NMST434 Exercise 12 - nonparametric regression and density estimation

Kernel ~~regression~~ density estimation

$X_1 .. X_n \overset{iid}{\sim} f(x)$ estimate $f(x)$ by $\hat f_n(x) = \frac{1}{n h_n}\sum_{i=1}^{n} k\left(\frac{x-X_i}{h_n}\right)$

for $k: \mathbb{R}\to[0,\infty)$ a symmetric density, $h_n \xrightarrow[n\to\infty]{} 0$, $n h_n \xrightarrow[n\to\infty]{} \infty$.

k kernel, h_n bandwidth bias $\hat f_n \to 0$ var $\hat f_n \to 0$

By theorem 17 then $\hat f_n(x) \xrightarrow{P} f(x)$ for all x. (under regularity assumptions)

In R: density function

kernels are taken in the form as in Remark 28. $\tilde k(x) = \sqrt{\mu_{2k}}\, k(\sqrt{\mu_{2k}}\, x)$

for $\mu_{2k} = \int y^2 k(y)\,dy = var(k)$

- for $k = \frac{1}{2} I[t\in[-1,1]]$ $\mu_{2k} = 1/3 \Rightarrow$ (in R) $\tilde k(x) = \frac{1}{\sqrt{3}\cdot 2} I\left[\frac{1}{\sqrt{3}} t \in [-1,1]\right]$

var Unif$[-1,1] = 1/3 \Rightarrow$ take $\tilde k \sim$ Unif$[-1,1]\cdot\sqrt{3}$

$= \frac{1}{2\sqrt{3}} I[t\in[-\sqrt{3},\sqrt{3}]]$ so that $\mu_{2\tilde k} = 1$

Bandwidth selection. usually based on asymptotic expansions

nrd - normal reference - assume Gaussianity of f Section 9.2.1

$h_n = 1{,}06\, n^{-1/5}\sigma$ estimated by $h_n = 1{,}06\, n^{-1/5}\min\left\{S_n, \frac{IQR_n}{1{,}34}\right\}$

nrd0 - should have a bit better robustness properties default choice

$h_n = 0{,}9\, n^{-1/5}\sigma$

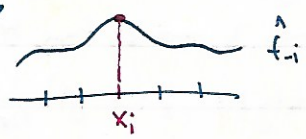
ucv - (unbiased CV) choose h that minimizes an unbiased estimator of

$MISE(\hat f_n) = \int E[\hat f_n(x) - f(x)]^2\,dx$ given by (Section 9.2.2) - up to a const.

$\mathcal{L}(h) = \int[\hat f_n(x)]^2\,dx - \frac{2}{n}\sum_{i=1}^{n} \hat f_{-i}(X_i)$ with $\hat f_{-i}$ the estimator without X_i

overall level of $\hat f_n$ smaller

$\hat f_{-i}(X_i)$ should be higher



bcv. - biased CV minimize the asymptotic ~~expression for~~ approximation to $MISE(\hat f_n)$ (Sect 9.3)

in that we have $R(f'') = \int[f''(x)]^2\,dx$ - the overall curvature of f.

estimator $R(\hat f_n'')$ overestimates $R(f'')$ $\rightarrow$ expansion leads to $R(\hat f'') - \frac{R(k'')}{n h_n^5}$

Boundary effects and mirror-reflection: if $X_i \geq 0$ a.s. the estimator $\hat f_n(x)$ does not take into account the support of X, i.e. $[0,\infty)$. Define

alleviates the

$\hat f_{MR,n}(x) := \begin{cases} \hat f_n(x) + \hat f_n(-x) & \text{if } x \geq 0 \end{cases}$

as. optimal bw (global) - Section 9.2 $h_n^{(opt)} = n^{-1/5}\left[\frac{R(k)}{R(f'')\,\mu_{2k}^2}\right]^{1/5}$

R script

Mathematics script

Multivariate kernel density estimation. $\binom{X_1}{Y_1} \dots \binom{X_m}{Y_m} \sim f_{X,Y}$

$$\hat f_m(x,y) := \frac{1}{m} \sum_{i=1}^{m} \frac{1}{|H|^{1/2}}\, k\left(H^{-1/2}\left[(x,y)^T - (x_i,y_i)^T\right]\right)$$

for k a bivariate symmetric density and H a symmetric, pos. def matrix

Bias reduction and higher order kernels. In the usual setting

$$E\hat f_m(x) = E\,\frac{1}{h} k\left(\frac{x-X}{h}\right) = \int_{\mathbb{R}} \frac{1}{h} k\left(\frac{x-y}{h}\right) f(y)\,dy = \int_{\mathbb{R}} L(x-y) f(y)\,dy$$

$L(\cdot) = \frac{1}{h} k\left(\frac{\cdot}{h}\right)$ is a density of $h.Z$ for $Z \sim k$ — convolution, density of $h.Z + X$

for h fixed and n growing, we estimate the density of $hZ+X$ (bias introducing term)

$$\hat f_m(x) = \frac{1}{m}\sum_{i=1}^{m} \underbrace{\frac{1}{h} k\left(\frac{x-X}{h}\right)}_{\text{iid}}$$

Bias: $E\hat f_m(x) \underset{\frac{x-y}{h}=t}{=} \int_{\mathbb{R}} k(t) f(x-th)\,dt = f(x)\underbrace{\int k(t)\,dt}_{1} - f'(x)h\underbrace{\int t k(t)\,dt}_{0 \text{ for a sym kernel}} +$

$$f(x-th) = f(x) + f'(x)(-ht)^1 + f''(x)\frac{(-ht)^2}{2!} + \dots + f^{(p)}(x)\frac{(-ht)^p}{p!} + R_m$$

$\dots + f^{(p)}(x)\frac{(-h)^p}{p!}\int (t)^p k(t)\,dt$ so typically bias $\hat f_m(x) = O(h^2)$

But if also $\int k(t)\,t^j = 0 \quad \forall j = 1,2,\dots,p-1$

and $\int k(t)\,t^p \neq 0$, we have bias $\hat f_m(x) = O(h^p)$

Though $\int t^2 k(t) = 0 \Rightarrow k \equiv 0$ if k is a symmetric density

and k would have to be chosen s.t. $k<0$ may appear.

$k(y) = \frac{1}{2}(3-y^2)\frac{1}{\sqrt{2\pi}}e^{-y^2/2}$ $p=4$ modification of of Gaussian kernel.

$m(x) = E[Y|X=x]$

Non-parametric regression $\binom{X_1}{Y_1} \dots \binom{X_m}{Y_m}$ iid $Y_i = m(X_i) + \varepsilon_i$ $E[\varepsilon|X] = 0$

estimate m by
$$\hat m(x) = \frac{\sum_{i=1}^{m}\frac{1}{h_m}k\left(\frac{x-X_i}{h_m}\right)\cdot Y_i}{\sum_{j=1}^{m}\frac{1}{h_m}k\left(\frac{x-X_j}{h_m}\right)} = \sum_{i=1}^{m} w_i Y_i$$

$w_i \geq 0$

$\sum w_i = 1$ ← $p=0$

more generally, estimate $m(x)$ by the intercept in regression

$$Y_i \sim \beta_0 + \sum_{j=1}^{p}\beta_j (X_i - x)^j \text{ weighted by } k\left(\frac{x-X_i}{h_m}\right)\frac{1}{h_m}$$

→ local polynomial regression

CV: $\min_h \frac{1}{m}\sum_{i=1}^{m}\left[Y_i - \hat m_{-i}(X_i)\right]^2$