HOMEWORK ASSIGNMENT 2

1 Data-driven bandwidth selection for histograms

We consider the task of unsupervised learning with $X_1, \dots, X_n \in [0,1]$ iid distributed according to P. It is assumed that P has a density with respect to the Lebesgue measure and this density, denoted by f is unknown. The goal is to estimate f by some estimator $\widehat{f}$ so that the Mean Integrated Squared Error :

$$\mathrm{MISE}_f(\widehat{f}) = \mathbf{E}\Big[\|\widehat{f} - f\|_2^2\Big], \qquad \Big(\|g\|_2^2 \triangleq \int_0^1 g^2(x)\,dx\Big)$$

be small.

We have seen in the lectures that the histogram estimator

$$\widehat{f}_{n,h}(x) = \frac{1}{h}\sum_{k=1}^{K} \widehat{p}_k \mathbb{1}_{C_k}(x), \qquad \text{where} \qquad \widehat{p}_k = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{C_k}(X_i)$$

has a risk equal to

$$\mathrm{MISE}_f(\widehat{f}_{n,h}) = \|f\|_2^2 + \frac{1}{nh} - \frac{n+1}{nh}\sum_{k=1}^{K} p_k^2, \qquad \text{where} \qquad p_k = \int_{C_k} f(x)\,dx.$$

Furthermore, we have shown that

$$\widehat{J}_n(h) = \frac{2}{(n-1)h} - \frac{n+1}{(n-1)h}\sum_{k=1}^{K} \widehat{p}_k^2$$

is an unbiased estimator of the quantity $J(h) = \mathrm{MISE}_f(\widehat{f}_{n,h}) - \|f\|_2^2$.

Based on this, we have proposed the following rule for selecting among M candidates $h_1, \dots, h_M$ of possible bandwidths :

$$\widehat{h} \in \arg\min_{h_1,\dots,h_M} \widehat{J}_n(h).$$

We have demonstrated that the resulting *adaptive* estimator $\widehat{f}_{n,\widehat{h}}$ satisfies the oracle inequality

$$\mathrm{MISE}_f(\widehat{f}_{n,\widehat{h}}) \le \min_{h\in h_1,\dots,h_M} \mathrm{MISE}_f(\widehat{f}_{n,h}) + 2\sqrt{M}\sup_{h>0} V_n(h)^{1/2},$$

where $V_n(h)$ is the variance of the zero-mean random variable

$$\xi_n(h) = \|\widehat{f}_{n,h} - f\|_2^2 - \|f\|_2^2 - \widehat{J}_n(h).$$

The aim of the present exercise is to find a relatively tight upper bound on $V_n(h)$. **In what follows, we assume that f is bounded and denote by $\|f\|_\infty$ the supremum of f.**

1. Show that $\xi_n(h) = \xi_n^{(1)}(h) + \xi_n^{(2)}(h)$, where

$$\xi_n^{(1)}(h) = \frac{2n}{(n-1)h}\sum_{k=1}^{K} (\widehat{p}_k^2 - \mathbf{E}[\widehat{p}_k^2]),$$

$$\xi_n^{(2)}(h) = \frac{2}{h}\sum_{k=1}^{K} p_k(p_k - \widehat{p}_k).$$

2. Let us set
$$Z_i = \frac{2}{h}\sum_{k=1}^{K} p_k(p_k - \mathbb{1}_{C_k}(X_i)), \qquad i = 1, \dots, n.$$
   (a) Check that $\xi_n^{(2)}(h) = (1/n)\sum_{i=1}^n Z_i$. What can you say about the dependency between the random variables $Z_1, \dots, Z_n$?
   (b) Show that the variance of Z_1 is upper-bounded by
$$\frac{4}{h^2}\sum_{k=1}^{K} p_k^3(1-p_k).$$
   (c) Deduce from previous questions an upper bound of the form
$$\mathrm{Var}[\xi_n^{(2)}(h)] \leq \frac{C\|f\|_\infty^2}{n}$$
   with a constant C to be made precise.
3. We study now the first term $\xi_n^{(1)}(h)$.
   (a) Let us denote $T = \{(i,j) \in \{1,\dots,n\}^2 \text{ s.t. } i < j\}$ and introduce the random variables
$$U_t = \sum_{k=1}^{K}\Big(\mathbb{1}_{C_k}(X_i)\mathbb{1}_{C_k}(X_j) - p_k^2\Big), \qquad \text{if} \qquad t = (i,j).$$
   Show that $\mathbf{E}[U_t] = 0$ for every t and that
$$\xi_n^{(1)}(h) = \frac{4}{nh(n-1)}\sum_{t\in T} U_t.$$
   (b) Let $t = (i,j)$ and $t' = (i',j')$ be elements of T such that all indices (i,i',j,j') are different. What is the value of the covariance between U_t and $U_{t'}$?
   (c) Check that for every $t = (i,j) \in T$:
$$\mathbf{E}\Big[\Big(\sum_{k=1}^{K}\mathbb{1}_{C_k}(X_i)\mathbb{1}_{C_k}(X_j)\Big)^2\Big] = \sum_{k=1}^{K}\mathbf{E}\Big[\Big(\mathbb{1}_{C_k}(X_i)\mathbb{1}_{C_k}(X_j)\Big)^2\Big] = \sum_{k=1}^{K} p_k^2.$$
   Compute the variance of U_t.
   (d) Let us define the set
$$\mathcal{W} = \Big\{(i,j,i',j') \in T\times T \quad \text{s.t.} \quad (i,j) \neq (i',j') \quad \text{and} \quad \{i,j\}\cap\{i',j'\} \neq \varnothing\Big\}.$$
   Check that the cardinality of $\mathcal{W}$ is less then $n(n-1)^2$.
   (e) Show that for every pair $(t,t') \in \mathcal{W}$, it holds that
$$\mathbf{E}[U_t U_{t'}] \leq \sum_{k=1}^{K} p_k^3 \leq h^2\|f\|_\infty^2.$$
   (f) Show that
$$\mathrm{Var}[\xi_n^{(1)}(h)] = \frac{16}{h^2n^2(n-1)^2}\Big\{\sum_{(t,t')\in\mathcal{W}}\mathbf{E}[U_tU_{t'}] + \sum_{t\in T}\mathbf{E}[U_t^2]\Big\}.$$
   Deduce from preceding questions that
$$\mathrm{Var}[\xi_n^{(1)}(h)] \leq \frac{16\|f\|_\infty^2}{n} + \frac{8\|f\|_\infty}{hn(n-1)}$$
4. Show that $\|f\|_\infty \geq 1$. Explain why it is reasonable to assume that $h > 1/n$ and show that
$$V_n(h)^{1/2} = (\mathrm{Var}[\xi_n(h)])^{1/2} \leq \frac{9\|f\|_\infty}{\sqrt{n}}.$$