*Séminaire Mathématique de Béjaïa (LaMOS)*
*Volume 1, 2003, pp. 37 - 40*

# 9

# Estimation de la courbe de régression de la moyenne par la méthode non paramétrique du noyau

S. ADJABI

Laboratoire de Modélisation et d'Optimisation des Systèmes (LAMOS)
Université de Béjaïa, Béjaïa 06000, Algérie
**Tél.** (213) 34 21 51 88,
email : adjabi@hotmail.com

**Résumé** Ce travail présente l'estimateur de la courbe de régression de la moyenne par la méthode du noyau de Parzen-Rosenblatt ainsi que ses propriétés. Cet estimateur noté $\mathbb{E}(X/Y)$ est utilisé pour estimer la dépendance de deux variables aléatoires $X$ et $Y$ quand on ne veut faire aucun à priori sur la loi de $X$ et de $Y$. Deux exemples sont présentés pour illustrer cet estimateur.

**Mots Clés :** Estimateur, noyau, densité de probabilité, paramètre de lissage.

## Introduction

On s'intéresse au problème de l'estimation de la relation éventuelle entre une variable aléatoire $Y$ et une variable explicative $X$.
La relation sera de la forme :

$$Y = f(x, \theta) + \varepsilon\,, \quad \varepsilon \text{ est l'erreur.}$$

- Si on suppose que la loi du couple $(X, Y)$ est gaussienne de $\mathbb{R}^2$, la relation est linéaire et on a :

$$Y = \theta_1 + \theta_2 X + \varepsilon$$

- Lorsqu'on ne veut faire sur le couple $(X, Y)$ aucune hypothèse permettant d'utiliser une méthode paramètrique, on approche alors $Y$ par $\mathbb{E}(Y/X = x)$. On obtient alors le modèle

$$Y = \mathbb{E}(Y/X = x) + \varepsilon = \int y f(y/x) dy + \varepsilon = \int y \frac{f(x, y)}{f(x)} dy + \varepsilon.$$

$f(x, y)$ et $f(x)$ sont, respectivement, les lois du couple $(X, Y)$ et $X$.
L'objectif étant, si $f(x, y)$ est inconnue, de l'estimer par la méthode non paramètrique du noyau.

## 9.1 Estimation de la loi du couple

L'estimateur $f_n(x, y)$ de la loi $f(x, y)$ du couple $(X, Y)$, par la méthode du noyau est donné par CACOULOS sous la forme :

$$f_n(x,y) = \frac{1}{nh^2}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) K\left(\frac{y-y_i}{h}\right).$$

$K$ étant le noyau telle que $\int K(y)dy = 1$ et $h$ paramètre de lissage dépend de $n$ et vérifiant $\lim_{n\to\infty} h = 0$ et $\lim_{n\to\infty} nh = \infty$.
Sous certaines conditions sur $K$ et sur $h$, cet estimateur de CACOULOS est consistant en moyenne quadratique, consistant uniformément presque sûrement et asymptotiquement gaussien.

## 9.2 Estimation de la courbe de régression de la moyenne

$$Y = \mathbb{E}(Y/X = x) + \varepsilon = y(x) + \varepsilon.$$

Une première approximation de $y(x)$ est $Moy(y_i,\, x_i \in V(x))$ où $V(x)$ est un voisinage de $x$ (moyenne mobile locale). Une version améliorée est la moyenne mobile locale pondérée, pour laquelle les observations dont l'abscisse est plus proche de $x$ ont un poids plus élevé que celles qui en sont éloignées. C'est l'estimateur de Nadaray-Watson dont la forme est

$$y_n(x) = \int_{-\infty}^{+\infty} y f_n(y/x)dy = \frac{\int y f_n(x,y)dy}{\int f_n(x,y)dy} = \frac{\sum_{i=1}^{n} y_i K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}.$$

$K$ étant le noyau, il doit vérifier : $\int_{\mathbb{R}} K(y)dy = 1$, et $\int_{\mathbb{R}} yK(y)dy = 0$.
$h$ est le paramètre de lissage vérifiant $\lim_{n\to\infty} h(n) = 0$ et $\lim_{n\to\infty} nh(n) = \infty$.
Cet estimateur est consistant ponctuellement en probabilité, asymptotiquement gaussien et consistant uniformément presque sûr.
**Remarque**
Cet estimateur de Nadaraya n'est pas le seul estimateur qui utilise la méthode du noyau. Priestley-Chao ont proposé des modifications qui en améliorent les qualités statistiques. Néanmoins, des études ont montré que le comportement de cet estimateur est assez similaire en pratique.
Estimateur de Nadaraya : $y_n(x) = \sum_{i=1}^{n} y_i w_i(x)\,,\quad$ avec $\quad w_i(x) = \frac{K\left(\frac{x-x_i}{h}\right)}{\sum_{i=1}^{n} K(\frac{x-x_i}{h})}$.
Estimateur de Priestley-Chao : $y_n(x) = \sum_{i=1}^{n} y_i w_i(x)\,,\quad$ avec $\quad w_i(x) = \frac{x_{i+1}-x_i}{h} K\left(\frac{x-x_i}{h}\right)$,
$i = 1\ldots, n$, $x \in [0,1]$, $x_0 = 0 < x_1 < \cdots < x_i < x_{i+1} \cdots < x_n < x_{n+1} = 1$
**Rapport de corrélation et comparaison avec une régression linéaire**
* Si le couple $(X,Y)$ est gaussien ; $Y_L = aX + \varepsilon_L$ et on a : $\hat{b} = \frac{Cov(X,Y)}{V(X)} = \rho\sqrt{\frac{V(Y)}{V(X)}}$ et $\hat{a} = \bar{X} - \hat{b}\bar{X}$. Le carré du coefficient de corrélation linéaire $\rho$ vérifie : $1 - \rho^2 = \frac{V(\varepsilon_L)}{V(Y_L)}$.
Par analogie, on définit pour la courbe de régression de la moyenne $Y_C = \mathbb{E}(Y/X) + \varepsilon_C$ un rapport de corrélation définit par $1 - \eta^2 = \frac{V(\varepsilon_C)}{V(Y_C)}$.
La mesure relative entre les deux régression est : $\eta^2 - \rho^2 = \frac{\mathbb{E}(Y_L - \mathbb{E}(Y/X))^2}{V(Y)}$.

**Application**

Le noyau utilisé est le noyau gaussien. L'estimateur de la courbe de régression de la moyenne s'écrit alors

$$y_n(x) = \frac{\sum_{i=1}^{n} y_i e^{-\frac{(x-x_i)^2}{2h}}}{\sum_{i=1}^{n} e^{-\frac{(x-x_i)^2}{2h}}}.$$

**Choix du paramètre de lissage**

Habbema et Herman ont estimé le paramètre de lissage $h$ par la méthode du Pseudo maximum de vraisemblance ou méthode du "Jackknif". La valeur $\hat{h}$ choisie pour $h$ à partir de l'échantillon est la valeur qui maximise

$$L(x_1, x_2, \ldots, x_n; h) = \prod_{j=1}^{n} \frac{1}{n-1} \sum_{i=1, i\neq j}^{n} K\left(\frac{x_i - x_j}{h}\right).$$

Généralement la procédure précédente d'estimation du paramètre de lissage ne différencie pas les régions de forte densité où la fonction densité risque d'être trop lissée (quand $h$ est trop grand ) et celles où les observations sont moins concentrées et où peuvent apparaître dans l'estimation si $h$ est trop faible des ruptures ou des pics non significatifs. Pour pallier à cette difficulté, on peut pour chaque observation $x_i$ moduler le paramètre $h$ en fonction de la concentration des observations autour de $x_i$.

$$h_i = hd(i), \quad h \text{ étant l'estimateur du } PL.$$

$d(i)$ est une fonction d'autant plus faible que la densité dans le voisinage du point est élevée.

$$d(i) = \frac{dd(i)}{\overline{dd}}, \quad dd(i) = \sqrt{\sigma^2_{i,k(n)}} \quad \overline{dd} = \sum_{i=1}^{k(n)} dd(i);$$
$$\sigma^2_{i,k} = \frac{1}{k(n)} \sum_{j=1}^{k(n)} (X_j - \overline{X})^2 \quad \text{et} \quad \overline{X} = \frac{1}{n} \sum_{j=1}^{k(n)} X_j.$$

La valeur de $k(n)$ proche de $\sqrt{n}$ donne d'assez bons résultats. Deux exemples sont donnés pour illustrer l'intérêt de l'estimation de la courbe de régression de la moyenne. Sur les deux figures, les observations sont représentées par le signe +, en gras est représentée la courbe de régression de la moyenne et en tirets la droite de régression linéaire. Sur l'exemple 1, la droite de régression linéaire s'écrit $Y = -0.565X + 25.01 + \varepsilon$ avec un coefficient de corrélation de $-0.3358$. Ce faible coefficient de corrélation signifie que la régression n'est pas linéaire, par conséquent la dépendance entre les deux variables est mieux représentée par la courbe de régression de la moyenne $\mathbb{E}(Y/X)$. Par contre, sur l'exemple 2, le coefficient de corrélation est égal à $-0.8101$, la droite de régression s'écrit $Y = -0.077X + 13.07 + \varepsilon$, le carré du coefficient de corrélation qui vaut 0.738 est sensiblement égal au carré du rapport de corrélation qui est égal à 0.738. par conséquent, la dépendance entre ces deux variables est linéaire.

## Références

1. S. Adjabi. Estimation non paramétrique de la fonction densité de probabilité par la méthode des noyaux. *Actes des premières journées de mathématiques appliquées, Rabat, Maroc, 15-17 juillet*, pages 626–631, 1992.

2. T. Cacoulos. Estimation of mulivariate density. *Ann. Math. Statist.*, 2 :179, 1996.
3. V.A. Epanechnikov. Non parametric estimation of a multivariate probability density. *Theory of probability and its applications*, 14 :141, 1964.
4. J.D.F. Habbema and J. Hermans. The allock package multigroup discriminant analysis programs based on direct density estimation. *Comp. Stat. Gordesch and P. naeve, physica, Verlag. - Wien,*, 1976.
5. E.A. Nadaraya. On estimating regression. *Theory of probability and its applications*, page 18, 1973.