TP 2 : Estimation de la R??gression

Evgenii Chzhen & Henry VONG

27 october 2015

Contents

Introduction

We simulate n = 100 couples of independent random variables $(X_i, Y_i),\ i = 1, ..., n$ where a sequence: $X_1, ..., X_n$ is i.i.d uniformly distributed on the interval $[0, 1]$, $\xi_1, ..., \xi_n$ are i.i.d random variables from standart gaussian distribution and

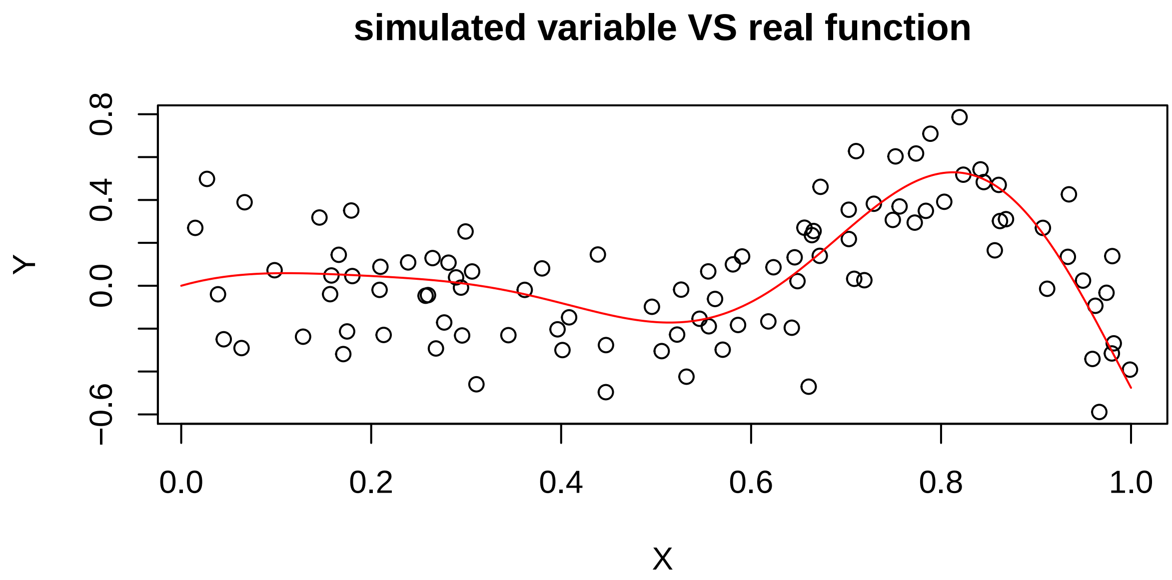
$$Y_i = f(X_i) + \sigma * \xi_i,\ \ \sigma = 0.2$$

and

$$f(x) = (x^2 2^{x-1} - (x - 0.5)^3) sin(10x)$$

1. Visualization

We plot a cloud of $(X_i, Y_i),\ i = 1, ..., n$ and the real function f on $[0, 1]$



2. Estimator by projection. Influence of N.

We consider a trigonometric base $\{\varphi_j\}_{j\geq 1}$ on the interval $[0,1]$:

$$\varphi_1(x) \equiv 1,$$

$$\varphi_{2k} = \sqrt{2}\cos 2\pi kx,$$

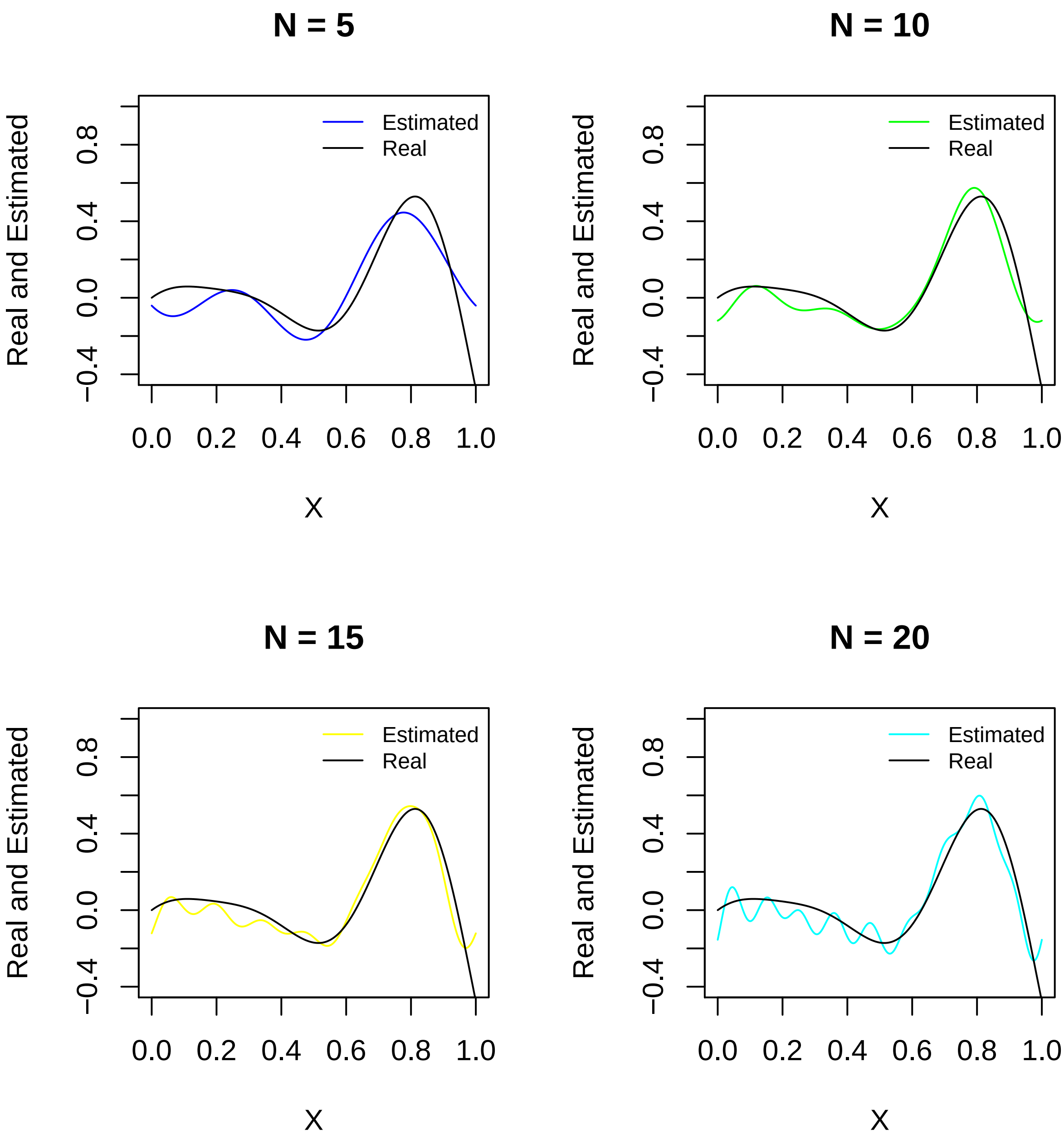
$$\varphi_{2k+1} = \sqrt{2}\sin 2\pi kx, \quad k = 1, 2, ...,$$

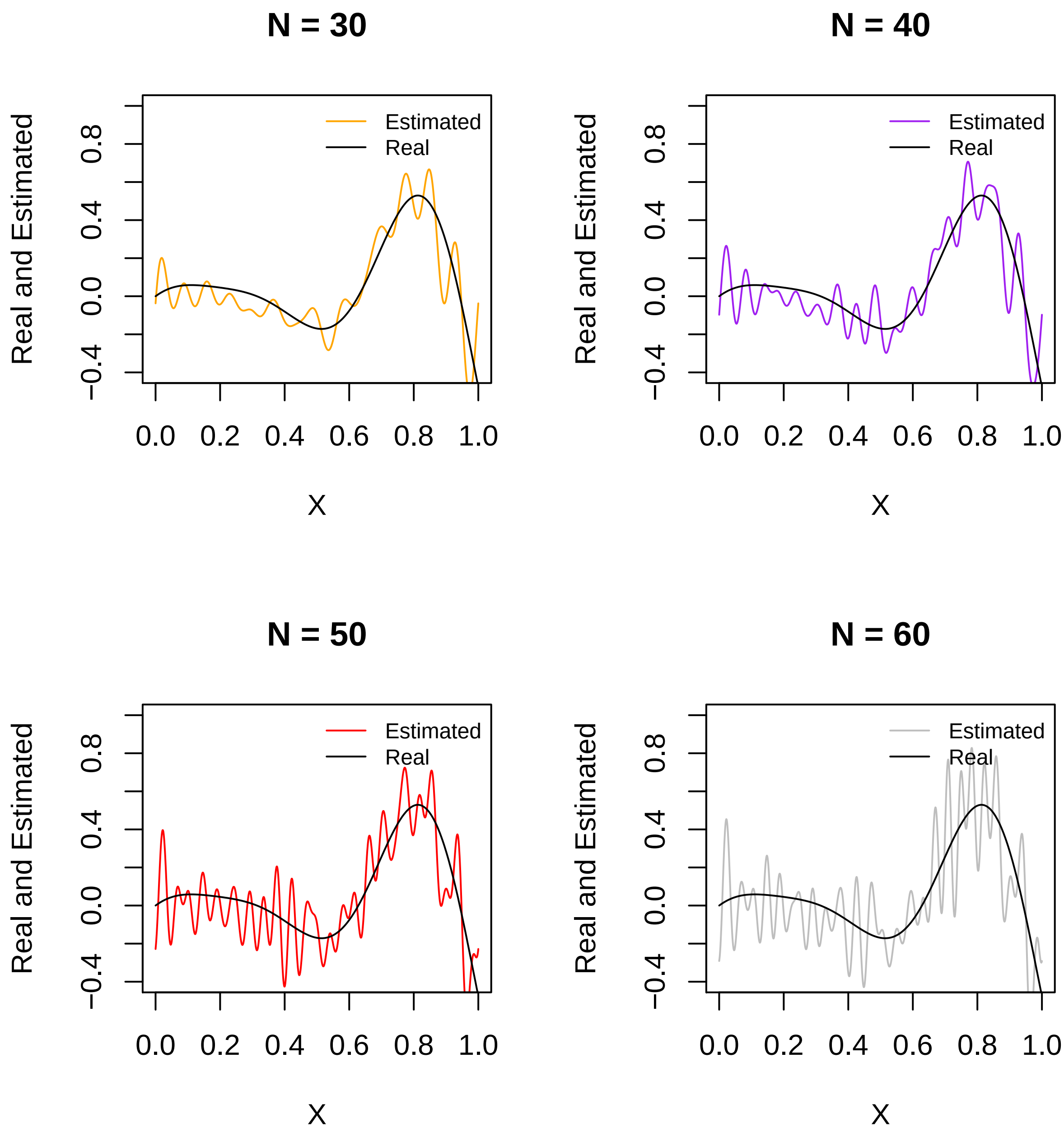
calculate estimators of Fourier coefficients :

$$\hat{\theta}_j = \frac{1}{n}\sum_{i=1}^{n} Y_i \varphi_j(X_i), \text{ for } j = 1, ..., 50.$$

After we consider an estimator by projection as it follows :

$$\hat{f}_{n,N} = \sum_{j=1}^{N} \hat{\theta}_j \varphi_j(x), \text{ for } N \in \{5, 10, 15, 20, 30, 40, 50, 60\}$$


N = 5
N = 10
N = 15
N = 20
Estimated
Real
Real and Estimated
X
−0.4
0.0
0.4
0.8
0.0
0.2
0.4
0.6
0.8
1.0



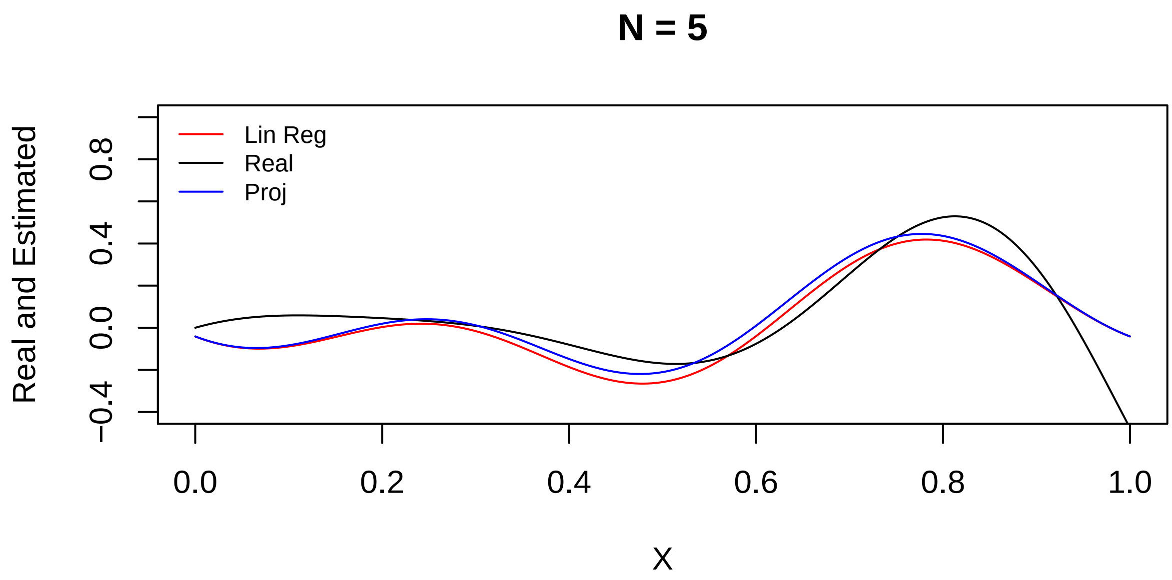
Corollary:One can notice that N = 5 and N = 10 are visualy more appropriate. The problem with a big numbers of N is that we overfit our estimator since we use too many projections.

3. Regression and Visualization of estimated function.

We consider $Z_j = (\varphi_j(X_1), ..., \varphi_j(X_n))^T$, we estimate $\beta = (\beta_1, ..., \beta_N)$ using a following linear model :

$$Y = \beta_1 \cdot Z_1 + ... + \beta_N \cdot Z_N + \xi$$

We plot two different estimator $\hat{f}_{n,N}$ and $\tilde{f}_{n,N} = \sum_{j=1}^{N} \hat{\beta}_j \varphi_j(x)$, for $N = 5$



4. Theory

We note $\mathbf{X} = (Z_1, ..., Z_N)$. If $\mathbf{X}^T\mathbf{X}/n = I_N$, therefore $\hat{\beta}_j = \left((\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T Y\right)_j = \frac{1}{n}\left(\mathbf{X}^T Y\right)_j = \hat{\theta}_j$ and finally $\tilde{f}_{n,N} = \hat{f}_{n,N}$.

5. Variance

For N $= 50$ estimated value of $\sigma^2 = 0.04$ is $\hat{\sigma}^2 = 0.0471989$

6. Minimization

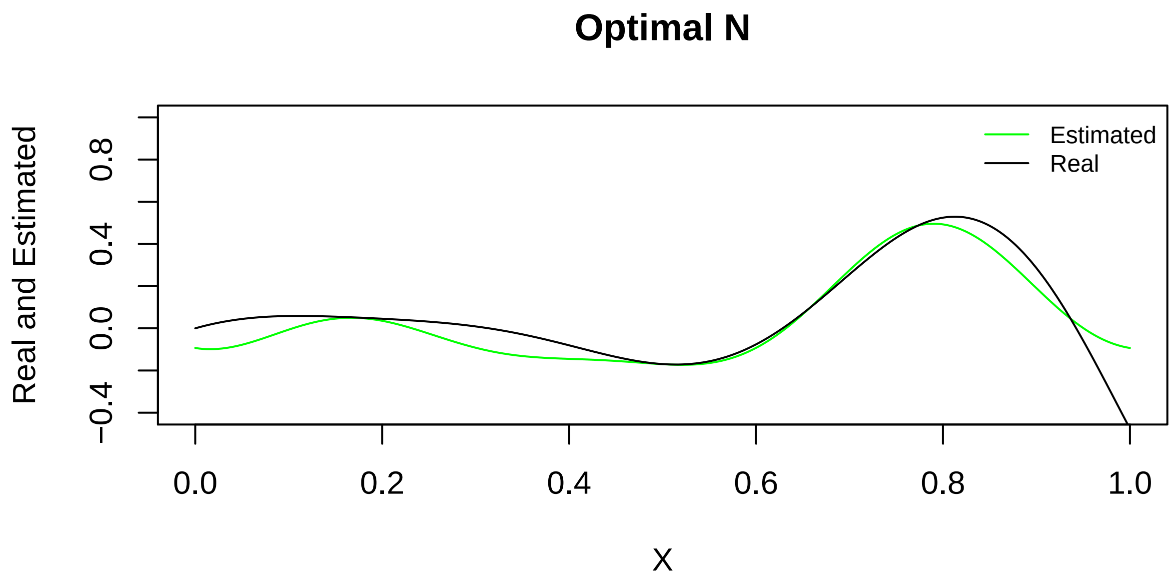
We observe an emperic loss of the estimator and obtain optimal N by minimization, see for instance (Tsybakov 2008, 59–61).

$$\hat{N} = \arg\min_{N=1,...,50}\left(||Y - \mathbf{X}\cdot\hat{\beta}||^2 - (n-2N)\hat{\sigma}^2\right)$$

We obtained that the optimal value is $\hat{N} = 7$.

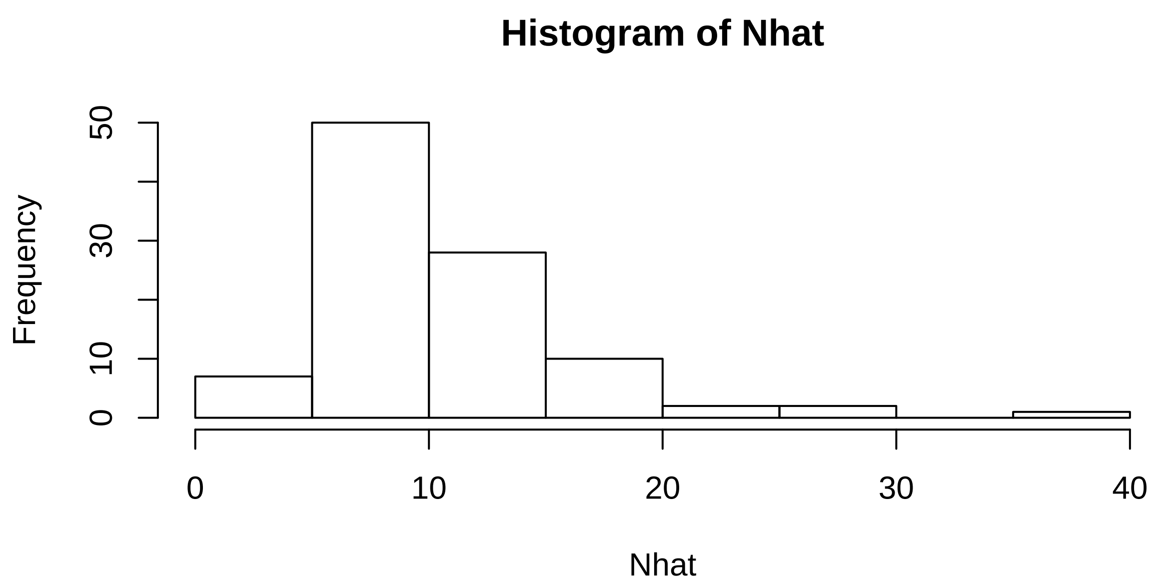
7. Visualization for optimal $\hat{N} = 7$

We compute estimation for obtained $\hat{N}$.



8. Histogramm for $\hat{N}$'s

In this section we consider $M = 100$ simulations of $n = 100$ observations from given function, for each simulation we find an optimal $\hat{N}_i$ for all $i = 1, ..., M$. We look at the histogramm of $\hat{N}_i$ for all $i = 1, ..., M$.



Corollary: One can notice that values of N between 5 and 12 are working in most cases.

References

Tsybakov, Alexandre B. 2008. *Introduction to Nonparametric Estimation*. Springer.