LOCALLY LINEAR REGRESSION

Consider a regression setting with training data (x_i, y_i), $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

and assume the relationship between x and y is <u>nonlinear</u>

Furthermore, suppose we are not willing to adopt a polynomial model, which we previously

(A) saw could be fit using ________ ________.

<u>Ideas</u>

- Local averaging

$$\hat{f}(x) = \text{avg. } y_i \text{ among } x_i$$
$$\text{s.t. } \|x - x_i\| < \delta$$

(B)

$$=$$

- Weighted local averaging

$$\hat{f}(x) = \text{weighted avg. of } y_i$$
$$\text{based on } \|x - x_i\|.$$

Known as the Nadayara-Watson estimate

$$=$$

Note: local averaging is the special case where

Smoothing Kernels

$k_\sigma(y)$ is called a smoothing kernel

Examples:

$$K_\sigma(y) = \frac{1}{\sigma^d} D\left(\frac{\|y\|}{\sigma}\right)$$

where

$$D(t) = \begin{cases} \frac{3}{4}(1-t^2) & \text{if } |t| \le 1 \\ 0 & \text{else} \end{cases} \qquad \boxed{\text{Epanechnikov}}$$

or

$$D(t) = \begin{cases} (1-t^3)^3 & \text{if } |t| \le 1 \\ 0 & \text{else} \end{cases} \qquad \boxed{\text{Tri-cube}}$$

or

$$D(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \qquad \boxed{\text{Gaussian}}$$

168 6. Kernel Methods

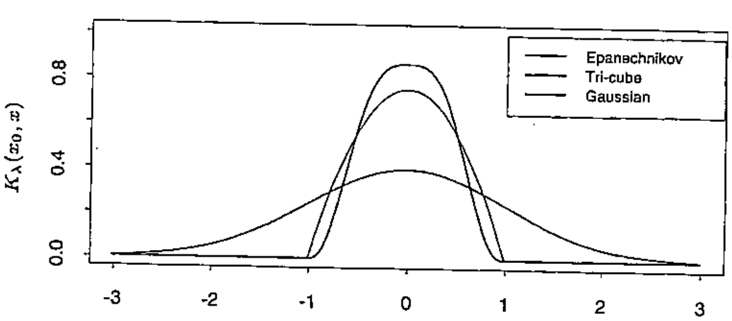


FIGURE 6.2. *A comparison of three popular kernels for local smoothing. Each has been calibrated to integrate to 1. The tri-cube kernel is compact and has two continuous derivatives at the boundary of its support, while the Epanechnikov kernel has none. The Gaussian kernel is continuously differentiable, but has infinite support.*

Unfortunately, local averaging suffers near the <u>boundaries</u> of the data set, and in other sparsely populated regions.

These problems can be alleviatied by ...

Locally Linear Regression

Here's the algorithm:

Input: $(x_i, y_i)_{i=1}^n$, $x_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$

$x \in \mathbb{R}^d$

Solve: $$\min_{\beta(x), \beta_0(x)} \sum_{i=1}^n K_\sigma(x_i - x)\left[y_i - \beta(x)^T x_i - \beta_0(x)\right]^2$$

$$\downarrow$$

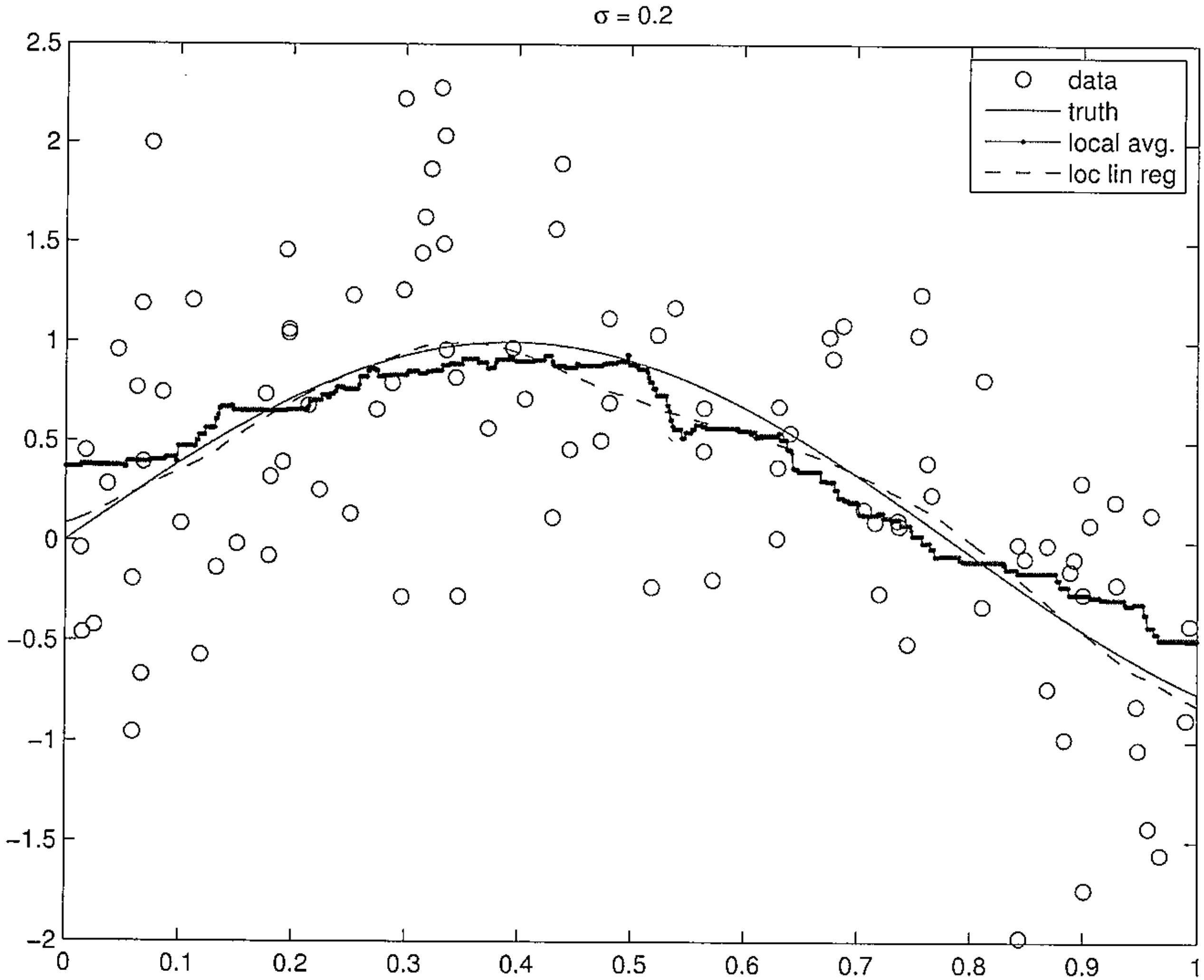
$$(\hat{\beta}(x), \hat{\beta}_0(x))$$

Output: $\hat{f}(x) = \hat{\beta}(x)^T x + \hat{\beta}_0(x)$

Note: A <u>weighted</u> <u>least</u> <u>squares</u> problem must be solved at each new x!

Relative to local averaging, LLR is

- smooth
- more accurate at boundaries



Weighted Least Squares

$$\sum_{i=1}^{n} w_i \left(y_i - \beta^T x_i - \beta_0 \right)^2$$

$$= (\underline{y} - A\theta)^T \cdot W \cdot (\underline{y} - A\theta)$$

where

$$\underline{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}, \quad \theta = \begin{bmatrix} \beta_0 \\ \beta \end{bmatrix}$$

$$A = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1d} \\ 1 & x_{21} & \cdots & x_{2d} \\ & \vdots & & \\ 1 & x_{n1} & \cdots & x_{nd} \end{bmatrix}, \quad W = \begin{bmatrix} w_1 & & & 0 \\ & w_2 & & \\ & & \ddots & \\ 0 & & & w_n \end{bmatrix}$$

How can we minimize

$$(\underline{y} - A\theta)^T W (\underline{y} - A\theta)$$

w.r.t. θ?

$$(\underline{y} - A\theta)^T W (\underline{y} - A\theta) = (\underline{y} - A\theta)^T W^{\frac{1}{2}} W^{\frac{1}{2}} (\underline{y} - A\theta)$$

$$= (W^{\frac{1}{2}}\underline{y} - W^{\frac{1}{2}}A\theta)^T (W^{\frac{1}{2}}\underline{y} - W^{\frac{1}{2}}A\theta)$$

$$= (\tilde{\underline{y}} - \tilde{A}\theta)^T (\tilde{\underline{y}} - \tilde{A}\theta), \qquad \tilde{y} = W^{\frac{1}{2}} y, \ \hat{A} = W^{\frac{1}{2}} A$$

$$= \| \tilde{\underline{y}} - \hat{A}\theta \|^2$$

(c) $\Rightarrow \hat{\theta} =$

$=$

Applying this to LLR, we have

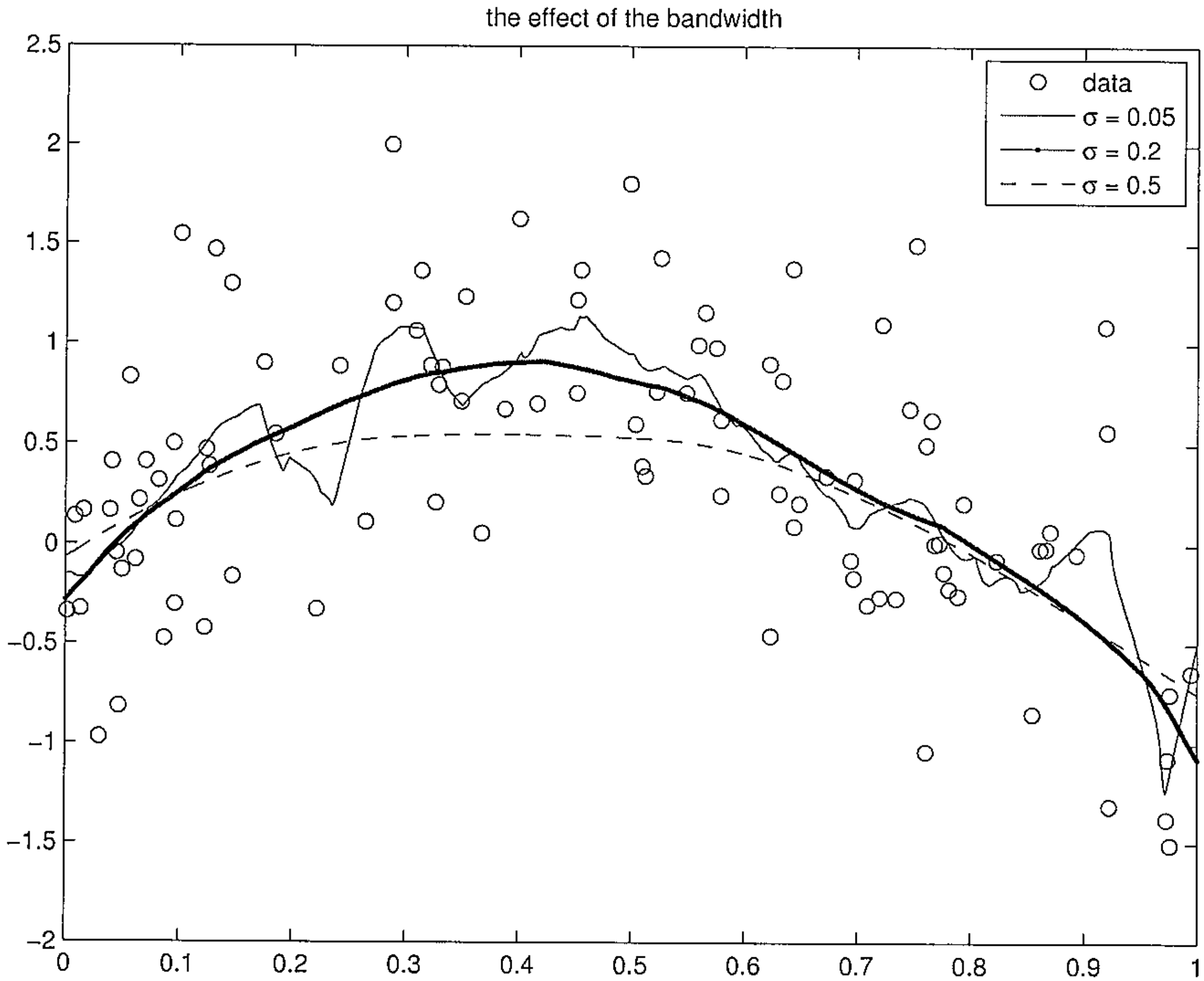
$$\hat{f}(x) = \hat{\theta}(x)^T \cdot \begin{bmatrix} 1 \\ x \end{bmatrix}$$

$$= \underline{y}^T \cdot W(x) A \, (A^T W(x) A)^{-1} \begin{bmatrix} 1 \\ x \end{bmatrix}$$

where

$W(x) =$

• bandwidth selection (bias-variance tradeoff)



$\boxed{\text{Key}}$ A. Least squares

B. $$\hat{f}(x) = \frac{1}{|\{i : \|x_i - x\| < \delta\}|} \sum_{i : \|x_i - x\| < \delta} y_i$$

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} k_\sigma(x - x_i) \cdot y_i}{\sum_{i=1}^{n} k_\sigma(x - x_i)}$$

$$k_\sigma(x - x_i) = \mathbb{1}_{\{\|x - x_i\| < \delta\}}$$

C. $$\hat{\theta} = (\tilde{A}^T \tilde{A})^{-1} \tilde{A}^T \tilde{\underline{y}}$$

$$= (A^T W A)^{-1} A^T W \underline{y}$$

$$W(x) = \begin{bmatrix} k_\sigma(x_1 - x) & & & 0 \\ & k_\sigma(x_2 - x) & & \\ & & \ddots & \\ 0 & & & k_\sigma(x_n - x) \end{bmatrix}$$