

Correction locale de l'estimateur à noyau de la densité d'une loi de probabilité*

Belkacem ABDOUS et Alain BERLINET

Université du Québec à Trois-Rivières et Université de Montpellier II

Key words and phrases: Kernel density estimators, Rao-Blackwellization, locally sufficient statistics.

AMS 1985 subject classification: 62G05.

RÉSUMÉ

L'estimateur à noyau de Parzen-Rosenblatt a tendance à dévier systématiquement de la vraie valeur de la densité au voisinage de certains points critiques du graphe. Pour remédier à cet inconvénient nous proposons une extension de la méthode de Rao-Blackwell s'appuyant sur des statistiques localement exhaustives. Nous étudions le risque quadratique de l'estimateur ainsi défini et démontrons sa convergence presque complète. L'estimateur mis en pratique donne à chaque observation suffisamment éloignée du point d'intérêt un poids fixe associé à l'estimateur global standard. Son intérêt est mis en évidence par des simulations.

ABSTRACT

The standard Parzen-Rosenblatt kernel density estimator is known to systematically deviate from the true value near critical points of the density curve. To overcome this difficulty, we extend the Rao-Blackwell method by using locally sufficient statistics: we define a new estimator and study its asymptotic behaviour. The interest of the method is shown by means of simulations.

1. INTRODUCTION

Soit \mathcal{P} un ensemble de probabilités absolument continues par rapport à la mesure de Lebesgue sur $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$, où $\mathcal{B}_{\mathbb{R}}$ désigne la tribu borélienne de \mathbb{R} . On considère un échantillon X_1, \dots, X_n issu d'une loi de probabilité Q inconnue appartenant à \mathcal{P} . On suppose que la densité de Q admet une version continue bornée f_Q . L'estimateur standard de Parzen-Rosenblatt de f_Q s'écrit sous la forme

$$f_n(x) = \frac{1}{nh} \sum_{j=1}^n K\left(\frac{x - X_j}{h}\right), \quad x \in \mathbb{R},$$

où K , appelé noyau, est une fonction réelle d'intégrale égale à 1, $h = h(n)$ est une suite de réels positifs convergeant vers 0. Le comportement de $f_n(x)$ a été largement étudié dans les dernières décennies; on pourra se reporter à Berlinet et Devroye (1990) pour une synthèse ou à Bosq et Lecoutre (1987) pour de plus amples détails. L'estimateur à noyau $f_n(x)$ présente souvent l'inconvénient de ne pas estimer convenablement la densité

*Le travail du premier auteur a été subventionné par le CRSNG du Canada et par le FIR de l'Université du Québec à Trois-Rivières.

au voisinage d'un mode, d'un "creux", d'un "plat" ou aux extrémités du support de la loi lorsque celui-ci est borné. Lorsque la courbure de la fonction f_Q est élevée au voisinage de x , le biais est lui aussi élevé. Il en est de même pour la variance lorsque la densité présente un mode voisin de x . Usuellement un estimateur basé sur une fenêtre h choisie par un critère global sur-estime la valeur de la densité au voisinage des minima et la sous-estime au voisinage des maxima. Les expériences numériques montrent même que la recherche des extrema de $f_n(x)$ considérée comme fonction de h à x fixé ne permet pas de surmonter ce problème sans dégrader sérieusement l'allure générale de f_n . Ceci prouve que l'utilisation de fenêtres dépendant du point d'estimation ne garantit pas une amélioration de l'estimateur de $f(x)$ si l'on veut conserver une information globale correcte à partir de f_n . On peut également intervenir sur l'estimateur en choisissant un noyau dont les premiers moments sont nuls. Sous des hypothèses adéquates ceci permet de réduire à la fois le biais et la variance (Berlinet 1993). Dans le présent article nous envisageons des corrections ponctuelles des estimateurs que nous justifions par des considérations théoriques dérivant de la Rao-Blackwellisation. Ces corrections peuvent s'appliquer à tout type d'estimateur de courbe ou de surface. Une approche heuristique du problème conduit à distinguer deux parties dans l'échantillon en fonction de la distance entre le point d'intérêt x et les observations. On se donne un intervalle U contenant x et on conserve la contribution de X_i à la valeur de $f_n(x)$ suivant que X_i appartient ou non à U . On obtient ainsi un nouvel estimateur de $f(x)$ de la forme

$$\frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{h} K \left(\frac{x - X_j}{h} \right) 1_U(X_j) + \phi_U(X_j) 1_{U^c}(X_j) \right\},$$

où la fonction ϕ_U devra être définie à partir de l'estimateur f_n de manière à conserver une information de nature globale sur l'estimation de $f(x)$. Les considérations théoriques de la section suivante montrent qu'un choix pertinent est celui qui fixe ϕ_U égale à la constante $m(f_n)$, valeur moyenne de la fonction $K((x - \cdot)/h)/h$ sur le complémentaire de U :

$$m(f_n) = \frac{\frac{1}{h} \int_{U^c} K \left(\frac{x - y}{h} \right) f_n(y) dy}{\int_{U^c} f_n(y) dy}.$$

On donne pour l'estimateur ainsi obtenu des résultats de convergence en probabilité et presque complète. L'étude de son comportement à distance finie permet la détermination pratique de l'intervalle U intervenant dans sa définition. Des simulations permettent de juger de son intérêt.

2. FORME DE L'ESTIMATEUR À NOYAU CONDITIONNÉ

La correction locale de $f_n(x)$ peut être réalisée grâce à la technique du conditionnement; technique introduite dans un cadre plus général par Berlinet (1984) et que nous particulierisons à l'estimateur à noyau $f_n(x)$. Sous la seule hypothèse $Q \in \mathcal{P}$, la connaissance de Q sur un sous-ensemble E de \mathbb{R} tel que $Q(E) < 1$ ne nous permet pas de déterminer f_Q , tandis que la connaissance de sa restriction à un voisinage U d'un réel fixé x entraîne celle de $f_Q(x)$. Le paramètre $f_Q(x)$ est dit localisé par U . Pour une définition plus générale, nous renvoyons à Bosq (1970) et Berlinet (1984). La conséquence de cette constatation sur l'estimation de $f_Q(x)$ au vu d'un échantillon X_1, \dots, X_n , est que les observations tombant dans ou à proximité de U fourniront l'essentiel de l'information apportée par

cet échantillon pour l'estimation de $f_Q(x)$. D'où l'idée de considérer la statistique S qui retient les observations qui tombent dans U et qui censure les autres, i.e. $S = (S_1, \dots, S_n)$, avec

$$S_i = \begin{cases} X_i & \text{si } X_i \in U \\ x_0 & \text{sinon, avec } x_0 \notin U. \end{cases}$$

Si on suppose que Q est connue sur U^c , le complémentaire de U , alors la statistique S a la remarquable propriété de résumer l'information que l'on peut avoir sur la restriction de Q à U (notée $Q|_U$), i.e. ce qui manque pour connaître entièrement la loi Q . Ceci revient à dire que S est localement exhaustive ou U^n -exhaustive: d'une façon générale, une statistique est dite U^n -exhaustive si elle est exhaustive au sens usuel dans le modèle $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathcal{P}(Q, U^c))^{\otimes n}$ où $\mathcal{P}(Q, U^c) = \{P \in \mathcal{P} : P|_{U^c} = Q|_{U^c}\}$. L'exhaustivité locale de S dans le modèle $(\mathbb{R}, \mathcal{B}_{\mathbb{R}}, \mathcal{P})^{\otimes n}$ est établie par Berlinet (1984). Par ailleurs, d'après le théorème de Rao-Blackwell (voir Lehmann 1983), pour améliorer un estimateur il suffit de le remplacer par son espérance conditionnelle par rapport à une statistique exhaustive.

Nous nous proposons d'étudier les effets du conditionnement par rapport à la statistique S sur l'estimateur à noyau $f_n(x)$. Pour simplifier, nous prenons pour U un intervalle symétrique autour de x , i.e. $U =]x - H, x + H[$, où H est un paramètre positif, dépendant du nombre d'observations n et dont le choix sera discuté dans la suite. Il sera facile d'étendre les considérations de cet article à un contexte beaucoup plus vaste. Notons f_P une version continue de la densité d'une loi P telle que $Q|_{U^c} = P|_{U^c}$ et $P(U^c) > 0$. La restriction $P|_{U^c}$ correspond à la partie supposée connue de Q . En pratique, cette probabilité proviendra en général d'une pré-estimation ou d'une information *a priori*; dans le cas d'une pré-estimation, la démarche ici envisagée revient à considérer que le pré-estimateur obtenu ne s'écarte pas trop de la véritable densité. Soit \mathcal{B}_S la tribu engendrée par S . D'après un résultat de Berlinet (1984), une version $\tilde{f}_n(x)$ de $E_{\mathcal{P}^{\otimes n}}^{\mathcal{B}_S}(f_n(x))$ est de la forme

$$\tilde{f}_n(x) = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{h} K \left(\frac{x - X_j}{h} \right) 1_U(X_j) + m(P) 1_{U^c}(X_j) \right\},$$

où

$$m(P) = \frac{1}{P(U^c)} \int_{U^c} \frac{1}{h} K \left(\frac{x - y}{h} \right) f_P(y) dy.$$

L'estimateur $\tilde{f}_n(x)$ peut être perçu, à son tour, comme un estimateur à noyau, avec un noyau de la forme

$$\tilde{K}(u) = K(u) 1_{\Omega}(u) + m(P) 1_{\Omega^c}(u),$$

où on a posé $\Omega =] - H/h, H/h[$.

Vu le contexte d'estimation locale en x que l'on s'est fixé, une étude globale de \tilde{f}_n aurait peu de sens. Ainsi, le fait que \tilde{K} ne soit pas toujours intégrable et de ce fait même que \tilde{f}_n ne soit pas une densité n'est pas un inconvénient pour notre étude. Pour ce qui est de l'amélioration obtenue en passant de $f_n(x)$ à $\tilde{f}_n(x)$, une application du théorème de Rao-Blackwell nous donne:

$$Q^{\otimes n}|_{(U^n)^c} = P^{\otimes n}|_{(U^n)^c} \Rightarrow R(Q, \tilde{f}_n(x)) \leq R(Q, f_n(x)),$$

où $R(Q, g_n(x)) = \int L(Q, g_n(x, x_1, \dots, x_n)) dQ^{\otimes n}(x_1, \dots, x_n)$ est le risque associé à $g_n(x)$; L étant une fonction de perte définie de $\mathcal{P} \times \mathbb{R}$ dans \mathbb{R}^+ telle que $\forall Q \in \mathcal{P}$, $L(Q, \cdot)$ soit convexe. Nous n'avons considéré ici qu'un risque ponctuel. Il est cependant facile de voir

que l'on peut obtenir de la même façon une amélioration de certains risques globaux. Par exemple lorsque la vraie densité est effectivement connue en dehors de U on a, pour tout sous-ensemble A de U et tout entier $p \geq 1$ l'inégalité

$$\mathbb{E} \int_A |\tilde{f}_n(y) - f_Q(y)|^p dy \leq \mathbb{E} \int_A |f_n(y) - f_Q(y)|^p dy.$$

Par la suite, on s'intéressera à la fonction de perte quadratique.

Par ailleurs, en pratique on dispose uniquement de l'estimateur f_n par lequel on remplacera f_p dans les formules donnant \tilde{f}_n . On conviendra de noter $f_n^*(x)$ le nouvel estimateur ainsi obtenu. On a donc

$$f_n^*(x) = \frac{1}{n} \sum_{j=1}^n \left\{ \frac{1}{h} K \left(\frac{x - X_j}{h} \right) 1_U(X_j) + m(f_n) 1_{U^c}(X_j) \right\},$$

avec

$$m(f_n) = \frac{\frac{1}{h} \int_{U^c} K \left(\frac{x-y}{h} \right) f_n(y) dy}{\int_{U^c} f_n(y) dy}.$$

Il est facile de vérifier que pour H suffisamment grand $f_n^*(x)$ coïncide avec $f_n(x)$ et que pour presque tout échantillon lorsque $H \rightarrow 0$, $f_n^*(x)$ converge vers

$$\mathbb{E}_{f_n}(f_n(x)) = \frac{1}{h} \int_{\mathbb{R}} K \left(\frac{x-y}{h} \right) f_n(y) dy.$$

Donc, lorsque H varie de 0 à $+\infty$, l'estimateur $f_n^*(x)$ varie de $g_n = f_n * K_h$ à f_n , où $*$ désigne l'opérateur de convolution et $K_h(\cdot) = K(\cdot/h)/h$. En fait, $f_n * K_h$ est aussi un estimateur de Parzen-Rosenblatt de noyau $K * K$. Par conséquent, en faisant varier H on obtient une infinité d'estimateurs à noyau f_n^* variant entre g_n et f_n . Ensuite, si le noyau K est positif et borné par $k \in]0, \infty[$ alors $|f_n^*(x)|$ est borné par k/h . On en déduit donc que $\underline{f}_n^*(x) = \inf_{H>0} f_n^*(x)$ et $\bar{f}_n^*(x) = \sup_{H>0} f_n^*(x)$ sont finis et vérifient:

$$\forall x \in \mathbb{R}, \quad \underline{f}_n^*(x) \leq f_n(x) \leq \bar{f}_n^*(x) \leq \frac{k}{h}. \quad (1)$$

Le théorème suivant donne le comportement de $f_n^*(x)$ en fonction du paramètre H :

THÉORÈME 1. *On suppose que K est une densité de probabilité paire décroissante sur $]0, +\infty[$. En tant que fonction de H , $f_n^*(x)(\cdot)$ est continue sur $]0, +\infty[$ sauf en chaque point $H_i = |x - X_i|$ en lequel elle est continue à gauche et présente un saut positif dont la valeur est*

$$\frac{1}{nh} \left(K \left(\frac{H_i}{h} \right) - \frac{\int_{U_i^c} K \left(\frac{x-y}{h} \right) f_n(y) dy}{\int_{U_i^c} f_n(y) dy} \right),$$

en notant $U_i =]x - H_i, x + H_i[$.

Si la fonction $f_n^*(x)(\cdot)$ admet les deux points de discontinuité successifs $H_i = |x - X_i|$ et $H_j = |x - X_j|$ alors elle est différentiable et décroissante sur $]H_i, H_j[$.

Posons $H_{(1)} = \min_i H_i$. Pour presque tout échantillon, $H_{(1)}$ est strictement positif et $f_n^*(x)(\cdot)$ est différentiable et décroissante sur $]0, H_{(1)}[$.

COROLLAIRE 1. SOUS les hypothèses du théorème 1, les bornes de $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$ sont atteintes sur $\{H_i : 1 \leq i \leq n\}$.

Démonstration du théorème 1. Soit $]L_1, L_2[$ un intervalle de \mathbb{R}^+ ne contenant pas de point de la forme $|x - X_i|$. En remarquant qu'une observation X_i appartient à $U =]x - H, x + H[$ si et seulement si $|x - X_i| < H$, on voit que lorsque H varie de L_1 à L_2 , les points X_i tombant dans U restent les mêmes. Soit $I = \{i : |x - X_i| \leq L_1\}$ et N le cardinal de I . Pour H variant dans $]L_1, L_2[$ la fonction $f_n^*(x)$ s'écrit

$$f_n^*(x) = \frac{1}{nh} \sum_{j \in I} K\left(\frac{x - X_j}{h}\right) + \frac{n - N}{n} m(f_n).$$

La variable H n'intervenant que linéairement dans les bornes des intégrales définissant le numérateur et le dénominateur de $m(f_n)$, on en déduit que $m(f_n)$ est dérivable sur $]L_1, L_2[$ et même continûment dérivable si K est continu. Le signe de la dérivée de $m(f_n)$ par rapport à H est celui de la quantité

$$\begin{aligned} D(H) &= \left\{ \frac{\partial}{\partial H} \int_{U^c} K\left(\frac{x-y}{h}\right) f_n(y) dy \right\} \int_{U^c} f_n(y) dy \\ &\quad - \left(\frac{\partial}{\partial H} \int_{U^c} f_n(y) dy \right) \int_{U^c} K\left(\frac{x-y}{h}\right) f_n(y) dy \\ &= \{f_n(x-H) + f_n(x+H)\} \left\{ \int_{U^c} K\left(\frac{x-y}{h}\right) f_n(y) dy - K\left(\frac{H}{h}\right) \int_{U^c} f_n(y) dy \right\}. \end{aligned}$$

Les hypothèses faites sur K entraînent la négativité de $D(H)$.

Soit $j \in \{1, \dots, n\}$ et J l'ensemble des indices i tels que $|x - X_i| < H_j$. Il est facile de voir que

$$f_n^*(x)(H_j + 0) = \frac{1}{nh} \sum_{i \in J} K\left(\frac{x - X_i}{h}\right) + \frac{1}{nh} K\left(\frac{H_j}{h}\right) + \frac{n-1-#J}{n} m(f_n)(H_j)$$

et que

$$f_n^*(x)(H_j - 0) = \frac{1}{nh} \sum_{i \in J} K\left(\frac{x - X_i}{h}\right) + \frac{n-#J}{n} m(f_n)(H_j).$$

On en déduit aisément la valeur et le signe du saut de $f_n^*(x)(\cdot)$ en H_j . \square

La figure 1 montre l'allure de $f_n^*(0)/f_Q(0)$ en fonction de H pour des échantillons de tailles respectives 5 et 100, issus de la loi $N(0, 1)$. La fenêtre h nécessaire à la construction de f_n étant choisie égale à $1.059n^{-\frac{1}{5}}$, valeur optimisant l'erreur quadratique intégrée asymptotique dans le cas d'une densité et d'un noyau $N(0, 1)$ (Bosq et Lecoutre 1987). La faible taille d'échantillon ($n = 5$) utilisée dans les figures 1 et 2 permet de mieux visualiser les conclusions du théorème 1 pour les fonctions représentées.

La répartition des points de discontinuité de $f_n^*(0)$ traduit celle de la distance des observations au point x . Pour une étude plus fine de la position des observations au voisinage de x (en particulier pour des problèmes de symétrie) on pourra considérer la décomposition de $f_n^*(0)$ en somme de deux fonctions faisant intervenir respectivement les observations tombant à gauche et à droite de x . On donne dans la figure 2 le comportement

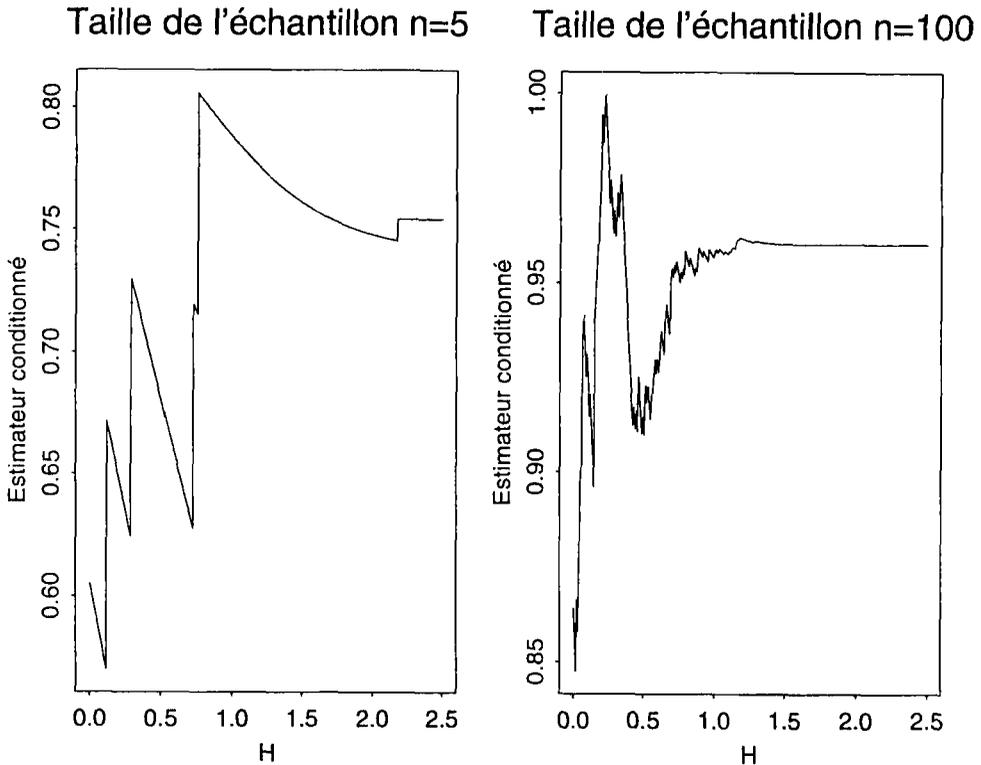


FIGURE 1: Comportement de $f_n^*(0)/f_Q(0)$ en fonction de H pour une loi $N(0,1)$ dans le cas d'un noyau gaussien et d'une fenêtre h asymptotiquement optimale.

de $f_n^*(0)/f_Q(0)$ en fonction des deux paramètres h et H . Comme on peut s'y attendre de par la définition de $f_n^*(x)$, on constate que plus h est petit plus l'écart entre $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$ est important.

Une question importante est bien sûr celle de la probabilité de la présence de $f_Q(x)$ dans l'intervalle $[\underline{f}_n^*(x), \bar{f}_n^*(x)]$. Les considérations théoriques associées sont délicates. Cependant, le fait que $f_n(x)$ vérifie (1) entraîne qu'il existe au moins un $H > 0$ pour lequel $|f_n^*(x) - f_Q(x)| < |f_n(x) - f_Q(x)|$.

Dans la section suivante, nous donnons quelques résultats théoriques sur le risque quadratique associé à $\tilde{f}_n(x)$ ainsi qu'un théorème de convergence presque complète. Ces résultats peuvent être étendus à $f_n^*(x)$ au prix de difficultés techniques supplémentaires qui ne sont pas détaillées ici (Abdous et Berlinet 1993).

3. COMPORTEMENT DE L'ESTIMATEUR CONDITIONNÉ

On suppose que les lois Q et P appartiennent à \mathcal{P} et que le noyau K est de carré intégrable par rapport à tout élément de \mathcal{P} . Dans toute la suite, λ désigne la mesure de Lebesgue sur \mathbb{R} . Nous posons $r = Q(U^c)/P(U^c)$ et lorsque la loi commune des observations est Q , nous notons E l'espérance de $\tilde{f}_n(x)$, V sa variance et R son risque quadratique. Les notations E_0 , V_0 et R_0 seront relatives à $f_n(x)$. Le théorème suivant compare les biais, les variances et les risques quadratiques associés à $\tilde{f}_n(x)$ et à $f_n(x)$.

Taille de l'échantillon n=5

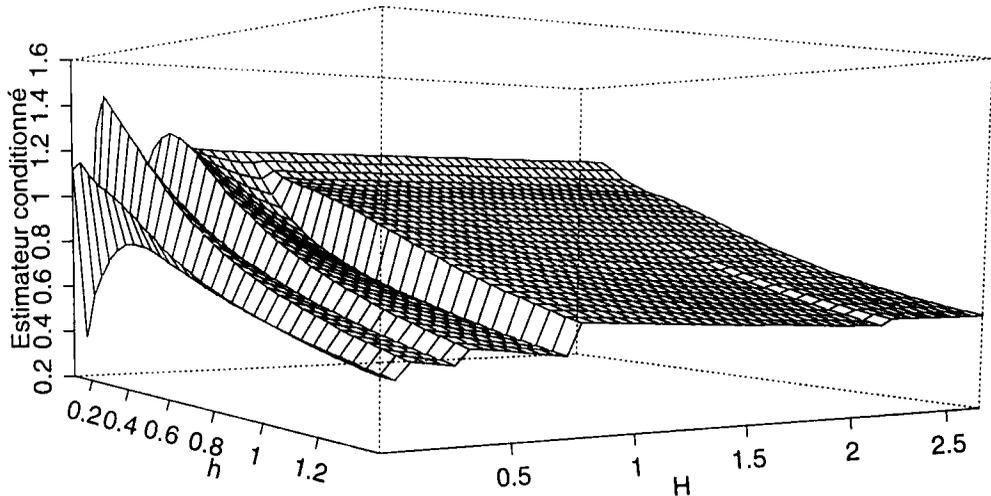


FIGURE 2: Comportement de $\hat{f}_n(0)/f_Q(0)$ en fonction de H et h pour une loi $N(0, 1)$ dans le cas d'un noyau gaussien.

THÉORÈME 2. (Comparaison des risques quadratiques).

(i) Les différences des biais et des variances sont données par:

$$\begin{aligned}
 E_0 - E &= Q(U^c)\{m(Q) - m(P)\}; \\
 V_0 - V &= \frac{1}{n} \left\{ \frac{1}{h} \int_{\Omega} K^2(y) f_Q(x - hy) dy - \{m(P)\}^2 Q(U^c) \right. \\
 &\quad \left. + Q(U^c)\{m(P) - m(Q)\} \right. \\
 &\quad \left. \times \left(Q(U^c)\{m(P) + m(Q)\} + 2 \int_{\Omega} K(y) f_Q(x - hy) dy \right) \right\} \\
 &\geq \frac{1}{n} \left\{ Q(U^c)\{m(Q) - m(P)\} \right. \\
 &\quad \left. \times \left(Q(U)\{m(P) + m(Q)\} + 2 \int_{\Omega} K(y) f_Q(x - hy) dy \right) \right\}
 \end{aligned}$$

(ii) Si $m(P) = m(Q)$ alors:

$$\begin{aligned}
 E_0 - E &= 0; \\
 0 \leq R_0 - R &= V_0 - V \\
 &= \frac{1}{n} \left(\frac{1}{h} \int_{\Omega} K^2(y) f_Q(x - hy) dy - Q(U^c)\{m(Q)\}^2 \right).
 \end{aligned}$$

Démonstration. (i): De l'indépendance des variables X_1, \dots, X_n , il résulte que

$$E = \int_{\Omega} K(y) f_Q(x - hy) dy + r \int_{\Omega} K(y) f_P(x - hy) dy, \tag{2}$$

$$nV = \frac{1}{h} \int_{\Omega} K^2(y) f_Q(x - hy) dy - \left(\int_{\Omega} K(y) f_Q(x - hy) dy \right)^2 + Q(U^c) m(P) \left(Q(U) m(P) - 2 \int_{\Omega} K(y) f_Q(x - hy) dy \right), \tag{3}$$

$$E_0 = \int K(y) f_Q(x - hy) dy,$$

$$nV_0 = \frac{1}{h} \int K^2(y) f_Q(x - hy) dy - \left(\int K(y) f_Q(x - hy) dy \right)^2.$$

D'où l'en déduit les expressions de $E - E_0$ et de $V - V_0$. D'autre part, l'inégalité de Schwarz entraîne que

$$\left\{ \int_{U^c} \frac{1}{h} K \left(\frac{x-y}{h} \right) f_Q(y) dy \right\}^2 \leq Q(U^c) \int_{U^c} \frac{1}{h^2} K^2 \left(\frac{x-y}{h} \right) f_Q(y) dy,$$

ce qui permet d'obtenir la minoration de $V - V_0$.

(ii): Immédiat à partir de (i) et du fait que le risque quadratique est la somme de la variance et du carré du biais. \square

REMARQUES.

(a) L'ensemble des éléments de \mathcal{P} pour lesquels on a diminution du risque quadratique semble être difficile à déterminer en toute généralité. Cependant, le théorème de Rao-Blackwell nous en fournit une partie qui est $\{P \in \mathcal{P} : Q^{\otimes n}|_{(U^n)^c} = P^{\otimes n}|_{(U^n)^c}\}$. Cette partie est élargie par le théorème 2 puisque $R \leq R_0$ sous la condition $m(P) = m(Q)$.

(b) L'hypothèse $m(P) = m(Q)$ entraîne que $R_0 \geq R$ et la quantité $R_0 - R$ reste invariante si on remplace P par P' telle que $m(P) = m(P')$. Si la restriction de K à U^c n'est pas constante Q -presque sûrement, alors $R_0 > R$ i.e. l'amélioration est stricte.

(c) Le théorème 2 n'est pas un résultat asymptotique, il est valide pour chaque n . D'autre part, sous la condition $m(P) = m(Q)$, la réduction du risque est un $O((nh)^{-1})$ et elle dépend de $m(Q)$. En pratique, cette réduction peut être estimée en remplaçant f_Q par f_n .

On se propose maintenant d'étudier le comportement asymptotique du biais et de la variance de $\tilde{f}_n(x)$.

THÉORÈME 3. (Comportement du biais). *Supposons que f_Q soit continue bornée.*

(i) Si $|K|$ est intégrable, $H = H(n) \rightarrow 0$, $h = h(n) \rightarrow 0$ et $H/h \rightarrow \alpha \in [0, \infty]$ lorsque $n \rightarrow \infty$, alors:

$$\lim_{n \rightarrow \infty} \mathcal{E}_Q(\tilde{f}_n(x)) = f_Q(x) \int_{]-\alpha, \alpha[} K(y) dy + f_P(x) \int_{]-\alpha, \alpha[^c} K(y) dy. \tag{4}$$

(ii) Si les versions $(f_Q)_{Q \in \mathcal{P}}$ admettent une dérivée d'ordre 4 en x , la fonction K est paire et satisfait $\int y^j K^j(y) dy < \infty$ pour $j \in \{1, 2\}$ et $i \in \{1, 2, 3, 4\}$, alors

$$\begin{aligned} \{\mathcal{E}_Q(\tilde{f}_n(x)) - f_Q(x)\}^2 &= \{I - f_Q(x)\}^2 \\ &\quad + h^2 \{I - f_Q(x)\} \{f_Q''(x) I_{21} + r f_P''(x) J_{21}\} \\ &\quad + \frac{h^4}{4} \{f_Q''(x) I_{21} + r f_P''(x) J_{21}\}^2 \end{aligned}$$

$$\begin{aligned}
& + \frac{1}{3} \{I - f_Q(x)\} \{f_Q^{(4)} I_{41} + r f_P^{(4)}(x) J_{41}\} \\
& + h^4 o(1),
\end{aligned} \tag{5}$$

où, $I_{kl} = \int_{\Omega} y^k K^l(y) dy$, $J_{kl} = \int_{\Omega} y^k K^l(y) dy$ pour $k = 0, \dots, 4$ et $l = 1, 2$ et $I = f_Q(x) I_{01} + r f_P(x) J_{01}$.

Démonstration. (i): Considérons l'expression de $\mathcal{E}_Q(\tilde{f}_n(x))$ fournie par (2), f_Q étant continue au voisinage de x , la fonction $k(y) f_Q(x - hy) 1_{\Omega}(y)$ converge λ -presque partout vers $f_Q(x) K(y) 1_{]-\alpha, \alpha[}(y)$ lorsque $n \rightarrow \infty$. Par ailleurs

$$|K(y) f_Q(x - hy) 1_{\Omega}(y)| \leq M |K(y)| \quad (\text{où } \sup_y f_Q(y) = M \in]0, \infty[).$$

Donc, une application du théorème de la convergence dominée donne

$$\lim_{n \rightarrow \infty} \int_{\Omega} K(y) f_Q(x - hy) dy = f_Q(x) \int_{]-\alpha, \alpha[} K(y) dy.$$

Par un raisonnement analogue, on obtient

$$\lim_{n \rightarrow \infty} \int_{\Omega} K(y) f_P(x - hy) dy = f_P(x) \int_{]-\alpha, \alpha[} K(y) dy.$$

Enfin, pour aboutir à (4), il suffit de remarquer que $H \rightarrow 0$ entraîne que $r = Q(U^c)/P(U^c) \rightarrow 1$.

(ii): Les conditions imposées à f_Q , f_P et K et une application de la formule de Taylor-Young à l'ordre 4 permettent d'obtenir aisément (5). \square

REMARQUES.

(a) Puisque $\int K(y) dy = 1$, nous avons

$$\lim_{n \rightarrow \infty} \{\mathcal{E}_Q(\tilde{f}_n(x)) - f_Q(x)\} = \{f_P(x) - f_Q(x)\} \int_{]-\alpha, \alpha[} K(y) dy,$$

par conséquent $\tilde{f}_n(x)$ sera asymptotiquement sans biais si $f_P(x) = f_Q(x)$ ou bien si $\int_{]-\alpha, \alpha[} K(y) dy = 0$. Dans le cas où $f_P(x) \neq f_Q(x)$, le biais asymptotique de $\tilde{f}_n(x)$ sera strictement plus petit $|f_P(x) - f_Q(x)|$ si et seulement si $|\int_{]-\alpha, \alpha[} K(y) dy| < 1$.

(b) Nous avons supposé que H dépend de n et converge vers 0 lorsque $n \rightarrow \infty$ car dans le cas contraire, il existe un réel $\beta > 0$ et une sous-suite $(H_{n_k})_k$ telle que $H_{n_k} \geq \beta$, $\forall k$. Ainsi, $\tilde{f}_n(x)$ et $f_n(x)$ sont très voisins puisque dans le cas où $\beta = +\infty$, U contient λ -p.s. toutes les observations, tandis que dans le cas où $\beta \in]0, \infty[$, la quantité $m(P)$ tend vers 0 lorsque $n_k \rightarrow \infty$, par conséquent, les observations qui tombent hors de U auront un poids voisin de 0, i.e. $\tilde{f}_n(x)$ reste voisin de $f_n(x)$.

THÉORÈME 4. (Comportement de la variance). *Si f_Q est continue bornée, K^2 est intégrable, et si $\lim_{n \rightarrow \infty} h = \lim_{n \rightarrow \infty} (nh)^{-1} = 0$ et $\lim_{n \rightarrow \infty} H/h = \alpha \in [0, \infty]$, alors on a $\lim_{n \rightarrow \infty} \text{Var } \tilde{f}_n(x) = 0$ et par conséquent $\lim_{n \rightarrow \infty} R = \{\lim_{n \rightarrow \infty} E - f_Q(x)\}^2$. Plus précisément*

$$\text{Var } \tilde{f}_n(x) = \frac{1}{nh} f_Q(x) \int_{\Omega} K^2(y) dy + o\left(\frac{1}{nh}\right).$$

Démonstration. Il suffit de reprendre l'expression de $\text{Var } \tilde{f}_n(x)$ donnée par (3) et de procéder de la même manière qu'au théorème 3. \square

Le risque quadratique asymptotique associé à $\tilde{f}_n(x)$ est plus difficile à étudier que celui de $f_n(x)$. Afin d'obtenir une expression en $o(1/nh + h^4)$ [comme c'est le cas pour $f_n(x)$], il faut imposer au noyau K de vérifier: $\int_{\Omega} K(y) dy = 1$ et $\int_{\Omega} y K(y) dy = 0$; ce qui présente l'inconvénient de devoir choisir un noyau qui dépend des quantités h et H . Il semble donc difficile d'utiliser le développement asymptotique de R afin de déterminer le pas h et le paramètre H asymptotiquement optimaux.

Dans ce qui suit, nous allons établir une condition suffisante pour que $\tilde{f}_n(x)$ converge presque complètement (p.c.) vers $\bar{f}_Q(x) = \lim_{n \rightarrow \infty} \mathcal{E}_Q(\tilde{f}_n(x))$.

THÉORÈME 5. *Supposons que f_Q et f_P soient continues au voisinage de x et qu'elles soient bornées. Si le noyau K est à variation bornée, si $\lim_{n \rightarrow \infty} H = \lim_{n \rightarrow \infty} h = 0$, $\lim_{n \rightarrow \infty} H/h = \alpha \in [0, \infty]$ et pour tout $\gamma > 0$, $\sum_{n \geq 1} \exp(-\gamma nh^2) < \infty$, alors $\tilde{f}_n(x)$ converge presque complètement vers $\bar{f}_Q(x)$, i.e.*

$$\forall \epsilon > 0, \quad \sum_{n \geq 1} \Pr(|\tilde{f}_n(x) - \bar{f}_Q(x)| \geq \epsilon) < \infty. \tag{6}$$

Démonstration. On a $\tilde{f}_n(x) - \bar{f}_Q(x) = \{\tilde{f}_n(x) - \mathcal{E}_Q(\tilde{f}_n(x))\} + \{\mathcal{E}_Q(\tilde{f}_n(x)) - \bar{f}_Q(x)\}$. Comme le second terme de cette somme converge vers 0 lorsque n tend vers l'infini, il suffit de montrer que $\tilde{f}_n(x) - \mathcal{E}_Q(\tilde{f}_n(x))$ converge presque complètement vers 0. On a presque sûrement:

$$\begin{aligned} |\tilde{f}_n(x) - \mathcal{E}_Q(\tilde{f}_n(x))| &= \left| \{F_n(x+H) - F_Q(x+H)\} \left\{ \frac{1}{h} K\left(-\frac{H}{h}\right) - m(P) \right\} \right. \\ &\quad \left. - \{F_n(x-H) - F_Q(x-H)\} \left\{ \frac{1}{h} K\left(\frac{H}{h}\right) - m(P) \right\} \right. \\ &\quad \left. - \int_U \{F_n(y) - F_Q(y)\} d\left\{ \frac{1}{h} K\left(\frac{x-y}{h}\right) \right\} \right| \\ &\leq \sup_z |F_n(z) - F_Q(z)| \frac{4k + V(K)}{h}, \end{aligned}$$

où F_n et F_Q désignent les fonctions de répartition empirique et théorique respectivement; $V(K)$ est la variation totale de K et $k = \sup_x |K(x)|$. On peut alors conclure en utilisant le fait que

$$\forall n \geq 1, \quad \forall \epsilon > 0, \quad P(\sup_z |F_n(z) - F_Q(z)| > \epsilon) \leq 2 \exp(-2n\epsilon^2). \quad \square$$

REMARQUE. La convergence presque complète de $\tilde{f}_n(x)$ peut être aussi obtenue sous la condition $nh/\log n \rightarrow \infty$, qui est moins forte que $\forall \gamma > 0, \sum_{n \geq 1} \exp(-\gamma nh^2) < \infty$. Cependant, nous avons opté pour cette dernière condition pour la brièveté de la preuve.

4. MISE EN PRATIQUE ET SIMULATIONS

Dans cette section on s'intéresse à la mise en pratique des estimateurs conditionnés. Pour se faire une idée de l'intérêt de l'estimateur $f_n^*(x)$, nous présentons quelques résultats de simulations. Nous avons simulé quatre lois de probabilité: $\mathbf{N}(0, 1)$, $\frac{1}{2}\{\mathbf{N}(-1.0, \frac{4}{9}) + \mathbf{N}(1.0, \frac{4}{9})\}$, $\frac{1}{2}\{\mathbf{N}(-1.5, \frac{1}{4}) + \mathbf{N}(1.5, \frac{1}{4})\}$ et l'exponentielle $f_Q(x) = e^{-x}1_{(x>0)}$. De chacune de ces lois, nous avons tiré 500 échantillons de taille 100. Nous avons utilisé le noyau

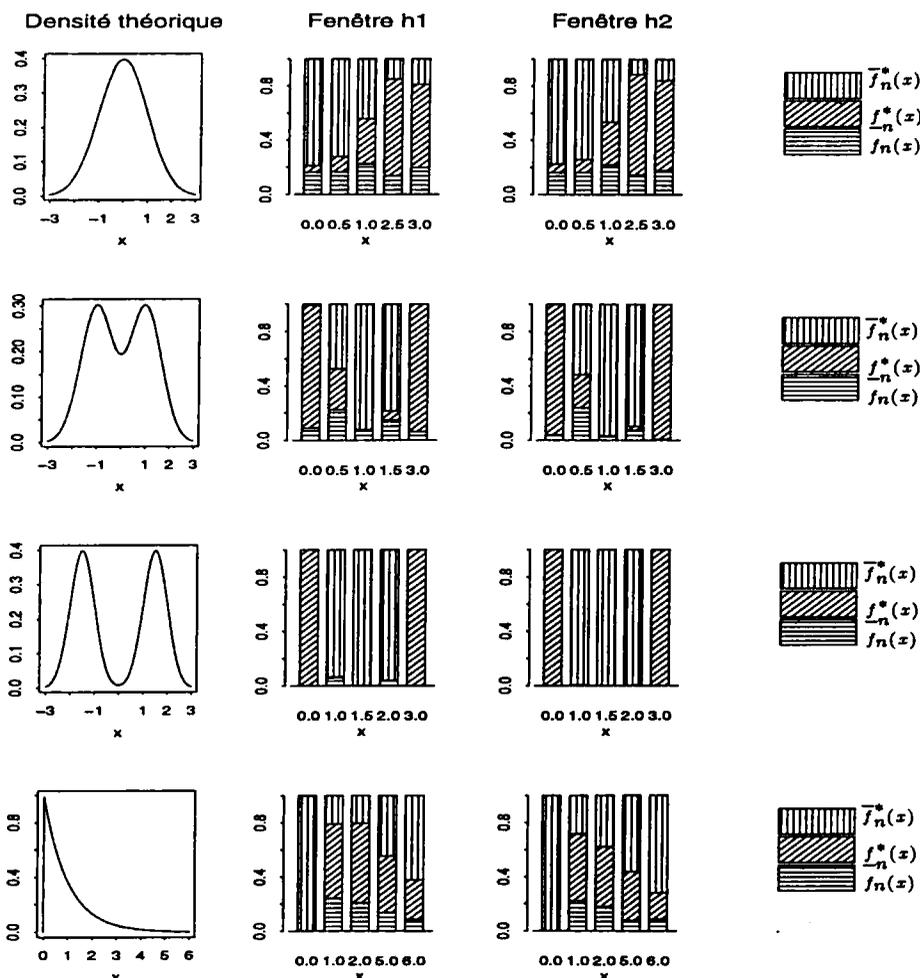


FIGURE 3: Graphiques des proportions p .

$N(0, 1)$ et les fenêtres suivantes: $h_1 = 1.059s_n n^{-\frac{1}{3}}$, où s_n^2 est la variance échantillonnale; et $h_2 =$ fenêtre de Hall *et al.* (1991). Pour chaque fenêtre et chaque échantillon, nous avons évalué les estimateurs $f_n(x)$, $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$, pour différentes valeurs de x . Ensuite pour chaque type d'estimateur, nous avons calculé la proportion de fois (p) que cet estimateur est plus proche de la vraie valeur de la densité $f_Q(x)$ que les deux autres estimateurs. Une première comparaison des estimateurs $f_n(x)$, $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$ peut être réalisée à l'aide des graphiques de la figure 3. En effet, afin de bien situer le point étudié x , nous donnons pour chaque loi simulée le graphe de sa densité, ensuite pour chacune des fenêtres h_1 et h_2 et chaque point x , nous présentons un diagramme des proportions p associées aux estimateurs $f_n(x)$, $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$. Nous constatons que dans une proportion très importante, l'estimateur $\bar{f}_n^*(x)$ est plus proche de $f_Q(x)$ que les deux autres estimateurs lorsque $f_Q(x)$ est grand. Il en est de même pour $\underline{f}_n^*(x)$ lorsque $f_Q(x)$ est faible et ce à l'exception de la loi exponentielle pour laquelle l'effet de bord joue un rôle spécifique. Enfin, pour juger de la qualité et de la variabilité des estimateurs $f_n(x)$,

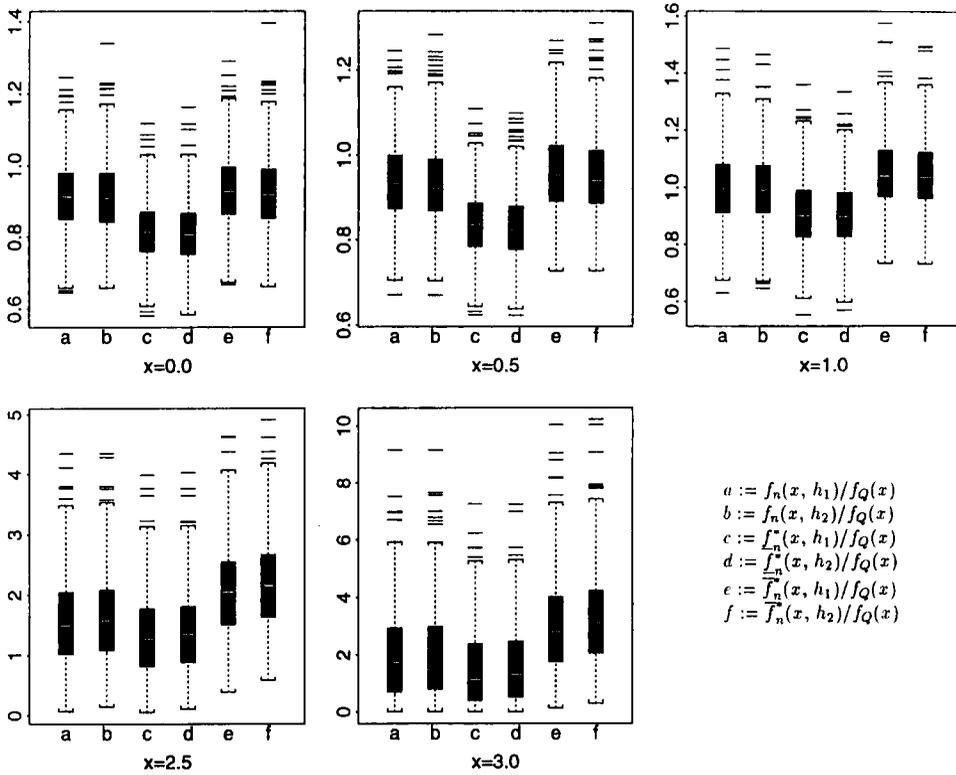


FIGURE 4: Loi $N(0, 1)$: diagrammes en boîte des estimateurs $f_n(x)$, $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$ divisés par $f_Q(x)$.

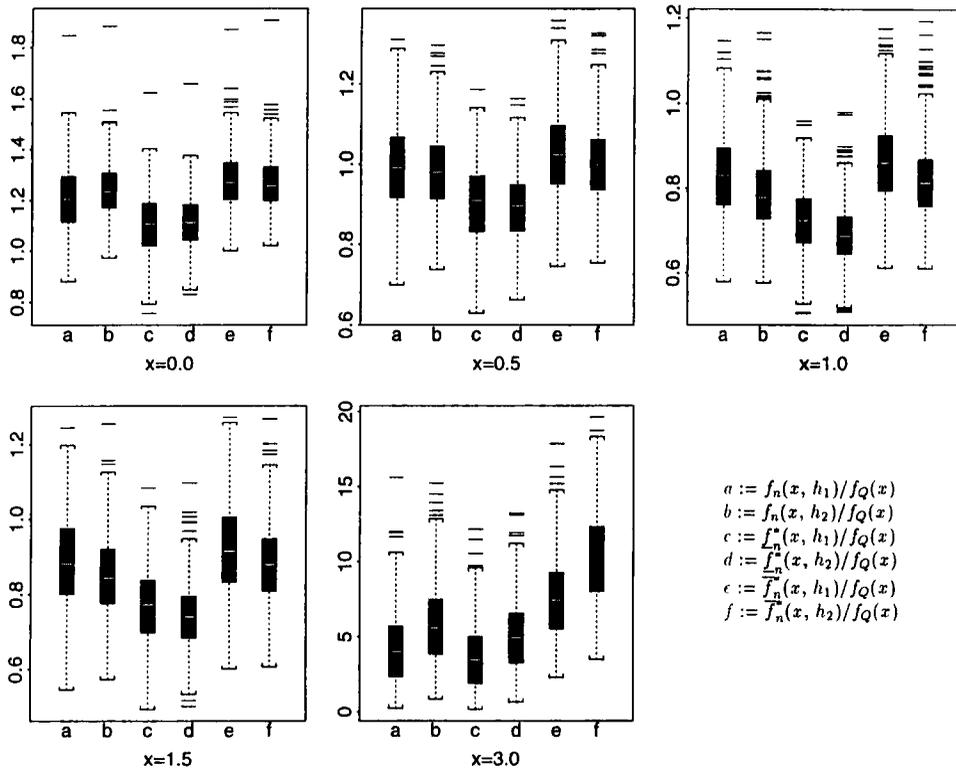


FIGURE 5: Loi $\frac{1}{2} \{N(-1.0, \frac{1}{3}) + N(1.0, \frac{1}{3})\}$: diagrammes en boîte des estimateurs $f_n(x)$, $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$ divisés par $f_Q(x)$.

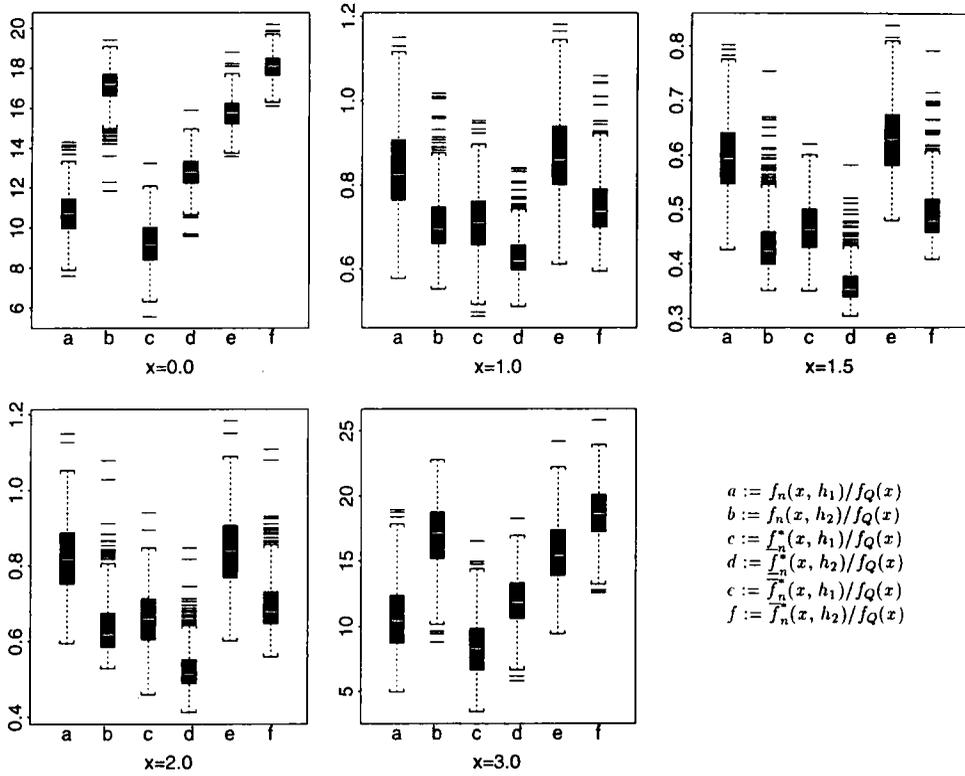


FIGURE 6: Loi $\frac{1}{2} \{N(-1.5, \frac{1}{4}) + N(1.5, \frac{1}{4})\}$: diagrammes en boîte des estimateurs $f_n(x)$, $\underline{f}_n^*(x)$ et $\overline{f}_n^*(x)$ divisés par $f_Q(x)$.

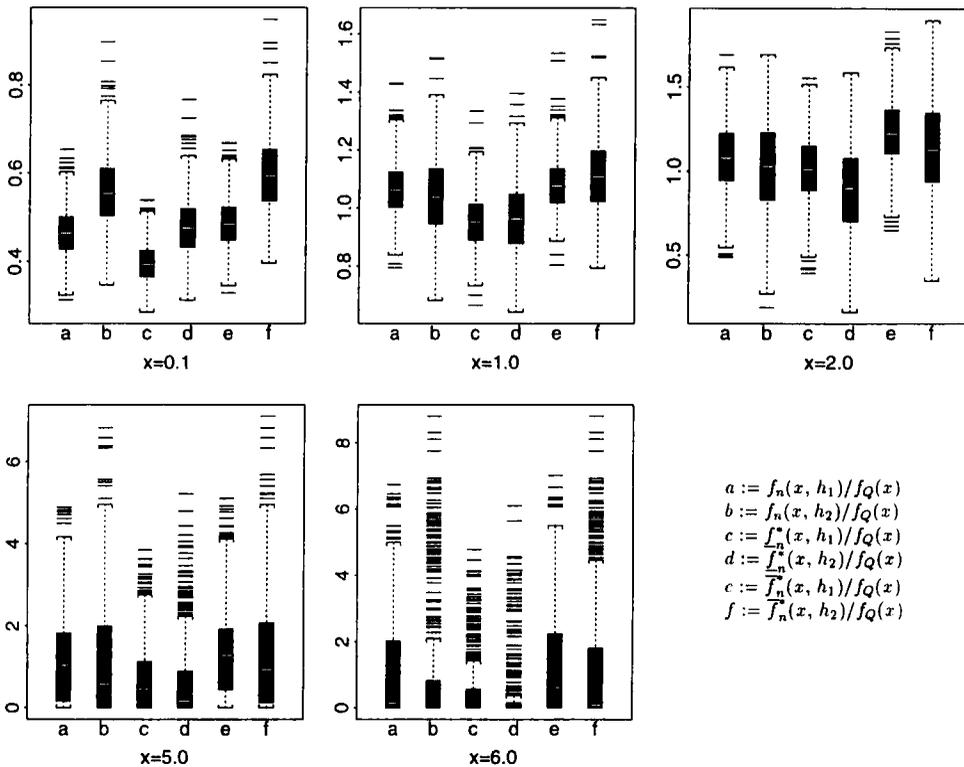


FIGURE 7: Loi exponentielle $f_Q(x) = e^{-x} 1_{(x>0)}$: diagrammes en boîte des estimateurs $f_n(x)$, $\underline{f}_n^*(x)$ et $\overline{f}_n^*(x)$ divisés par $f_Q(x)$.

$\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$, nous avons construit pour chaque loi et chaque point x les diagrammes en boîte (donnant les quartiles et les valeurs extrêmes) des 500 valeurs de $f_n(x, h_1)$, $f_n(x, h_2)$, $\underline{f}_n^*(x, h_1)$, $\underline{f}_n^*(x, h_2)$, $\bar{f}_n^*(x, h_1)$ et $\bar{f}_n^*(x, h_2)$ divisées par $f_Q(x)$ (figures 4 à 7). Les estimateurs $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$ sont plus performants que $f_n(x)$ aux voisinages des minima et maxima respectivement. Les qualités des trois estimateurs se détériorent lorsque x est situé dans la queue de la loi étudiée, phénomène inhérent à la taille d'échantillon retenue.

L'ensemble des résultats précédents et les considérations de l'introduction conduisent à proposer une méthode automatique de choix entre $f_n(x)$, $\underline{f}_n^*(x)$ et $\bar{f}_n^*(x)$. On privilégiera la valeur $\bar{f}_n^*(x)$ au voisinage des modes de f_n , et la valeur $\underline{f}_n^*(x)$ au voisinage des "creux" ou des "plats" de f_n . Cette méthode peut être utilisée de manière interactive ou automatique en analysant les graphes de f_n et de ses deux premières dérivées.

REMERCIEMENTS

Nous remercions les arbitres dont les critiques constructives ont permis d'améliorer la présentation de cet article.

REFERENCES

- Abdous, B., et Berlinet, A. (1993). Conditioning multivariate density estimates: Theoretical aspects. Rapport de recherche n° 94-01. Unité de Biométrie, Université Montpellier II.
- Berlinet, A. (1984). Propriétés locales d'un paramètre. Application à l'estimation. Note aux *C. R. Acad. Sci. Paris*, 298 (14), 345-348.
- Berlinet, A. (1993). Hierarchies of higher order kernels. *Probab. Theory Related Fields*, 94, 489-504.
- Berlinet, A., et Devroye, L. (1990). Estimation d'une densité: Un point sur la méthode du noyau. *Statist. Anal. Données*, 14, 1-32.
- Bosq, D. (1970). Contribution à la théorie de l'estimation fonctionnelle. Thèse de Doctorat de l'Université Pierre et Marie Curie (Paris VI).
- Bosq, D., et Lecoutre, J.P. (1987). *Théorie de l'Estimation Fonctionnelle*. Economica, Paris.
- Hall, P., Sheather, S.J., Jones, M.C., et Marron, J.S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika* 78, 263-269.
- Lehmann, E.L. (1983). *Theory of Point Estimation*. Wiley, New York.

Received 1 October 1992

Revised 15 June 1993

Accepted 1 November 1993

Université du Québec à Trois-Rivières
Département de mathématiques et d'informatique
Trois-Rivières, Québec
Canada G9A 5H7

Université Montpellier II
Unité de Biométrie
9, Place Pierre Viala
34060 Montpellier Cedex
France