Lecture 5

Last time: Density estimation

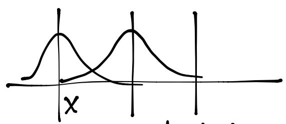
Data $x_1, \dots, x_n$ (one dimensional)

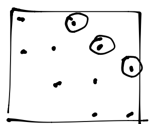
$$\hat{f}_h(x) = \underbrace{\frac{1}{nh}}_{\text{bandwidth} \to \text{controls smoothness}} \sum_{i=1}^n \underbrace{K\left(\frac{x - x_i}{h}\right)}_{\text{kernel}}$$

$K(x) \geq 0$, $K(x) = K(-x)$

$$\int_{-\infty}^{\infty} K(x)\,dx = 1$$

Extension to bivariate and higher





- bivariate kernel $K(x_1, x_2) = K(\underset{\sim}{x})$

$K(\underset{\sim}{x}) \geq 0$ $\quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(\underset{\sim}{x})\,d\underset{\sim}{x} = 1$

e.g. $K(\underset{\sim}{x}) = K_1(x_1) K_2(x_2)$

univariate kernels

Given data $\underset{\sim}{x}_1 = \begin{pmatrix} x_{11} \\ x_{21} \end{pmatrix}$, $\underset{\sim}{x}_2 = \begin{pmatrix} x_{12} \\ x_{22} \end{pmatrix}$, $\underset{\sim}{x}_i = \begin{pmatrix} x_{1i} \\ x_{2i} \end{pmatrix}$

define $\hat{f}_H(\underset{\sim}{x}) = \frac{1}{n|H|} \sum_{i=1}^n K(H^{-1}(\underset{\sim}{x} - \underset{\sim}{x}_i))$

H: bandwidth matrix

In practice, $K(\underset{\sim}{x}) = K(x_1, x_2) = K_1(x_1) K_2(x_2)$

e.g. K_1, K_2 Gaussian densities

$H = \begin{pmatrix} h_1 & 0 \\ 0 & h_2 \end{pmatrix}$ or $H = \begin{pmatrix} h & 0 \\ 0 & h \end{pmatrix}$

$$\hat{f}_H(\underset{\sim}{x}) = \frac{1}{n h_1 h_2} \sum_{i=1}^n K_1\left(\frac{x_1 - x_{1i}}{h_1}\right) K_2\left(\frac{x_2 - x_{2i}}{h_2}\right)$$

- extends trivially to $p \geq 3$
- p-variate kernel $K(\underset{\sim}{x}) = K(x_1, \dots, x_p)$

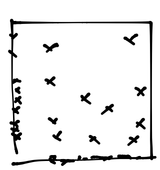
$K(\underset{\sim}{x}) \geq 0$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} K(\underset{\sim}{x})\,d\underset{\sim}{x} = 1$$

$$\hat{f}_H(\underset{\sim}{x}) = \frac{1}{n|H|} \sum_{i=1}^n K(H^{-1}(\underset{\sim}{x} - \underset{\sim}{x}_i))$$

H: $p \times p$ bandwidth matrix

What can go wrong? (As $p \uparrow$)
- choice of bandwidth parameters become more complicated
- "curse of dimensionality"



- as p increases, points become more sparse within p dimensional space

Example: Sps iid r.v. $\underset{\sim}{X}_1, \underset{\sim}{X}_2, \cdots, \underset{\sim}{X}_p$ with density $f(\underset{\sim}{x}) = f(x_1, \cdots, x_p)$
Define $A = [x_1 - h, x_1 + h] \times [x_2 - h, x_2 + h] \times \cdots \times [x_p - h, x_p + h]$
$=$ cube in $\mathbb{R}^p$ with lengh of edges $= 2h$

$N = N(A) = \sum_{i=1}^n I(\underset{\sim}{X}_i \in A) = \#$ of observations in cube $A \sim Bin(n, P(\underset{\sim}{X}_i \in A))$
where if h is small, $P(\underset{\sim}{X}_i \in A) \doteq \underbrace{(2h)^p}_{\text{volume of cube}} \underbrace{f(\underset{\sim}{x})}_{\text{density}}$

So $E[N(A)] \doteq n f(\underset{\sim}{x}) (2h)^p \longrightarrow N(A)$ becomes smaller as p increases

Implications

true density function

Look at $MSE[\hat{f}_H(\underset{\sim}{x})] = E[(\hat{f}_H(\underset{\sim}{x}) - f(\underset{\sim}{x}))^2]$
$= Var[\hat{f}_H(\underset{\sim}{x})] + \underbrace{[E[\hat{f}_H(\underset{\sim}{x})] - f(\underset{\sim}{x})]^2}_{\text{bias}}$

For a given density $\hat{f}(\underset{\sim}{x})$, can find bandwidth parameters to minimize MSE.
Optimal $MSE(\hat{f}_H(\underset{\sim}{x})) = \dfrac{C(f)}{n^{4/(4+p)}}$ $\longrightarrow$ some function of f over a constant

e.g. $n = 100$, $p = 1$, $n^{-4/(4+p)} = n^{-4/5} = 0.025$
$p = 10$, $n^{-4/(4+p)} = n^{-4/14} = 0.268$
$n = 10000$, $p = 10$, $n^{-4(4+p)} = n^{-4/14} = 0.037$
Bottom line: Try to avoid kernel density estimation for large p!

What can you do if p is large?
Projection pursuit — simplistic formulation
Idea: $\underset{\sim}{X} \sim f(\underset{\sim}{x}) \longrightarrow$ suppose there are projections/vectors $\underset{\sim}{a}_1, \cdots, \underset{\sim}{a}_p$
s.t. $\underset{\sim}{a}_1^T \underset{\sim}{X}, \underset{\sim}{a}_2^T \underset{\sim}{X}, \cdots, \underset{\sim}{a}_p^T \underset{\sim}{X}$ are independent with densities $f_1, \cdots, f_p$
Then $f(\underset{\sim}{x}) = |A| \prod_{k=1}^p f_k(\underset{\sim}{a}_k^T \underset{\sim}{x})$ where $A = \begin{pmatrix} \underset{\sim}{a}_1^T \\ \vdots \\ \underset{\sim}{a}_p^T \end{pmatrix}$
$\parallel$
$f(x_1, \cdots, x_p)$

Program: Given data $\underset{\sim}{x}_1, \cdots, \underset{\sim}{x}_n$
① Find projections $\hat{\underset{\sim}{a}}_1, \cdots, \hat{\underset{\sim}{a}}_p$ so that $\{\underset{\sim}{a}_1^T \underset{\sim}{x}_i\}, \cdots, \{\underset{\sim}{a}_p^T \underset{\sim}{x}_i\}$ are independent
② Estimate $f_1, \cdots, f_p$ using the projections

$\hat{\underset{\sim}{a}}_1^T \underset{\sim}{x}_i$ $\qquad$ $\hat{\underset{\sim}{a}}_p^T \underset{\sim}{x}_i$

Issues: ① Independence assumption is very strong!
② How to measure independence (lack of dependence) between $\{\hat{\underset{\sim}{a}}_1^T \underset{\sim}{x}_i\}, \cdots \{\hat{\underset{\sim}{a}}_p^T \underset{\sim}{x}_i\}$

- Projection pursuit density estimation
  - more complicated
  - $\hat{f}_m(\underset{\sim}{x}) = \hat{f}_0(\underset{\sim}{x}) \prod_{k=0}^m \hat{P}_k(\hat{\underset{\sim}{a}}_k^T \underset{\sim}{x})$
$= \hat{f}_{m-1}(\underset{\sim}{x}) \hat{P}_m(\hat{\underset{\sim}{a}}_k^T \underset{\sim}{x})$ ← new projection

new adjustment