BIOSTAT 527

4/5/23

Lecture 4

KDE
selecting h
by CV

K-NN density estimation

HW 1 TB posted

OH 4-5 Tue
B-321

CV → notes + 6
posted

HTF:
Ch 6.1-3, 6.6
7.1-3, 7.10

Lecture III – Kernel and k-NN density estimation

Marina Meilă
mmp@stat.washington.edu

Department of Statistics
University of Washington

STAT/BIOST 527
Spring 2023

Outline

1. Kernels
2. Kernel density estimators
3. Choosing h by Cross-Validation and the Bias-Variance trade-off
4. The k-Nearest Neighbor density estimator

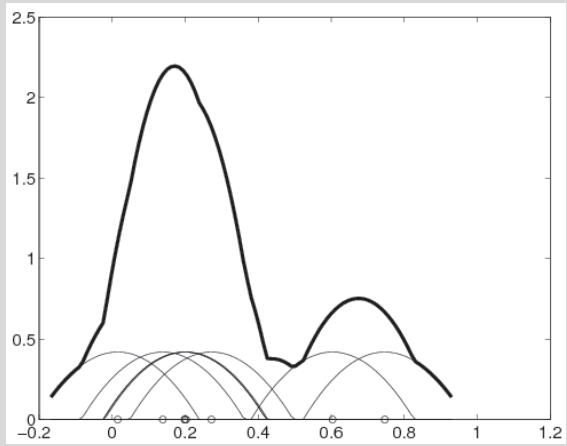
Reading AoNPS Ch.: 4.1-4.3, 6.1-6.3, HTF Ch.:

All of NP Statistics

Kernel density estimators

Data $\mathcal{D} = \{x^{1:n}\}$ ← sample indices

$p_x = ?$



$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} = x$$

dim

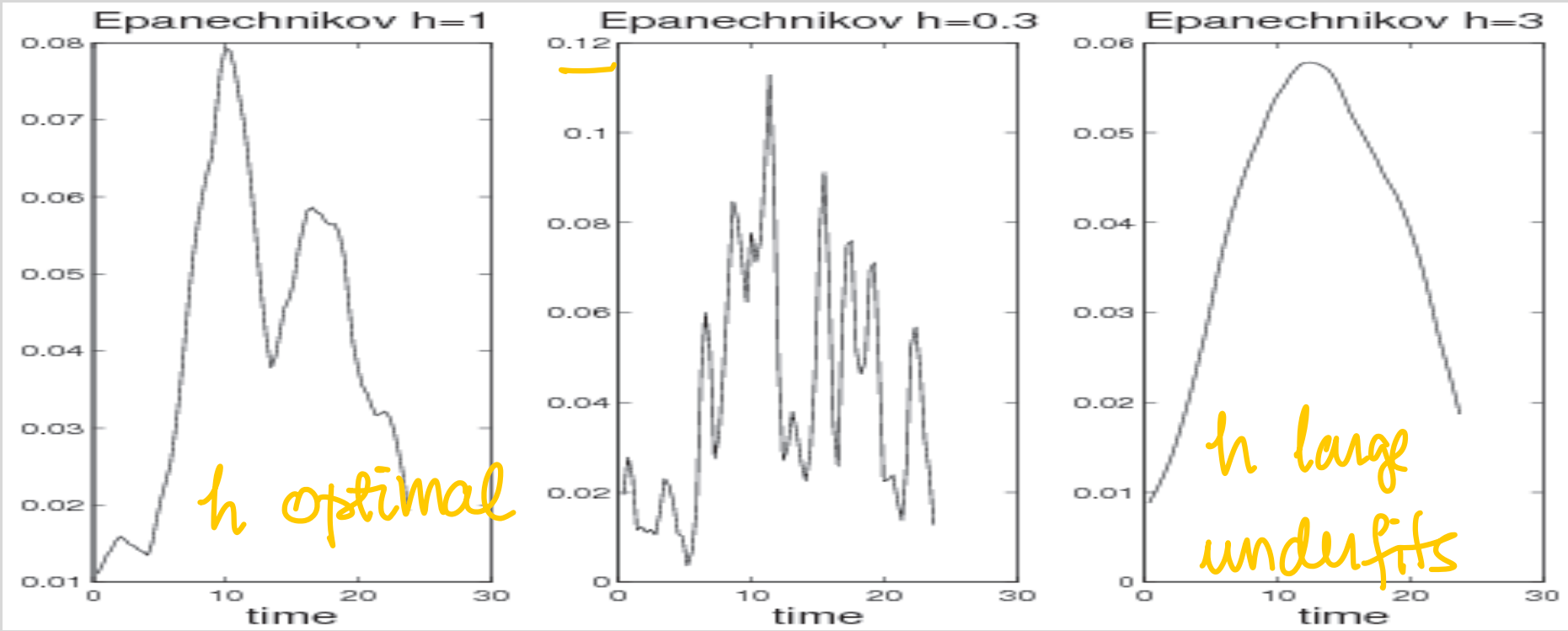
Model family

$b(\,)$ $[\equiv K(\,)]$ kernel

h kernel width

$$\hat{p}_x(x) = \frac{1}{nh} \sum_{i=1}^{n} b\left(\frac{x - x^i}{h}\right)$$

Choosing h


Epanechnikov h=1
Epanechnikov h=0.3
Epanechnikov h=3
time
time
time
h optimal
h large
underfits
for this n
and this n
h small
overfit

The Bias-Variance trade-off and CV

with separate $\mathcal{D}^V$ validation set

train on $\mathcal{D}$, $|\mathcal{D}| = n$

validate on $\mathcal{D}^V$, $|\mathcal{D}^V| = n'$

compute $l(\mathcal{D}^V \mid \mathcal{D}, h)$ → log-likelihood

K-fold CV

$\mathcal{D}$ split into $\mathcal{D}^1, \dots \mathcal{D}^K$ equal size partition

$\mathcal{D}^{-k} = \mathcal{D} \setminus \mathcal{D}^k$

for $k = 1:K$

- estimate $\hat{p}_X$ on $\mathcal{D}^{-k}$
- $l_{cv}^{(k)}(\mathcal{D}^k \mid \mathcal{D}^{-k}, h)$

$\bar{l}_{cv}(h)$

$\bar{l}_{cv} = \frac{1}{K}\sum_k l_{cv}^{(k)}$, std $\bar{l}_{cv}$

The Bias-Variance trade-off

$$\ell_{CV}^{(k)}(\mathcal{D}^k \mid \mathcal{D}^{-k}, h) = \frac{K}{n} \sum_{i \in \mathcal{D}^k} \ell(x^i \mid \mathcal{D}^{-k}, h)$$

$$\ell(x \mid \mathcal{D}, h) = \ln \hat{p}_x(x \mid \mathcal{D}, h)$$

$$|\mathcal{D}^{(k)}| = \frac{n}{K}$$

$$\bar{\ell}_{CV}(h) = \frac{1}{K} \frac{K}{n} \sum_{k=1}^{K} \sum_{i \in \mathcal{D}^k} \ell(x^i \mid \mathcal{D}^{-k}, h)$$

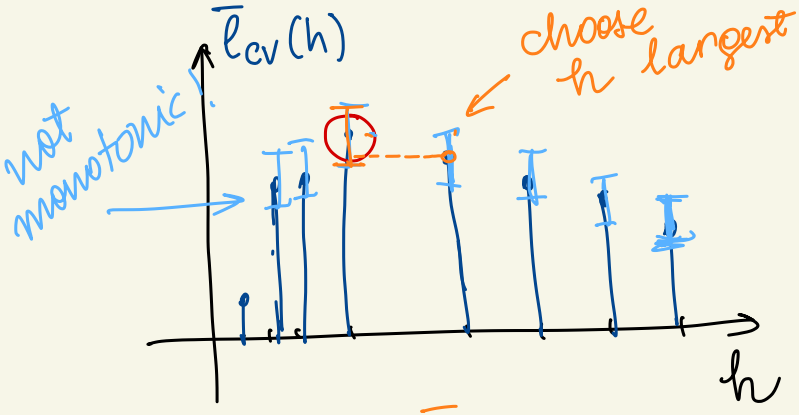
$i \in \mathcal{D}^{(k)}$

$\mathcal{D}^{-k}$: fold k

Ex: how many operations ? i evaluations of $b(\cdot)$

for every h

Choose $h^{opt} = \operatorname{argmax}_h \bar{\ell}_{CV}(h)$



if std $\bar{l}_{cv}(h)$ large

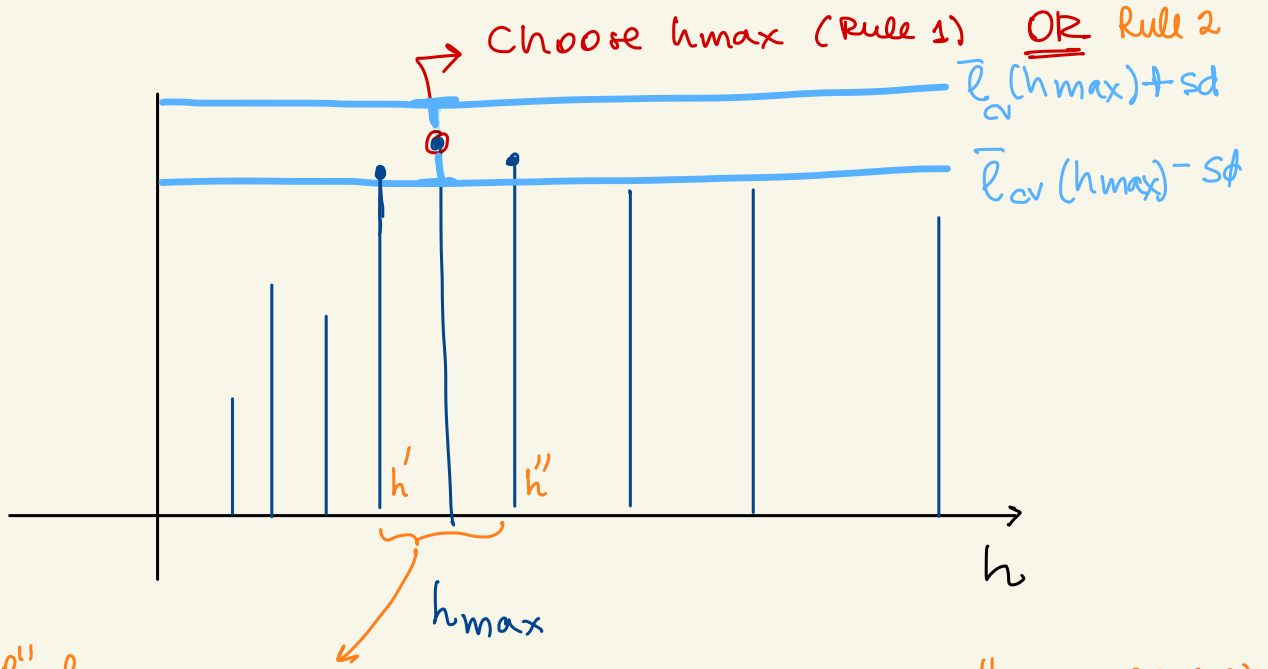
so that $\bar{l}_{cv}(h) + std(h) > \bar{l}_{cv}(h_{max}) - std(h_{max})$

"as long as $\pm$ std intervals overlap"

$K = n$ or K large $\Rightarrow$ std $\bar{l}_{cv}(h)$ large

K small $\rightarrow$ too little data for training

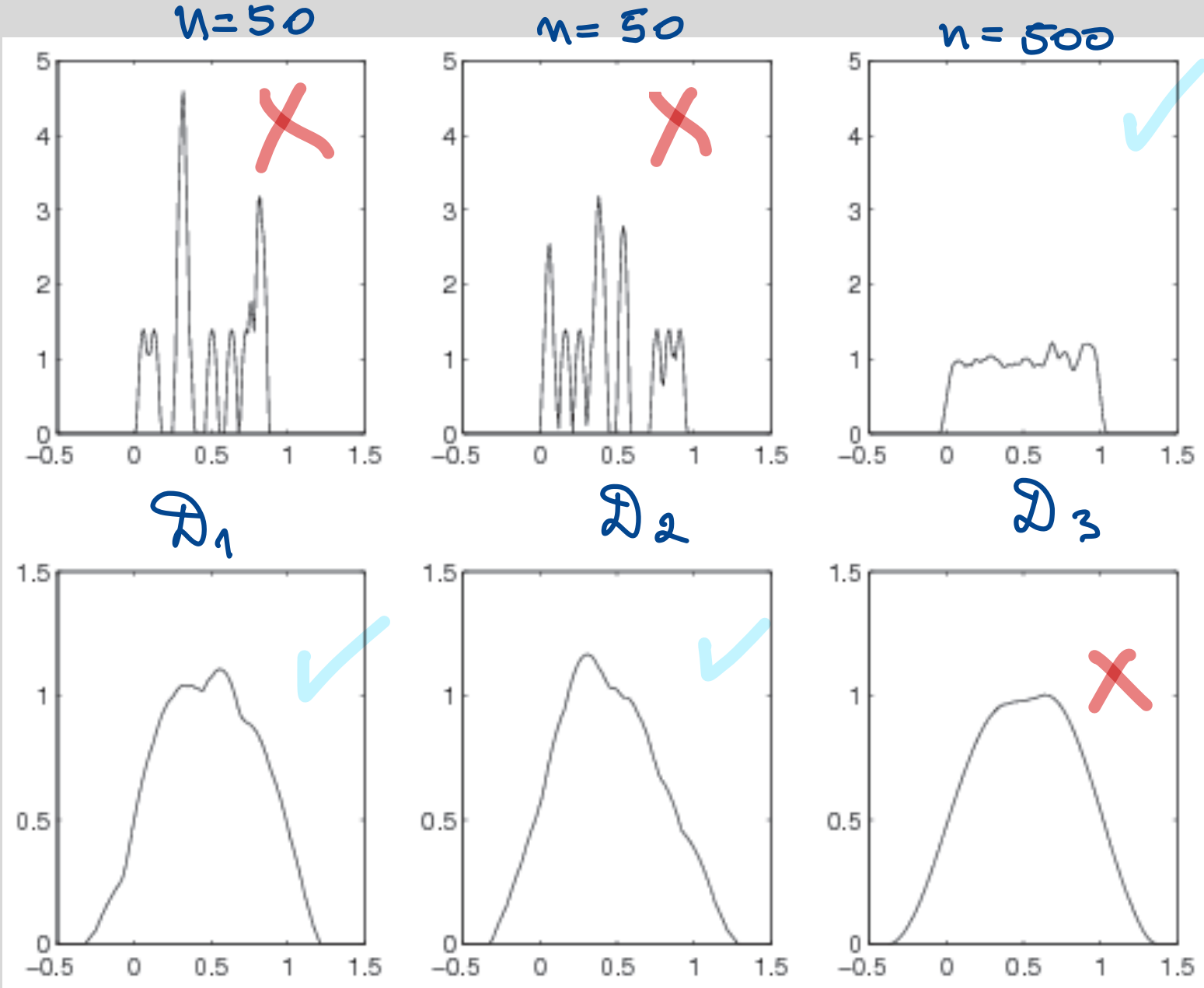
Rule of thumb { $n > 100$ $K = 5 \cdots 10$
$n <$ $K = n$ Leave One Out (LOO)



h', h'', h_{max} have $\bar{\ell}_{cv}$ in interval → choose largest = h'' (Rule 2)

$h' < h_{max} < h''$ ⇒ smoothest model

The Bias-Variance trade-off



The k-Nearest Neighbor density estimator

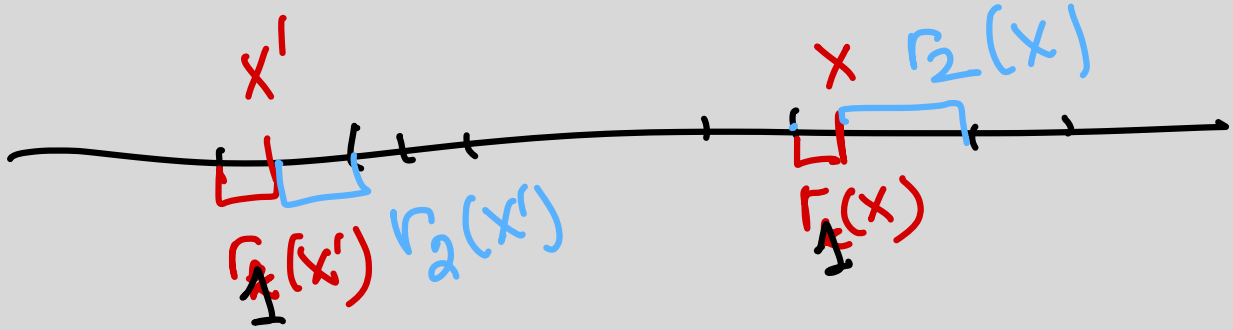
$$\hat{p}_{\text{kNN}}(x) = \frac{k}{n} \frac{1}{\omega_d r_k(x)^d} \quad (1)$$

distance to k-th NN of x

where

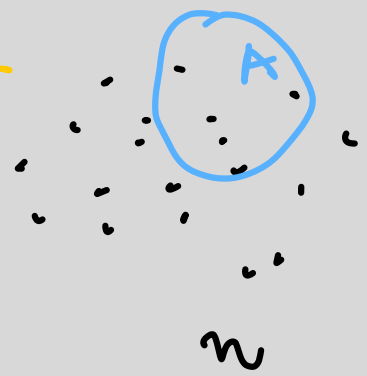
- n is sample size, $\mathcal{D} = \{x^1, \ldots x^n\}$
- $r_k(x)$ is distance to k-th nearest neighbor of x in $\mathcal{D}$
- d is the dimension of the data
- $\omega_d = \frac{\pi^{d/2}}{\Gamma(1+d/2)}$ is the volume of the unit ball in d dimensions
  - $d = 1$, $\omega_1 = 2$
  - $d = 2$, $\omega_2 = \pi$
  - $d = 3$, $\omega_3 = \frac{4}{3}\pi$

$\omega_d r^d$ = Volume Ball (x, r)



The k-Nearest Neighbor density estimator – Motivation

$$\hat{p}_{\text{kNN}}(x) = \frac{k}{n}\frac{1}{\omega_d r_k(x)^d}$$



Idea

- let $B(x, r)$ be the ball of radius r centered at x
- probability of any set approximated by the empirical distribution

$$Pr[B(x,r)] \approx \frac{\#\text{ data points in } B(x,r)}{n} = \frac{|\mathcal{D} \cap B(x,r)|}{n} \tag{2}$$

- density at x approximated by

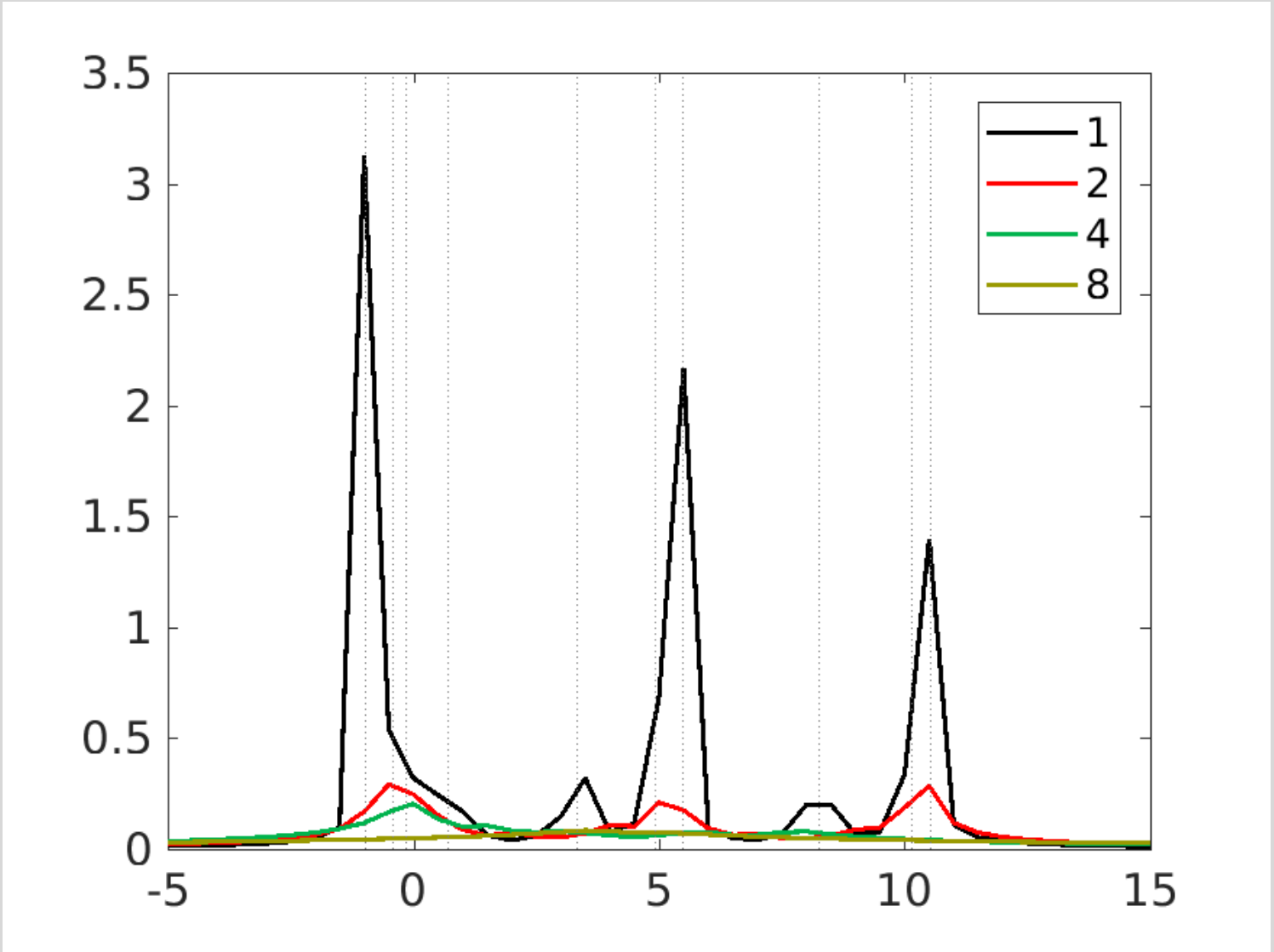
$$\hat{p}_{\text{kNN}}(x) \approx \frac{Pr[B(x,r)]}{\text{Vol}\, B(x,r)} \tag{3}$$

- $\text{Vol}\, B(x,r) = r^d \omega_d$
- if $r = r_k(x)$, then #data points in $B(x,r) = k$

. . . putting it all together we get (1)

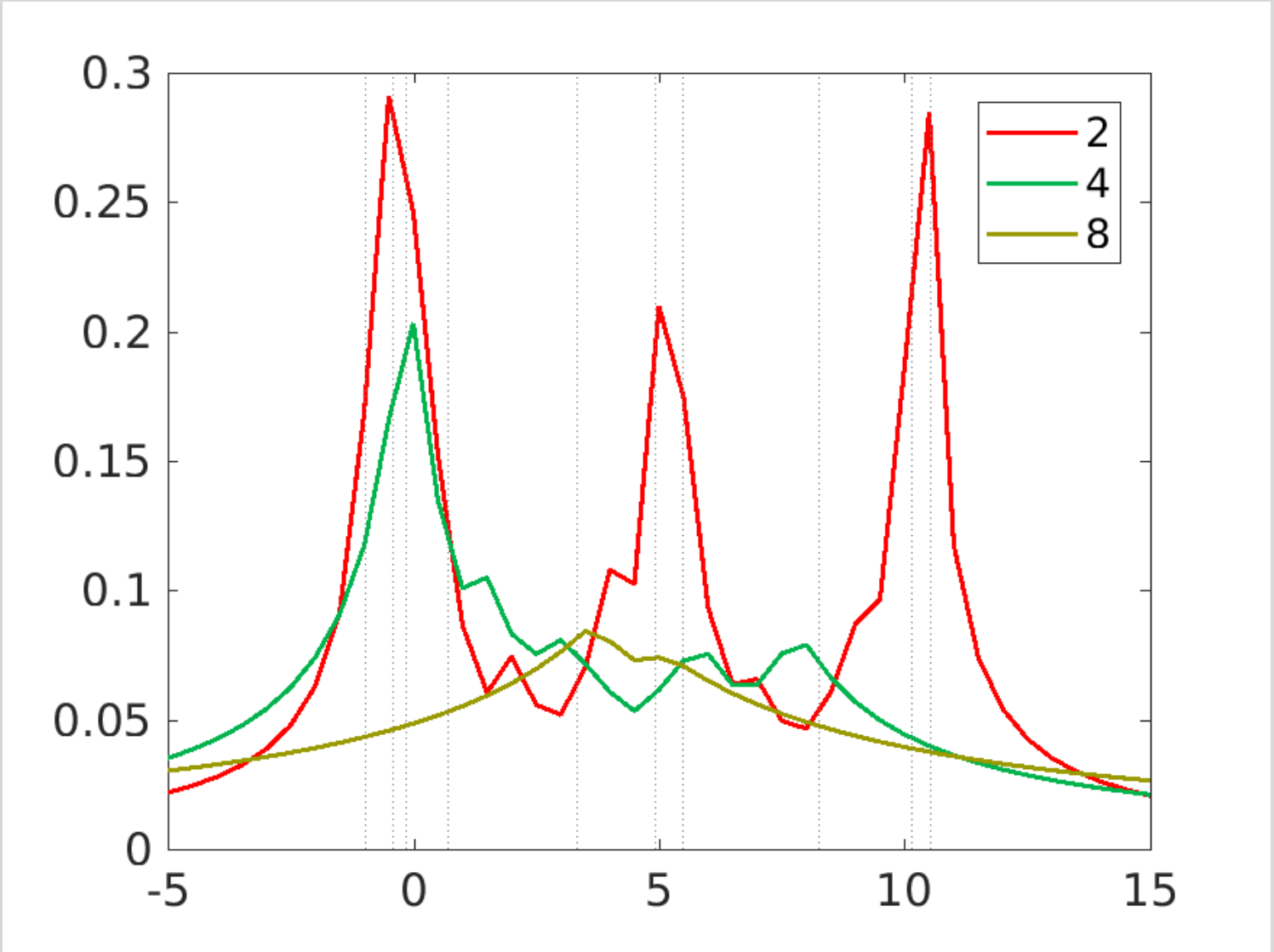
smoothing parameter k { $k=1$ Var ↑ bias ↓ ; k large bias ↑ var ↓

k-NN density estimator for $n = 10$ points



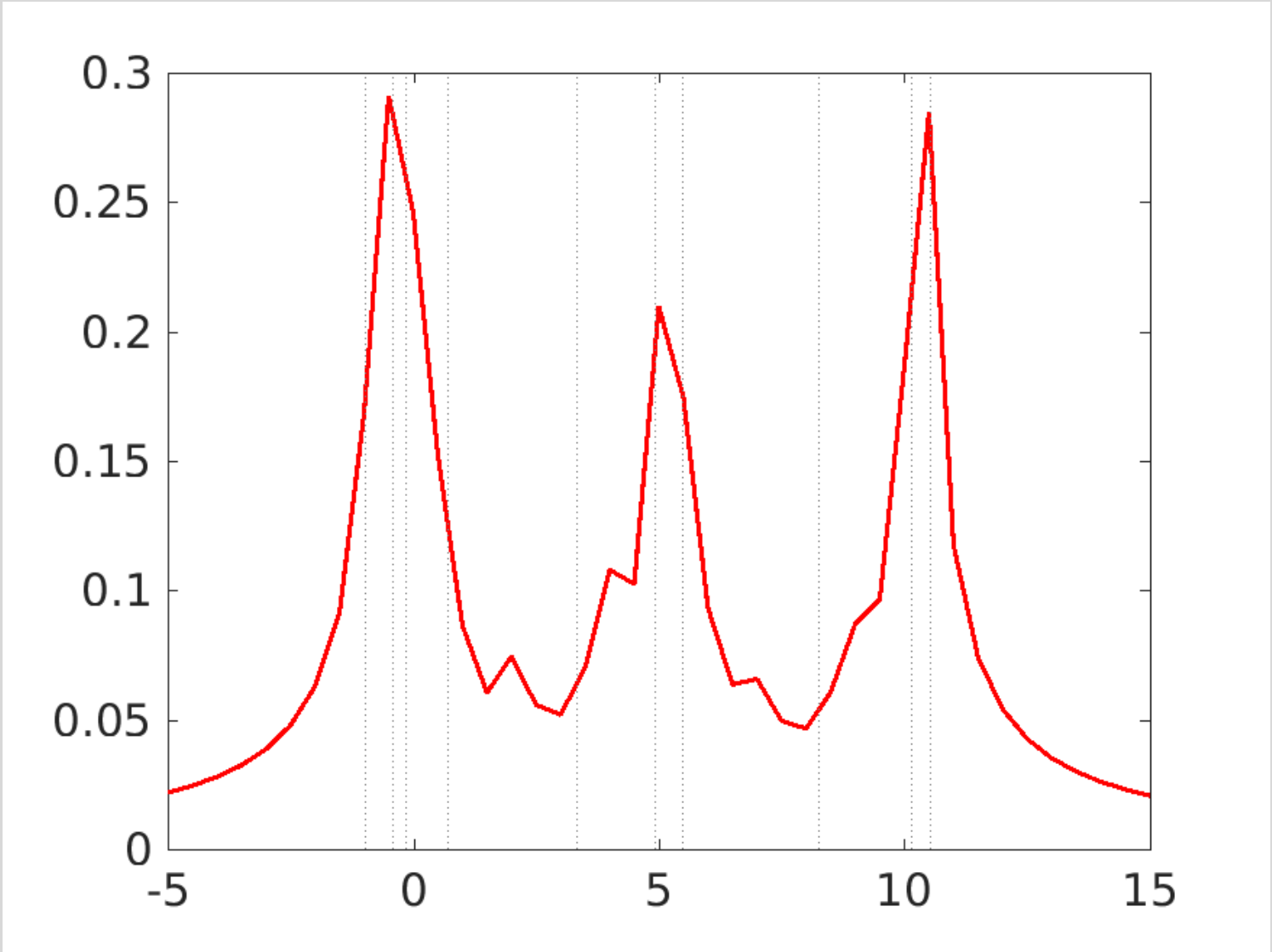
Note that when $d = 1$, $\int_{\mathbb{R}} \hat{p}_{\text{kNN}}(x)dx = \infty$!!!!!!

k-NN density estimator for $n = 10$ points



Note that when $d = 1$, $\int_{\mathbb{R}} \hat{p}_{\text{kNN}}(x)dx = \infty$!!!!!!

k-NN density estimator for $n = 10$ points



only $k = 2$

Note that when $d = 1$, $\int_{\mathbb{R}} \hat{p}_{\text{kNN}}(x)dx = \infty$!!!!!!

Convergence rates

- for $d = 1$: $k^* \sim n^{4/5}$ MSE$\sim n^{-4/5}$

1) $\frac{k^*}{n} \to 0$ for $n \to \infty$

increases slowly

KR $d = 1$ $\quad MSE \sim n^{-\frac{4}{5}}$

$h \sim n^{-\frac{1}{5}}$

KDE — " —

$n^{\frac{1}{d+4}}$

$$E_{P_X}\left[\left(\hat{P}_X - P_X\right)^2 \middle| n\right] = MISE$$