Non-parametric density estimation.

$\hookrightarrow$ histogram

Consider iid r.v.s. $\{X_i\}_{i=1}^{n} \sim \underset{\text{unknown}}{p(x)}$

Assumptions

For simplicity, we consider $X_i \in [0,1]$.

$p(x)$ is non-zero only within $[0,1]$.

$|p'(x)| \leq L$ bounded first derivative

estimator

histogram partition $[0,1]$ into M equispaced bins

$$B_1 = [0, 1/M), \; B_2 = \left[\frac{1}{M}, 2/M\right) \cdots$$

$$B_{M-1} = \left[\frac{M-2}{M}, \frac{M-1}{M}\right), \; B_M = \left[\frac{M-1}{M}, 1\right)$$

For any value $x \in B_\ell$, the density estimate given by the histogram is

$$\hat{p}_n(x) = \frac{\#\text{ of obs. with } B_\ell}{n \text{ binwidth}}$$

$$= M \frac{\sum_{i=1}^{n} I(X_i \in B_\ell)}{n} \qquad I \rightarrow \text{indicator function}$$

$$E[\hat{p}_n(x)] = M \, E\left(\sum_{i=1}^{n} \frac{I(X_i \in B_\ell)}{n}\right)$$

$$= M P(X_i \in B_\ell) \rightarrow \text{using true density}$$

$$= M \int_{(\ell-1)/M}^{\ell/M} p(u)\,du = M\left[F_X(\ell/M) - F_X\left(\frac{\ell-1}{M}\right)\right]$$

$F_X \hookrightarrow$ true CDF

$$E(\hat{p}_n(x)) = \frac{F_X(l/m) - F_X\left(\frac{l-1}{m}\right)}{\frac{l}{m} - \frac{l-1}{m}} = p(x^*)$$

Taylor Series exp

$$f(x) = f(x_0) + (x - x_0) f'(x^*)$$

where $x^* \in [x_0, x]$

$$F_X(l/m) = F_X\left(\frac{l-1}{m}\right) + \left(\frac{l}{m} - \frac{l-1}{m}\right) p(x^*)$$

$p \rightarrow$ derivative of F_X

$$x^* \in \left[\frac{l-1}{m}, \frac{l}{m}\right)$$

$$E(\hat{p}_n(x)) = p(x^*)$$

$$\text{bias}(\hat{p}_n(x)) = E(\hat{p}_n(x)) - p(x)$$

$$= p(x^*) - p(x)$$

$$\left(p(x^*) = p(x) + (x^* - x) p'(x^{**}) \quad x^{**} \in [x^*, x] \right)$$

$$\text{bias}(\hat{p}_n(x)) = p'(x^{**})(x^* - x)$$

$$\leq |p'(x^{**})| \, \underline{|x^* - x|}$$

$$\leq L \times 1/m = L/m$$

As $L \uparrow$, bias $\uparrow$ As $M \uparrow$, bias drops.

$$\mathrm{Var}(\hat{p}_n(x)) = \mathrm{Var}\left[\frac{M}{n}\sum_{i=1}^{n} I(X_i \in B_l)\right]$$

$$= \frac{M^2}{n^2}\sum_{i=1}^{n} \mathrm{Var}\left[I(X_i \in B_l)\right]$$ due to indep of $\{X_i\}_{i=1}^{n}$

$\left(I(X_i \in B_l) \longrightarrow \text{Bernoulli R.V. with parameter } \underline{P(X_i \in B_l)} \right)$

$$= \frac{M^2}{\cancel{n^2}} \times \cancel{n} \times \underline{P(X_i \in B_l)}\,(1 - P(X_i \in B_l))$$

$$P(X_i \in B_l) = p(x^*)/M$$

$$\mathrm{Var}(\hat{p}_n(x)) = \frac{M^2}{n}\,\frac{p(x^*)}{M}\left(1 - \frac{p(x^*)}{M}\right)$$

$$\leq \frac{M}{n}\,p(x^*) + \frac{\cancel{M^2}}{n\,\cancel{M^2}}\,p^2(x^*)$$

$$= \frac{M}{n}\,p(x^*) + \frac{p^2(x^*)}{n}$$

Var. increases with M and drops with n.

$$\underset{(\hat{p}_n(x))}{\mathrm{MSE}} \leq \mathrm{bias}^2 + \mathrm{var}$$

$$= \frac{L^2}{M^2} + \frac{M}{n}\,p(x^*) + \frac{p^2(x^*)}{n}.$$

$$\partial \mathrm{MSE}/\partial M = \frac{L^2(-2)}{M^3} + \frac{p(x^*)}{n} = 0$$

$$\frac{p(x^*)}{n} = \frac{2L^2}{m^3}$$

$$M = \left(\frac{2L^2}{p(x^*)} n\right)^{1/3} \longrightarrow O(n^{1/3})$$

$$\text{binwidth} = \frac{1}{m} = O(n^{-1/3})$$

Exact value of M is not computable because $p(x^*)$ is not known.

Plug in the optimal M into upper bound for MSE.

$$MSE_{opt} \le O(n^{-2/3}) + \frac{n^{-1/3}}{n} p(x^*) + p^2(x^*)/n$$

$$= O(n^{-2/3})$$

A hist. approaches the true density at the error rate $O(n^{-2/3})$ in terms of MSE as long as # of bins i.e. M is $O(n^{1/3})$.