Lecture III – Kernel and k-NN density estimation

Marina Meilă
mmp@stat.washington.edu

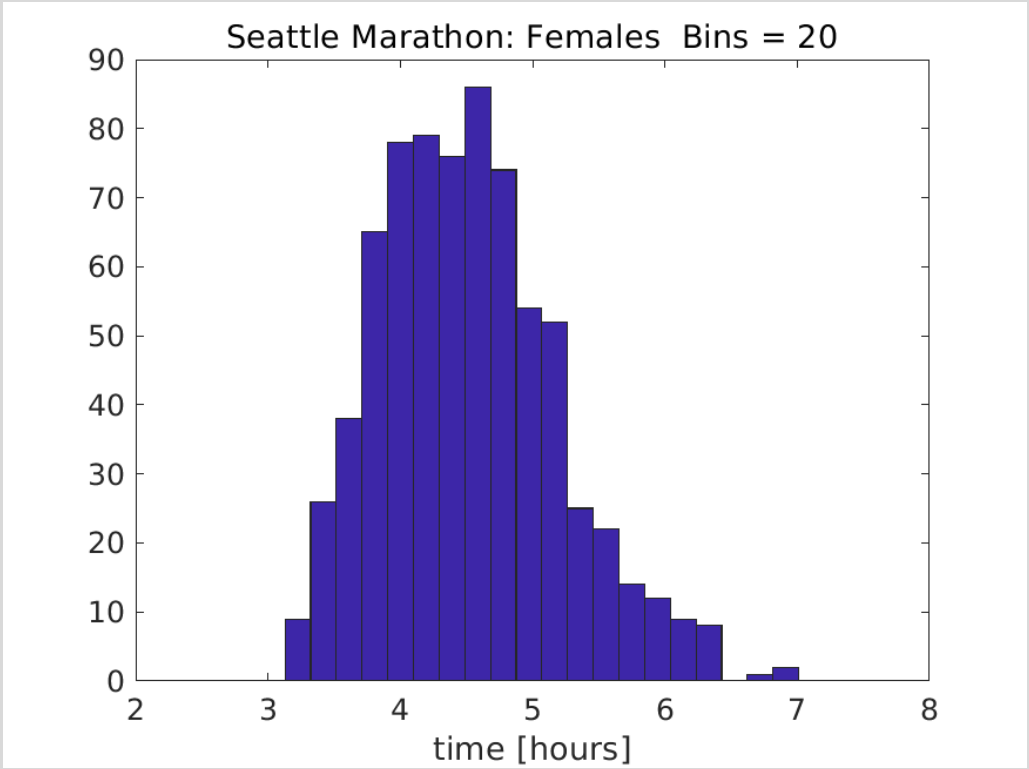
Department of Statistics
University of Washington

STAT/BIOST 527
Spring 2023

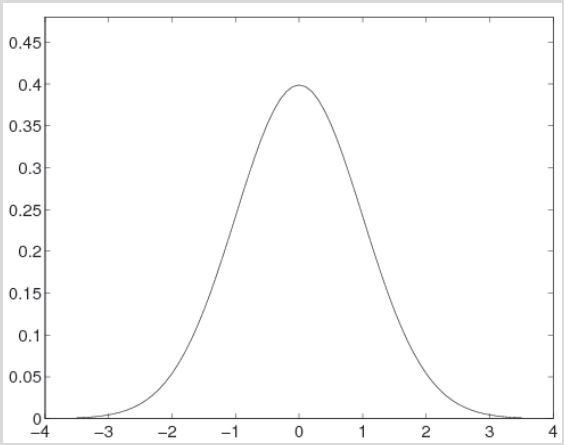
Outline

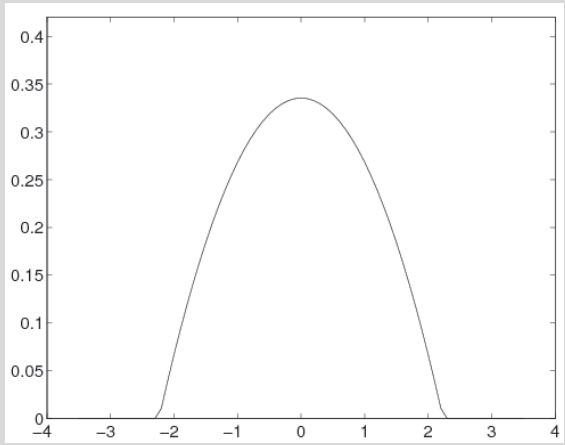
Reading AoNPS Ch.: 4.1-4.3, 6.1-6.3, HTF Ch.:

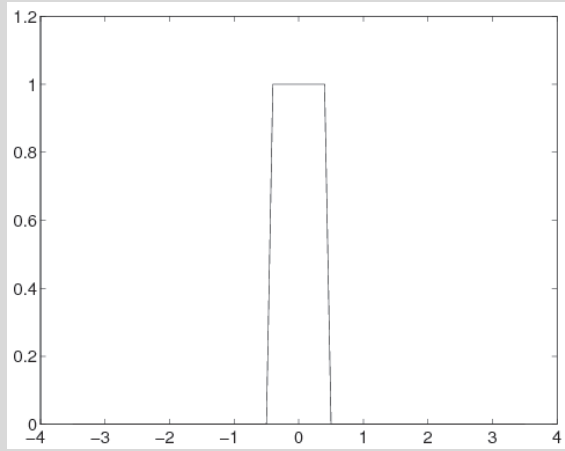
Arbitrary shaped densities



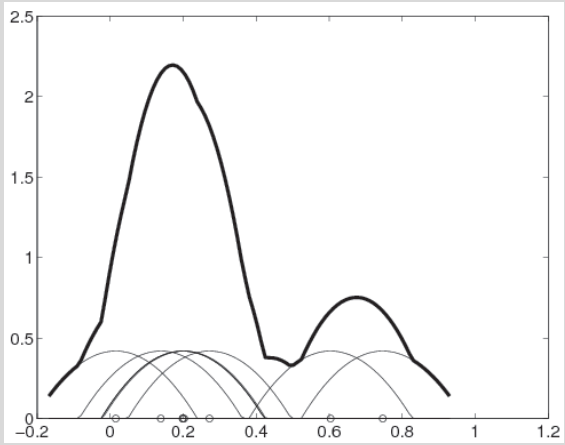
Kernels


0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4 −3 −2 −1 0 1 2 3 4

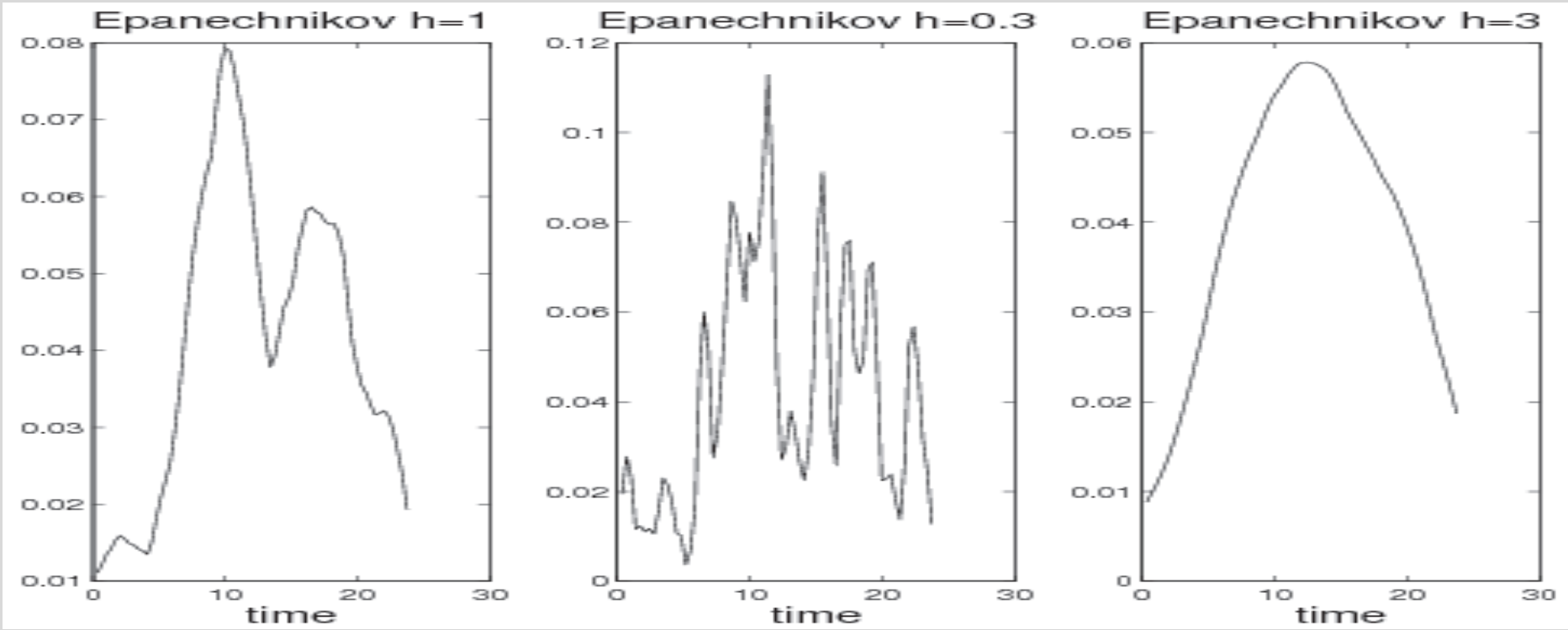

0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
−4 −3 −2 −1 0 1 2 3 4


1.2
1
0.8
0.6
0.4
0.2
0
−4 −3 −2 −1 0 1 2 3 4

Kernel density estimators



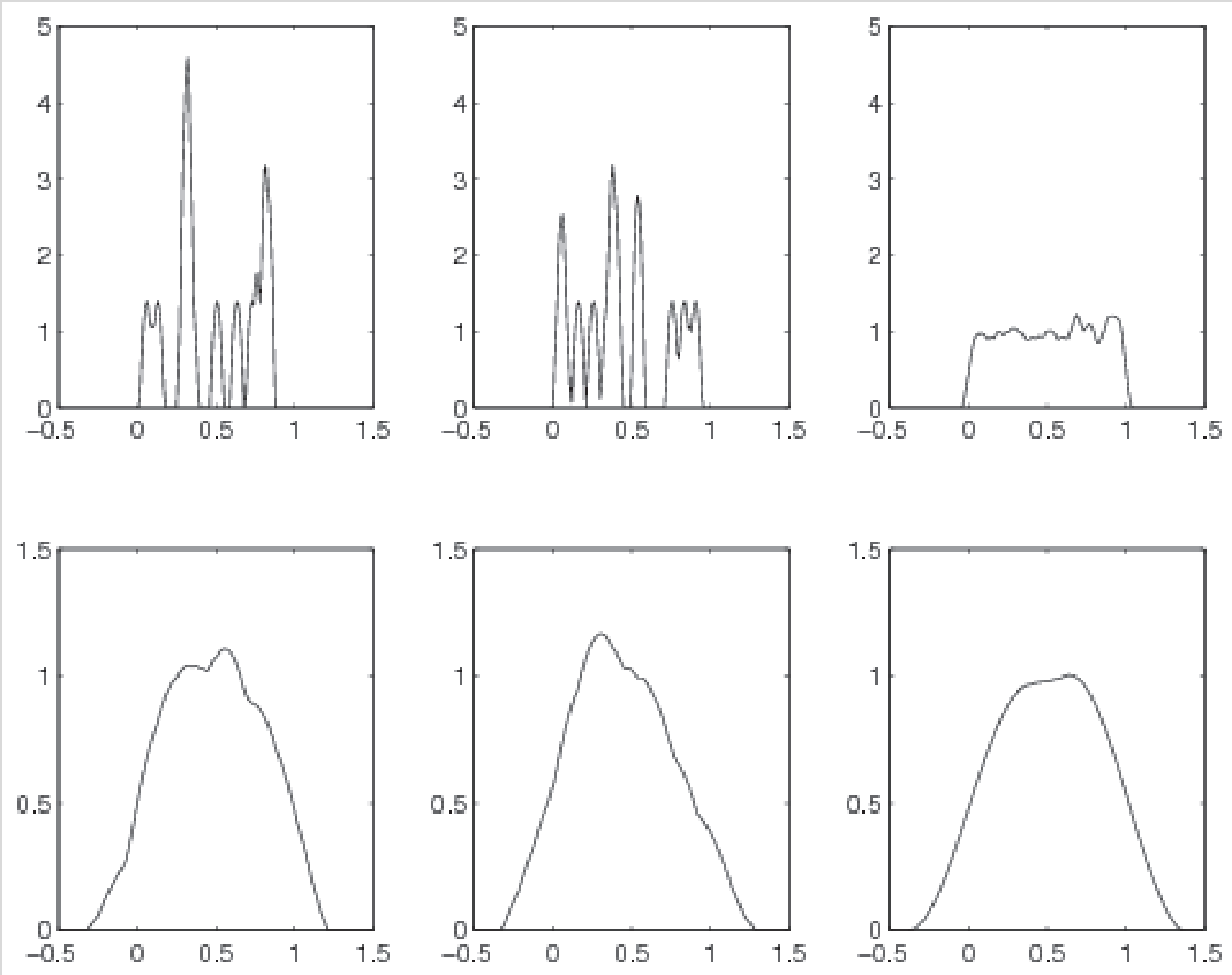
Choosing h


Epanechnikov h=1
0.08
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
10
20
30
time
Epanechnikov h=0.3
0.12
0.1
0.08
0.06
0.04
0.02
0
0
10
20
30
time
Epanechnikov h=3
0.06
0.05
0.04
0.03
0.02
0.01
0
0
10
20
30
time

The Bias-Variance trade-off

Choosing h

The Bias-Variance trade-off



The k-Nearest Neighbor density estimator

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

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
  - $d = 1$, $\omega_1 = 2$
  - $d = 2$, $\omega_2 = \pi$
  - $d = 3$, $\omega_3 = \frac{4}{3}\pi$

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

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

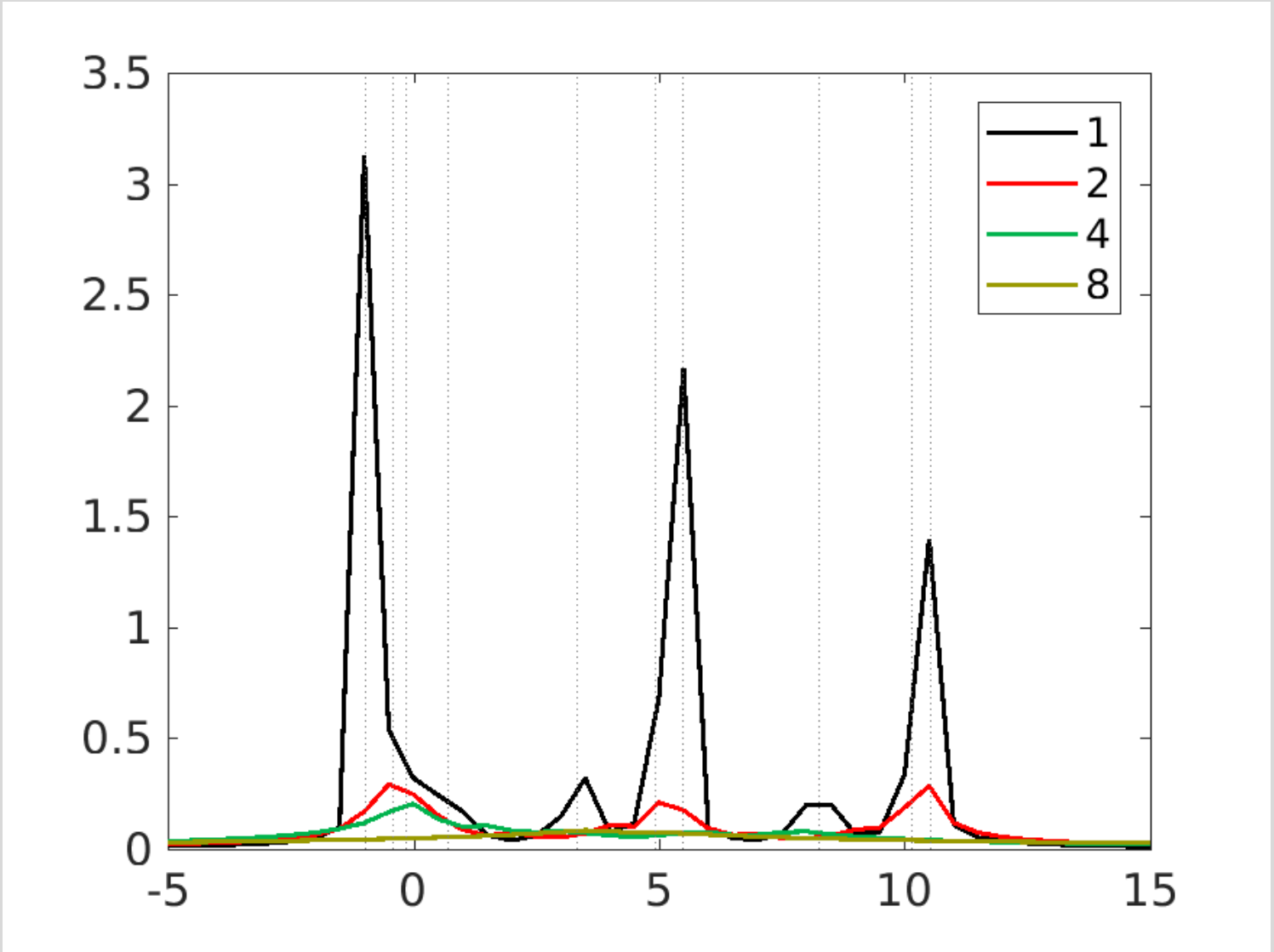
- density at x approximated by

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

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

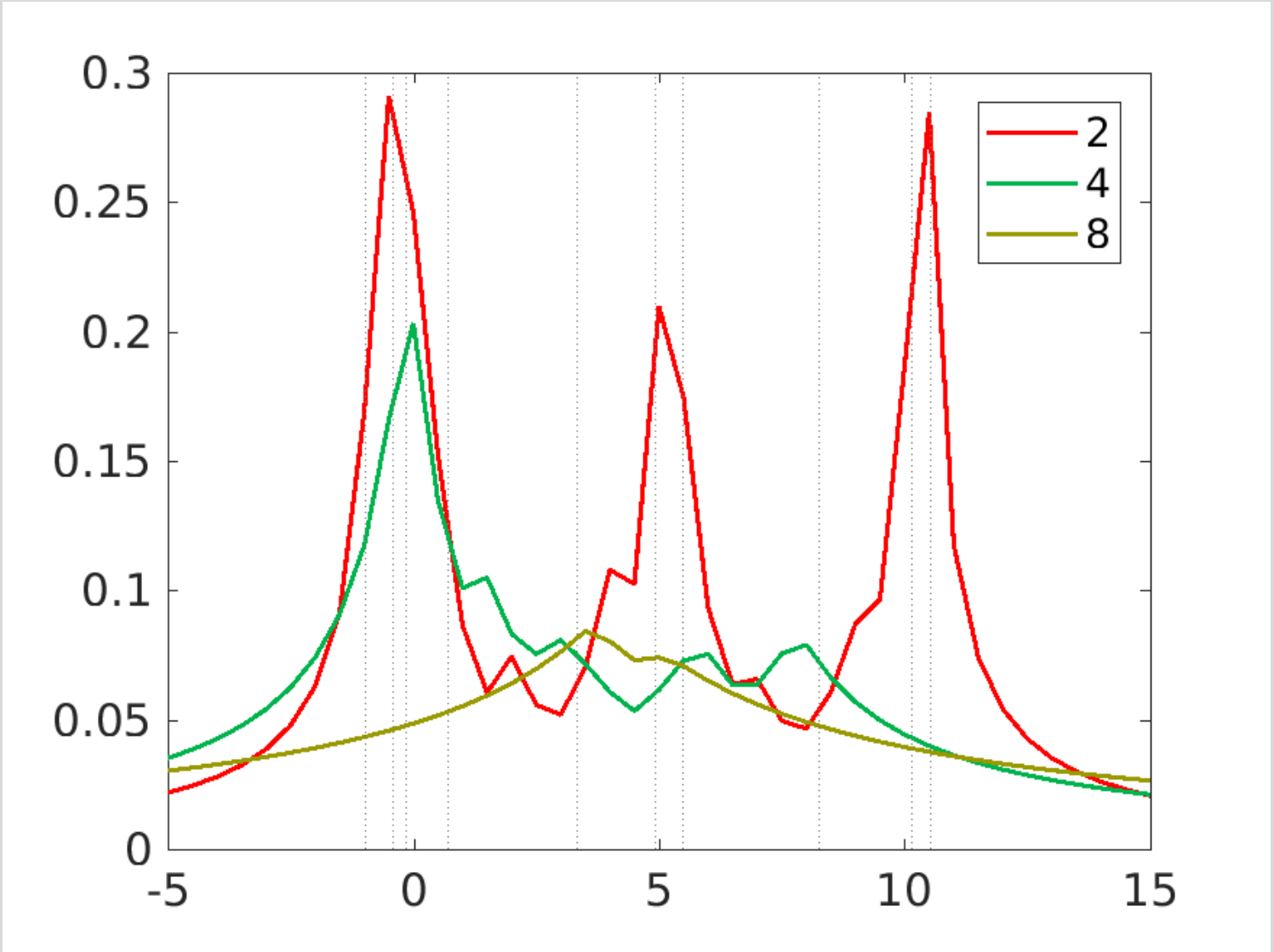
  . . . putting it all together we get (1)

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



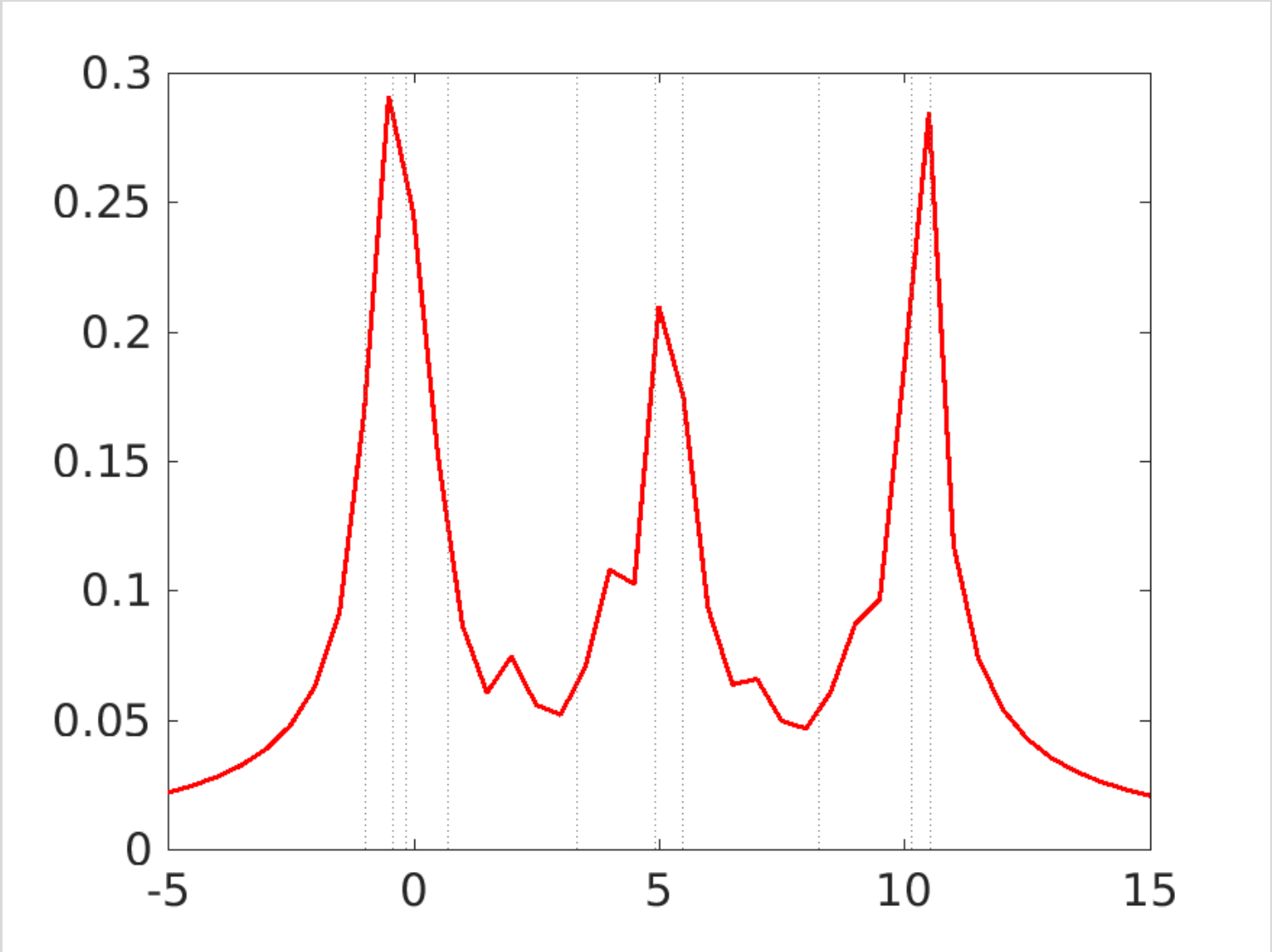
Note that when $d = 1$, $\int_{\mathbb{R}} \hat{p}_{\text{kNN}}(x)dx = \infty$!!!!!!

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$\sim n^{-4/5}$