Analyzing the bias and variance of KDE

Bias:

$E[\hat{f}_n(x)]$

$$\mathbb{E}[K_h(x,X)] - f(x) \approx \frac{1}{2}\sigma_k^2 h^2 f''(x).$$

Variance:

$$\mathbb{V}[\widehat{f}_n(x)] \approx \frac{f(x)\int K^2(x)\,dx}{n\,h_n}.$$

Risk:

20.14 Theorem. *Under weak assumptions on f and K,*

$$R(f,\widehat{f}_n) \approx \frac{1}{4}\sigma_K^4 h^4 \int (f''(x))^2 + \frac{\int K^2(x)dx}{nh} \tag{20.22}$$

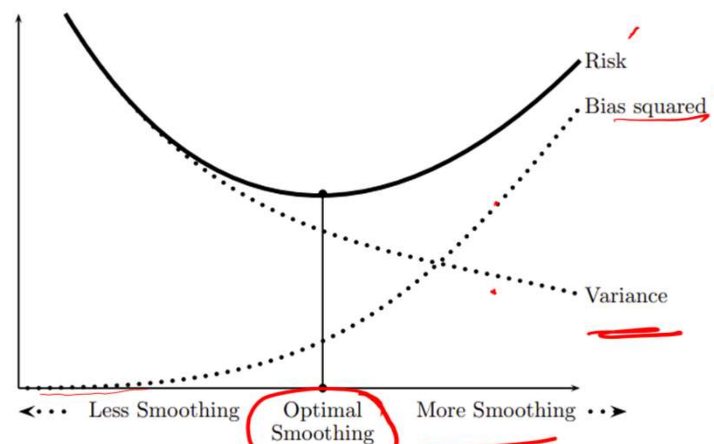
where $\sigma_K^2 = \int x^2 K(x)dx$. The optimal bandwidth is

$$h^* = \frac{c_1^{-2/5} c_2^{1/5} c_3^{-1/5}}{n^{1/5}} \tag{20.23}$$

where $c_1 = \int x^2 K(x)dx$, $c_2 = \int K(x)^2 dx$ and $c_3 = \int (f''(x))^2 dx$. With this choice of bandwidth,

$$\boxed{R(f,\widehat{f}_n) \approx \frac{c_4}{n^{4/5}}}$$

for some constant $c_4 > 0$.



h

h_{opt}

Proof for Bias

$E[\hat{f}_n(x)] - f(x)$

$$E[\hat{f}_n(x_0)] = \int_{x_1} \cdots \int_{x_n} \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_0 - X_i}{h}\right) f_1(x_1) \cdots f(x_n)\, dx_1 \cdots dx_n$$

$[x_1, \ldots, x_i \text{ are i.i.d}]$

$$= \frac{n}{n} \int_x K\left(\frac{x_0 - X}{h}\right) f(x) \frac{dx}{h}$$

Let $u = \dfrac{X - x_0}{h}$ h is small

$$= \int_u K(u)\, \underbrace{f(x_0 + uh)}\, du$$

$$= \int K(u)\left[f(x_0) + uhf'(x_0) + \frac{u^2h^2}{2} f''(x_0) + O(h^3)\right]$$

$$= f(x_0) \underbrace{\int_u K(u)\,du}_{=1} + hf'(x_0) \underbrace{\int_u uK(u)\,du}_{=0} + \frac{h^2}{2} f''(x_0) \underbrace{\int_u u^2K(u)\,du}_{\mu_K} + O(h^3)$$

$$f(x_0) + \frac{h^2}{2} f''(x_0)\mu_K$$

where $\mu_K = \int_u u^2 K(u)\,du$

Proof for Variance

$$\text{Variance}(\hat{f}_n(x_0)) = Var\left(\frac{1}{nh}\sum_i k\left(\frac{X_i - x_0}{h}\right)\right)$$

$$= \frac{1}{nh^2} Var_x\left(k\left(\frac{x-x_0}{h}\right)\right) = E[k^2] - \underbrace{E[k]^2}$$

$$\leq \frac{1}{nh^2}\int_{-\infty} k\left(\frac{x-x_0}{h}\right)^2 f(x)\,dx$$

$\vdots$

$$\underline{\frac{1}{nh} f(x_0)\sigma_k^2} + O(h^2) \text{ where } \sigma_k^2 = \int_u k^2(u)\,du$$

$$\text{Risk}(\hat{f}_n(x_0), f(x)) = \text{Bias}(\hat{f}_n(x_0))^2 + \text{Var}(\hat{f}_n(x_0))$$

$$\simeq \frac{1}{4} h^4 f''(x_0)^2 \mu_k^2 + \frac{1}{nh} f(x_0) \sigma_k^2$$

$\boxed{\text{optimal } \tilde{h}}$

$$\min_{\underline{h}} \Big[\underbrace{\frac{1}{4} h^4 f''(x_0)^2 \mu_k^2 + \frac{1}{nh} f(x_0) \sigma_k^2}_{\tilde{R}} \Big]$$

$$\frac{\partial \tilde{R}}{\partial h} = h^3 f''(x_0)^2 \mu_k^2 - \frac{1}{nh^2} f(x_0) \sigma_k^2$$

h_{opt} :
$$h_{opt}^3 f''(x_0)^2 \mu_k^2 - \frac{1}{n h_{opt}^2} f(x_0) \sigma_k^2 = 0$$

$$\Rightarrow \quad h_{opt}(x_0) = \left[\frac{f(x_0) \sigma_k^2}{n f''(x_0)^2 \mu_k^2} \right]^{1/5} \longleftarrow$$

Evaluate $\tilde{R}$ at h_{opt} $\equiv \dfrac{C}{n^{4/5}} \equiv O(n^{-4/5})$

fn of gold $f(x)$ which is unknown.

Contrast with square error of maximum likelihood estimate of mean parameter

eg: μ of $N(\mu, \sigma^2)$

$$\text{Risk}(\hat{\mu}_n, \mu) \equiv \frac{\sigma^2}{n} \equiv O(n^{-1})$$

Convergence analysis of histogram density estimator

$$\text{Bias}[\hat{f}_n(x_0)] \leq |f'(x)|\left(\frac{1}{M}\right) = |f'(x)|h \qquad h = \frac{1}{M}$$

$$\text{Var}[\hat{f}_n(x_0)] = \frac{f(x)}{nh} - \frac{f(x)^2}{n} \qquad L = \max_x |f'(x)|$$

$$\tilde{R} = h^2L^2 + \frac{f(x)}{nh} - \frac{f(x)^2}{n}$$

Find h_{opt}

$$\frac{\partial \tilde{R}}{\partial h} = 2hL^2 - \frac{f(x)}{nh^2} \Rightarrow h_{opt} = \left(\frac{f(x)}{2nL^2}\right)^{1/3}$$

Evaluate $\tilde{R}$ at h_{opt}:

$$\tilde{R}_{opt} = O(n^{-2/3})$$

Risk reduces at the rate $n^{2/3}$.

Comparing Histogram and KDE

- Risk of histogram reduces at rate $O(n^{-2/3})$

- Risk of KDE reduces at rate $O(n^{-4/5})$

Summary of estimation methods

$\underline{\underline{D}}$

1) Parametric methods.
   Assumed a fixed functional form of density $f(x;\theta)$
   Estimated θ
   - Maximum likelihood estimation
   - Bayesian estimation.
2) Empirical CDF.
3) Non-parametric density
   - Histograms
   - KDE.