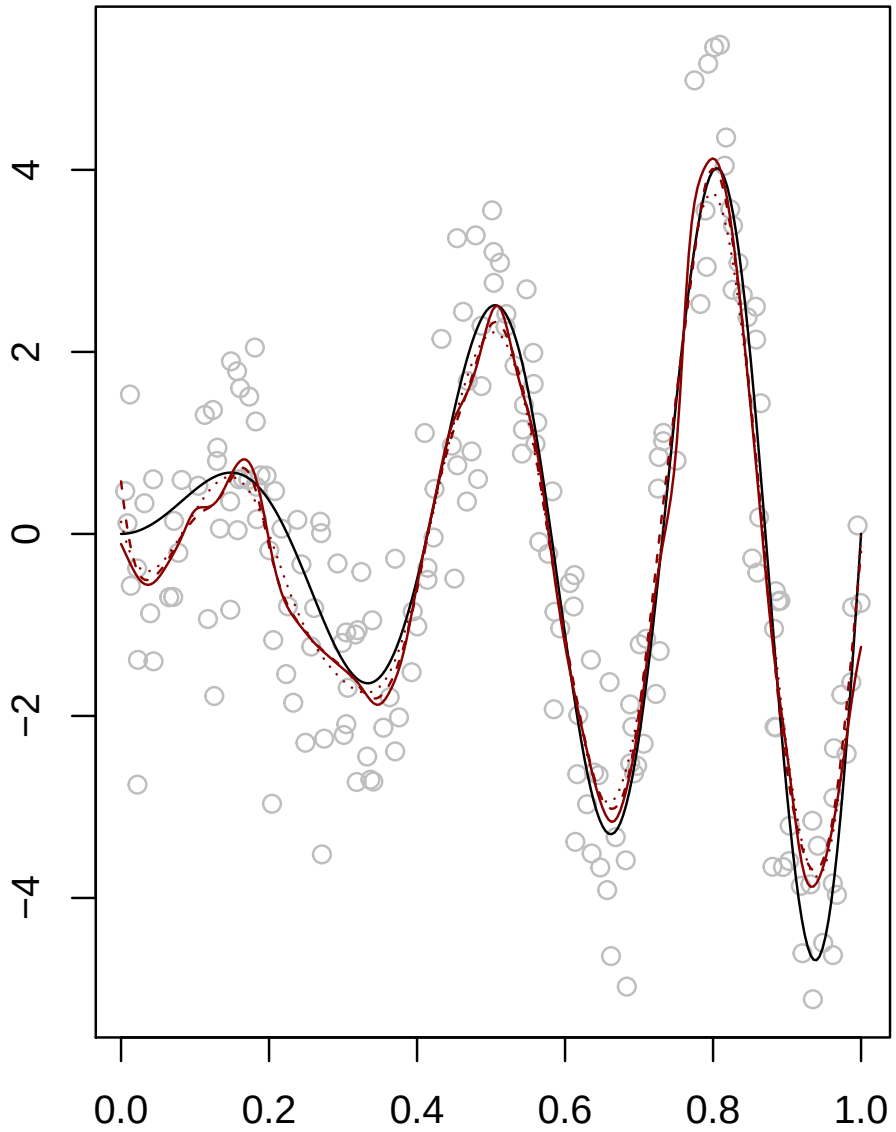
1. Find matrices $\mathbf{U}_x$ and $\mathbf{K}_x$ and the vector $\mathbf{Y}$ such that

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

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

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


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

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

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

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

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

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

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

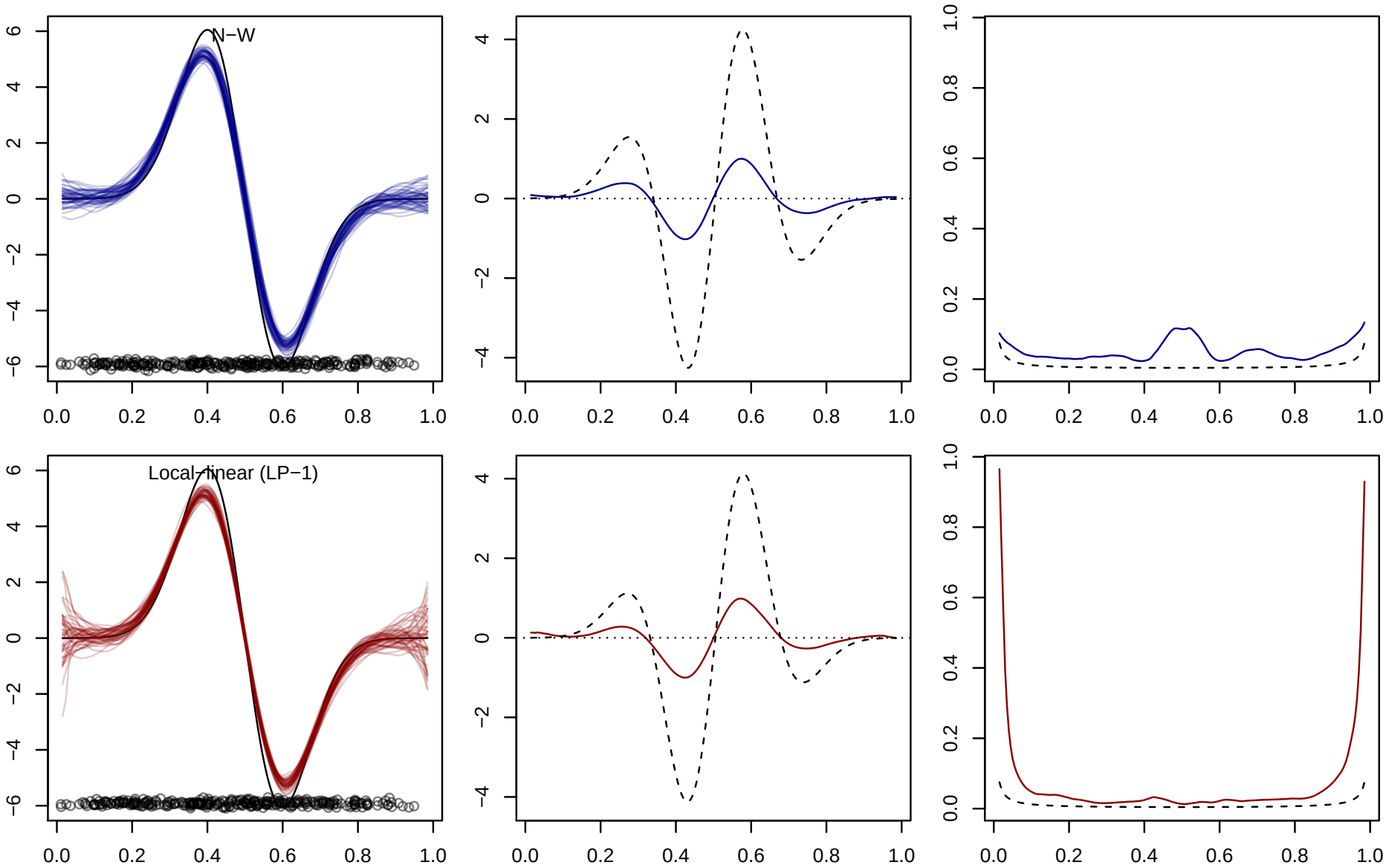
See Wasserman book [3].

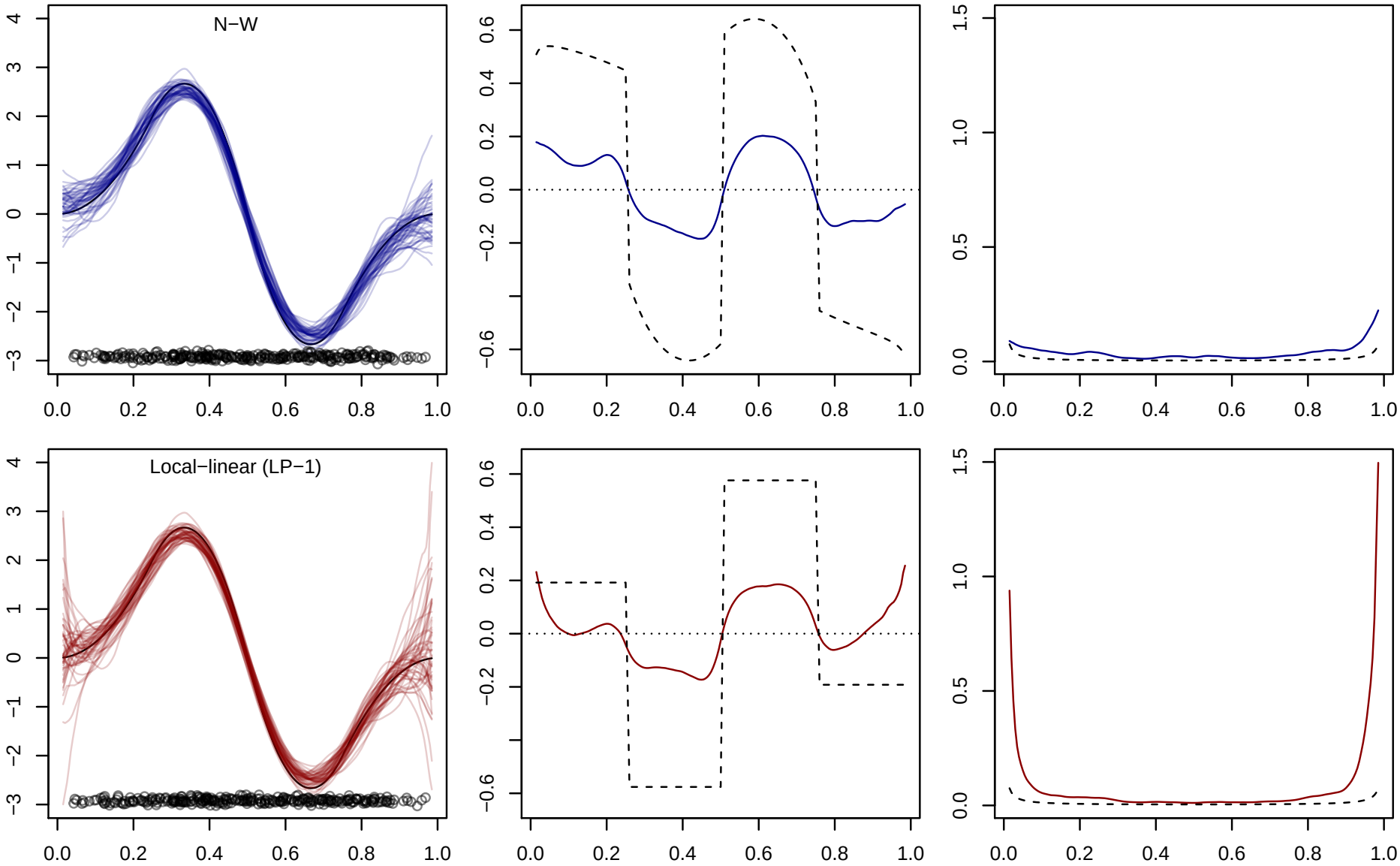
Discuss: Insights from above. Curvature, design bias, optimal rate, etc.

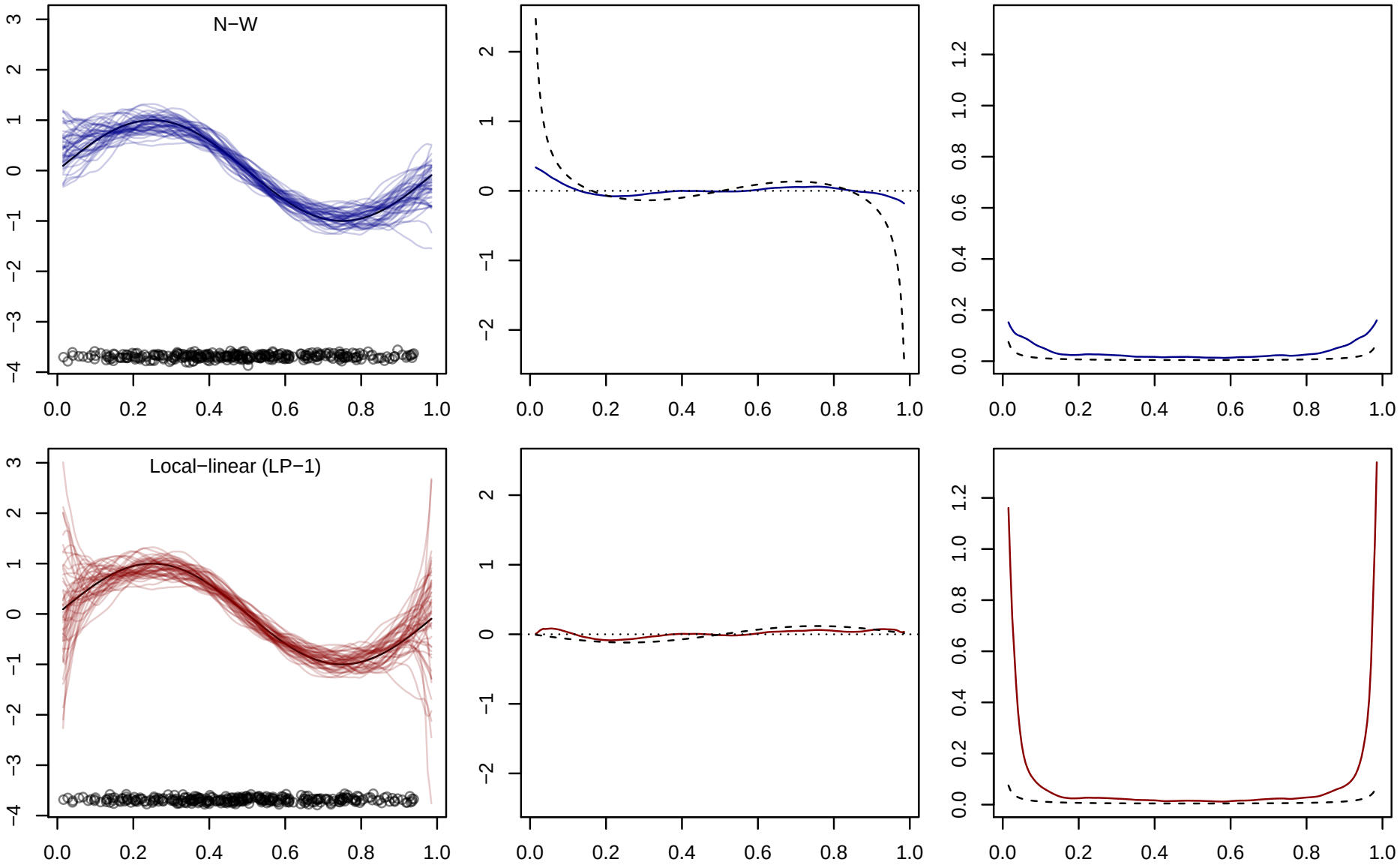
Illustration: In the next four slides, we have

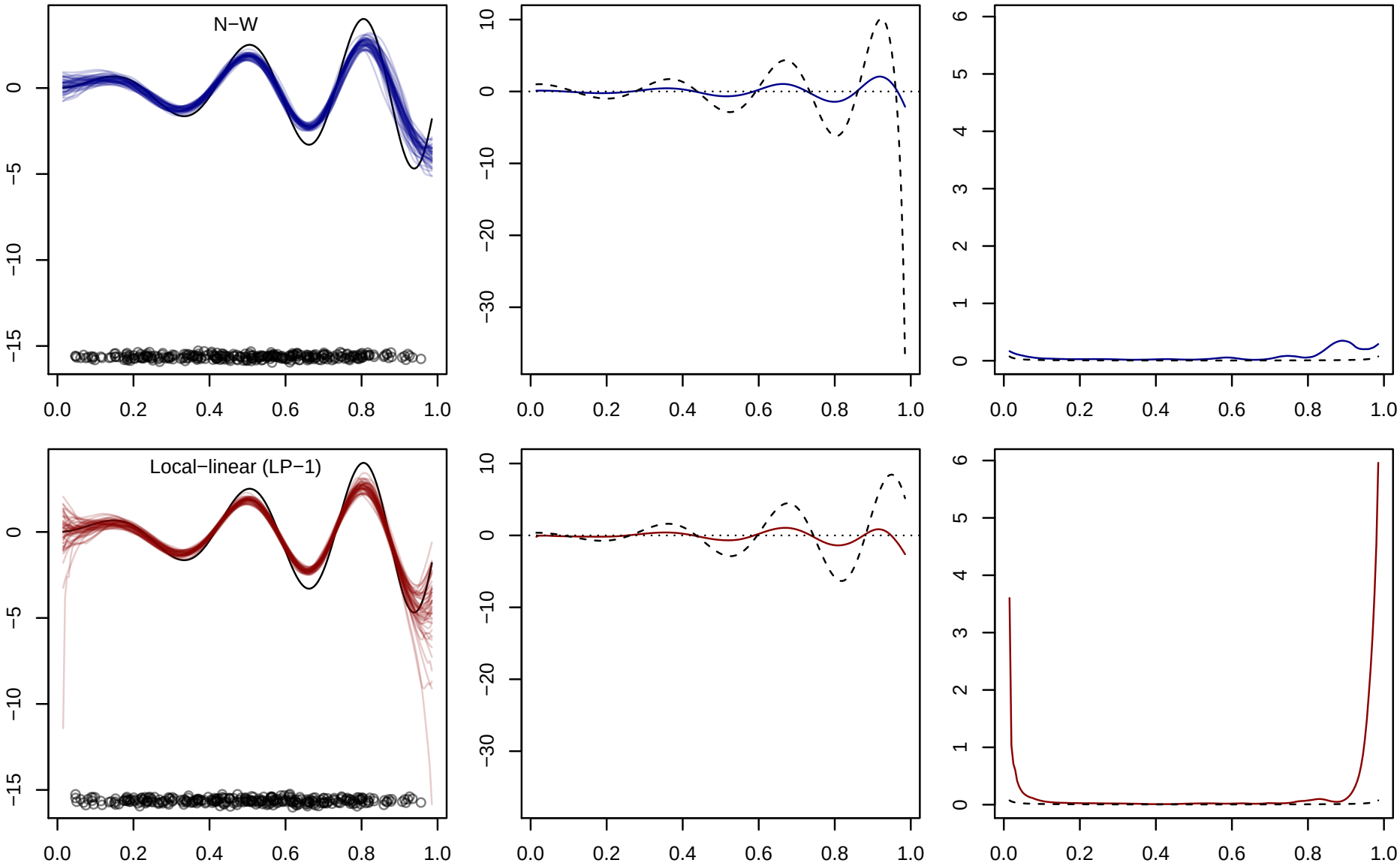
- $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Beta}(2, 2)$.
- $n = 300$, $h = 0.1$.
- N-W and local linear (LP-1) estimators fit on 50 simulated data sets.
- Four different choices of the function m.

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


N−W
Local−linear (LP−1)


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.