Exam (Solutions)

Advanced Statistics

Ex 1

1) $\Phi_X(t) = \mathbb{E}[e^{itX}]$

The estimator $\hat{\Phi}_n(t)$ is obtained by replacing X's distribution in the expectation above by the empirical distribution $\hat{F}_n = \frac{1}{n}\sum_{k=1}^{n} \delta_{X_k}$ based on the sample $X_1, \dots, X_n$

$$\mathbb{E}_{X \sim \hat{F}_n}[e^{itX}] = \frac{1}{n}\sum_{k=1}^{n} e^{itX_k} = \hat{\Phi}_n(t)$$

2) Since the X_k's are copies of X, we have by linearity of the expectation:

$$\mathbb{E}[\hat{\Phi}_n(t)] = \frac{1}{n}\sum_{k=1}^{n} \mathbb{E}[e^{itX_k}] = \Phi_X(t),$$

which proves that $\hat{\Phi}_n(t)$ is an unbiased estimator of $\Phi_X(t)$.

As the X_k's are independent, we have:

$$\mathrm{Var}(\hat{\Phi}_n(t)) = \frac{1}{n^2}\sum_{k=1}^{n} \mathrm{Var}(e^{itX_k})$$

$$= \frac{1}{n}\mathrm{Var}(e^{itX}) = \frac{1}{n}(1 - |\Phi_X(t)|^2),$$

3) As $\hat{\Phi}_n(t)$ is unbiased (cf 2)), its quadratic risk coincides with its variance and is of order $O(1/n)$.

Ex 2

1) For $k \in \{1, \dots, m\}$, set

$$p_k = \mathbb{P}\{X \in [\frac{k-1}{m}, \frac{k}{m})\} = \int_{\frac{k-1}{m}}^{\frac{k}{m}} f(x)\,dx .$$

Observe that $\mathbb{E}[\hat{p}_k] = p_k$. The bias of $\hat{f}_{h,n}(x)$ is:

$$\mathbb{E}[\hat{f}_{h,n}(x)] - f(x) = \sum_{k=1}^{m} \left(\frac{1}{h} p_k - f(x)\right) \mathbb{I}\{x \in [\frac{k-1}{m}, \frac{k}{m})\},$$

so that, if, $x \in [\frac{k-1}{m}, \frac{k}{m})$, the bias is equal to:

$$p_k/h - f(x) .$$

If $x \in [\frac{k-1}{m}, \frac{k}{m})$, the variance of $\hat{f}_{h,n}(x) = \frac{\hat{p}_k}{h}$ is $\frac{1}{h^2} \operatorname{Var}(\hat{p}_k) = \frac{1}{nh^2} p_k(1-p_k)$ because the X_i's are independent copies of X.

We have

$$R(h) = \int_{x=0}^{1} \mathbb{E}[(\hat{f}_{h,n}(x) - f(x))^2]\,dx$$

$$= \int_{x=0}^{1} \left(\operatorname{Var}(\hat{f}_{h,n}(x)) + (\mathbb{E}[\hat{f}_{h,n}(x)] - f(x))^2\right) dx$$

$$= \int_{x=0}^{1} \left(\operatorname{Var}(\hat{f}_{h,n}(x)) + \mathbb{E}[\hat{f}_{h,n}(x)]^2 - 2f(x)\mathbb{E}[\hat{f}_{h,n}(x)] + f(x)^2\right) dx$$

$$= \sum_{k=1}^{m} \left(\frac{p_k(1-p_k)}{nh} + \frac{p_k^2}{h} - \frac{2p_k^2}{h}\right) + \|f\|_2^2$$

As $\sum_{k=1}^{m} p_k = +1$, we have:

$$hJ(h) = \frac{1}{n} - \frac{n+1}{n} \sum_{1}^{\hat{}} p_k^2.$$

In addition $\frac{n}{n-1}(\hat{p}_k - \hat{p}_k^2)$ is an unbiased estimator of the sequence $p_k - p_k^2$: in

$$p_k - p_k^2 = \frac{n}{n-1} \mathbb{E}[\hat{p}_k - \hat{p}_k^2].$$ Since $\sum_1 \hat{p}_k = +1$,

we have $1 - \sum p_k^2 = \frac{n}{n-1}\left(1 - \mathbb{E}[\sum \hat{p}_k^2]\right)$

$$\Rightarrow \mathbb{E}[h\hat{J}(h)] = \frac{2}{n-1} - \frac{n+1}{n-1} \mathbb{E}[\sum \hat{p}_k^2]$$

$$= \frac{1}{n} - \frac{n+1}{n} \sum_1 p_k^2 = hJ(h)$$

$\hat{J}(h)$ is an unbiased estimator of $J(h)$.

Ex3

1) Observe that: $\forall k \neq l$, $\int \varphi_k \varphi_l = 0$ and $\int \varphi_k^2 = +1$. The φ_k's form an orthonormal collection. The space V_h being spanned by the φ_k's, the latter thus form an orthonormal basis of V_h.

$$f_h(x) = \sum_1^m \langle f, \varphi_k \rangle \varphi_k(x) = \frac{1}{h} \sum_1^k p_k \mathbb{I}\{x \in [\tfrac{k-1}{m}, \tfrac{k}{m})\}$$

$f_h(x)$ is the expectation of $\hat{f}_{h,n}(x)$

2) Clearly, $\|f - f_h\|_2^2 = \int_0^1 (f(x) - f_h(x))^2 dx$

$$\leq \sup_{x \in [0,1]} (f(x) - f_h(x))^2 \leq \max_k \sup_{x \in [\frac{k-1}{m}, \frac{k}{m})} (f(x) - f_h(x))^2$$

If $x \in [\frac{k-1}{m}, \frac{k}{m})$,

$$f(x) - f_h(x) = \frac{1}{h}\int_{\frac{k-1}{m}}^{\frac{k}{m}} (f(x) - f(t))\,dt. \text{ Hence,}$$

$$|f(x) - f_h(x)| \leq \frac{1}{h}\int_{\frac{k-1}{m}}^{\frac{k}{m}} |f(x) - f(t)|\,dt$$

$$\leq \frac{C}{h} \times \frac{1}{m} \times \left|\frac{k}{m} - \frac{k-1}{m}\right|^{\alpha} = C m^{-\alpha}$$

$$\Rightarrow \|f - f_h\|_2^2 \leq C^2 m^{-2\alpha}$$

3) The integrated quadratic risk is

$$R(h) = \int_0^1 \mathrm{Var}(\hat{f}_{h,n}(x))\,dx + \|f_h - f\|_2^2$$

$$\leq \sum_{k=1}^{m} \frac{p_k(1-p_k)}{nh} + C^2 m^{-2\alpha}$$

$$\leq \frac{\sum_1^m p_k}{nh} + C^2 m^{-2\alpha} = \frac{m}{n} + C^2 m^{-2\alpha}$$

The order of magnitude of the upper bound is minimum for $\frac{m}{n} \sim m^{-2\alpha}$. $\Leftrightarrow$

$m \sim n^{\frac{1}{1+2\alpha}}$. As m is an integer, one may propose $m = \lfloor n^{1/1+2\alpha} \rfloor$ (integer part).

Ex 4

1) As ε_1 is independent from X_1, we almost-surely have:

$$\mathbb{E}[Y_1^2 | X_1] = \mathbb{E}[\sigma^2(X_1)\varepsilon_1^2 | X_1] = \sigma^2(X_1)\,\mathbb{E}[\varepsilon_1^2 | X_1]$$

$$= \sigma^2(X_1)\,\mathbb{E}[\varepsilon_1^2] = \sigma^2(X_1).$$

2) As F is bijective, one may write: $\forall t \in [0,1]$,

$$\mathbb{P}(F(X_1) \leq t) = \mathbb{P}(X_1 \leq F^{-1}(t)) = F(F^{-1}(t)) = t$$

$\Rightarrow F(X_1)$ is uniformly distributed on $[0,1]$.

Since the (X_i, Y_i)'s are i.i.d, we have

$$\mathrm{Var}(\hat{\ell}_h(x)) = \frac{1}{n\,h^2} \times \mathrm{Var}\left(Y_1^2 K\left(\frac{F(X_1)-x}{h}\right)\right)$$

And

$$\mathrm{Var}\left(Y_1^2 K\left(\frac{F(X_1)-x}{h}\right)\right) \leq \mathbb{E}\left[Y_1^4 K^2\left(\frac{F(X_1)-x}{h}\right)\right]$$

$$= \mathbb{E}\left[\sigma^4(X_1)\varepsilon_1^4 \cdot K^2\left(\frac{F(X_1)-x}{h}\right)\right]$$

$$= m_4\,\mathbb{E}\left[\sigma^4(X_1)\,K^2\left(\frac{F(X_1)-x}{h}\right)\right]$$

because ε_1 and X_1 are independent.

$$\Rightarrow \mathrm{Var}\left(Y_1^2 K\left(\frac{F(X_1)-x}{h}\right)\right) \leq m_4 C^2\,\mathbb{E}\left[K^2\left(\frac{F(X_1)-x}{h}\right)\right]$$

$$= C^2 m_4 \int K^2\left(\frac{u-x}{h}\right)du = C^2 m_4 h \int K^2(t)\,dt$$

which proves the desired bound.

3) We have:

$$\mathbb{E}[\hat{\ell}_h(x)] = \frac{1}{h}\mathbb{E}\left[Y_1^2 K\left(\frac{F(X_1)-x}{h}\right)\right]$$

$$= \frac{1}{h}\mathbb{E}\left[\xi_1^2 \sigma^2(X_1) K\left(\frac{F(X_1)-x}{h}\right)\right]$$

$$= \frac{1}{h}\mathbb{E}[\xi_1^2]\,\mathbb{E}\left[\ell(F(X_1)) K\left(\frac{F(X_1)-x}{h}\right)\right]$$

$$= \frac{1}{h}\int_{u=0}^{1} \ell(u) K\left(\frac{u-x}{h}\right) du$$

$$= (\ell * K_h)(x) \quad \text{with} \quad K_h(t) = \frac{1}{h}K\left(\frac{t}{h}\right).$$

4) Reproducing the argument of Proposition 1.35 (Taylor expansion of ℓ at x of order 3), we have:

$$\mathbb{E}[\hat{\ell}_h(x)] - \ell(x) = \frac{h^3}{6}\int \ell'''(x) u^3 K(u)\, du$$

Hence, we can take: $\beta = 3$, $K = \frac{C_K M}{6}$.

5) The pointwise quadratic risk is

$$\mathrm{var}(\hat{\ell}_h(x)) + (\mathbb{E}[\hat{\ell}_h(x)] - \ell(x))^2$$

$$\leq \frac{c^2 m_4 \int K^2(t)dt}{nh} + K^2 h^6$$

Balancing the two terms, one finds $h \sim n^{-1/7}$

When F is known,

$\hat{\ell}_h(F(x))$ is a natural estimator

of $\sigma^2(x) = \ell(F(x))$

When $F(\cdot)$ is unknown, one can

replace it by $\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \le t\}$

(the empirical cdf) and consider

the (plug-in) estimator $\hat{\ell}_h(\hat{F}_n(x))$.