Nonasymptotic statistic - Final exam

December 14, 2015. 9 am - 12 am.

Calculators and documents are not allowed. This examination paper has **two** pages.

Part 1. Parametric statistics (S. Pergamenchtchikov)

1. Simple nonrandom regression (4 points)

1. Give the definition for the simple nonrandom regression model with unknown parameter λ.
2. Show that the least square estimator $\widehat{\lambda}_n$ is optimal in the mean square accuracy sense in the class of all unbiased linear estimators, i.e. for any $n \geq 1$
$$\mathbf{E}\,(\widehat{\lambda}_n - \lambda)^2 \leq \mathbf{E}\,(\widetilde{\lambda}_n - \lambda)^2\,,$$
where $\widetilde{\lambda}_n$ is a linear estimator constructed on the n observations.

2. Sequential estimation (4 points)

Let us $(y_j)_{j\geq 1}$ be the first order autoregressive process defined as
$$y_j = \lambda y_{j-1} + \xi_j\,, \quad y_0 = 0\,.$$
Here λ is an unknown constant parameter, $(\xi_j)_{j\geq 1}$ is i.i.d. sequence of random variables with $\mathbf{E}\,\xi_j = 0$ and $\mathbf{E}\,\xi_j^2 = 1$. We set $\mathcal{F}_0 = \{\emptyset, \Omega\}$ and $\mathcal{F}_j = \sigma\{y_1, \ldots, y_j\}$ for any $j \geq 1$.

Let for some $H > 0$
$$\tau_H = \inf\{n \geq 1\,:\, \sum_{j=1}^{n} y_{j-1}^2 \geq H\}\,.$$

1. Show that $\mathbf{P}(\tau_H < \infty) = 1$ for any $H > 0$.
2. Give the definition for the stopping times. Show that for any $H > 0$ the moment τ_H is the stopping time with respect to $(\mathcal{F}_j)_{j\geq 0}$.
3. Write the sequential estimator for the parameter λ.

3. Tests (2 points)

Let us $(X_j)_{j\geq 1}$ be an i.i.d. sequence of random variables having a density.

1. Write the test problem in the sequential setting.
2. Write the Wald test.

Part 2. Nonparametric statistics (G. Chagny)

In the sequel, we denote by $X_1, \ldots, X_n$ a sample of independent and identically distributed real random variables with unknown density f with respect to the Lebesgue measure on $\mathbb{R}$ and unknown cumulative distribution function F.

Course notion (1 point).

We observe $(X_i)_{i=1,\ldots,n}$ a sample of independent and identically distributed variables with unknown density $f \in L^2(I)$ ($I \subset \mathbb{R}$ an interval). Give the definition of the estimator for f built with the projection method on a subset $S_D \subset L^2(I)$ spanned by an orthonormal basis $\{\varphi_1, \ldots, \varphi_D\}$, (with $\varphi_j : I \to \mathbb{R}$).

Exercise. Bivariate density estimation, with kernel methods (9 points).

We observe n couples of real random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ independent and identically distributed with unknown density $f : \mathbb{R}^2 \to \mathbb{R}$ with respect to the Lebesgue measure on $\mathbb{R}^2$.

For any $h = (h_1, h_2) \in (\mathbb{R}_+^*)^2$, we consider the estimator $\hat{f}_h$ of f defined, for any $(x, y) \in \mathbb{R}^2$, by

$$\hat{f}_h(x, y) = \frac{1}{nh_1h_2} \sum_{i=1}^{n} Q\left(\frac{x - X_i}{h_1}, \frac{y - Y_i}{h_2}\right),$$

where $Q : \mathbb{R}^2 \to \mathbb{R}$ is a kernel function on $\mathbb{R}^2$: it is an integrable function which satisfies $\iint_{\mathbb{R}^2} Q(u, v)dudv = 1$. We assume that

$$\|Q\|_2^2 := \iint_{\mathbb{R}^2} Q^2(u, v)dudv < \infty, \quad \iint_{\mathbb{R}^2} |Q(u, v)||v|dudv < \infty, \quad \iint_{\mathbb{R}^2} Q(u, v)|u|^{1/2}dudv < \infty.$$

We consider the pointwise quadratic risk on $(x_0, y_0) \in \mathbb{R}^2$ for $\hat{f}_h$,

$$R_{(x_0,y_0)}(\hat{f}_h, f) = \mathbb{E}\left[\left(\hat{f}_h(x_0, y_0) - f(x_0, y_0)\right)^2\right].$$

For any $(u, v) \in \mathbb{R}^2$, we introduce the notations

$$Q_h(u, v) = \frac{1}{h_1h_2} Q\left(\frac{u}{h_1}, \frac{v}{h_2}\right).$$

For any two-variables functions f_1 and f_2, we denote by $\star$ the convolution product between f_1 and f_2, as soon as it exits,

$$(f_1 \star f_2)(x, y) = \iint_{\mathbb{R}^2} f_1(x - u, y - v)f_2'u, v)dudv, \quad (x, y) \in \mathbb{R}^2.$$

1. Justify that, if $K : \mathbb{R} \to \mathbb{R}$ is a kernel on $\mathbb{R}$, then the function defined by $Q(u, v) = K(u)K(v)$, $(u, v) \in \mathbb{R}^2$ is a kernel on $\mathbb{R}^2$.
2. Prove that $R_{(x_0,y_0)}(\hat{f}_h, f) = \left(\mathbb{E}\left[\hat{f}_h(x_0, y_0)\right] - f(x_0, y_0)\right)^2 + \mathrm{Var}(\hat{f}_h(x_0, y_0))$.
3. **(a)** Prove that $\mathbb{E}\left[\hat{f}_h(x_0, y_0)\right] = Q_h \star f(x_0, y_0)$.

   (b) Assume that f is bounded by $M > 0$ on $\mathbb{R}^2$. Compute the variance of the estimator, and prove the following upper-bound:
   $$\mathrm{Var}(\hat{f}_h(x_0, y_0)) \leq \frac{M\|Q\|_2^2}{nh_1h_2}.$$
4. We also assume that f satisfies the following property.
   $$\forall((u, v), (u', v')) \in (\mathbb{R}^2)^2, \quad |f(u, v) - f(u', v')| \leq |u - u'|^{1/2} + |v - v'|.$$
   Then, prove that
   $$\left(\mathbb{E}\left[\hat{f}_h(x_0, y_0)\right] - f(x_0, y_0)\right)^2 \leq C(h_1 + h_2^2),$$
   where C is a constant that only depends on Q.
5. **(a)** Let g be the function defined by $g(u, v) = u + v^2 + 1/(nuv)$, $(u, v) \in (\mathbb{R}_+^*)^2$. Justify that g admits on $(\mathbb{R}_+^*)^2$ a global minimum, and compute its value.

   (b) Deduce from the previous questions that the convergence rate of $\hat{f}_h$, is $n^{-1/5}$ for the pointwise quadratic risk.
6. We denote by f_X the marginal density of X_1.

   (a) How can be computed f_X from f.

   (b) Propose an estimator $\hat{f}_X$ of f_X as a function of $\hat{f}_h$ defined above.

   (c) Assume that Q can be written $Q(u, v) = K(u)K(v)$, $(u, v) \in \mathbb{R}^2$, with K a kernel on $\mathbb{R}$. How can $\hat{f}_X$ be written in that case?