MASTER 2 – Statistics & Econometrics
EXAM – November 15, 2016 (11h00 – 13h00)
Nonparametric models

Q1 Let $\widehat{f}_n$ be a kernel density estimator based on a random sample $\{X_1, \ldots, X_n\}$, where the underlying density is f.

a) What is the formula of $\widehat{f}_n(x)$?

b) Show that $\widehat{f}_n(\cdot)$ is a probability density function when the underlying kernel $K(\cdot)$ is symmetric.

c) Given that

$$\begin{aligned}
\mathrm{MSE}(\widehat{f}_n(x)) &= \mathbb{E}\left(\widehat{f}_n(x) - f(x)\right)^2 \\
&= \frac{1}{4} h_n^4 \left(f''(x)\right)^2 \tau^4 + \frac{f(x)}{nh_n} \int K^2(y)dy + o\left(h_n^4 + \frac{1}{nh_n}\right)
\end{aligned}$$

with $\tau^2 = \int y^2 K(y)dy$, compute the optimal bandwidth by choosing Gaussian kernel.

Q2 We have tried to obtain a direct kernel density estimate for the Ariege population data (*'pop09.txt'*), with bandwidth chosen by Normal Scale Rule. The resulting estimate appeared clearly incorrect (much smaller than the naive histogram estimator).

a) What was causing this problem ?

b) How did we obtain a reasonable-looking estimate of the density ? Do you know another way to fix the problem ?

Q3 Let $r(x) = \mathbb{E}(Y|X = x)$ be the conditional mean of Y given $X = x$, and let $F(y|x) = \mathbb{P}(Y \leq y|X = x)$ be the conditional distribution function of Y given $X = x$.

a) By making use of kernel density estimators, show how to get kernel estimators for both $r(x)$ and $F(y|x)$.

b) When estimating the regression mean $r(x)$, why it is not optimal to interpolate the observation points (X_i, Y_i) ?

Q4 Prove that the space of natural cubic splines, with k knots, has dimension k.

Q5 Suppose $n \geq 2$, and that $\tilde{g}$ is the natural cubic spline interpolant to the values $y_1, \ldots, y_n$ at points $x_1, \ldots, x_n$ with $a < x_1 < \cdots < x_n < b$. This is a natural spline with a knot at every x_i. Let g be any other twice continuously differentiable function on $[a, b]$ that interpolates the n pairs, *i.e.*, $g(x_i) = y_i$ for $i = 1, \ldots, n$.

a) Let $h(x) = g(x) - \tilde{g}(x)$. Use integration by parts and the fact that $\tilde{g}$ is a natural cubic spline to show that

$$\int_a^b \tilde{g}''(x)h''(x)dx = 0.$$

b) Hence show that

$$\int_a^b [g''(x)]^2 dx \geq \int_a^b [\tilde{g}''(x)]^2 dx.$$

c) Consider the penalized least squares problem

$$\min_{g \in S[a,b]} \left(\sum_{i=1}^{n} [y_i - g(x_i)]^2 + \lambda \int_a^b [g''(x)]^2 dx \right),$$

where $\lambda > 0$ and $S[a, b]$ denotes the space of all 'smooth' functions g on $[a, b]$ that have two continuous derivatives. Use (b) to argue that the minimizer must be a cubic spline with knots at each of the x_i.