NONASYMPTOTIC STATISTIC - FINAL EXAM

December 17, 2015. 1 pm - 4 pm.

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

Part 1. Parametric statistics (S. Pergamenchtchikov)

1. Simple regression (6 points).

1. Give the definition for the simple regression model.
2. Assuming that in the simple regression model the noise distribution is Gaussian. Construct the test to check if $a_1 = 1$ or not with some fixed confidence level $0 < \alpha < 1$.

2. Multiple regression (4 points).

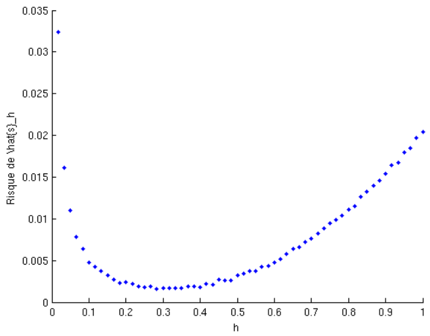
1. Give the definition of the multiple regression model.
2. We consider a multiple regression model of order 3. Construct the least square estimator for the parameters: a_1, a_3 and $a_2 + 2a_3$.

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.

3. Course notions (3 points).

1. Recall the definition of the kernel estimator $\hat{f}_h$ of the density f, associated to a kernel $K : \mathbb{R} \to \mathbb{R}$ and a bandwidth parameter $h > 0$.
2. Let $x_0 \in \mathbb{R}$ and denote by $MSE_{x_0}(\hat{f}_h) = \mathbb{E}[(\hat{f}_h(x_0) - f(x_0))^2]$ the mean squared error at the point x_0. Write (without proving it) an upper bound for $MSE_{x_0}(\hat{f}_h)$, and the assumptions required to obtain it.
3. In the figure below, the $MSE_{x_0}(\hat{f}_h)$ is plotted with respect to the value of the parameter h (this is the result of simulation experiments). Could you explain the shape of the curve? How can the statistician choose the bandwidth h?



4. Projection estimators for the cumulative distribution function (7 points).
We first describe the notations that are used in the sequel.

Norm. For $A = \mathbb{R}$ or $A = [0;1]$, we denote by $\|.\|_{L^2(A)}$ the usual norm of $L^2(A)$, that is $\|g\|_{L^2(A)} = (\int_A g^2(x)dx)^{1/2}$, for $g \in L^2(A)$.

Projection subspace. Let $D > 0$ be an integer and, for any $j \in \{1, \dots, D\}$,

$$\varphi_j(x) = \sqrt{D}\mathbf{1}_{[\frac{j-1}{D}, \frac{j}{D}[}(x), \; x \in [0;1].$$

Let $S_D = \mathrm{Span}\{\varphi_1, \dots, \varphi_D\}$, and $\Pi_{S_D} F$ the orthogonal projection of F onto S_D. It is recalled that $\Pi_{S_D} F = \sum_{j=1}^{D} \theta_j \varphi_j$, with $\theta_j = \int_0^1 F(x)\varphi_j(x)dx$.

Estimator. We consider the following estimator for F:

$$\hat{F}_D = \sum_{j=1}^{D} \hat{\theta}_j \varphi_j, \text{ with } \hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} \varphi_j(x) \mathbf{1}_{X_i \le x} dx, \; j \ge 1.$$

1. The aim of this question is to study the integrated error of $\hat{F}_D$, defined by

$$MISE(\hat{F}_D) = \mathbb{E}\left[\|\hat{F}_D - F\|^2_{L^2([0;1])}\right].$$

   (a) Calculate $\mathbb{E}[\hat{\theta}_j]$, for $j \in \{1, \dots, D\}$. Conclude that $\mathbb{E}[\hat{F}_D(x)] = \Pi_{S_D} F(x)$ for $x \in [0,1]$.

   (b) Justify that

$$MISE(\hat{F}_D) = \|F - \Pi_{S_D} F\|^2_{L^2([0;1])} + \mathbb{E}\left[\|\hat{F}_D - \Pi_{S_D} F\|^2_{L^2([0;1])}\right].$$

   (c) Prove that

$$\forall j \in \{1, \dots, D\}, \; \left(\hat{\theta}_j - \theta_j\right)^2 \le \int_{\frac{j-1}{D}}^{\frac{j}{D}} \left(\frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \le x} - F(x)\right)^2 dx.$$

   (d) Conclude that

$$MISE(\hat{F}_D) \le \|F - \Pi_{S_D} F\|^2_{L^2([0;1])} + \frac{1}{n}.$$

2. What can be concluded about the best choice of the dimension D? Do you know another estimator for the cumulative distribution function F that permits to justify this phenomenon?

3. For $g \in \mathbb{L}^2(\mathbb{R})$, we define the following contrast function

$$\gamma_n(g) = \|g\|^2_{L^2(\mathbb{R})} - \frac{2}{n}\sum_{i=1}^{n} \int_{\mathbb{R}} g(x)\mathbf{1}_{X_i \le x} dx.$$

   (a) Prove that $\mathbb{E}[\gamma_n(g)] = \|g - F\|^2_{L^2(\mathbb{R})} - \|F\|^2_{L^2(\mathbb{R})}$, for any $g \in \mathbb{L}^2(\mathbb{R})$. Deduce that γ_n suits well to estimate the cumulative distribution function F.

   (b) Calculate $\tilde{F}_D = \arg\min_{g \in S_D} \gamma_n(g)$, and conclude that $\tilde{F}_D = \hat{F}_D$.