Johns Hopkins Engineering

625.464 Computational Statistics

Kernel Estimators : Choice of Bandwidth part 1

Module 11 Lecture 11C

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Kernel Estimators

to estimate f from $X_1, \dots, X_n \sim f$

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

- K is the kernel function
- h is the bandwidth

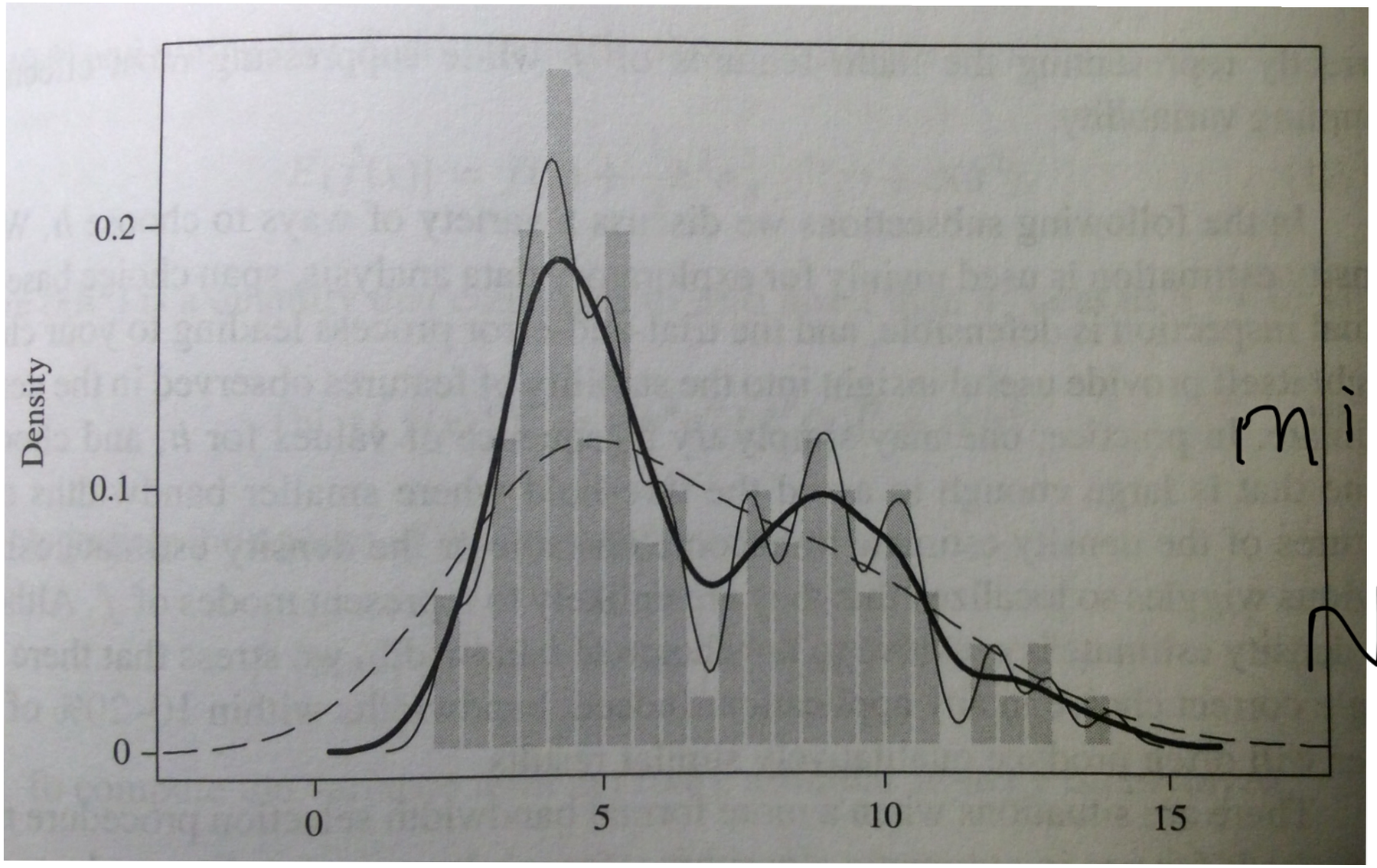
Choice of Bandwidth h

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

- h is too small $\Rightarrow$ $\hat{f}$ wiggly / false modes and high variance.

- h is too big $\Rightarrow$ lose features / higher bias.

Bandwidth Example



mixture of $N(4,1)$ and $N(9,4)$

$h = 1.875$ (– – –)

$h = .3$ (——)

$h = .625$ (▬▬)

How to choose h?

$$MISE(\hat{f}) = IV(\hat{f}) + ISB(\hat{f})$$

$$O(n^{-4/5}) = AMISE(\hat{f}) + \text{error}$$

cond: $h \to 0$ as $n \to \infty$

So we want to minimize the $AMISE(\hat{f})$.

If K is a symm. cont prob density function w/ mean 0 and var $0 < \sigma_K^2 < \infty$, then the AMISE is minimized when

$$h = \left(\frac{R(K)}{n \sigma_K^2 R(f'')} \right)^{1/5}$$

$O(n^{-1/5})$

where $R(g) = \int g^2(z)\,dz$ is the roughness of g.

Method 1: Cross Validation

- think of $\hat{f}$ as a function of h
- want to optimize quality $Q(h)$
- using $x_1, \dots, x_n$ to find $\hat{f}$ and to calculate $\hat{Q}(h)$ can cause overfitting
- instead to evaluate $\hat{Q}$ at x_i, we omit x_i when gen. the est.

$$\hat{f}_{-i}(x_i) = \frac{1}{h(n-1)} \sum_{j \neq i} \left(\frac{x_i - x_j}{h} \right)$$

Method 1 : Cross Validation

Options for bandwidth selection

① $\widehat{Q}(h)$ is the pseudo-likelihood function

$$PL(h) = \prod_{i=1}^{n} \hat{f}_{-i}(X_i)$$

maximize w.r.t bandwidth h

② Unbiased Cross validation (UCV)

Minimize $ISE(h) = \int \hat{f}^2(x)dx - 2E[\hat{f}(x)] + \int f(x)dx$

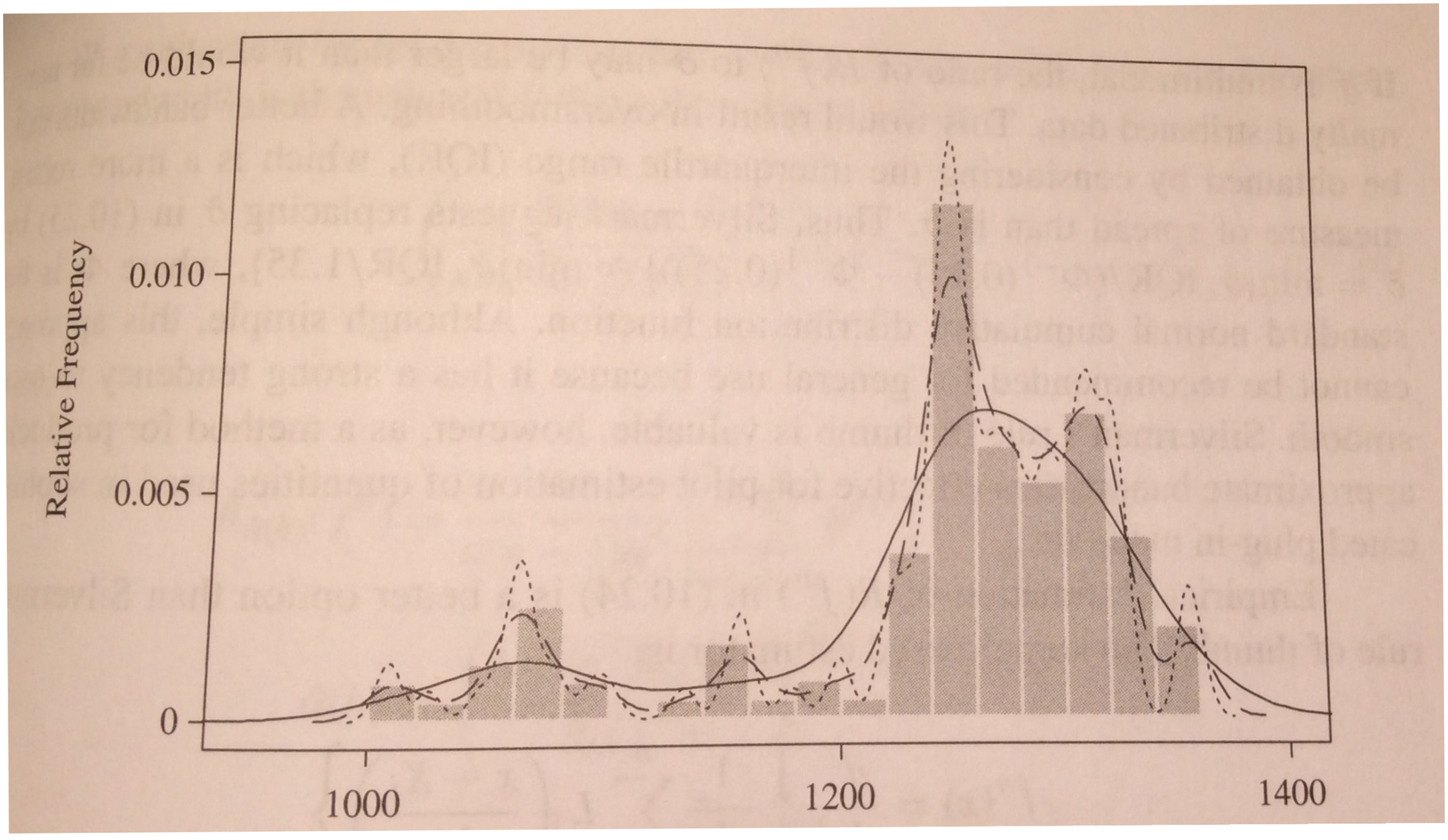
$$= R(\hat{f}) - 2E[\hat{f}(x)] + R(f)$$

$$\approx R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} f_i(x_i) + R(f)$$

constant

Choose h to minimize $UCV(h) = R(\hat{f}) - \frac{2}{n}\sum_{i=1}^{n} f_{-i}(X_i)$

Cross Validation Method Example



— — — — $PL(h)$ $h = 9.75$

..... $UCV(h)$ $h = 5.08$

——— $h = 26.51$