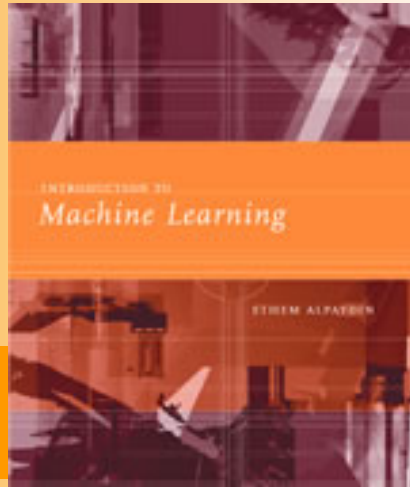




Lecture Slides for

INTRODUCTION TO

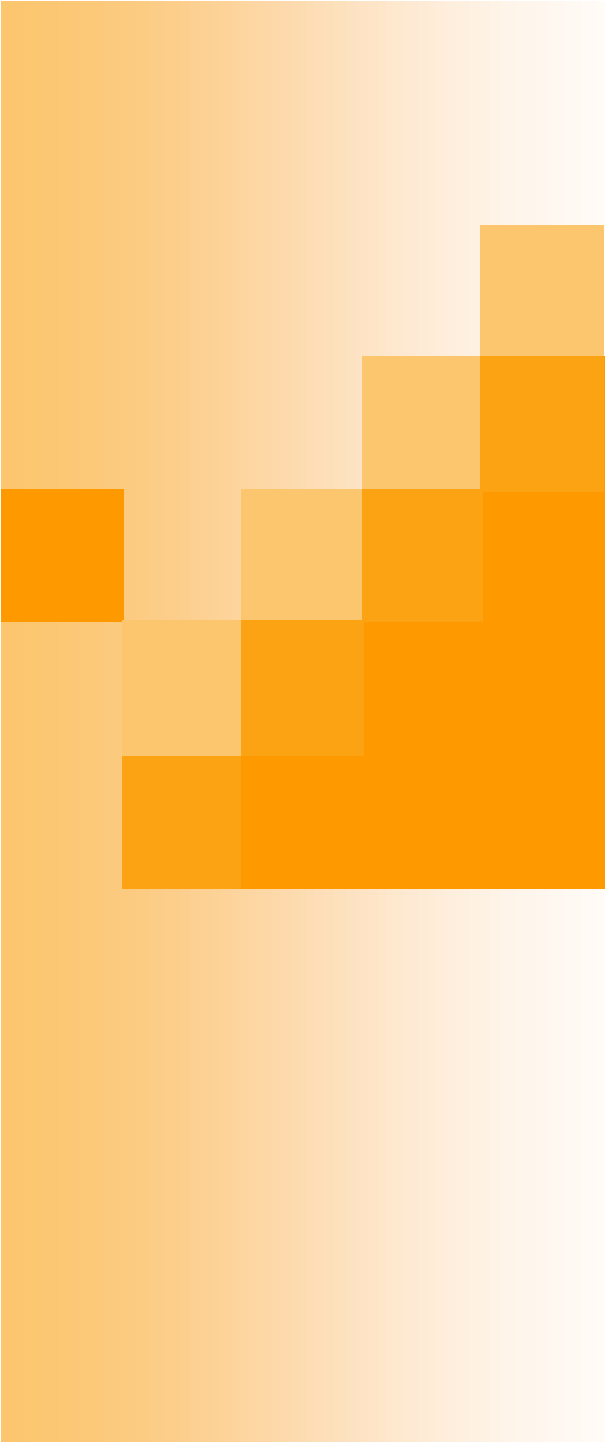
Machine Learning

ETHEM ALPAYDIN

© The MIT Press, 2004

alpaydin@boun.edu.tr

<http://www.cmpe.boun.edu.tr/~ethem/i2ml>



CHAPTER 8:

Nonparametric Methods



Nonparametric Estimation

- Parametric (single global model), semiparametric (small number of local models)
- Nonparametric: Similar inputs have similar outputs
- Functions (pdf, discriminant, regression) change smoothly
- Keep the training data; “let the data speak for itself”
- Given x , find a small number of **closest** training instances and **interpolate** from these
- Aka lazy/memory-based/case-based/instance-based learning



Density Estimation

- Given the training set $\mathcal{X}=\{x^t\}_t$ drawn iid from $p(x)$
- Divide data into bins of size h

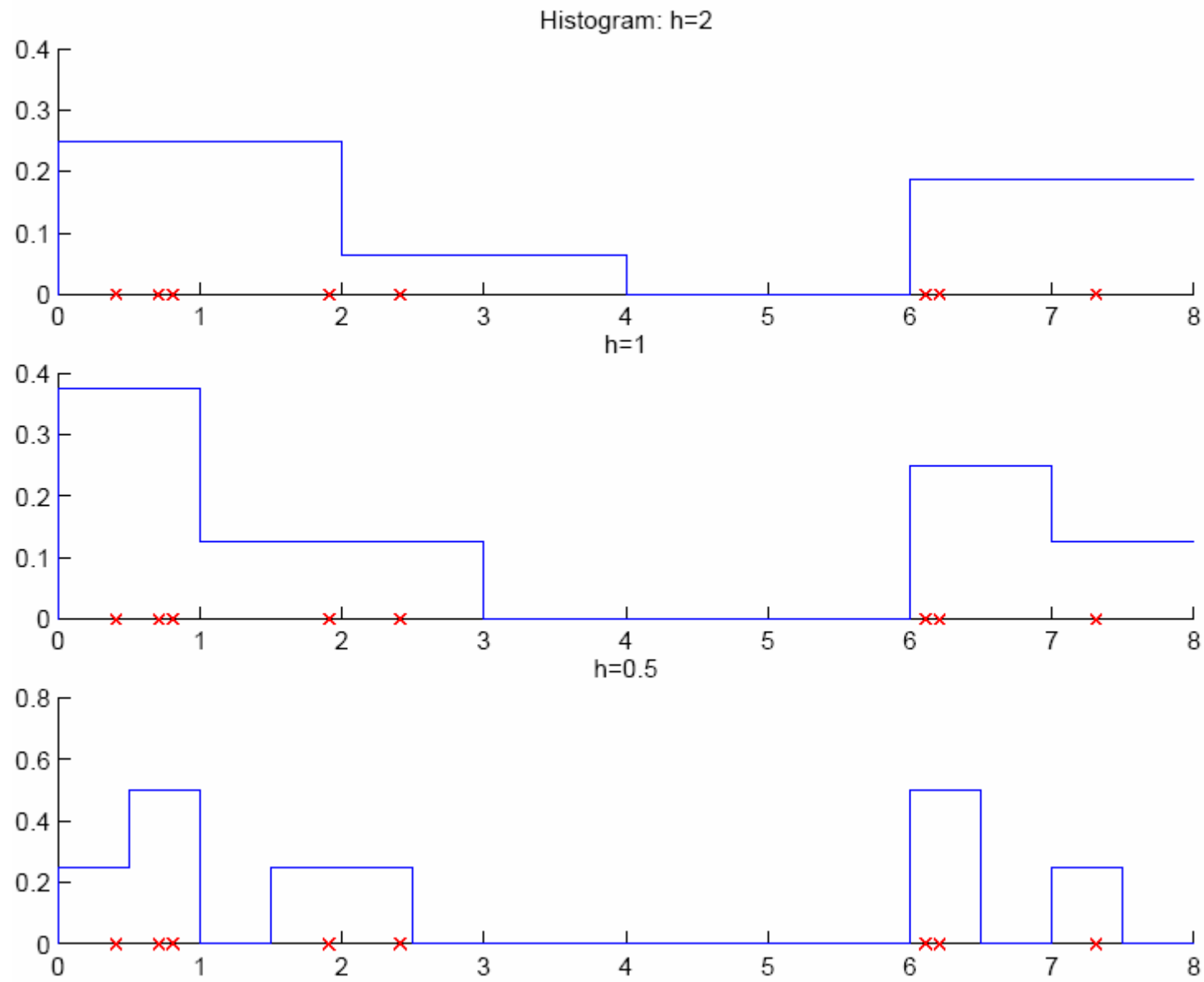
- **Histogram:**
$$\hat{p}(x) = \frac{\#\{x^t \text{ in the same bin as } x\}}{Nh}$$

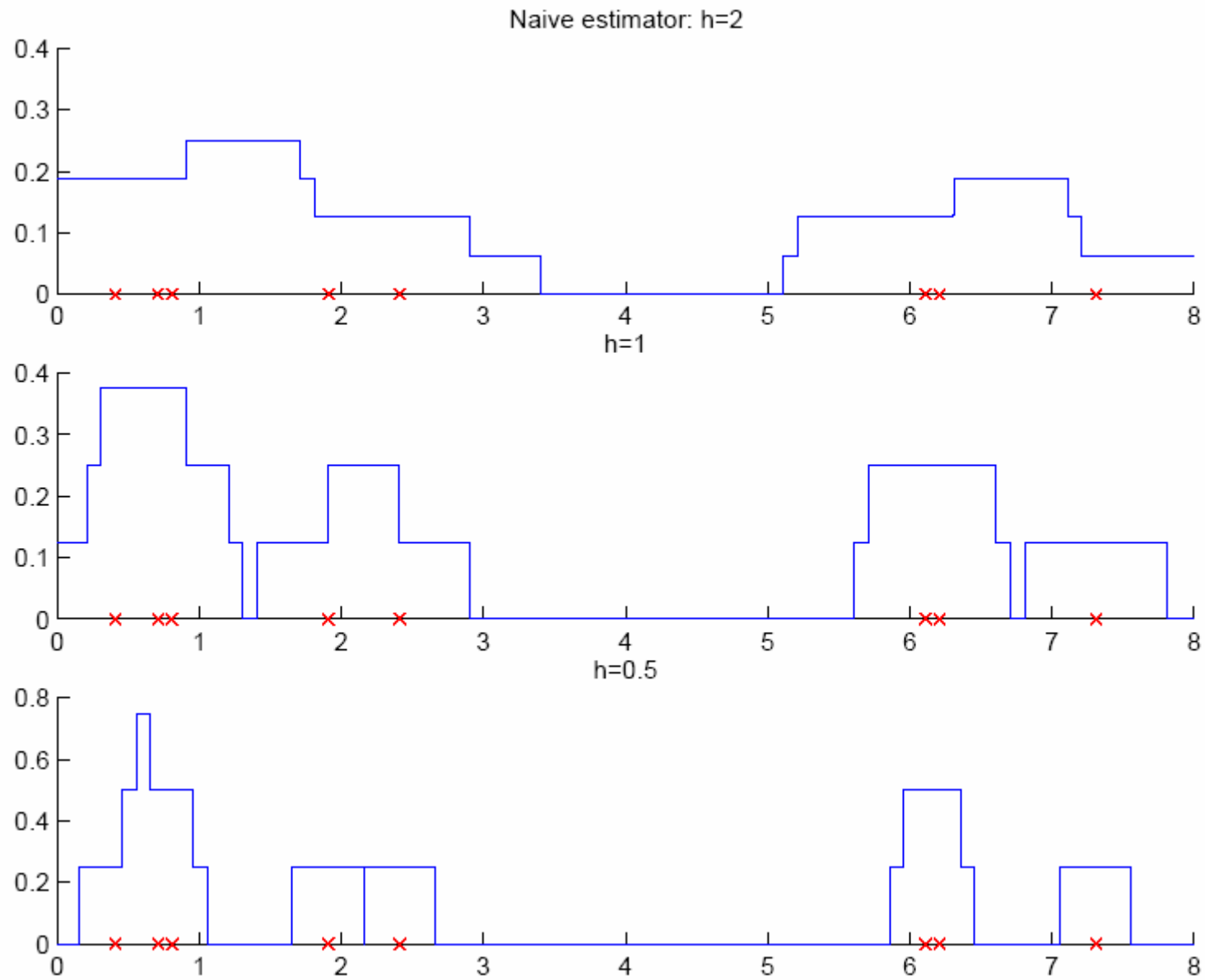
- **Naive estimator:**

$$\hat{p}(x) = \frac{\#\{x - h < x^t \leq x + h\}}{2Nh}$$

or

$$\hat{p}(x) = \frac{1}{Nh} \sum_{t=1}^N w\left(\frac{x - x^t}{h}\right) \quad w(u) = \begin{cases} \frac{1}{2} & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$







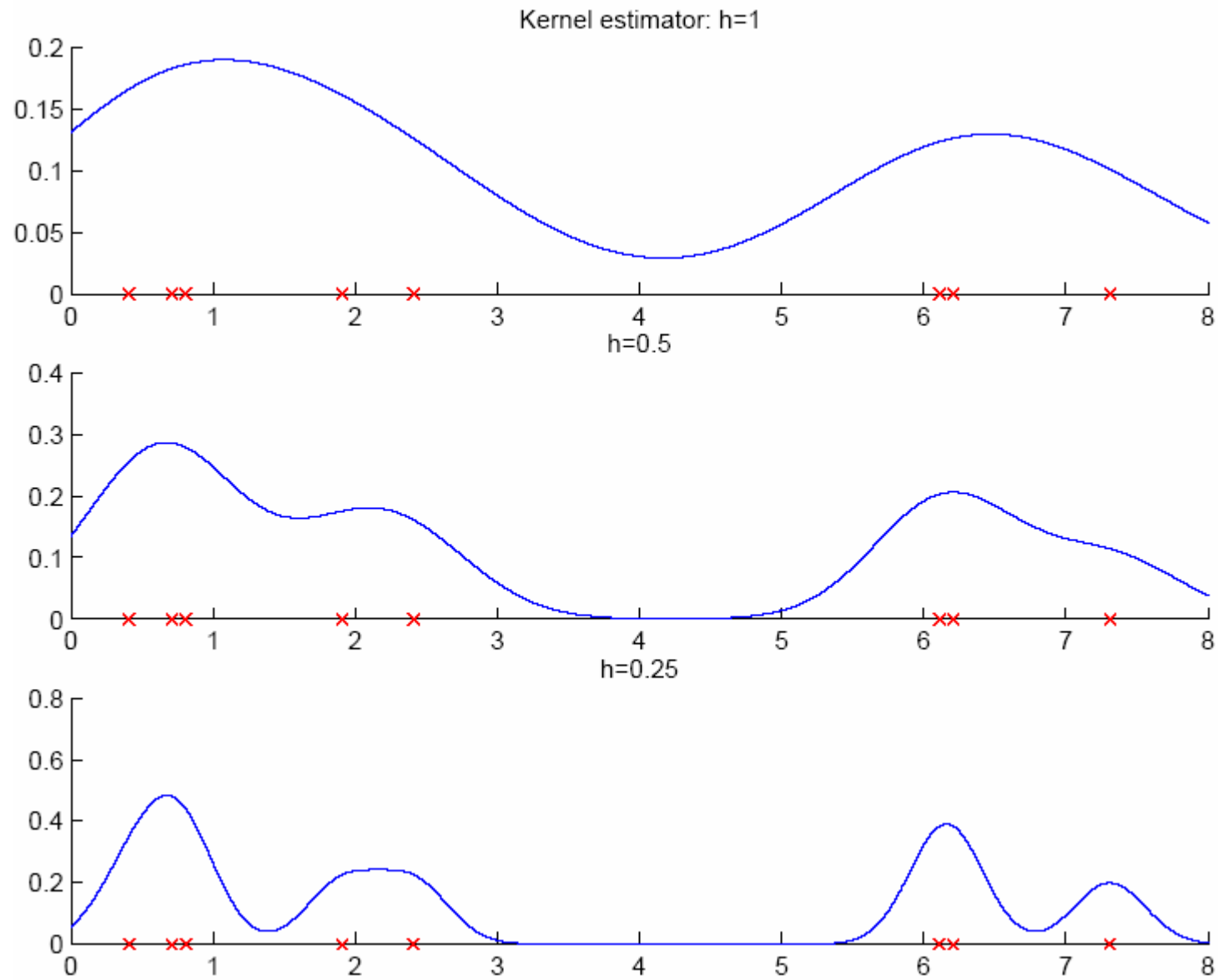
Kernel Estimator

- Kernel function, e.g., Gaussian kernel:

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{u^2}{2} \right]$$

- Kernel estimator (Parzen windows)

$$\hat{p}(x) = \frac{1}{Nh} \sum_{t=1}^N K \left(\frac{x - x^t}{h} \right)$$



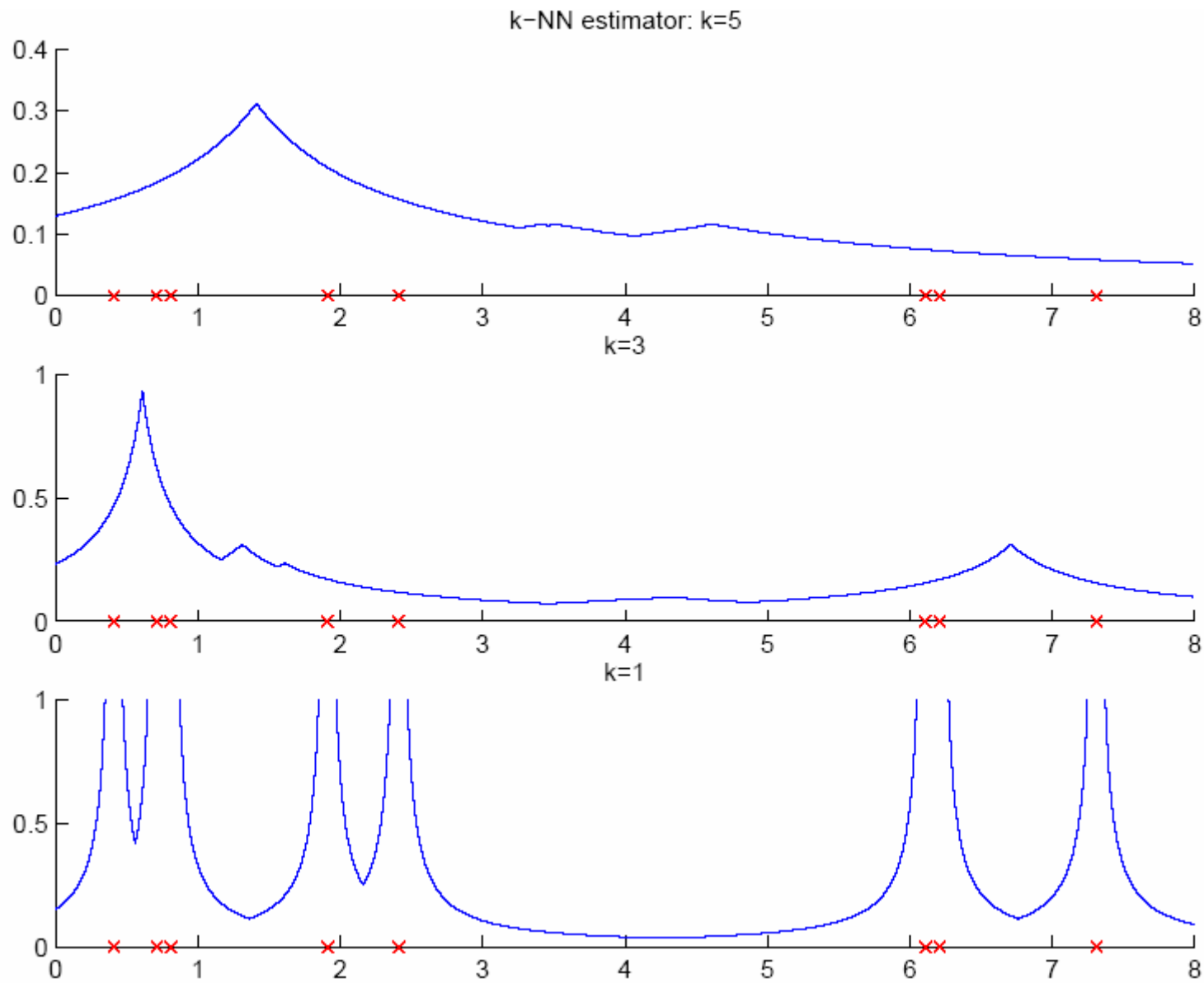


k-Nearest Neighbor Estimator

- Instead of fixing bin width h and counting the number of instances, fix the instances (neighbors) k and check bin width

$$\hat{p}(x) = \frac{k}{2Nd_k(x)}$$

$d_k(x)$, distance to k th closest instance to x





Multivariate Data

- Kernel density estimator

$$\hat{p}(\mathbf{x}) = \frac{1}{Nh^d} \sum_{t=1}^N K\left(\frac{\mathbf{x} - \mathbf{x}^t}{h}\right)$$

Multivariate Gaussian kernel

spheric $K(\mathbf{u}) = \left(\frac{1}{\sqrt{2\pi}}\right)^d \exp\left[-\frac{\|\mathbf{u}\|^2}{2}\right]$

ellipsoid $K(\mathbf{u}) = \frac{1}{(2\pi)^{d/2} |\mathbf{S}|^{1/2}} \exp\left[-\frac{1}{2} \mathbf{u}^T \mathbf{S}^{-1} \mathbf{u}\right]$

Nonparametric Classification

- Estimate $p(\mathbf{x}|C_i)$ and use Bayes' rule
- Kernel estimator

$$\hat{p}(\mathbf{x}|C_i) = \frac{1}{N_i h^d} \sum_{t=1}^N K\left(\frac{\mathbf{x} - \mathbf{x}^t}{h}\right) r_i^t \quad \hat{P}(C_i) = N_i / N$$

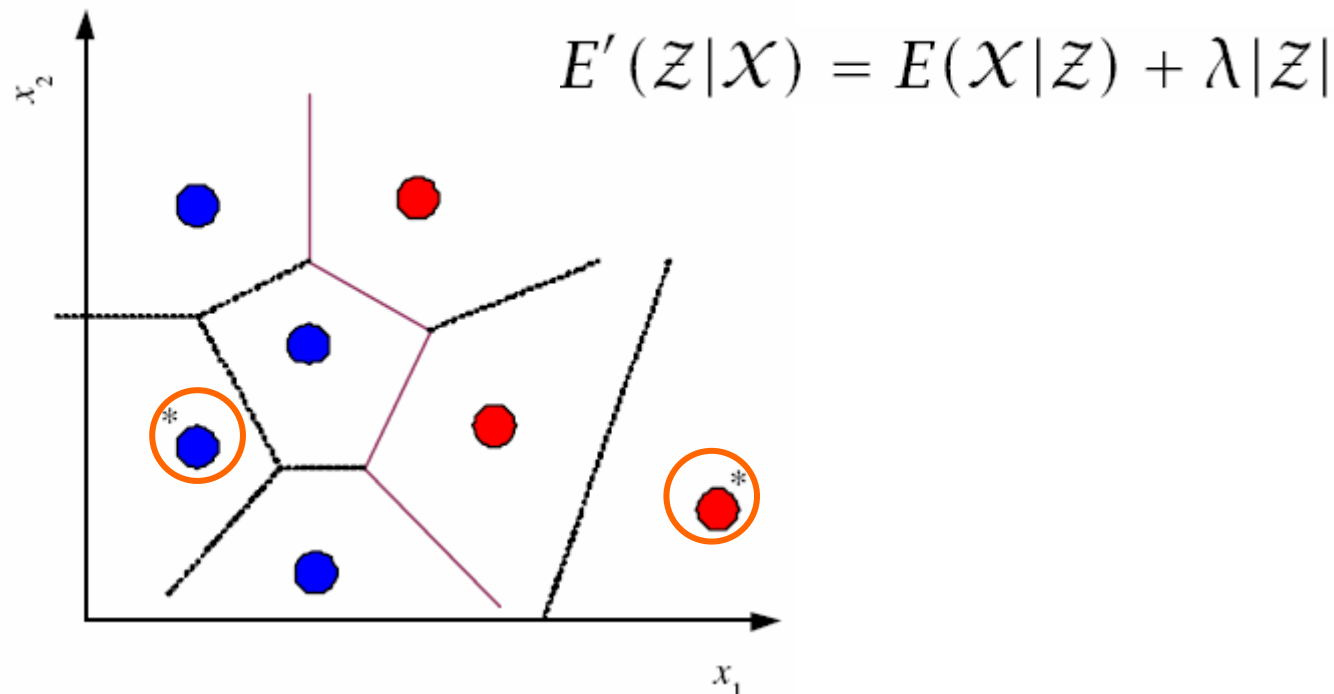
$$g_i(\mathbf{x}) = \hat{p}(\mathbf{x}|C_i) \hat{P}(C_i) = \frac{1}{N h^d} \sum_{t=1}^N K\left(\frac{\mathbf{x} - \mathbf{x}^t}{h}\right) r_i^t$$

- k -NN estimator

$$\hat{p}(\mathbf{x}|C_i) = \frac{k_i}{N_i V^k(\mathbf{x})} \quad \hat{P}(C_i|\mathbf{x}) = \frac{\hat{p}(\mathbf{x}|C_i) \hat{P}(C_i)}{\hat{p}(\mathbf{x})} = \frac{k_i}{k}$$

Condensed Nearest Neighbor

- Time/space complexity of k -NN is $O(N)$
- Find a subset of $\mathcal{X}$ that is small and is accurate in classifying $\mathcal{X}$ (Hart, 1968)



Condensed Nearest Neighbor

- Incremental algorithm: Add instance if needed

$\mathcal{Z} \leftarrow \emptyset$

Repeat

For all $\mathbf{x} \in \mathcal{X}$ (in random order)

Find $\mathbf{x}' \in \mathcal{Z}$ s.t. $\|\mathbf{x} - \mathbf{x}'\| = \min_{\mathbf{x}^j \in \mathcal{Z}} \|\mathbf{x} - \mathbf{x}^j\|$

If $\text{class}(\mathbf{x}) \neq \text{class}(\mathbf{x}')$ add $\mathbf{x}$ to $\mathcal{Z}$

Until $\mathcal{Z}$ does not change



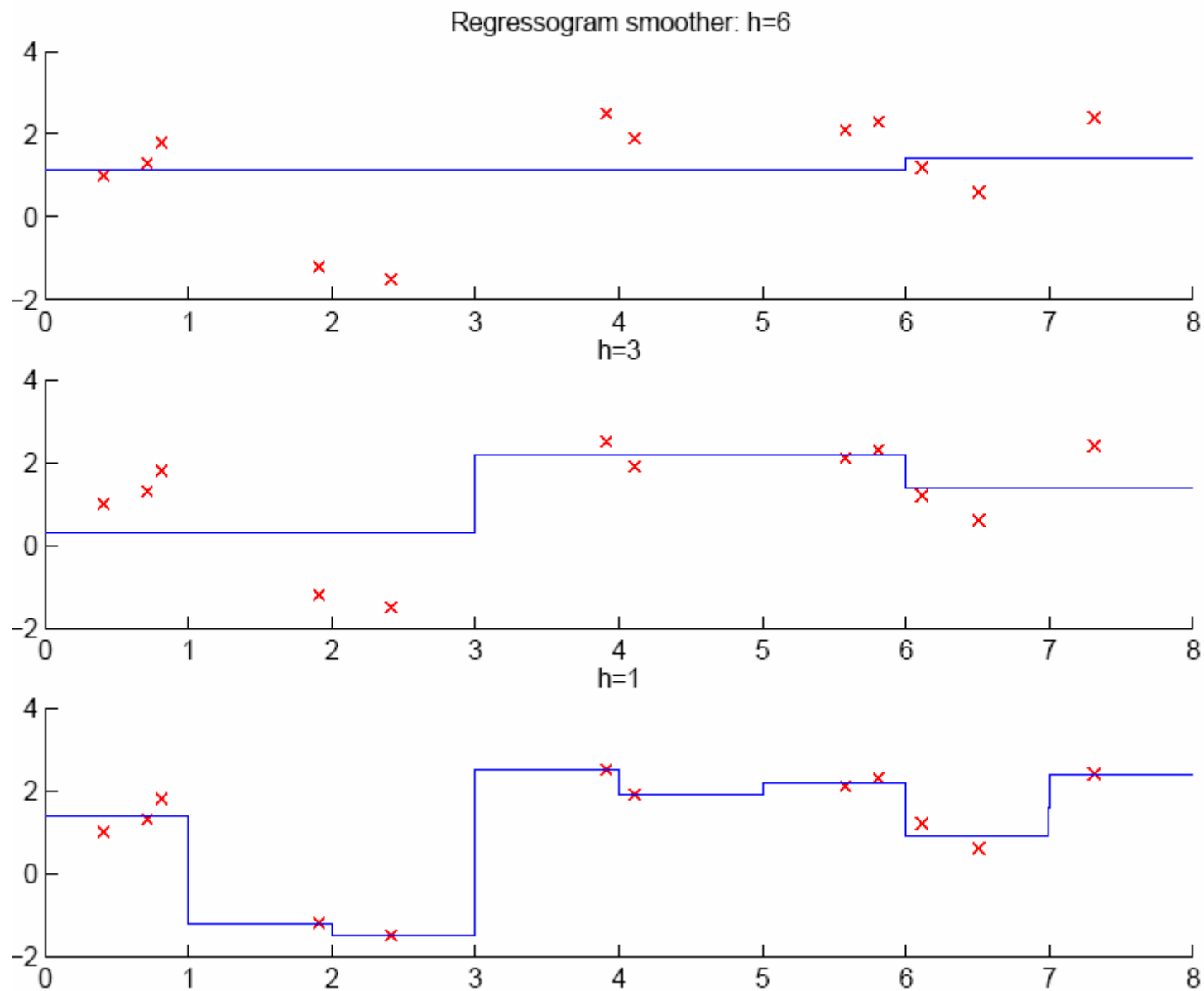
Nonparametric Regression

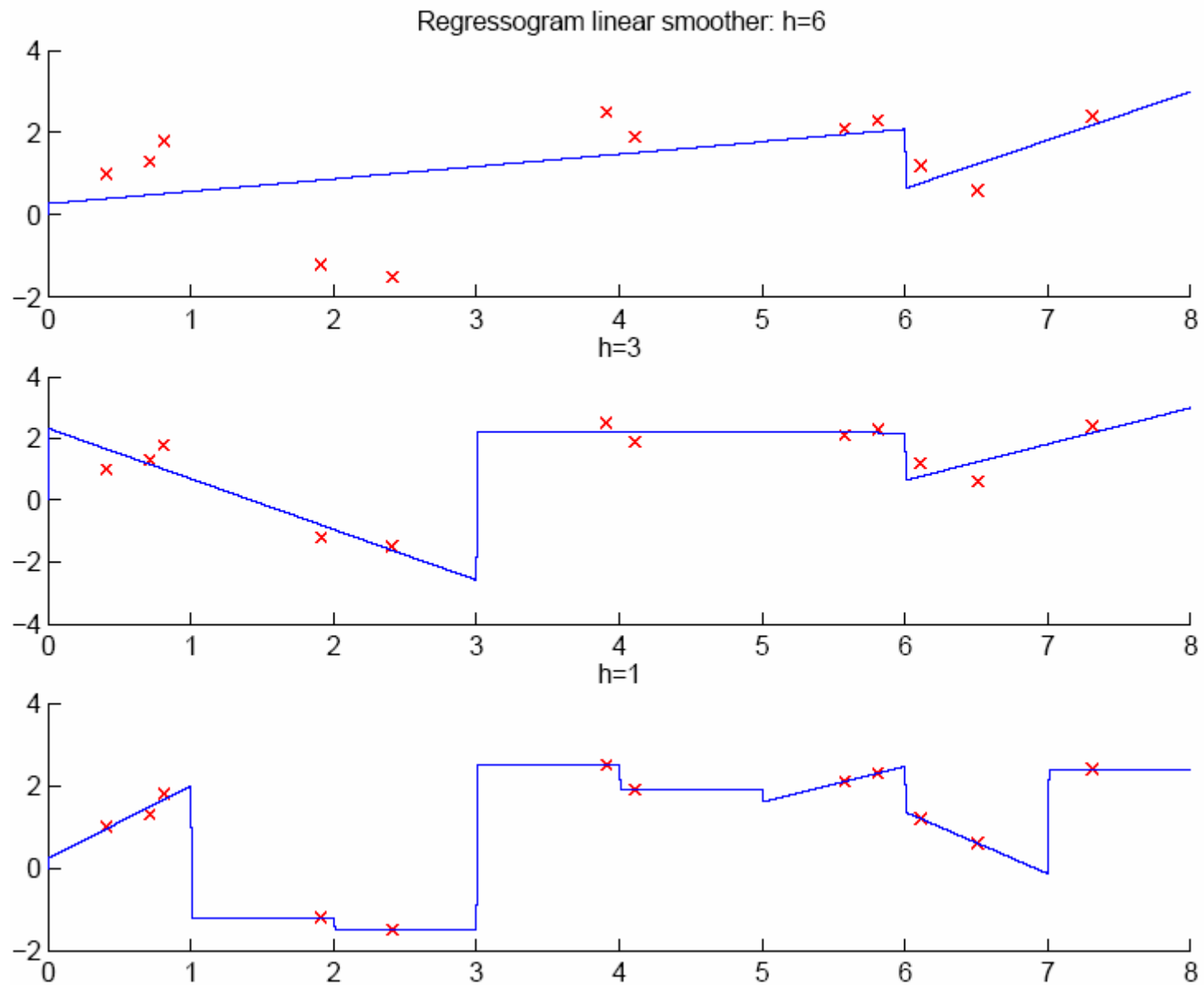
- Aka smoothing models
- Regressogram

$$\hat{g}(x) = \frac{\sum_{t=1}^N b(x, x^t) r^t}{\sum_{t=1}^N b(x, x^t)}$$

where

$$b(x, x^t) = \begin{cases} 1 & \text{if } x^t \text{ is the same bin with } x \\ 0 & \text{otherwise} \end{cases}$$







Running Mean/Kernel Smoother

- Running mean smoother

$$\hat{g}(x) = \frac{\sum_{t=1}^N w\left(\frac{x-x^t}{h}\right) r^t}{\sum_{t=1}^N w\left(\frac{x-x^t}{h}\right)}$$

where

$$w(u) = \begin{cases} 1 & \text{if } |u| < 1 \\ 0 & \text{otherwise} \end{cases}$$

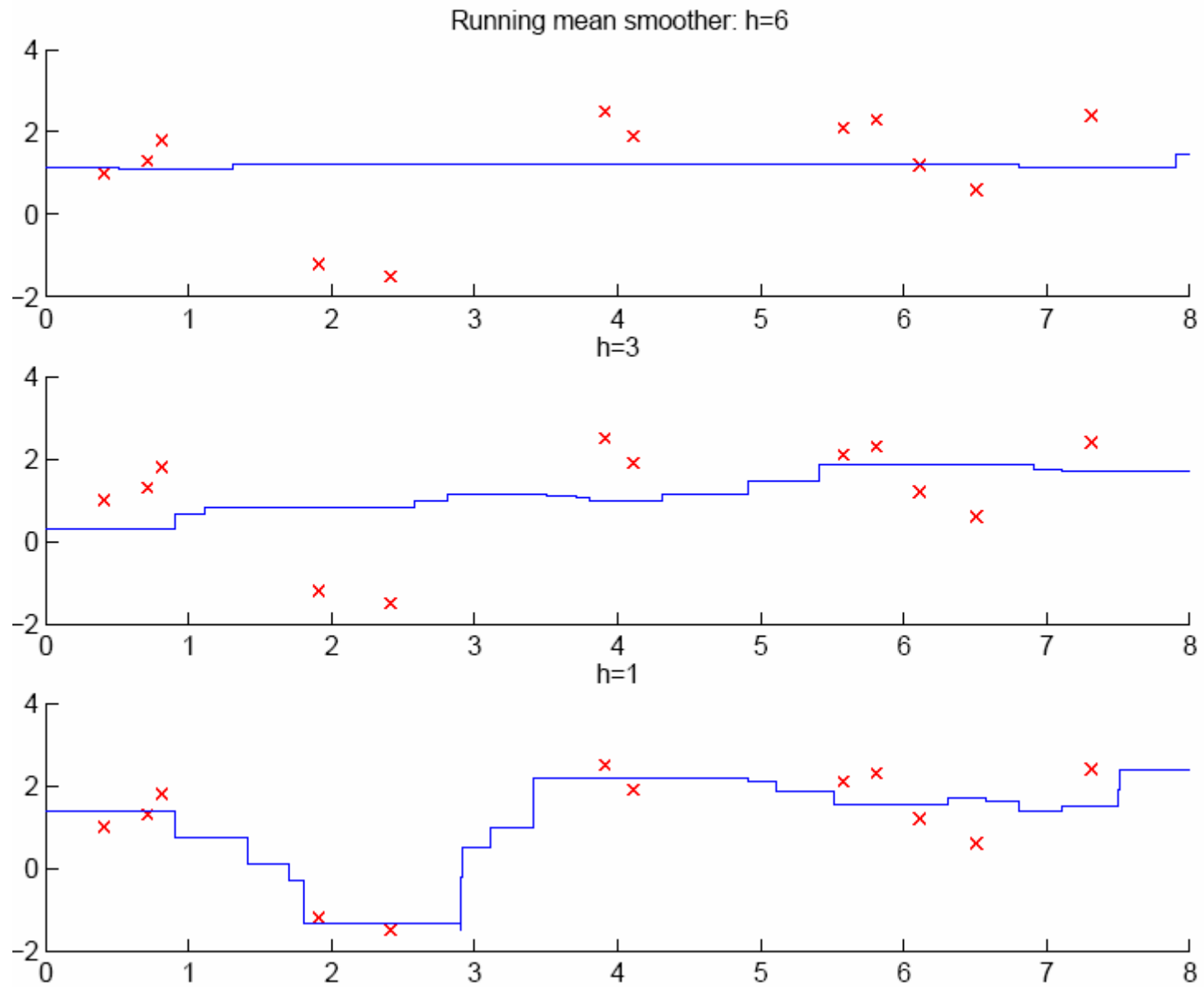
- Running line smoother

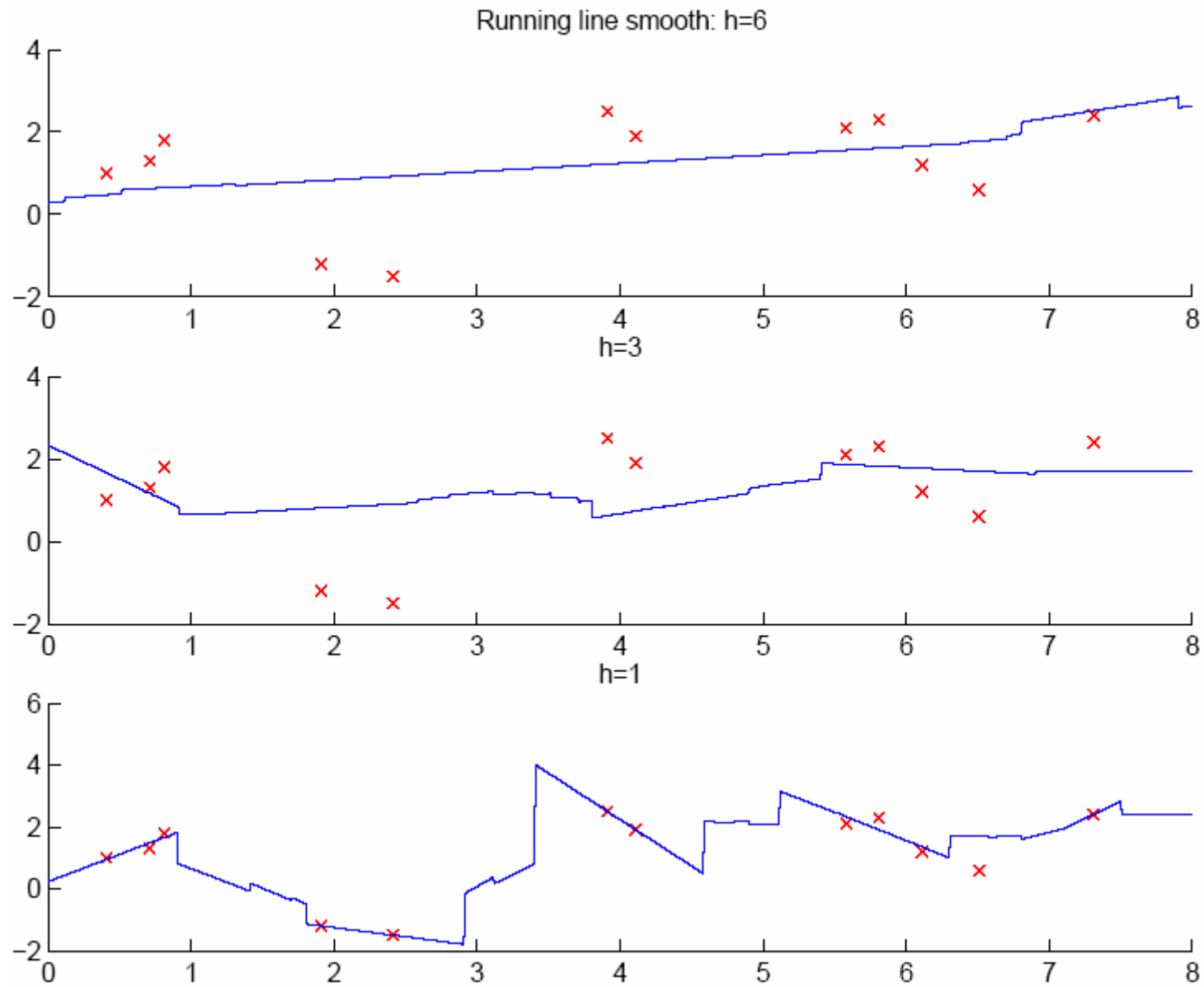
- Kernel smoother

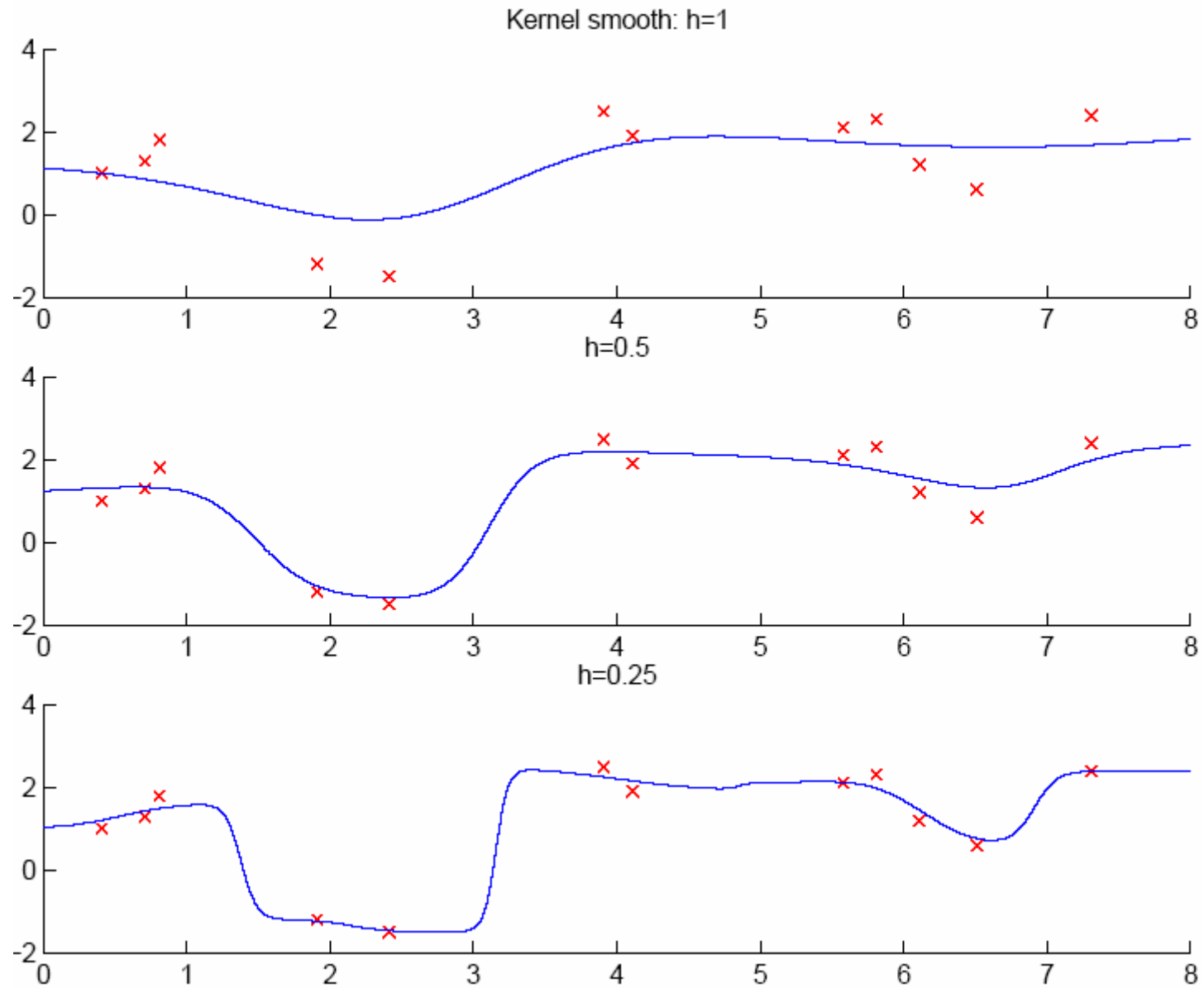
$$\hat{g}(x) = \frac{\sum_t K\left(\frac{x-x^t}{h}\right) r^t}{\sum_t K\left(\frac{x-x^t}{h}\right)}$$

where $K(\cdot)$ is Gaussian

- Additive models (Hastie and Tibshirani, 1990)









How to Choose k or h ?

- When k or h is small, single instances matter; bias is small, variance is large (undersmoothing): High complexity
- As k or h increases, we average over more instances and variance decreases but bias increases (oversmoothing): Low complexity
- Cross-validation is used to finetune k or h .