

SÉLECTION D'HISTOGRAMMES MODIFIÉS ITÉRÉS

Laurent ROUVIÈRE

*Institut de Recherche Mathématique de Rennes
UMR CNRS 6625, Laboratoire de Statistique
Université Rennes 2-Haute Bretagne, Campus Villejean
Place du Recteur Henri Le Moal, CS 24307, 35043 Rennes Cedex, France*

laurent.rouviere@uhb.fr

Résumé

Les histogrammes modifiés sont des estimateurs de la densité connus pour posséder de bonnes propriétés de convergence au sens des critères de la théorie de l'information. Ces estimateurs sont construits à partir d'une densité de référence g . Dans ce travail, nous envisageons les histogrammes modifiés comme des systèmes dynamiques fonctionnels qui, à une densité g donnée, associent une trajectoire d'estimateurs $\{B^p g\}_{p \geq 0}$. Berlinet et Biau [3] ont montré, sous certaines hypothèses, que cette trajectoire devient presque sûrement stationnaire. La famille d'estimateurs $\{B^p g, p \geq 0\}$ est alors de cardinal fini. Nous présentons deux méthodes permettant de sélectionner automatiquement un estimateur à l'intérieur de cette famille. Ces deux procédures sont ensuite utilisées pour tenter d'améliorer les performances d'estimateurs à noyau de la densité.

Mots-clés – Estimation de densité, estimation non paramétrique, système dynamique.

Classification AMS 2000 : 62G07.

Abstract

Modified histograms are density estimates known to have good consistency properties according to several information theoretic criteria. These estimates are defined from a reference density g . In this paper, modified histograms are viewed as a dynamical system in a functional space which associates to each g a trajectory $\{B^p g\}_{p \geq 0}$. Under some assumptions, Berlinet et Biau [3] have proved that this trajectory is almost surely stationary. Consequently, the family $\{B^p g, p \geq 0\}$ is finite. We study two methods to select automatically an estimate within this family. We then apply these algorithms to try to improve performances of kernel density estimates.

Keywords – Density estimation, nonparametric estimation, dynamical system.

AMS 2000 Classification: 62G07.

1 Introduction

Plaçons-nous sur l'espace mesurable $(\mathbb{R}, \mathcal{B})$, où \mathcal{B} représente la tribu borélienne de \mathbb{R} , et désignons par f et g deux densités de probabilité définies par rapport à la mesure de Lebesgue. On rappelle que la distance L_1 et l'information de *Kullback-Leibler* (ou *entropie relative*) entre f et g sont respectivement définies par :

$$\|f - g\|_1 = \int |f - g| \quad \text{et} \quad D(f, g) = \begin{cases} \int f \log \frac{f}{g} & \text{si } f \ll g \\ \infty & \text{sinon.} \end{cases}$$

Il est bien connu (voir par exemple Kullback [15]) que ces deux quantités sont liées par l'inégalité dite de *Pinsker* :

$$\|f - g\|_1 \leq 2D(f, g).$$

Cette relation implique que l'information de Kullback-Leibler induit une topologie plus forte sur l'espace des densités de probabilité que celle associée à la distance L_1 .

Dans de nombreux domaines de la statistique tels que la compression de données, les réseaux de télécommunications, les problèmes de classification ou encore les réseaux de neurones (voir Berlinet, Vajda et van der Meulen [7]), la convergence L_1 peut se révéler insuffisante et on lui préfère alors la convergence définie par l'information de Kullback-Leibler. Cependant, trouver des estimateurs de la densité convergeant au sens de l'entropie relative peut soulever quelques difficultés. Remarquons, par exemple, que dans le cas de l'estimateur histogramme usuel \hat{f}_n , la quantité $D(f, \hat{f}_n)$ peut être infinie avec une probabilité non nulle.

Afin de pallier cette difficulté, Barron, Györfi et van der Meulen [2] ont montré qu'il était possible de construire un estimateur f_n de f convergeant presque sûrement en entropie relative et en entropie relative moyenne. Cet estimateur, initialement proposé par Barron [1], est appelé *histogramme modifié*. Il est défini à partir du n -échantillon i.i.d. X_1, \dots, X_n d'une variable aléatoire X de \mathbb{R} possédant f (inconnue) comme densité commune de la manière suivante :

- Soit g une densité connue (dite *densité de référence*) associée à la loi de probabilité ν_g (dite *mesure de référence*).
- Soit ℓ un entier tel que $1 \leq \ell$ et soit $h = 1/\ell$.
- Considérons une partition de \mathbb{R} , $P = \{A_1, \dots, A_\ell\}$ telle que $\nu_g(A_i) = h$, $i = 1, \dots, \ell$.
- Alors, en notant $a_n = 1/(nh + 1)$, on définit l'histogramme modifié f_n par :

$$f_n(x) = (1 - a_n) \frac{\mu_n(A(x))}{h} g(x) + a_n g(x) = \frac{n\mu_n(A(x)) + 1}{nh + 1} g(x), \quad (1)$$

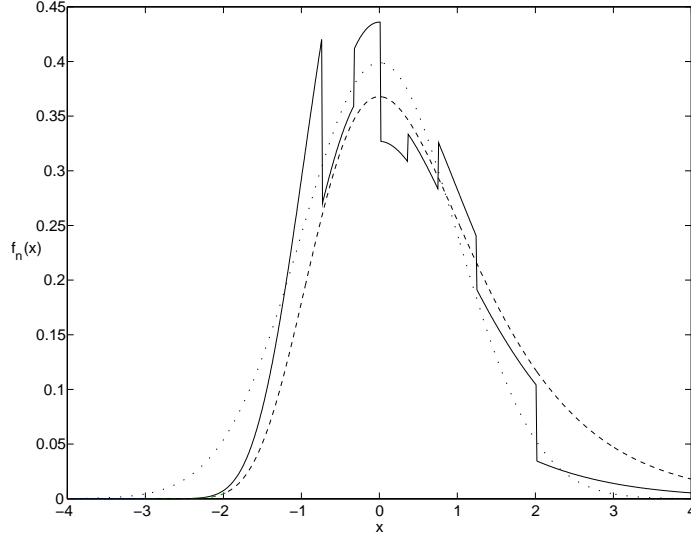


FIG. 1 – Histogramme modifié (trait continu) d’une densité gaussienne $\mathcal{N}(0, 1)$ (pointillés). La densité de référence est une densité de Gumbel (tirets), $n = 100$, $\ell = 8$.

où μ_n désigne la mesure empirique associée à l’échantillon X_1, \dots, X_n et $A(x) = A_i$ si $x \in A_i$.

La première expression de f_n présente cet estimateur comme un mélange entre l’estimateur histogramme usuel $\mu_n(A(x))g(x)/h$ (relatif à la mesure ν_g) et la densité de référence $g(x)$, d’où le nom d’histogramme modifié. La seconde écriture montre que cet estimateur est en fait construit comme une déformation par morceaux de la densité de référence g . Sur chaque cellule A_i de la partition, le coefficient multiplicateur de $g(x)$ vaut

$$\frac{n\mu_n(A_i) + 1}{nh + 1} = \frac{n\mu_n(A_i) + 1}{n\nu_g(A_i) + 1}.$$

Ainsi, sur les classes où la mesure empirique est supérieure à la mesure de référence, on corrige la densité g en la “déformant vers le haut” ; à l’inverse lorsque la mesure empirique est inférieure à la mesure de référence, on déforme g vers le bas (voir Figure 1).

Lorsque le nombre de classes ℓ et l’échantillon sont fixés, l’estimateur (1) peut être vu comme une fonctionnelle admettant en entrée la densité de référence g et fournissant en sortie une nouvelle densité de probabilité (estimateur de f) que l’on peut noter $B_\ell g$. Puisque $B_\ell g$ est une densité de probabilité, nous pouvons alors la considérer comme densité de référence de l’histogramme modifié : cela conduit à un nouvel estimateur de f , noté $B_\ell^2 g$, qui peut à son tour devenir densité de référence et ainsi de suite... Ce processus

itératif nous amène à considérer le *système dynamique* B_ℓ défini de la façon suivante :

$$\begin{aligned} B_\ell &: \mathcal{D} \rightarrow \mathcal{D} \\ g &\mapsto B_\ell g, \end{aligned} \tag{2}$$

où \mathcal{D} , l'espace d'états, représente l'ensemble des densités de probabilité sur \mathbb{R} .

On dira qu'une densité g de \mathcal{D} est *stationnaire* pour B_ℓ si elle vérifie l'équation $B_\ell g = g$. Une trajectoire $\{B_\ell^p g\}_{p \geq 0}$ contenant une telle densité est dite stationnaire. Dans ce cas, le système n'évolue plus après un nombre *fini* d'itérations. Le système dynamique (2) a été étudié par Berlinet et Biau [3] qui ont montré la stationnarité de la trajectoire $\{B_\ell^p g\}_{p \geq 0}$ sous certaines hypothèses. Ceci engendre, sous ces mêmes hypothèses, que la famille d'estimateurs

$$\{B_\ell^p g : p \geq 0\} \tag{3}$$

est de cardinal fini. Dans ce travail, nous prolongeons les idées de Berlinet et Biau [3] en étudiant deux méthodes (une pour le critère L_1 , l'autre pour le critère de Kullback-Leibler) permettant de sélectionner automatiquement un estimateur à l'intérieur de la famille (3).

La suite de ce travail se divise en trois parties. Dans la première partie (Section 2) nous rappelons les principaux résultats obtenus par Berlinet et Biau [3] relatifs au système dynamique (2). Puis, dans une deuxième partie (Section 3), nous présentons les deux procédures de sélection. Enfin, dans la dernière partie, nous utilisons les méthodes présentées précédemment pour tenter d'améliorer les performances d'un estimateur à noyau de la densité (Section 4).

2 Etude du système dynamique

Dans cette partie, nous rappelons les principaux résultats concernant le système dynamique (2). Pour plus de détails et pour les preuves de ces résultats, nous renvoyons le lecteur à Berlinet et Biau [3].

Nous commençons par une remarque élémentaire : une densité g est stationnaire pour B_ℓ si et seulement si $\mu_n(A_i) = h$ pour $i = 1, \dots, \ell$. Ceci ne peut se produire que si ℓ (le nombre de classes) est un diviseur de n (le nombre d'observations), ce que nous supposons dans la suite*. Pour tout $p \in \mathbb{N}$, on note :

- $q_1^p, q_2^p, \dots, q_{\ell-1}^p$ les quantiles d'ordre ih ($i = 1, \dots, \ell - 1$) de la densité $B_\ell^p g$ et on désigne par $\{A_1^p, \dots, A_\ell^p\}$ la partition associée, *i.e.*,

$$\{A_1^p, \dots, A_\ell^p\} = \{] - \infty, q_1^p],]q_1^p, q_2^p], \dots,]q_{\ell-1}^p, \infty[\};$$

*Si ℓ n'est pas un diviseur de n , une stratégie consisterait à choisir un entier k le plus petit possible tel que ℓ divise $n - k$. On ne considérerait alors que les $n - k$ premières observations.

- $\alpha_1^p, \dots, \alpha_\ell^p$ les coefficients associés à la partition $\{A_1^p, \dots, A_\ell^p\}$, *i.e.*,

$$\alpha_i^p = \frac{n\mu_n(A_i^p) + 1}{nh + 1}, \quad i = 1, \dots, \ell.$$

Avec ces notations, les densités B_ℓ^{p+1} et B_ℓ^p sont liées par la relation :

$$B_\ell^{p+1}g = \sum_{i=0}^{\ell-1} \alpha_{i+1}^p \mathbf{1}_{[q_i^p, q_{i+1}^p]} B_\ell^p g,$$

où $\mathbf{1}_A$ désigne la fonction indicatrice de l'ensemble A . Par conséquent, l'étude de la trajectoire $\{B_\ell^p g\}_{p \geq 0}$ se réduit à l'étude des suites $\{(q_1^p, \dots, q_{\ell-1}^p)\}_{p \geq 0}$ et $\{(\alpha_1^p, \dots, \alpha_\ell^p)\}_{p \geq 0}$.

Effectuons, lorsque $\ell \geq 3$, l'hypothèse suivante :

HYPOTHÈSE (\mathcal{H}) : pour tout $i = 2, \dots, \ell - 1$, on a pour p assez grand

$$q_i^p > X_{\left(\frac{in}{\ell} - 1\right)},$$

où $X_{(1)}, \dots, X_{(n)}$ désigne le vecteur des statistiques d'ordre des observations X_1, \dots, X_n .

Berlinet et Biau [3] ont alors démontré le théorème suivant :

Théorème 2.1 *Supposons que ℓ divise n et que l'hypothèse (\mathcal{H}) soit vérifiée dès que $\ell \geq 3$. Alors chaque suite $\{q_i^p\}_{p \geq 0}$ ($i = 1, \dots, \ell - 1$) devient presque sûrement stationnaire après un nombre fini d'itérations.*

On en déduit facilement le corollaire suivant :

Corollaire 2.1 *Sous les hypothèses du Théorème 2.1 :*

- Chaque suite $\{\alpha_i^p\}_{p \geq 0}$ ($i = 1, \dots, \ell$) atteint la valeur 1 après un nombre fini d'itérations.
- La suite de densités $\{B_\ell^p g\}_{p \geq 0}$ est presque sûrement stationnaire.

Parmi les nombreuses simulations que nous avons effectuées, nous n'avons jamais rencontré une situation où la convergence n'ait pas lieu, ce qui laisse sous-entendre que la condition (\mathcal{H}) est vérifiée pour une grande famille de densités g . Ainsi, lorsque le nombre de classes ℓ divise le nombre d'observations n , les suites de quantiles $\{q_i^p\}_{p \geq 0}$ et de coefficients $\{\alpha_i^p\}_{p \geq 0}$ deviennent stationnaires après un nombre fini d'itérations (voir Figure 2). Le système dynamique B_ℓ engendre alors naturellement une famille d'estimateurs $\{B_\ell^p g : p \geq 0\}$ de cardinal *fini*. On notera P_ℓ ce cardinal.

Bien que le choix de la densité de référence n'affecte le comportement asymptotique de l'histogramme modifié qu'au travers des constantes (voir par exemple Berlinet et Brunel [5]), ce choix se révèle crucial à distance finie (voir Berlinet, Biau et Rouvière [4]). Sur

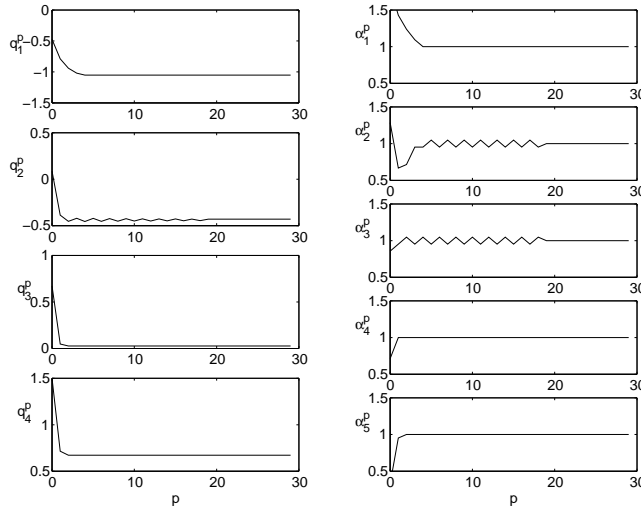


FIG. 2 – Evolution des quantiles (gauche) et des coefficients (droite) obtenus avec une densité initiale g de Gumbel. La densité à estimer est une gaussienne $\mathcal{N}(0, 1)$, $n = 100$, $\ell = 5$, $n/\ell = 20$.

l'exemple de la Figure 4 nous voyons que les erreurs L_1 commises par les premiers itérés sont très élevées (proches de 2) en comparaison aux derniers itérés (proches de 0.5). On peut alors se poser le problème du choix d'un estimateur à l'intérieur de cette famille. Une première idée consiste à choisir comme estimateur de f la densité stationnaire. Comme nous le montre la Figure 3, ce choix est en effet particulièrement efficace lorsque la densité de référence est "éloignée" de la densité à estimer. En revanche, toujours sur ce même exemple, l'erreur L_1 minimale n'est pas commise par la densité stationnaire mais par le cinquième itéré (voir Figure 4).

Nous nous posons ainsi le problème de sélectionner dans la famille finie $\{B_\ell^p g, p \geq 0\}$, un estimateur particulier $B_\ell^{p_0} g$ tel que

$$p_0 \in \operatorname{argmin}_{p \geq 0} \{\|f - B_\ell^p g\|_1\} \quad \text{ou} \quad p_0 \in \operatorname{argmin}_{p \geq 0} \{D(f, B_\ell^p g)\}.$$

Le problème auquel nous sommes confrontés est qu'en pratique la densité f est inconnue. Par conséquent nous ne sommes pas à même de calculer les erreurs $\|f - B_\ell^p g\|_1$ et $D(f, B_\ell^p g)$. Il nous faut donc définir une stratégie permettant de choisir automatiquement un estimateur dans la famille $\{B_\ell^p g : p \geq 0\}$ à partir de l'échantillon X_1, \dots, X_n .

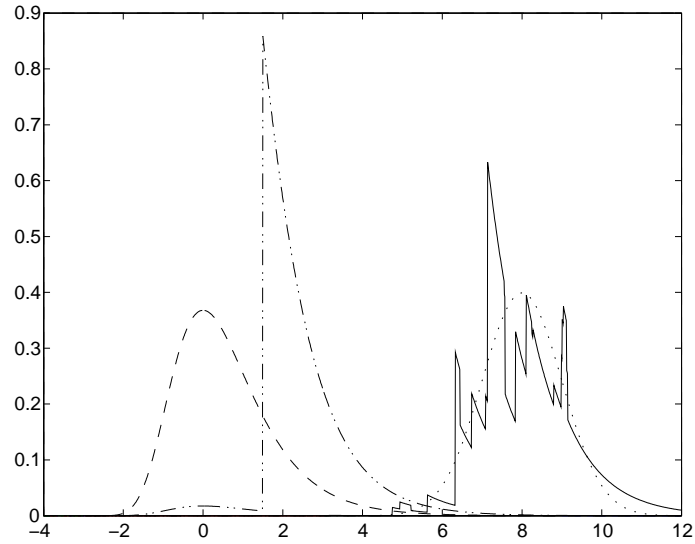


FIG. 3 – Comparaison du premier itéré (points-tirets) avec l'estimateur stationnaire (trait plein) d'une densité gaussienne $\mathcal{N}(8, 1)$ (pointillés). La densité de référence est une densité de Gumbel (tirets), $n = 100$, $\ell = 5$, $n/\ell = 20$.

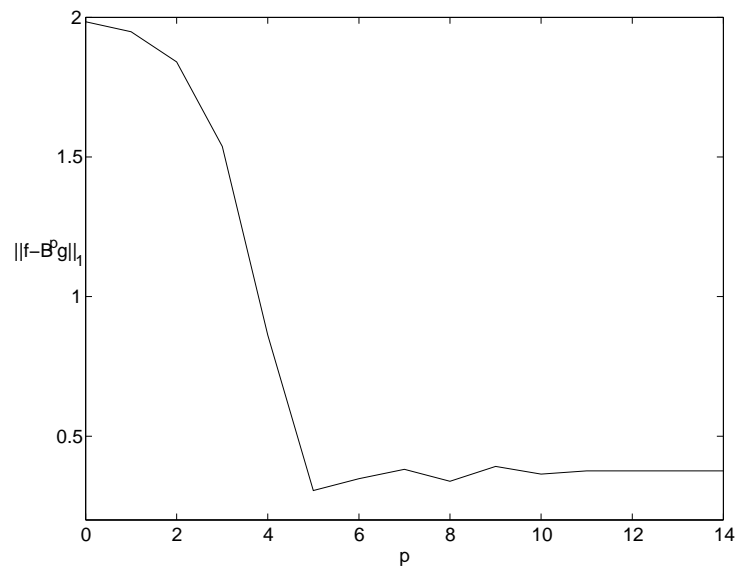


FIG. 4 – Erreurs L_1 commises par les 15 premiers itérés de l'exemple de la Figure 3.

3 Les procédures de sélection

Nous présentons dans cette section deux algorithmes (un pour le critère L_1 , l'autre pour le critère de Kullback-Leibler) permettant de sélectionner automatiquement un estimateur de la densité dans une famille d'estimateurs candidats. Le modèle mathématique se présente de la manière suivante. Soit g une densité de référence fixée. Nous n'effectuons aucune hypothèse sur cette densité. En outre, ce peut être un estimateur de la densité cible f dépendant des observations X_1, \dots, X_n . Pour chaque entier ℓ et p , on note $f_{n,\theta}$ le p -ième histogramme modifié itéré (défini par (1) et (2)) construit à partir de ℓ classes, $\theta = (\ell, p)$. Etant donné l'ensemble

$$\Theta = \{(\ell, p), \ell \text{ divise } n \text{ et } p \geq 0\}, \quad (4)$$

nous proposons maintenant deux procédures permettant de sélectionner un estimateur particulier dans la famille (finie) $\{f_{n,\theta}, \theta \in \Theta\}$.

3.1 Critère L_1

Dans un ouvrage récent, Devroye et Lugosi [14] explorent de nouvelles méthodes permettant de sélectionner automatiquement les paramètres d'estimateurs de la densité. Nous présentons l'algorithme de sélection dans un contexte général. Soit $m < n$ un entier qui partage l'échantillon en deux sous-échantillons. On notera dans cette partie $\mathcal{F}_\Theta = \{f_{n-m,\theta}, \theta \in \Theta\}$ une famille d'estimateurs candidats construits sur les $n - m$ premières observations et paramétrée par $\theta \in \Theta$.

Rappelons à ce stade que le critère L_1 possède une signification claire en terme de probabilité grâce à l'égalité dite de Scheffé [18]

$$\int |f_{n-m,\theta} - f| = 2 \sup_{B \in \mathcal{B}} \left| \int_B f_{n-m,\theta} - \int_B f \right|,$$

le second terme étant égal à deux fois la *distance en variation totale* entre les mesures associées à $f_{n-m,\theta}$ et f . Le principe de la méthode combinatoire développée par Devroye et Lugosi [14] consiste à minimiser un critère empirique du genre :

$$\sup_{A \in \mathcal{A}} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right|, \quad (5)$$

où μ_m désigne la mesure empirique associée à l'échantillon X_{n-m+1}, \dots, X_n et \mathcal{A} représente une classe d'ensembles judicieusement choisie. Remarquons d'emblée que si \mathcal{A} désigne la tribu borélienne de \mathbb{R} , alors la quantité (5) à optimiser est constamment égale à 1. Bien entendu, cela ne se produit plus si on considère une classe d'ensembles plus petite : on minimise alors un critère "plus petit" que la distance en variation totale.

Devroye et Lugosi [14], en s'inspirant des travaux de Yatracos [23], ont montré dans leur ouvrage qu'il suffit de considérer la classe d'ensembles (appelée *classe de Yatracos* associée à Θ)

$$\mathcal{A}_\Theta = \left\{ \{x : f_{n-m,\theta}(x) > f_{n-m,\theta'}(x)\} : (\theta, \theta') \in \Theta^2 \right\}.$$

Ainsi, en notant

$$\Delta(f_{n-m,\theta}) = \sup_{A \in \mathcal{A}_\Theta} \left| \int_A f_{n-m,\theta} - \mu_m(A) \right|,$$

la procédure de sélection consiste à choisir une densité $f_{n-m,\theta}$ dans la famille \mathcal{F}_Θ qui minimise

$$\Delta(f_{n-m,\theta}) + \frac{1}{n},$$

le terme $1/n$ visant simplement à assurer l'existence d'une telle densité. Le candidat sélectionné est appelé *estimateur de la distance minimum* et sera noté f_n . Devroye et Lugosi [14] ont montré qu'un tel estimateur vérifie l'inégalité oracle suivante :

$$\int |f_n - f| \leq 3 \inf_{\theta \in \Theta} \int |f_{n-m,\theta} - f| + 4\Delta(f) + \frac{3}{n}. \quad (6)$$

Le terme $\inf_{\theta \in \Theta} \int |f_{n-m,\theta} - f|$ représente la plus petite erreur qui puisse être commise lorsque l'on approche f par un élément de $\{f_{n-m,\theta}, \theta \in \Theta\}$. Evidemment, la valeur de ce terme d'erreur optimale, qui dépend de la cible f , nous est inconnue. Heuristiquement, l'inégalité (6) signifie donc que l'erreur commise par l'estimateur f_n ne dépasse pas trois fois l'erreur minimum sur la classe plus un terme résiduel, $\Delta(f)$, qu'il va falloir s'attacher à contrôler. Ce contrôle peut être effectué via un détour par la théorie de Vapnik et Chervonenkis [22] sur la convergence uniforme de la mesure empirique.

Rappelons que le coefficient de pulvérisation $\mathcal{S}_{\mathcal{A}_\Theta}(m)$ d'un ensemble de m points par la classe d'ensembles \mathcal{A}_Θ est défini par

$$\mathcal{S}_{\mathcal{A}_\Theta}(m) = \max_{x_1, \dots, x_m \in \mathbb{R}} \text{Card}\{\{x_1, \dots, x_m\} \cap A : A \in \mathcal{A}_\Theta\}.$$

Dit autrement, le coefficient de pulvérisation n'est autre que le nombre maximum de sous-ensembles de m points pouvant être obtenus à l'aide de recouvrements par des ensembles de \mathcal{A}_Θ . Des arguments de nature combinatoire (voir Vapnik et Chervonenkis [22]) montrent que

$$\mathbf{E}\{\Delta(f)\} \leq 2\mathbf{E}\left\{\sqrt{\frac{\log 2\mathcal{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\}.$$

Cette majoration, combinée avec l'inégalité (6), nous conduit à l'inégalité suivante

$$\mathbf{E}\left\{\int |f_n - f|\right\} \leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} + 8\mathbf{E}\left\{\sqrt{\frac{\log 2\mathcal{S}_{\mathcal{A}_\Theta}(m)}{m}}\right\} + \frac{3}{n} \quad (7)$$

qui est centrale dans les travaux de Devroye et Lugosi [14]. On remarquera que ce résultat est *non-asymptotique* (nul besoin d'un passage à la limite pour avoir de l'information) et qu'il ne nécessite *aucune hypothèse de régularité* sur la densité cible f . Il nous semble que ces faits sont suffisamment rares pour mériter d'être soulignés. La seule difficulté, qui réside dans le calcul de $\mathcal{S}_{\mathcal{A}_\Theta}(m)$, est désormais de nature combinatoire.

Dans le contexte qui est le nôtre, nous rappelons que $f_{n-m,\theta}$ désigne le p -ième histogramme modifié itéré construit à partir de ℓ classes et des $n - m$ premières données. Afin que la famille $\mathcal{F}_\Theta = \{f_{n-m,\theta} : \theta \in \Theta\}$ soit finie, nous cherchons à sélectionner θ dans l'ensemble

$$\Theta = \{\theta = (\ell, p) : \ell \text{ divise } n - m \text{ et } p \geq 0\}.$$

Nous rappelons que P_ℓ désigne le cardinal de la famille d'estimateurs itérés construits à partir de ℓ classes. Ainsi, en notant $P_{n,m} = \max_{\ell: \ell \text{ divise } n-m} P_\ell$, on a

$$\text{Card}\{\mathcal{F}_\Theta\} = \sum_{\ell: \ell \text{ divise } n-m} P_\ell \leq P_{n,m}(n - m).$$

Il est également facile de voir que lorsque la famille d'estimateurs \mathcal{F}_Θ est finie, on a

$$\mathcal{S}_{\mathcal{A}_\Theta}(m) \leq \text{Card}\{\mathcal{A}_\Theta\} \leq \text{Card}\{\mathcal{F}_\Theta\}(\text{Card}\{\mathcal{F}_\Theta\} - 1) \leq P_{n,m}^2(n - m)^2.$$

On déduit alors de (7) l'inégalité suivante pour l'estimateur de la distance minimum

$$\begin{aligned} \mathbf{E}\left\{\int |f_n - f|\right\} &\leq 3 \inf_{\theta \in \Theta} \mathbf{E}\left\{\int |f_{n-m,\theta} - f|\right\} \\ &\quad + 8\mathbf{E}\left\{\sqrt{\frac{\log 2 + 2 \log(P_{n,m}(n - m))}{m}}\right\} + \frac{3}{n}. \end{aligned}$$

3.2 Critère de Kullback-Leibler

Dans un travail récent, Berlinet et Brunel [5] ont montré que le critère de validation croisée au sens de l'information de Kullback-Leibler possédait de bonnes propriétés pour sélectionner le nombre de classes des histogrammes modifiés univariés. Nous élargissons ici ce critère à la sélection de θ dans

$$\Theta = \{\theta = (\ell, p) : \ell \text{ divise } n \text{ et } p \geq 0\}.$$

Nous rappelons que l'information de Kullback-Leibler entre la densité à estimer f et l'estimateur $f_{n,\theta}$ s'écrit

$$\begin{aligned} D(f, f_{n,\theta}) &= \int f \log f - \int f \log f_{n,\theta} \\ &= \mathbf{E}\left\{\log f(X)\right\} - \mathbf{E}\left\{\log f_{n,\theta}(X) \mid X_1, \dots, X_n\right\}. \end{aligned}$$

La première quantité est l'opposée de l'entropie de f que nous supposons finie. Bien entendu, elle ne dépend pas de θ et n'intervient pas dans la minimisation de l'erreur. Quant au second terme, qui dépend de la densité f inconnue, il peut être approché par :

$$-\frac{1}{n} \sum_{i=1}^n \log f_{n,\theta}^i(X_i)$$

où $f_{n,\theta}^i$ désigne l'estimateur $f_{n,\theta}$ amputé de la i -ème observation, *i.e.*,

$$f_{n,\theta}^i(x) = \frac{n\mu_n^i(A(x)) + 1}{nh + 1} g(x) \quad \text{avec} \quad \mu_n^i(A(x)) = \frac{1}{n-1} \sum_{j \neq i} \mathbf{1}_{\{X_j \in A(x)\}}.$$

Le critère de validation croisée consiste alors à choisir θ dans Θ qui minimise

$$-\frac{1}{n} \sum_{i=1}^n \log f_{n,\theta}^i(X_i). \tag{8}$$

Remarquons que ce critère peut être aussi interprété comme l'estimateur de validation croisée du maximum de vraisemblance. En effet, choisir θ qui minimise (8) équivaut à maximiser

$$\prod_{i=1}^n f_{n,\theta}^i(X_i),$$

qui apparaît comme un estimateur de la vraisemblance.

La méthode de validation croisée au sens de l'information de Kullback-Leibler révèle un certain nombre de faiblesses pour les estimateurs non paramétriques classiques tels que les histogrammes ou les estimateurs à noyau (Brunel [10], Chapitre 1). Les résultats relatifs à ces estimateurs nécessitent des hypothèses précises sur la densité à estimer (notamment sur le comportement des queues de distribution de f). En revanche, les travaux de Berlinet et Brunel [5] ont mis en évidence l'intérêt de cette méthode dans le contexte des histogrammes modifiés. Ces auteurs ont, en particulier, montré que ce critère de validation croisée est asymptotiquement optimal pour sélectionner le nombre de classes ℓ de ces estimateurs.

4 Application aux estimateurs à noyau

Dans cette partie, nous appliquons les méthodes de sélection présentées précédemment au problème suivant. Soit un n -échantillon i.i.d. X_1, \dots, X_n issu d'une variable aléatoire X admettant f comme densité commune. Etant donné K une fonction réelle, positive,

d'intégrale 1 sur \mathbb{R} et h un entier strictement positif, on définit l'estimateur g_n associé au noyau K et à la fenêtre (ou paramètre de lissage) h par

$$g_n(x) = \frac{1}{nh} \sum_{i=1}^n K\left(\frac{x - X_i}{h}\right). \quad (9)$$

Nous proposons, dans cette section, d'utiliser cet estimateur comme densité initiale du système dynamique (2). Pour chaque entier ℓ et p , on notera $f_{n,\theta}$ le p -ième histogramme modifié construit à partir de ℓ classes ($\theta = (\ell, p)$) et de la densité initiale (9). Nous adaptons, dans un premier temps, les deux procédures de sélection présentées dans la section précédente pour choisir un estimateur $f_{n,\theta}$ particulier. Puis, dans un second temps, nous comparons les performances de l'estimateur sélectionné à celles de l'estimateur à noyau.

L'estimateur à noyau g_n est flexible, dans la mesure où il laisse à l'utilisateur une grande latitude non seulement dans le choix du noyau K , mais encore dans le choix du paramètre réel h . Lorsqu'on se limite aux noyaux K positifs, les vitesses de convergence varient peu en fonction de K et les critères essentiels du choix du noyau sont alors la régularité de la courbe à obtenir d'une part, la simplicité et la vitesse de calcul d'autre part. C'est pour cette dernière raison que, dans cette étude, nous nous limiterons à un noyau gaussien.

En revanche, le choix du paramètre de lissage h se révèle crucial aussi bien pour la précision locale que globale de l'estimateur g_n . A ce jour, de nombreuses méthodes de sélection automatique ont été proposées, testées et comparées (voir par exemple Marron [16], Berlinet et Devroye [6]). Nous reprenons ici deux de ces méthodes que nous résumons brièvement dans le paragraphe suivant.

4.1 Les estimateurs à noyau considérés

Plug-in L_2 : si $h \rightarrow 0$ et $nh \rightarrow \infty$ lorsque $n \rightarrow \infty$, alors sous des hypothèses standard concernant la régularité de la densité f , la fenêtre qui minimise l'erreur quadratique intégrée moyenne est de la forme :

$$h_{pi} = \frac{\int K^2}{(\int t^2 K)^2} \left(\int f''^2 \right)^{-1/5} n^{-1/5} \quad (10)$$

(voir Bosq et Lecoutre [9], Simonoff [20], Tsybakov [21]). Outre sa nature asymptotique, la largeur de la fenêtre optimale dépend de la densité cible f au travers du paramètre $\int f''^2$ et ne peut donc être utilisée telle quelle dans les calculs. Une façon classique de remédier à ce dernier problème consiste à remplacer la quantité $\int f''^2$ par un estimateur approprié. Cette approche conduit à un ensemble de méthodes que l'on a coutume de regrouper sous le vocable général de *méthodes plug-in*, et qui ont fait l'objet d'une recherche active (voir

par exemple Nadaraya [17], Sheater et Jones [19] ou encore Biau [8]). L'approche *plug-in* que nous utilisons ici suggère de choisir la valeur de $\int f''^2$ associée à la densité normale

$$\frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{x^2}{2\sigma^2}\right),$$

où σ^2 sera estimé par la variance empirique $S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$, avec $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$. Pour ce choix, lorsque K est le noyau gaussien, la fenêtre (10) s'écrit alors

$$h_{pi} \simeq 1.059 \frac{S_n}{n^{1/5}},$$

(voir Deheuvels [12], Deheuvels et Hominal [13]). On notera $g_{nh_{pi}}$ l'estimateur à noyau associé à cette fenêtre. Bien entendu, même si n est suffisamment grand, rien d'un point de vue théorique ne nous assure de bonnes performances de l'estimateur associé à la fenêtre (10) pour les critères L_1 et de Kullback-Leibler. Cependant, les méthodes *plug-in* L_2 étant souvent utilisées, il n'est pas incohérent d'étudier leurs performances pour des critères autres que le critère L_2 .

La méthode du double noyau : dans la méthode du double noyau (Berlinet et Devroye [6]), on considère deux noyaux différents K et L dont les fonctions caractéristiques associées ne coïncident sur aucun voisinage ouvert de l'origine. Le noyau K sera le noyau gaussien et on notera g_{nh}^1 l'estimateur associé à ce noyau et à la fenêtre h . Comme suggéré par Berlinet et Devroye [6], nous considérons comme second noyau le noyau polynomial L défini par

$$L(x) = \begin{cases} \frac{7-31x^2}{4} & \text{si } |x| \leq 1/2 \\ \frac{x^2-1}{4} & \text{si } 1/2 \leq |x| \leq 1 \\ 0 & \text{si } 1 \leq |x| \end{cases}$$

et nous désignons par g_{nh}^2 l'estimateur associé à ce nouveau noyau L et à la fenêtre h . Le paramètre de lissage h_{dn} est sélectionné en minimisant la distance L_1 entre g_{nh}^1 et g_{nh}^2 (voir Figure 5), c'est-à-dire

$$h_{dn} = \operatorname{argmin}_{h>0} \int |g_{nh}^1 - g_{nh}^2|.$$

Berlinet et Devroye [6] ont étudié les propriétés de l'estimateur sélectionné, en montrant notamment que le résultat de convergence suivant est vérifié :

$$\lim_{n \rightarrow \infty} \mathbf{E} \left\{ \int |g_{nh_{dn}}^1 - f| \right\} = 0.$$

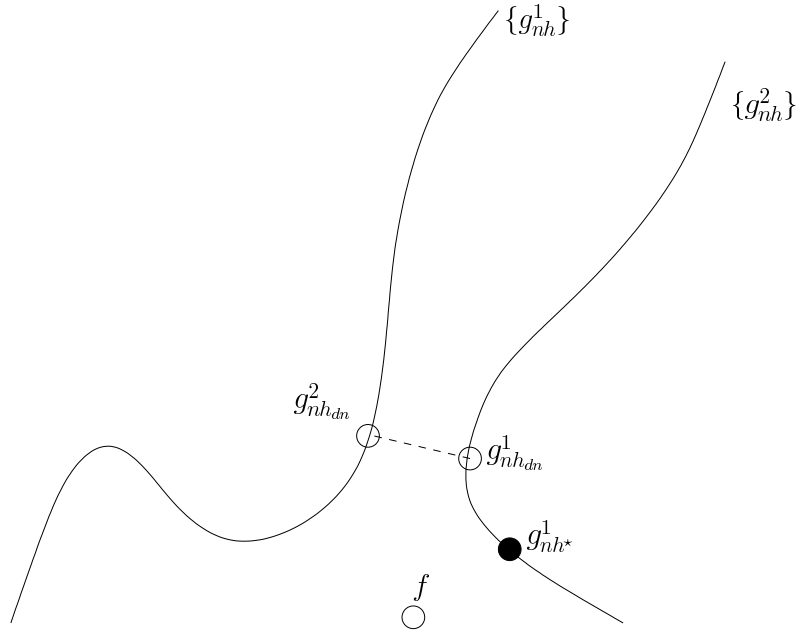


FIG. 5 – Deux familles d'estimateurs de la densité dans l'ensemble de toutes les densités. Le paramètre de lissage sélectionné minimise la distance L_1 entre g_{nh}^1 et g_{nh}^2 .

4.2 Les densités tests

Afin d'illustrer les performances des méthodes de sélection présentées, nous utilisons 9 densités tests :

- D1. La densité uniforme $U_{[0,1]}$ sur $[0, 1]$;
- D2. La densité gaussienne $\mathcal{N}(0, 1)$;
- D3. La densité $f(x) = 1/(2\sqrt{x})$ sur $[0, 1]$;
- D4. La densité du mélange $0.5U_{[0,1]} + 0.5\mathcal{N}(1, 1)$;
- D5. La densité d'un mélange de deux lois gaussiennes $\frac{1}{3}\mathcal{N}(-20, \frac{1}{4}) + \frac{2}{3}\mathcal{N}(0, 1)$;
- D6. La densité d'un mélange de trois lois uniformes $0.5U_{[0,1]} + 0.25U_{[0,0.1]} + 0.25U_{[0.9,1]}$;
- D7. La densité de $S(X + 0.1)$ où S est la variable aléatoire *signe* qui vaut 1 avec probabilité 0.5 ou -1 avec probabilité 0.5 et X a pour densité $f(x) = 4(1 - x^{1/3})$ sur $[0, 1]$.
- D8. La densité d'un mélange de six lois gaussiennes

$$\begin{aligned} & \frac{32}{63}\mathcal{N}\left(-\frac{31}{21}, \frac{32}{63}\right) + \frac{32}{63}\mathcal{N}\left(\frac{17}{21}, \frac{16}{63}\right) + \frac{8}{63}\mathcal{N}\left(\frac{41}{21}, \frac{8}{63}\right) \\ & + \frac{4}{63}\mathcal{N}\left(\frac{53}{21}, \frac{4}{63}\right) + \frac{2}{63}\mathcal{N}\left(\frac{59}{21}, \frac{2}{63}\right) + \frac{1}{63}\mathcal{N}\left(\frac{62}{21}, \frac{1}{63}\right); \end{aligned}$$

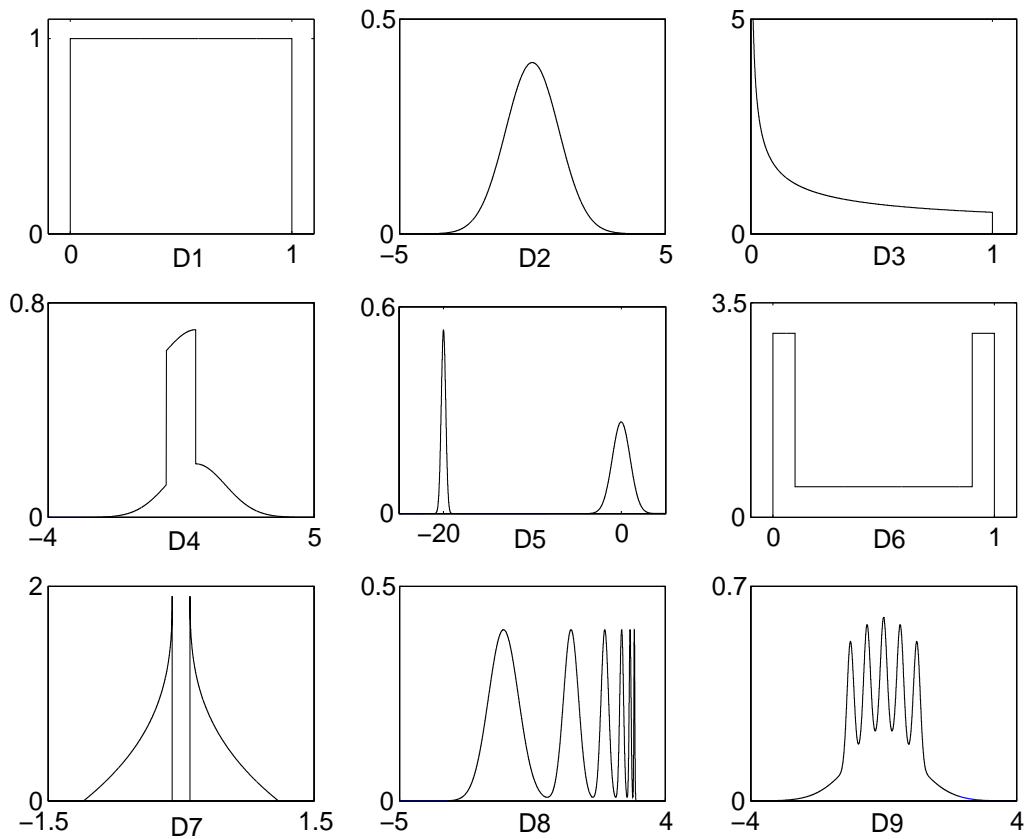


FIG. 6 – Les 9 densités tests.

D9. La densité d'un autre mélange de six lois gaussiennes

$$\frac{1}{2}\mathcal{N}(0, 1) + \frac{1}{10}\mathcal{N}(-1, 0.1) + \frac{1}{10}\mathcal{N}(-0.5, 0.1) \\ + \frac{1}{10}\mathcal{N}(0, 0.1) + \frac{1}{10}\mathcal{N}(0.5, 0.1) + \frac{1}{10}\mathcal{N}(1, 0.1).$$

Ces neuf densités sont représentées sur la Figure 6. Afin de rendre l'étude la plus pertinente possible, nous avons choisi des densités présentant différents aspects. Ainsi, nous remarquons que :

- les densités D2, D5, D8 et D9 sont lisses ;
- la densité D3 possède un pic "infini" à l'origine ;
- les densités D4, D6 et D7 possèdent des discontinuités ;
- les densités D5, D6, D7, D8 et D9 sont multimodales.

Pour chaque densité, nous avons simulé 50 échantillons et nous avons appliqué les deux procédures de sélection décrites précédemment.

Dans le cas de la méthode combinatoire (critère L_1), les échantillons sont de taille $n = 500$. Nous considérons les 300 premières observations pour construire les estimateurs et les 200 dernières pour sélectionner les paramètres, c'est-à-dire que $m = 200$. Nous utilisons les estimateurs à noyau $g_n = g_{nh_{p_i}}$ et $g_n = g_{nh_{d_n}}$ comme densité de référence des histogrammes modifiés et nous noterons $f_{n-m,\theta}$ le p -ième itéré construit à partir de ℓ classes, $\theta = (\ell, p)$.

Nous appliquons la méthode combinatoire présentée en Section 3.1, pour sélectionner un estimateur dans la famille

$$\{f_{n-m,\theta}, \theta \in \Theta\}$$

avec

$$\Theta = \{\theta = (\ell, p), \ell \in \{5, 10, 15\} \text{ et } p \geq 0\}. \quad (11)$$

On désignera par f_n l'estimateur sélectionné (*estimateur de la distance minimum*) et par $L_1(f_n)$ l'erreur L_1 commise par cet estimateur. Nous avons représenté dans le Tableau 1 :

- Les erreurs L_1 moyennes (sur les 50 répétitions) commises par les estimateurs à noyau $g_{nh_{p_i}}$ (colonne 2) et $g_{nh_{d_n}}$ (colonne 5).
- Les erreurs L_1 moyennes commises par les estimateurs de la distance minimum lorsque la densité initiale est l'estimateur à noyau $g_{nh_{p_i}}$ (colonne 3) et lorsque la densité initiale est l'estimateur à noyau $g_{nh_{d_n}}$ (colonne 6).
- Les écarts relatifs **E.R.** entre les erreurs L_1 des estimateurs à noyau et des histogrammes modifiés, *i.e.*,

$$\mathbf{E.R.} = \frac{L_1(g_n) - L_1(f_n)}{L_1(g_n)}.$$

f	$L_1(g_{n,h_{p_i}})$	$L_1(f_n)$	E.R.	$L_1(g_{n,h_{d_n}})$	$L_1(f_n)$	E.R.
D1	0.1799	0.2057	-0.1434	0.2120	0.2172	-0.0245
D2	0.0747	0.0868	-0.1620	0.1717	0.1482	0.1369
D3	0.3651	0.2956	0.1904	0.3381	0.2884	0.1470
D4	0.2226	0.2052	0.0782	0.2222	0.1986	0.1062
D5	1.2047	0.3538	0.7063	0.4625	0.3164	0.3159
D6	0.7508	0.3064	0.5919	0.3605	0.3182	0.1173
D7	0.3136	0.2783	0.1126	0.3980	0.2550	0.3593
D8	0.5465	0.3671	0.3283	0.2757	0.2732	0.0098
D9	0.3393	0.2755	0.1880	0.4601	0.2660	0.4219

TAB. 1 – La méthode combinatoire pour la sélection de θ .

Pour l'approche par validation croisée (critère de Kullback-Leibler) présentée en Section 3.2, les échantillons sont de taille $n = 150$. On notera f_n^{cv} l'estimateur sélectionné à

l'intérieur de la famille

$$\{f_{n,\theta}, \theta \in \Theta\}$$

où Θ désigne l'espace des paramètres (11). On notera également $KL(f_n^{cv})$ l'erreur au sens de l'information de Kullback-Leibler commise par l'estimateur f_n^{cv} . Les résultats sont présentés dans le Tableau 2.

f	$KL(g_{n,h_{pi}})$	$KL(f_n^{cv})$	E.R.	$KL(g_{n,h_{dn}})$	$KL(f_n^{cv})$	E.R.
D1	0.1125	0.1218	-0.0827	0.1163	0.1844	-0.5856
D2	0.0187	0.0273	-0.4599	0.0613	0.0491	0.1990
D3	0.3600	0.1851	0.4858	0.3789	0.1803	0.5241
D4	0.0762	0.0813	-0.0669	0.0921	0.0815	0.1151
D5	1.3177	0.1382	0.8951	0.3827	0.1066	0.7215
D6	0.5582	0.1683	0.6985	0.2289	0.1514	0.3386
D7	0.2033	0.1275	0.3728	0.4190	0.1494	0.6434
D8	0.3231	0.2132	0.3401	0.1585	0.1600	-0.0095
D9	0.0936	0.0976	-0.0427	0.1684	0.1017	0.3961

TAB. 2 – Le critère de validation croisée pour la sélection de θ .

Nous observons tout d'abord que les performances des estimateurs à noyau varient suivant la densité à estimer. Ces estimateurs se comportent en effet très bien pour les densités uniformes et gaussiennes (D1 et D2). Ils sont en revanche moins performants pour les cibles plus complexes (D3, D5, D6 et D8 par exemple). Les résultats de ces simulations mettent également en relief le problème bien connu du choix de la fenêtre. En effet, la performance de la procédure de sélection de la fenêtre de l'estimateur à noyau varie en fonction de la densité à estimer : la méthode plug-in donne de meilleurs résultats pour les densités D2, D7 et D9 tandis que la méthode du double noyau est plus performante pour les modèles D5, D6 et D8. Ces disparités peuvent être corrigées, dans une certaine mesure, en sélectionnant un histogramme modifié itéré. Nous remarquons en effet que les erreurs (L_1 ou de Kullback-Leibler) commises par les histogrammes modifiés sélectionnés ne diffèrent guère suivant la procédure de sélection du paramètre de lissage de l'estimateur à noyau. De plus, à l'exception des modèles D1 et D2 (pour lesquels l'estimateur à noyau se comporte très bien), les performances des histogrammes modifiés sélectionnés sont meilleures que celles des estimateurs à noyau. Cette amélioration devient même très significative pour les modèles D5, D6, D7, D8 et D9 (voir Figure 7 et Figure 8). D'une certaine manière, ces remarques nous amènent à considérer les histogrammes modifiés comme un outil potentiel permettant d'améliorer, ou tout au moins de corriger dans certaines situations, les performances des estimateurs à noyau.

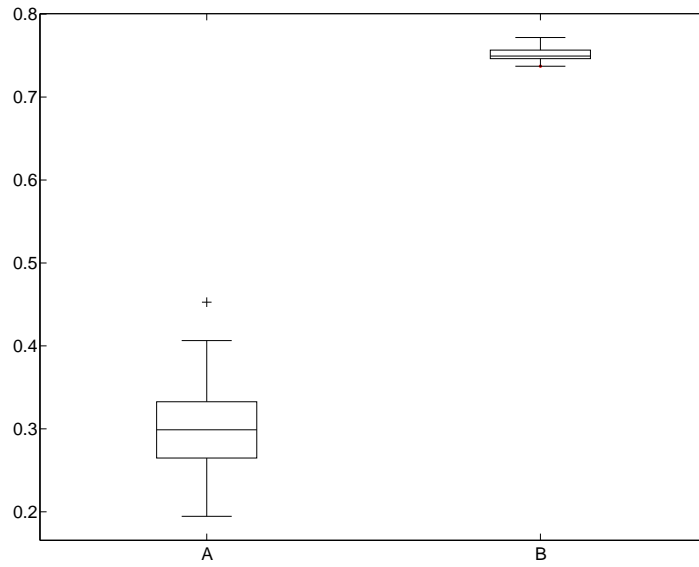


FIG. 7 – Boxplots (pour les 50 répétitions) des erreur L_1 commises par les histogrammes modifiés sélectionnés par la méthode combinatoire (A) et les estimateurs à noyau $g_{n,h_{pi}}$ (B) pour le modèle D6.

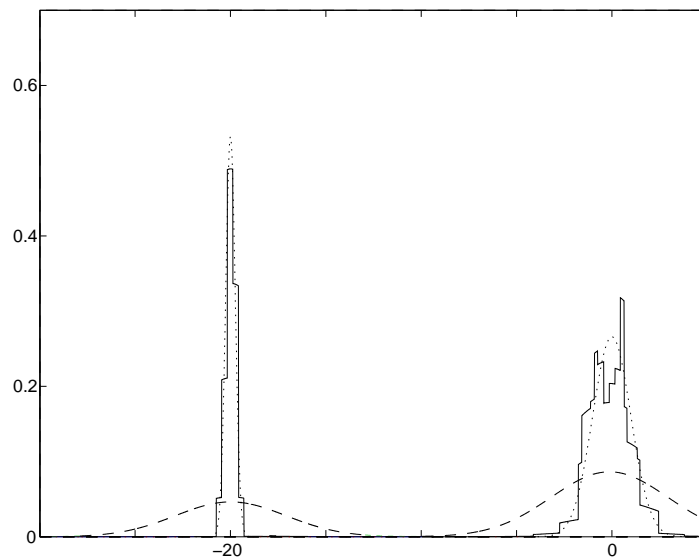


FIG. 8 – Densité D5 (pointillés), estimateur à noyau $g_{n,h_{pi}}$ (tirets), histogramme modifié sélectionné par la méthode combinatoire (trait plein).

Remerciements. Qu'il me soit permis de remercier Henri Caussinus, Editeur en Chef du Journal de la SFDS, ainsi que deux relecteurs anonymes : leurs commentaires, leurs critiques et leurs suggestions constructives m'ont permis d'améliorer la qualité de ce travail.

Références

- [1] A. R. Barron. The convergence in information of probability density estimators. Dans *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe : Japan, 19-24 Juin, 1988.
- [2] A. R. Barron, L. Györfi et E. C. van der Meulen. Distribution estimation consistent in total variation and in two types of information divergence. *IEEE Transactions on Information Theory*, 38 :1437–1454, 1992.
- [3] A. Berlinet et G. Biau. Iterated modified histograms as dynamical systems. *Journal of Nonparametric Statistics*, 16 :385–401, 2004.
- [4] A. Berlinet, G. Biau et L. Rouvière. Parameter selection in modified histogram estimates. *Statistics*, 39 :91–105, 2005.
- [5] A. Berlinet et E. Brunel. Cross-validated density estimates based on Kullback-Leibler information. *Journal of Nonparametric Statistics*, 16 :493–513, 2003.
- [6] A. Berlinet et L. Devroye. A comparison of kernel density estimates. *Publications de l'Institut de Statistique de l'Université de Paris*, 38 :3–59, 1994.
- [7] A. Berlinet, I. Vajda et E.C. van der Meulen. About the asymptotic accuracy of Barron density estimates. *IEEE Transactions on Information Theory*, 44 :999–1009, 1998.
- [8] G. Biau. Estimateurs à noyau itérés : synthèse bibliographique. *Journal de la Société Française de Statistique*, 1 :41–67, 1999.
- [9] D. Bosq et J.P. Lecoutre. *Théorie de l'Estimation Fonctionnelle*. Economica, Paris, 1987.
- [10] E. Brunel. *Sur l'estimation de la densité et de la fonction de hasard : Estimateurs à noyaux et de Barron, critère de Kullback, applications*. Thèse de Doctorat, Université Montpellier II, 1999.
- [11] P. Deheuvels. Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité. *Comptes Rendus Mathématiques de l'Académie des Sciences de Paris*, 278 :1217–1220, 1974.
- [12] P. Deheuvels. Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée*, 25 :5–42, 1977.
- [13] P. Deheuvels et P. Hominal. Estimation automatique de la densité. *Revue de Statistique Appliquée*, 28 :25–55, 1980.

- [14] L. Devroye et G. Lugosi. *Combinatorial Methods in Density Estimation*. Springer–Verlag, New York, 2001.
- [15] S. Kullback. A lower bound for discrimination in terms of variation. *IEEE Transactions on Information Theory*, 13 :126–127, 1967.
- [16] J.S. Marron. Automatic smoothing parameter selection : a survey. *Empirical Economics*, 13 :187–208, 1988.
- [17] E.A. Nadaraya. On the integral mean square error of some nonparametric estimates for the density function. *Theory of Probability and its Applications*, 19 :133–141, 1974.
- [18] H. Scheffé. A useful convergence theorem for probability distributions. *Annals of Mathematical Statistics*, 18 :434–458, 1947.
- [19] S.J. Sheater et M.C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society*, B53 :683–690, 1991.
- [20] J. Simonoff. *Smoothing Methods in Statistics*. Springer, New-York, 1996.
- [21] A.B. Tsybakov. *Introduction à l'estimation non-paramétrique*. Springer, 2004.
- [22] V.N. Vapnik et A.Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16 :264–280, 1971.
- [23] Y.G. Yatracos. Rates of convergence of minimum distance estimators and Kolmogorov's entropy. *The Annals of Statistics*, 13 :768–774, 1985.