Johns Hopkins Engineering

625.464 Computational Statistics

Choice of Kernel and Multidimensional Estimators

Module 11 Lecture 11E

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

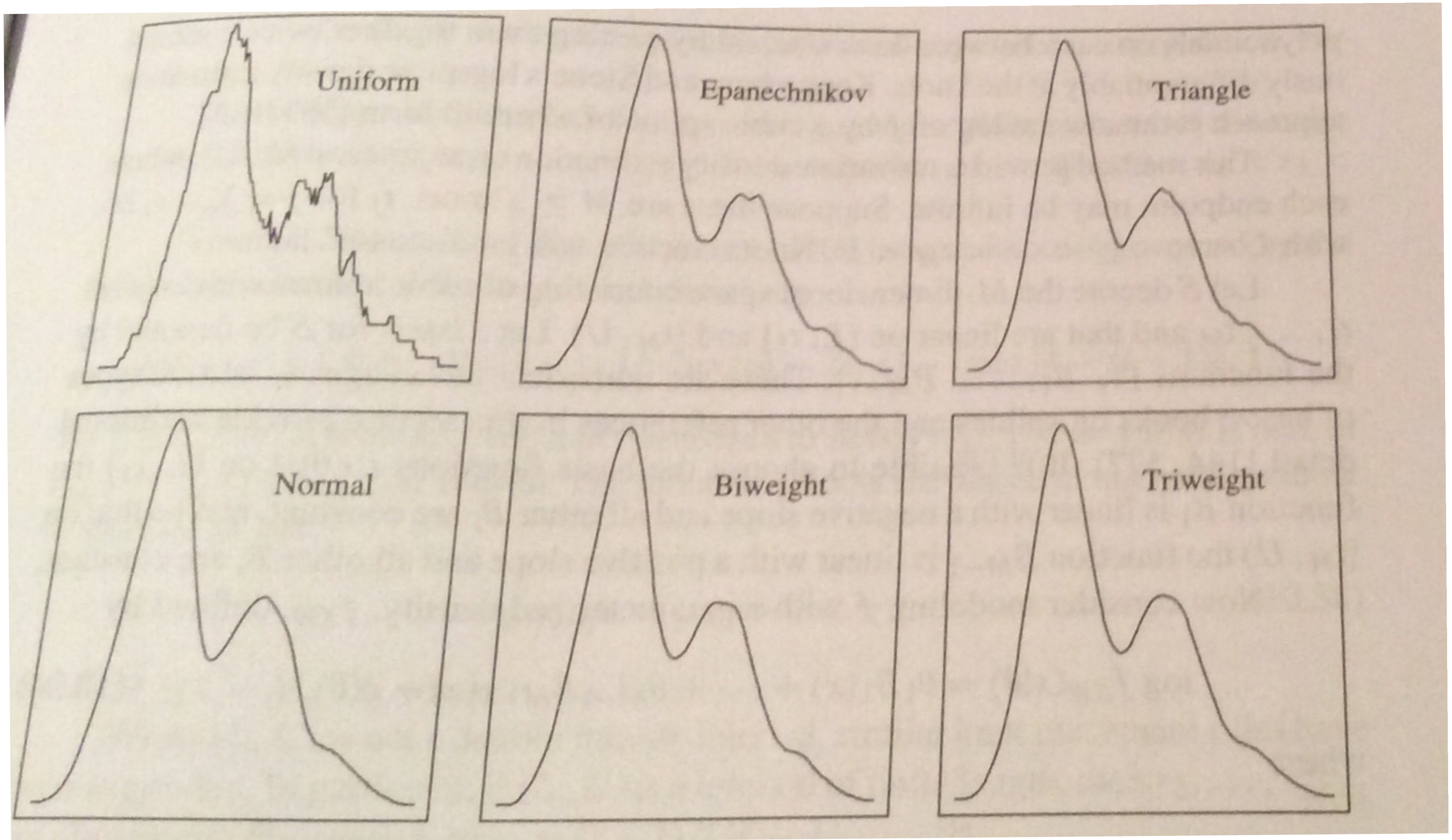
Choice of Kernel

$$\hat{f}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

Epanechnikov showed that if we minimize the AMISE w.r.t. K we obtain a minimum when

$$K^*(z) = \begin{cases} \frac{3}{4}(1 - z^2) & \text{if } |z| < 1 \\ 0 & \text{o.w} \end{cases}$$

Kernel Choice Example



$h = .69$

Possible Kernels

Name	$K(z)$	$R(K)$	$\delta(K)$	RE
Normal	$\exp\{-z^2/2\}/\sqrt{2\pi}$	$1/(2\sqrt{\pi})$	$(1/(2\sqrt{\pi}))^{1/5}$	1.051
Uniform	$\frac{1}{2}$	$\frac{1}{2}$	$\left(\frac{9}{2}\right)^{1/5}$	1.076
Epanechnikov	$\left(\frac{3}{4}\right)(1-z^2)$	$\frac{3}{5}$	$15^{1/5}$	1.000
Triangle	$1-\lvert z\rvert$	$\frac{2}{3}$	$24^{1/5}$	1.014
Biweight	$\left(\frac{15}{16}\right)(1-z^2)^2$	$\frac{5}{7}$	$35^{1/5}$	1.006
Triweight	$\left(\frac{35}{32}\right)(1-z^2)^3$	$\frac{350}{429}$	$\left(\frac{9450}{143}\right)^{1/5}$	1.013

if using a kernel other than the normal you need to multiply $K(z)$ by $1_{\{|x|<1\}}$

Rescaling

Suppose h_K and h_L are opt. (AMISE) for sym, mean 0, pos variance L & K

then

$$\frac{h_K}{h_L} = \frac{\delta(K)}{\delta(L)} \quad \text{where } \delta(K) = \left(\frac{R(K)}{\sigma_K^4}\right)^{1/5}$$

$$h_L = \frac{h_K\,(\delta(L))}{\delta(K)}$$

Multivariate Methods

Suppose we wish to est. f based on iid samples $X_i = (X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ip})^T$ of a p dim random variable.

Method (1): Histogram

Multivariate Methods

Method (2): Kernel Est. Product Kernel

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \prod_{j=1}^{p} \frac{1}{h_j} K\left(\frac{x_j - x_{ij}}{h_j}\right)$$

where $K(z)$ is a univariate Kernel, and h_j are fixed bandwidths for each $j = 1, \ldots, p$

$$h_i = \left(\frac{4}{n(p+2)}\right)^{1/(p+4)} \hat{\sigma}_i$$

$\hookrightarrow$ est of s.d along ith coord.

Multivariate Methods

Method ③: Nearest Neighbor.

$$\hat{f}(x) = \frac{K}{n V_p d_K(x)^p}$$

K-th n.n. est

where

- $d_K(x)$ = Euclid distance from x to the Kth nearest obs. point.
- V_p is the volume of the unit sphere in p dim. $V_p = \pi^{p/2} / \Gamma(p/2+1)$
- p - dim of the data

Problems estimating in higher dimensions

ex/ Want to est. p-variate normal w/optimal rel. m.s.e = .0289

p	n
1	30
2	180
3	806
5	17,400
10	112,000,000
15	2,190,000,000,000