Johns Hopkins Engineering

625.464 Computational Statistics

Introduction to Density Estimation
Orthogonal Series and Histogram Estimators

Module 11 Lecture 11A

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Density Estimation

$X_1, \dots X_n$ } iid obs from density f on D and we

need $\hat{f}$ with

- $\hat{f}(x) \geq 0 \ \forall x \in D$
- $\int_D \hat{f}(x)\,dx = 1$

hoping to find $\hat{f}$ with:

- small error (MSE)
- $E[\hat{f}_n(x)] \longrightarrow f(x) \ \forall x \in D$ as $n \to \infty$

If f is a Parametric Density...

$f_{x|\theta}$

- MLE
- MOM
- logspline
- fitting by matching quantiles
- mixtures

We will assume not parametric

Nonparametric Density Estimation

1. Orthoganal Series Est.
2. Histogram Estimators
3. Kernel Estimators

Orthogonal Series Estimators

Recall,

$$\hat{f}(x) = \frac{1}{n} \sum_{K=0}^{i} \sum_{i=1}^{n} g_K(x_i) g_K(x)$$

where g_i is an orthoganal series.

① # of terms has a major effect and more is not necess. better.

② $\hat{f}$ may not be smooth & may have infinite variance

③ Convergence rate (to f) is ind of dim $\therefore$ may be a good candidate for multivariate problems.

④ Most commonly Fourier & Hermite series used.

Histograms Estimators

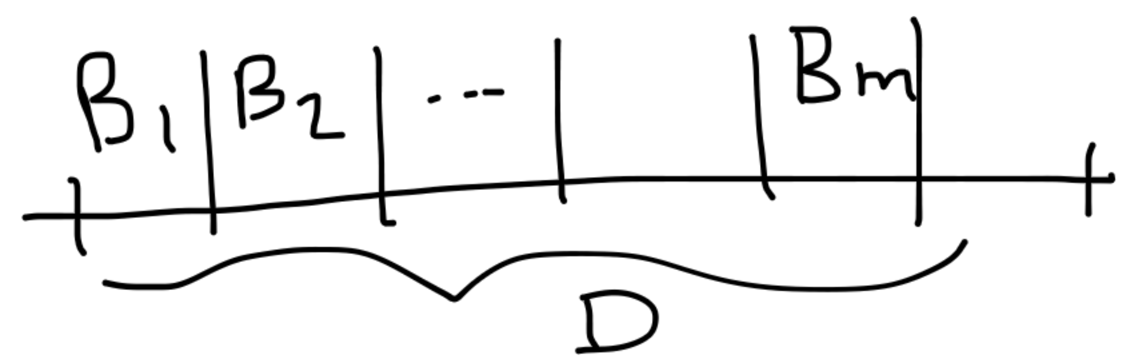
A histogram is a piecewise constant density estimator.

What is $\hat{f}(x)$? Consider how we construct the histogram.

- Assume the support D is finite.
- Construct a fixed partition of D using m nonoverlapping bins B_k

ie $B_j \cap B_i = \emptyset \ \forall j \neq i$ and

$$D = \bigcup_{k=1}^{m} B_k$$



Histogram Estimators

D is partitioned into m bins B_K.

- Let V_K be the volume of bin B_K.
  - one-dim just the length
  - often equal.
- Let n_K be the # of obs in B_K

$$n_K = \sum_{i=1}^{n} I(X_i \in B_K)$$

- The proportion of obs in B_K is $\hat{P}_K = \frac{n_K}{n}$
- The probability content of the bin is

$$P_K = \int_{B_K} f(u)\,du$$

Histogram Estimators

The histogram estimator of f

$$\hat{f}_n(x) = \begin{cases} \hat{p}_1/v_1 & x \in B_1 \\ \hat{p}_2/v_2 & x \in B_2 \\ \vdots & \vdots \\ \hat{p}_m/v_m & x \in B_m \end{cases}$$

$$\hat{f}_n(x) = \sum_{k=1}^{m} \frac{\hat{p}_k}{v_k} I(x \in B_k) = \sum_{k=1}^{m} \frac{n_k}{n v_k} I(x \in B_k)$$

Comments on Histogram Estimators

① $\hat{f}(x) \geq 0 \quad \forall x \in D$

② $\int_D \hat{f}(x)\,dx = \sum_{K=1}^{m} \frac{\hat{p}_K}{V_K} \cdot V_K = \sum_{K=1}^{m} \frac{n_K}{n\cancel{V_K}} \cdot \cancel{V_K} = \frac{1}{n} \sum_{K=1}^{m} n_K$

$= 1$

③ $E[\hat{f}(x)] = \frac{p_K}{V_K}$ for $x \in B_K$

④ $Var[\hat{f}(x)] = \frac{p_K(1-p_K)}{n\,V_K^2}$ for $x \in B_K$

⑤ Under certain cond. you can bound var/bias mse...

⑥ Can easily be extended to multivariate case
- not limited to boxes
- problem becomes choice of size & # of bins