*Séminaire Mathématique de Béjaïa (LaMOS)*
*Volume 3, 2005, pp. 59 - 63*

# 9

# Correction de l'effet de bord dans la méthode du noyau

A. BARECHE[1]

Laboratoire de Modélisation et d'Optimisation des Systèmes LAMOS
Université de Béjaïa 06000, Algerie.
email : aicha_barecheyahoo.fr

**Résumé** Dans ce travail, nous discutons le choix du paramètre de lissage dans l'estimation fonctionnelle de la densité de probabilité par la méthode du noyau. Nous présentons des méthodes pour la correction des effets de bord dans le cas où la densité $f$ présente des discontinuités au bord de son support. Un exemple numérique basé sur la simulation est donné pour illustrer ces différentes approches.

**Mots clés :** Méthode du noyau, Effet de bord, Fenêtre, Cross-validation, Plug-in.

## 9.1 Introduction

Etant donné un échantillon $X_1, ..., X_n$ d'une distribution de fonction densité $f$, la méthode non paramétrique d'estimation de densité la plus commune et utilisée est l'estimateur à noyau de Rosenblatt [5] :

$$\hat{f}_h(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{9.1}$$

$K$ est une fonction densité symétrique appelée noyau et $h$ est appelé paramètre de lissage (fenêtre).

Dans la pratique, la critique étape dans l'estimation densité est le choix de la fenêtre $h$, qui contrôle le lissage de l'estimateur densité. Ce dernier problème a été largement étudié et plusieurs méthodes ont été proposées. La plupart parmi elles supposent que $f$ est une fonction lisse à travers la droite réelle $\mathbb{R}$. Relativement, peu est connu sur le cas où $f$ présente des discontinuités au bord de son support. Il est bien connu que l'effet de bord cause plusieurs difficultés quand des procédures désignées à estimer des densités lisses sont appliquées à des densités avec des discontinuités [4, 7, 2]. L'usage d'une large fenêtre introduit un large biais au niveau ou autour du bord et une petite fenêtre augmente dramatiquement la variation.

Plusieurs approches pour manier les problèmes causés par les effets de bord dans l'estimation des densités ont été proposées.

## 9.2 Choix de la fenêtre

Le problème du choix de la fenêtre a été largement étudié. Les méthodes proposées dans la littérature peuvent être divisées en deux grandes classes :

### 9.2.1 Méthodes classiques

Ce sont plus ou moins des extensions naturelles des méthodes utilisées dans la modélisation paramétrique. Les plus connues sont :

- La validation croisée (cross-validation) [3].
- La méthode "least square cross-validation" [6, 1].

### 9.2.2 Méthodes plug-in

Le biais d'un estimateur densité $\hat{f}$ s'écrit en fonction de la densité inconnue $f$ et est souvent approximé par les développements en série de Taylor. Un estimateur pilote de $f$ est alors injecté pour dériver un estimateur du biais et par la suite un estimateur de l'erreur quadratique moyenne intégrée. Le $h$ optimal minimise cette dernière mesure estimée. Diverses méthodes sont proposées, nous citons :

- La méthode de Sheather-Jones [9].
- La cross-validation biaisée [8].
- La méthode "Rule of thumb" [10].

## 9.3 Correction de l'effet de bord

Les difficultés causées par les effets de bord peuvent être expliquées à travers l'estimateur (9.1). Ce dernier est en fait un "passe-filtre" qui passe l'information des composantes de basses fréquences et supprime le bruit aux fréquences élevées. Pour des densités lisses, l'information sur $f$ est concentrée aux basses fréquences et l'estimateur (9.1) peut efficacement passer l'information sans passer beaucoup de bruit.
Cependant, quand $f$ a des discontinuités, les composantes de hautes fréquences contiennent toujours de l'information considérable sur $f$. Ainsi, le "passe-filtre" va soit supprimer beaucoup d'information ou passer beaucoup de bruit dans le but de passer la plupart de l'information. L'idée essentielle des procédures proposées est d'utiliser les composantes de hautes

fréquences pour estimer les effets de bord et ajuster par la suite ces derniers en soustrayant ou ajoutant une fonction à la fonction cumulative empirique avant d'appliquer les procédures désignées à estimer des densités lisses. Plusieurs méthodes ont été proposées, nous citons :

### 9.3.1 Noyaux bornés

Beaucoup d'auteurs considèrent l'usage des noyaux bornés autour des bords. Nous citons ici, le noyau de Rice [4] :

$$w(x) = \frac{3}{4}(-x^2 + 1), \quad |x| \leq 1 \tag{9.2}$$

qui est appliqué pour l'intérieur, et le noyau :

$$u_q(x) = (ax + b)w(x), \quad q > -1 \tag{9.3}$$

appliqué autour du bord, où a,b sont des constantes satisfaisant :

$$\int_{-q}^{1} u_q(x)dx = 1 \quad \text{et} \quad \int_{-q}^{1} xu_q(x)dx = 0 \tag{9.4}$$

### 9.3.2 pseudo-données

Plusieurs auteurs proposent de créer quelques pseudo-données en reflétant les données autour du bord ; les procedures pour estimer les fonctions lissées sont alors appliquées aux données améliorées pour choisir la fenêtre et estimer la densité.

Le principal problème avec ces approches est que les noyaux utilisés autour du bord sont différents de celui utilisé pour l'intérieur ; alors la simple structure de l'estimateur (9.1) peut être perdue facilement. Nous citons ici, la méthode de Cowling-Hall [2] :
Les pseudo-données sont crées comme suit :

$$X_{(-i)} = -5X_{(\frac{i}{3})} - 4X_{(\frac{2i}{3})} + \frac{10}{3}X_{(i)}, \quad i = 1, \ldots, n \tag{9.5}$$

où $X_{(0)} = 0$ et $X_{(i)}$ sont les statistiques d'ordre. Les auteurs appliquent le noyau :

$$K(x) = (\frac{3}{4})(-x^2 + 1), \quad |x| < 1 \tag{9.6}$$

aux pseudo-données crées pour obtenir l'estimateur densité. Pour le choix de la fenêtre, ils utilisent la méthode "least square cross-validation" [6, 1].

### 9.3.3 Symétrisation ou effet mémoire

En 1985, Schuster [7] suggère de créer une image mémoire des données de l'autre coté du bord et appliquer ensuite l'estimateur (9.1) à l'ensemble des données et leur reflection. $f(x)$ est alors estimée, pour $x \geq 0$, comme suit :

$$f(x) = \frac{1}{nh} \sum_{i=1}^{n} [K(\frac{x - X_i}{h}) + K(\frac{x + X_i}{h})] \tag{9.7}$$

## 9.4 Application-Simulation

Pour voir l'intérêt de la correction de l'effet de bord dans l'application de la méthode du noyau, on fait appel à la simulation et à la programmation sous Matlab 6.5. On génère un échantillon aléatoire de loi exponentielle de paramètre 1 et de taille $n = 100$. La densité de cet échantillon est représentée sur la figure (Fig. 9.1) par la courbe bleue. Dans un premier temps, on estime la densité $f$ par la méthode du noyau en utilisant l'estimateur à noyau donné dans la formule (9.1), avec $K$ le noyau normal et le paramètre de lissage $h$ est choisi par la méthode "rule of thumb" de Silverman [10]. La densité estimée est représentée sur la figure (Fig. 9.1) par la courbe verte. Dans un second temps, on estime la densité $f$ en prenant en considération l'effet du bord en utilisant la symétrisation de schuster [7] donnée dans la formule (9.7). La densité estimée est représentée sur la figure (Fig. 9.1) par la courbe rouge.

On voit bien, d'après la figure (Fig. 9.1), qu'en prenant en considération la correction de l'effet de bord, on améliore l'estimation de la densité par la méthode du noyau.

## 9.5 Conclusion

Dans ce travail, nous avons porté, dans un premier temps, un accent sur le choix du paramètre de lissage dans l'estimation fonctionnelle de la densité de probabilité par la méthode du noyau. Dans un deuxième temps, on s'est intéressé au cas où $f$ présente des discontinuités au bord de son support. Plusieurs approches ont été citées pour remédier aux problèmes causés par les effets de bord dans l'estimation des densités par la méthode du noyau. Un exemple pratique basé sur la simulation est donné comme illustration.

## Références

1. A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. *Biometrika.*, 71 :353–360, 1984.

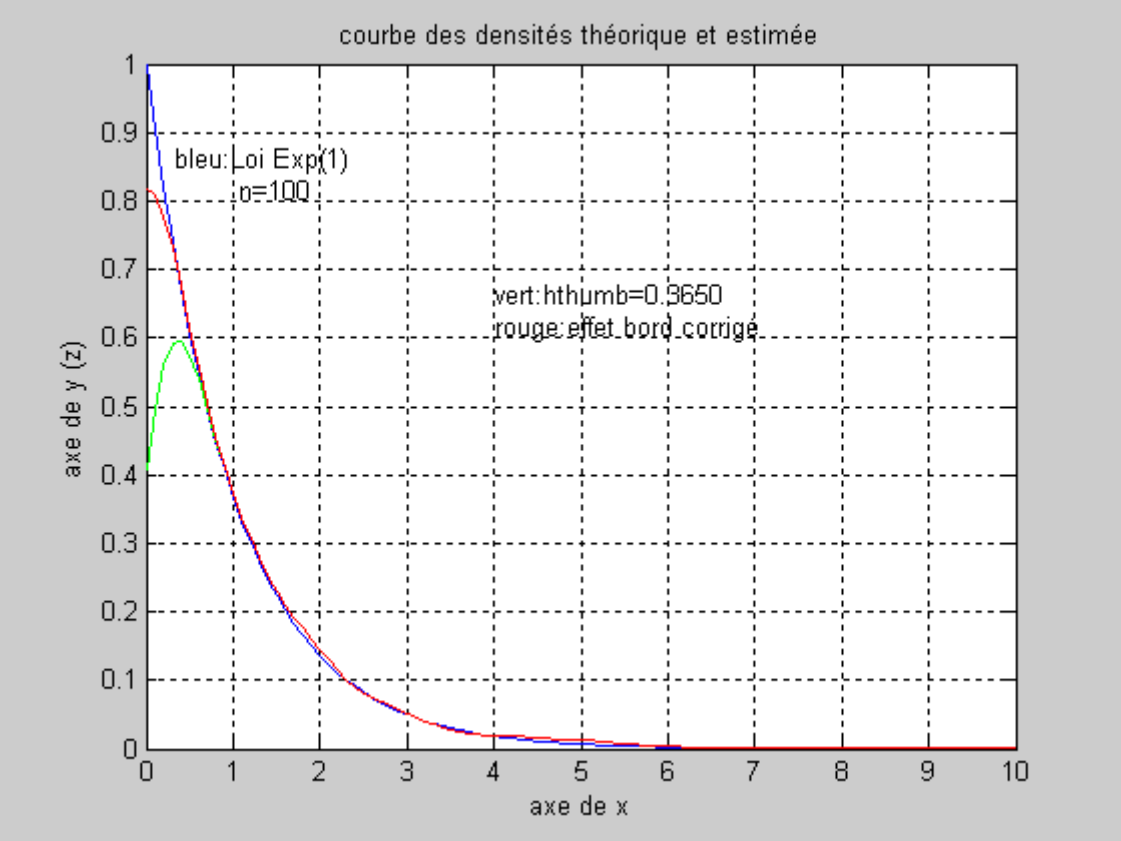


**Figure 9.1.** courbes des densités théorique et estimée


2. A. Cowling, P. Hall. On pseudodata methods for removing boundary effects in kernel density estimation. *J. Roy. Statist. Soc. Ser*, B 58 :551–563, 1996.
3. J.D.F. Habbema, J. Hermans, and K. Van Der Broek. A stepwise discriminant analysis program using density estimation. In COMPSTAT 1974, *Proceedings in Computational Statistics, Vienna (G. Bruckman ed.)*. Physica, Heidelberg, 101–110, 1974.
4. J. Rice. Boundary modification for kernel regression. *Commun. Statist. Theory Meth.*, 13 :893–900, 1984.
5. M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.*, 27 :832–837, 1956.
6. M. Rudemo. Empirical choice of histograms and kernel density estimators. *Scand. J. Statist.*, 9 :65–78, 1982.
7. E.F. Schuster. Incorporating support constraints into nonparametric estimation of densities. *Commun. Statist. Theory Meth.*, 14 :1123–1136, 1985.
8. D.W. Scott, G.R. Terrell. Biased and unbiased cross-validation in density estimation. *J. Amer. Statist. Assoc.*, 82 :1131–1146, 1987.
9. S.J. Sheather, M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *J. R. Statist. Soc. Ser.*, B 53 :683–690, 1991.
10. B.W. Silverman. Density Estimation for Statistics and Data Analysis. estimation. Chapman and Hall, London, 1986.