Extension de l'algorithme plug-in pour l'optimisation du paramètre de lissage de l'estimateur du noyau-difféomorphisme

Molka Troudi[1], Faouzi Ghorbel[2]

1. *Institut des Hautes Études Commerciales*
   Carthage Présidence, 2016, Tunis, Tunisie
   molkaghorbel@gmail.com

2. *Laboratoire CRISTAL, Pôle GRIFT*
   École Nationale des Sciences de l'Informatique
   Université La Manouba, 2010, La Manouba, Tunisie
   Faouzi.ghorbel@ensi.rnu.tn

RÉSUMÉ. L'estimateur noyau-difféomorphisme est une généralisation de l'estimateur à noyau permettant d'estimer les densités en tenant compte de leur support naturel. Le recours à un changement de variable approprié permet de limiter significativement le phénomène de Gibbs. Cependant, la qualité de l'estimation est tributaire de la valeur du pas qui doit être ajusté. Dans cet article, nous nous focalisons sur l'algorithme plug-in pour l'optimisation du pas. Ainsi, nous proposons une extension de cet algorithme itératif à l'estimateur du noyau difféomorphisme. Après un aperçu des théorèmes de convergence de la méthode du noyau-difféomorphisme et la présentation de l'algorithme proposé, la mesure de l'écart quadratique moyen intégré de quelques densités semi-bornées et bornées simulées puis ré-estimées permet de mettre en évidence l'intérêt de cette approche.

ABSTRACT. The kernel-diffeomorphism estimate is a generalization of the kernel estimate taking into account of the natural support of estimated densities. Using a suitable change of variables can significantly limit the Gibbs phenomenon. The quality of the estimate depends on the value of the bandwidth which must be adjusted. In this article, we focus on the plug-in algorithm to optimize the bandwidth. Thus, we propose to extend it to the kernel-diffeomorphism estimate. The Mean Integrated Square Error of simulated and re-estimated bounded and semi-bounded densities highlight the interest of this approach.

MOTS-CLÉS : estimateur non paramétrique, estimateur noyau-difféomorphisme, paramètre de lissage, support borné, algorithme plug-in.

KEYWORDS: non parametric estimate, kernel diffeomorphism estimate, smoothing parameter, bounded distributions, plug-in algorithm.

DOI:10.3166/TS.31.321-338

Extended abstract

The kernel-diffeomorphism probability density functions estimator is a generalization of the kernel probability density functions estimator. It is adapted to estimate the densities taking into account their natural support. Indeed, a simple change of variable leads to a better estimate which limits significantly the Gibbs phenomenon. However, the quality of the estimate depends on the value of the smoothing parameter, which must be adjusted. In this article, we focus on plug-in algorithm to optimize the smoothing parameter. Thus, we propose an extension of this algorithm to the kernel diffeomorphism estimator. As a first step, we recall the principles and theorems of convergence of the kernel-diffeomorphism estimate. An asymptotic study estimates the optimal value of the smoothing parameter by minimizing the mean integrated square error (MISE). Thus, this estimator is expressed in two entities associated with the probability density f to estimate, $M_\phi(K)$ et $J_\phi(f)$. The implementation of the generalized plug-in-algorithm presents further difficulties compared to conventional plug-in-algorithm. Indeed, in this algorithm, $M(K)$ depends only on the choice of kernel while, in case of the generalized plug-in algorithm, $M_\phi(K)$ is related to the unknown density which must be approached throughout the iterations. Furthermore, $J_\phi(f)$ is a function of f and f' in addition to f'' which leads to an increased complexity of the generalized plug-in algorithm. The proposed iterative algorithm is subsequently tested on some bounded or semi-bounded simulated densities. Measuring the MISE enables to highlight the benefits of this approach for densities with support information.

In the case of semi-bounded densities, the results show a significantly improved estimate. This is not as obvious for densities with bounded support. Indeed, despite a significant limitation of Gibbs effect, the MISE is more important because of the disruption of the estimated density suggesting a value lower than the optimal smoothing parameter value. This divergence could be explained by the accumulation of errors in the estimation of the various entities involved in the estimation of $J_\phi(f)$. To remedy this problem, the optimal bandwidth is adjusted empirically by varying the powers of $J_\phi(f)$ entity in the analytical expression of the optimal bandwidth. Thanks to this fitting, the generalized plug-in algorithm provides a significant improvement in the estimation of the densities having a bounded support.

1. Introduction

Il est bien connu qu'une estimation fiable des densités de probabilité permet d'améliorer les performances des systèmes technologiques. À titre d'exemple, nous pouvons citer dans le domaine des nouvelles technologies, la quantification scalaire basée sur l'estimation des densités de probabilité des paramètres de codage du signal et des images ainsi que le hachage, tâche essentielle en indexation des données.

De même, l'application des règles de classification qui sont à la base de la conception des systèmes de reconnaissance de formes sont souvent tributaires de la qualité de l'estimation de la densité de probabilité mélange ainsi que des densités de probabilité conditionnelles. Ces différents exemples de statistiques répondent à des distributions dont les supports peuvent être bornés ou semi-bornés. L'estimation de ce type de densités de probabilité présente des problèmes de convergence aux bords, le phénomène de Gibbs. Pour remédier à ce problème, plusieurs auteurs ont développé des méthodes d'estimation des densités de probabilité tenant compte d'informations sur le support. Parmi ces méthodes, citons la méthode des fonctions orthogonales (Hall, 1982) et la méthode du noyau difféomorphisme (Saoudi et *al.*, 1994 ; 1997 ; Ghorbel, 2011). Cette dernière qui est une généralisation de la méthode du noyau, se base sur un changement de variable approprié qui permet d'estimer les densités de probabilité sur leur support naturel et de limiter l'effet du phénomène de Gibbs.

Cependant, la sélection d'une valeur appropriée pour le paramètre de lissage est nécessaire pour garantir une bonne qualité de l'estimation. La majorité des méthodes développées ont été proposées pour la sélection du pas optimal au sens de l'écart quadratique moyen intégré (EQMI) pour la méthode du noyau conventionnelle, les principales étant la méthode de la validation croisée et ses variantes (Bowman, 1984 ; Hall et *al.*, 1992), les méthodes plug-in (Hall et Marron, 1992 ; Troudi et *al.*, 2008 ; Troudi, 2009), la méthode des contrastes (Mugadi et Ahmad, 2004) et plus récemment une méthode développée par Botev, basée sur un processus de diffusion linéaire (Botev et *al.*, 2011). Dans cet article, nous proposons d'étendre l'algorithme plug-in à la méthode du noyau difféomorphisme. La minimisation de l'écart quadratique moyen intégré (EQMI) par une étude asymptotique permet d'exprimer analytiquement le paramètre de lissage optimal selon deux entités $J_\Phi(f)$ et $M_\Phi(K)$ liées à la densité de probabilité f à estimer. Ainsi, la mise en œuvre de l'algorithme plug-in généralisé présente des difficultés supplémentaires par rapport à l'algorithme plug-in conventionnel puisque $M_\Phi(K)$ n'est plus une constante à déterminer analytiquement ou numériquement comme pour l'algorithme plug-in et doit être approchée au fil des itérations car elle dépend également de la densité de probabilité à estimer. De même $J_\Phi(f)$ est fonction de f et de f'en plus de f''.

La section 2 présente de manière sommaire la méthode du noyau classique ainsi que les principales méthodes d'optimisation du paramètre de lissage. La section 3 est dédiée à la présentation de l'estimateur incluant une étude de la convergence ainsi qu'une étude asymptotique. L'expression analytique du paramètre de lissage optimal permet de généraliser l'algorithme plug-in et de l'étendre à l'estimateur noyau difféomorphisme. Ainsi, l'algorithme plug-in généralisé est présenté tout au long de la section 4. Puis, dans la section 5, une étude comparative est menée grâce à quelques distributions simulées et permet de mettre en évidence l'intérêt de l'algorithme proposé. Une conclusion et quelques perspectives font l'objet d'une dernière section.

2. Estimateurs à noyau

L'estimateur à noyau introduit par Rozenblatt (1956) puis développé par Parzen (1962) permet d'estimer de manière non paramétrique une densité de probabilité *f* à partir d'un échantillon $\left(X_1; X_2; ....; X_N\right)$ suivant la loi parente de *X*. Ainsi, l'estimée de f_X notée par f_N est déduite selon l'expression suivante :

$$\hat{f}_N\left(x\right) = \frac{1}{Nh_N}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h_N}\right) \tag{1}$$

avec *K* une densité de probabilité appelée noyau. L'image d'un réel *x* par *f* est estimée en sommant la contribution des différents éléments de l'échantillon X_i et en normalisant par l'entité h_N appelée « paramètre de lissage » ou plus simplement « pas ». La qualité de l'estimation dépend de l'échantillon et de la valeur sélectionnée pour le paramètre de lissage. L'étude de la convergence de l'estimateur à noyau (Deheuvels et Hominal, 1980) a permis d'exprimer le paramètre de lissage optimal noté h_N^* comme suit :

$$h_n^* = n^{-\frac{1}{5}}.\left(J\left(f\right)\right)^{-\frac{1}{5}}.\left(M\left(K\right)\right)^{\frac{1}{5}} \tag{2}$$

Cette expression théorique du paramètre de lissage optimal au sens de l'EQMI (écart quadratique moyen intégré) est fonction de la taille de l'échantillon *N*, de l'intégrale du noyau choisi élevé au carré *M*(*K*) et de l'intégrale de la dérivée seconde élevée au carré de $\hat{f}_N$ notée $J\left(f\right)$.

La valeur de *M*(*K*), liée au noyau choisi, est facilement estimée analytiquement ou numériquement. Par contre, l'entité *J*(*f*) découle directement de la densité inconnue f, fonction à estimer.

Cette problématique a fait l'objet d'un nombre important de travaux dont l'objectif est d'approcher la valeur optimale du paramètre de lissage en minimisant l'erreur quadratique moyenne intégrée (EQMI). Nous citons ci-dessous les principales méthodes de la littérature du domaine :

– les méthodes Rule of thumb (rot) (Silvermann, 1986; Terrel, 1990; Hardle, 1991) ;

– les méthodes cross-validation (Bowman, 1984 ; Hall et Marron, 1987 ; Scott et Terrel, 1987 ; Hall et Marron, 1991).

– la méthode Plug-in (Hall et Marron, 1987 ; Park et Marron, 1990 ; Troudi *et al.*, 2008 ; Troudi, 2009) ;

– la méthode des contrastes (Mugadi et Ahmad, 2004) ;

– la méthode de diffusion linéaire (Botev *et al.*, 2010).

3. Estimateur du noyau difféomorphisme

Les densités de probabilité à support borné ou semi-borné présentent des difficultés d'estimation en raison de la nature particulière de leur support. L'effet de Gibbs en particulier est observé aux bords lorsque ces densités sont estimées par la méthode du noyau ou la méthode des fonctions orthogonales. Le recours à un changement de variable par un C1-difféomorphisme noté φ de]*a*; *b*[dans R a permis de limiter de manière importante cet artéfact et de mieux estimer les densités de probabilité avec informations sur le support. Dans la présente section, nous proposons de rappeler la méthode du noyau difféomorphisme et ses théorèmes de convergence bien détaillés dans (Saoudi et *al.*, 1994 ; 1997).

Rappelons l'expression de l'estimateur du noyau difféomorphisme :

$$\hat{f}_N(x) = \frac{|\phi'(x)|}{Nh_N} \sum_{i=1}^{N} K\left(\frac{\phi(x) - \phi(X_i)}{h_N} \right) \tag{3}$$

où φ est un C1 - difféomorphisme qui a pour limite l'infini lorsque *x* tend vers *a* ou vers *b*.

Nous présentons ci-dessous deux exemples de C1 – difféomorphisme :

$$D_1 : \begin{cases} \varphi_{a,b} :]\, a, b\, [\ \to R \\ \qquad x \quad \to Log\left\{ (x-a)/(b-x) \right\} \end{cases}$$

$$D_2 : \begin{cases} \Psi_{a,b} :]a, b[\to R \\ \qquad x \quad \to tg\left\{ \alpha(\beta - x) \right\} \end{cases}$$

avec $\alpha = \pi/(b-a)$ et $\beta = (a+b)/2$

Dans (Saoudi et *al.*, 1994), une étude montre que les difféomorphismes de type logarithmique permettent une meilleure estimation des densités de probabilité en question.

3.1. Étude de la convergence de la méthode du noyau-difféomorphisme

Nous commençons par rappeler l'expression analytique de l'espérance de l'estimateur du noyau-difféomorphisme :

$$E\left[\hat{f}_N(x) \right] = \frac{|\phi'(x)|}{Nh_N} \sum_{i=1}^{N} E\left[K\left(\frac{\phi(x) - \phi(X_i)}{h_N} \right) \right] \tag{4}$$

Le changement de variable $y = \frac{\phi(x) - \phi(u)}{h_N}$ permet d'exprimer cette espérance de la manière suivante :

$$E\left[\hat{f}_N(x)\right] = \left|\phi'(x)\right| \int_R K(y) \underbrace{fo\phi^{-1}\left(\phi(x) - uh_N\right)\left|\left(\phi^{-1}\right)'_{(\phi(x)-yh_N)}\right|}_{g(x,y)} dy \tag{5}$$

Le développement de l'expression analytique de la variance aboutit quant à elle à l'expression suivante :

$$\begin{aligned} \text{var}\left[\hat{f}_N(x)\right] &= \frac{\left|\phi'(x)\right|^2}{N}\left\{E\left[\frac{1}{h_N^2}K^2\left(\frac{\phi(x)-\phi(X_1)}{h_N}\right)\right]\right\} - \frac{1}{N}\left\{E\left[\hat{f}_N(x)\right]\right\}^2 \\ &= \frac{\left|\phi'(x)\right|^2}{N}\int_a^b \frac{1}{h_N^2}K^2\left(\frac{\phi(x)-\phi(u)}{h_N}\right) f(u)\,du - \frac{1}{N}\left\{E\left[\hat{f}_N(x)\right]\right\}^2 \end{aligned} \tag{6}$$

En ayant recours au même changement de variable que pour l'espérance, la variance de l'estimateur s'écrit :

$$\text{var}\left[\hat{f}_N(x)\right] = \frac{\left|\phi'(x)\right|^2}{Nh_N}\int_R K^2(y) \underbrace{fo\phi^{-1}\left(\phi(x) - uh_N\right)\left|\left(\phi^{-1}\right)'_{(\phi(x)-yh_N)}\right|}_{g(x,y)} dy - \frac{1}{N}\left\{E\left[\hat{f}_N(x)\right]\right\}^2$$

Il est connu que l'erreur quadratique moyenne s'écrit en fonction de l'espérance et de la variance de $\hat{f}_N(x)$ selon l'expression suivante :

$$\begin{aligned} E\left[\left|\hat{f}_N(x) - f(x)\right|^2\right] &= E\left[\left|\hat{f}_N(x)\right|^2\right] - 2f(x)E\left[\hat{f}_N(x)\right] + f^2(x) \\ &= \text{var}\left[\hat{f}_N(x)\right] + \left\{E\left[\hat{f}_N(x)\right]\right\}^2 - 2f(x)E\left[\hat{f}_N(x)\right] + f^2(x) \end{aligned} \tag{7}$$

En remplaçant les espérance et variance de $\hat{f}_N(x)$ par les expressions (6) et (7) de la section précédente et en posant $g(x) = f \circ \Phi^{-1}\left(\Phi(x) - uh_N\right)\left|\left(\Phi^{-1}\right)'_{(\Phi(x)-yh_N)}\right|$, l'EQM devient :

$$E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right]=\left\{\frac{\left|\phi'(x)\right|^2}{Nh_N}\int_R K^2(y)\,g(x,y)\,dy\right\}-\frac{1}{N}\left\{E\left[\hat{f}_N(x)\right]\right\}^2$$
$$+\left\{E\left[\hat{f}_N(x)\right]-f(x)\int_R K(y)\,dy\right\}^2$$
$$=\frac{\left|\phi'(x)\right|^2}{Nh_N}\left\{\int_R K^2(y)\,g(x,y)\,dy-h_N\left[\int_R K(y)\,g(x,y)\,dy\right]^2\right\}$$
$$+\left\{\int_R K(y)\left(\left|\phi'(x)\right|g\ \ (x,y)-f(x)\right)dy\right\}^2$$

Cela revient à l'exprimer en fonction de trois entités $A_N(x)$, $B_N(x)$ et $C_N(x)$ avec :

$$A_N(x)=\frac{\left|\phi'(x)\right|^2}{Nh_N}\int_R K^2(y)\,g(x,y)\,dy \tag{8}$$

$$B_N(x)=\left\{\int_R K(y)\left[\left|\phi'\right|g(x,y)-f(x)\right]dy\right\}^2 \tag{9}$$

$$C_N(x)=\frac{\left|\phi'(x)\right|^2}{N}\left\{\int_R K(y)\,g(x,y)\,dy\right\}^2 \tag{10}$$

$\left(\phi^{-1}\right)'$ étant borné sur R, f étant bornée sur]a; b[et K^2 intégrable sur R, le théorème de convergence de Lebesgue nous permet, pour N élevé, d'exprimer l'EQM par :

$$E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right]=\frac{\left|\phi'(x)\right|f(x)}{Nh_N}\int_R K^2(y)\,dy-\frac{f^2(x)}{N}+o(h_N) \tag{11}$$

En raison de la continuité de ϕ' sur R et de f sur]a; b[, cet estimateur converge vers l'EQMI pour tout compact de]a; b[. Afin d'obtenir la convergence au sens de l'écart quadratique moyen intégré, $\phi'(f)$ doit être intégrable sur]a; b[car,

$$\int_a^b E\left[\left|\hat{f}_N(x)-f(x)\right|^2\right]dx=\frac{1}{Nh_N}\int_a^b\left|f(x)\phi'(x)\right|dx\int_R K^2(y)\,dy$$
$$-\frac{1}{N}\int_a^b f^2(x)\,dx+(b-a)o(h_N) \tag{12}$$

3.2. Étude asymptotique

Le développement de Taylor de la fonction H_y définie dans le voisinage de $\phi(x)$ par :

$$\phi(x) \xrightarrow{H_y} fo\phi^{-1}\left(\phi(x)-yh_N\right)\left|\left(\phi^{-1}\right)'\left(\phi(x)-yh_N\right)\right| \tag{13}$$

implique qu'il existe un nombre positif θ inférieur à 1 tel que :

$$\begin{aligned} H_y\left(\phi(x)-yh_N\right) &= H_y\left(\phi(x)\right)-yh_N H_y'\left(\phi(x)\right)+ \\ \frac{y^2h_N^2}{2}H_y''\left(\phi(x)\right)-\frac{y^3h_N^3}{6}H_y'''\left(\phi(x)-\theta yh_N\right) \end{aligned} \tag{14}$$

Les approximations suivantes sont déduites à partir du calcul des dérivées successives de la fonction H_y dans $\phi(x)$:

$$A_N(x) \approx \frac{\left|\phi'(x)\right| f(x)}{Nh_N} M(K) \tag{15}$$

$$\text{avec } M(K) = \int_R K^2(y)\,dy \tag{16}$$

$$B_N(x) = \frac{h_N^4}{4\left[\phi'(x)\right]^8} F^2(x) \tag{17}$$

avec :

$$F(x) = \left[f(x)\left[3\phi''(x)^2 - \phi'(x)\phi'''(x)\right]\right] - 3f'(x)\phi'(x)\phi''(x) + f''(x)\left[\phi'(x)\right]^2 \tag{18}$$

$$C_N(x) = \frac{\left[f(x)\right]^2}{N} \tag{19}$$

Après étude asymptotique l'EQMI s'exprime par :

$$\begin{aligned} D^2\left(\hat{f}_N, f\right) &= \int_R \left[A_N(x)+B_N(x)-C_N(x)\right]dx \\ &\approx \frac{M(K)}{Nh_N}\int_R \left|\phi'(x)\right| f(x)\,dx + \frac{h_N^4}{4}\int_R \frac{F^2(x)}{\left[\phi'(x)\right]^8}dx \end{aligned} \tag{20}$$

Dans le cas où M_ϕ et J_ϕ existent, ils s'expriment par :

$$M_\phi(K) = M(K)\int_R \left|\phi'(x)\right| f(x)\,dx \tag{21}$$

$$J_\phi(f) = \int_R \frac{F^2(x)}{\left[\phi'(x)\right]^8}\, dx \tag{22}$$

La valeur optimale de h_N notée h_N^* peut ainsi être déduite par minimisation de l'EQMI. On obtient :

$$h_N^* = \left[M_\phi(K)\right]^{\frac{1}{5}} \left[J_\phi(f)\right]^{-\frac{1}{5}} N^{-\frac{1}{5}} \tag{23}$$

4. Algorithme plug-in généralisé

Dans cette section, nous proposons une extension de l'algorithme plug-in pour l'adapter à l'estimateur noyau-difféomorphisme (algorithme1).

Algorithme 1. Plug-in généralisé

1: ***pluginDiffGen*** (*X : Vecteur des données réelles, a,b : limites du support naturel de la*
2: *densité*)
3: {
4: *Calcul M(K)*
5: *Initialisation J(f)*
6: *Calcul* h_N
7: h_N*Estime* ← h_N
8: *Estimation f par la méthode du noyau*
9: **répéter**
10 : h_N*(k-1)* ← h_N
11: *Estimer* M_ϕ*(K)*
10: *Estimer f ', f '', J*$_\phi$*(f) et* h_N
11: *Estimer f par l'estimateur noyau-difféomorphisme*
12: **jusqu'à** $\frac{\left|h_N(k-1) - h_N\right|}{h_N} < 0.01$
13: **fin boucle}**

Ainsi, l'algorithme plug-in classique devient un cas particulier de l'algorithme plug-in généralisé puisqu'il suffit de choisir le difféomorphisme identité pour le retrouver. L'expression analytique du paramètre de lissage optimal par l'algorithme Plug-in généralisé pour la méthode du noyau difféomorphisme présente une complexité accrue puisque $M_\phi(K)$ dépend de *f* la densité de probabilité inconnue et $J_\phi(f)$ dépend de *f*, f' et f'' alors que pour l'algorithme plug-in conventionnel $M_{Id}(K)$ est une constante et $J_{Id}(f)$ ne dépend que de f''.

5. Étude comparative des estimateurs à noyau et noyau-difféomorphisme avec ajustement du paramètre de lissage

Cette section est dédiée à l'évaluation des performances de l'estimateur noyau-difféomorphisme avec ajustement du paramètre de lissage par l'algorithme plug-in généralisé. Trois distributions sont simulées puis étudiées dans le sens de la convergence des estimateurs :

– D1 : Distribution de type semi bornée (loi exponentielle de moyenne 1);

– D2 : Distribution bornée (loi uniforme entre]0; 0:1[);

– D3 : Distribution bornée (loi beta de paramètre (2,2));

Afin de mener une étude comparative entre l'estimateur à noyau et l'estimateur noyau-difféomorphisme, la densité de chacune de ces distributions est estimée avec les deux méthodes. Une représentation graphique permet de visualiser la quasi absence du phénomène de Gibbs lorsque l'estimateur noyau-difféomorphormisme avec optimisation du pas par l'algorithme plug-in généralisé est utilisé. Une évaluation plus objective est menée en comparant les EQMI générés par les deux estimations.

La figure 1 représente l'estimation de la densité D1 par la méthode du noyau avec ajustement du pas par l'algorithme Plug-in alors que la figure 2 représente l'estimation de cette densité par l'estimateur noyau-difféomorphisme avec ajustement du pas par l'algorithme plug-in généralisé. L'atténuation du phénomène de Gibbs dans la figure 2 comparativement à la figure 1 est remarquable. En effet, l'estimation de la densité de probabilité déborde de son support naturel dans la figure 1 (phénomène de Gibbs). Par ailleurs, des perturbations influant à la hausse la valeur de l'EQMI sont également observées sur la figure 1 alors que sur la figure 2 l'estimation de la densité est presque parfaite. Ces observations sont confirmées par les valeurs des EQMI reportées dans le tableau 1.

L'absence du phénomène de Gibbs est également observée pour les distributions bornées D2 et D3 lorsque l'estimation est réalisée par la méthode du noyau-difféomorphisme avec estimation du pas optimal par l'algorithme plug-in-généralisé (figures 5 et 6) comparativement aux résultats obtenus par l'estimateur du noyau classique (figures 3 et 4). Cependant, en comparant les valeurs d'EQMI des distributions uniformes et bêta simulées puis ré-estimées par les deux méthodes, nous observons que l'estimation par la méthode du noyau conventionnel est meilleure au sens de l'EQMI (tableau 2). En effet, dans le cas des distributions à support borné, des perturbations sont observées au niveau du lissage de la densité de probabilité estimée. Elles ont pour effet d'augmenter la distance entre les densités réelles et les densités estimées malgré l'atténuation du phénomène de Gibbs.

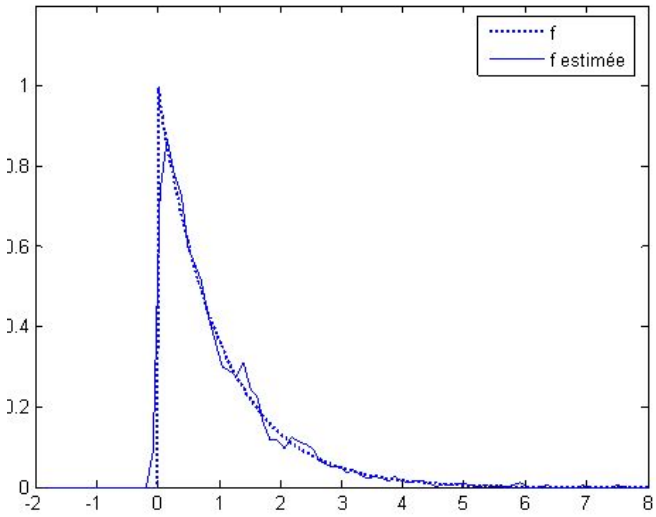

Figure 1. Estimation de la densité de probabilité d'une loi exponentielle par la méthode du noyau classique avec ajustement du pas par l'algorithme Plug-in

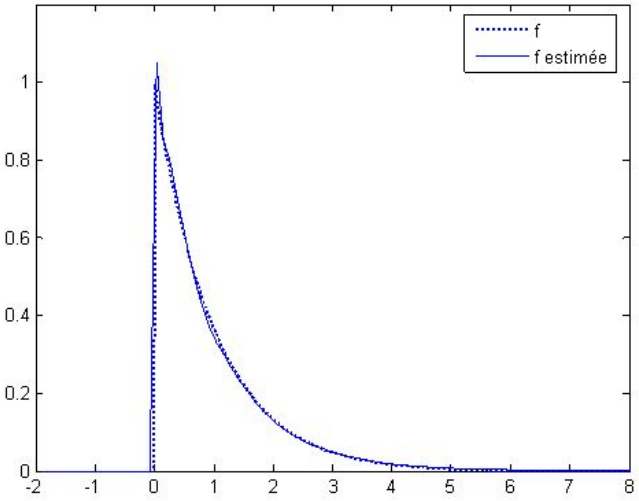

Figure 2. Estimation de la densité de probabilité d'une loi exponentielle par la méthode du noyau-difféomorphisme avec ajustement du pas par l'algorithme Plug-in généralisé

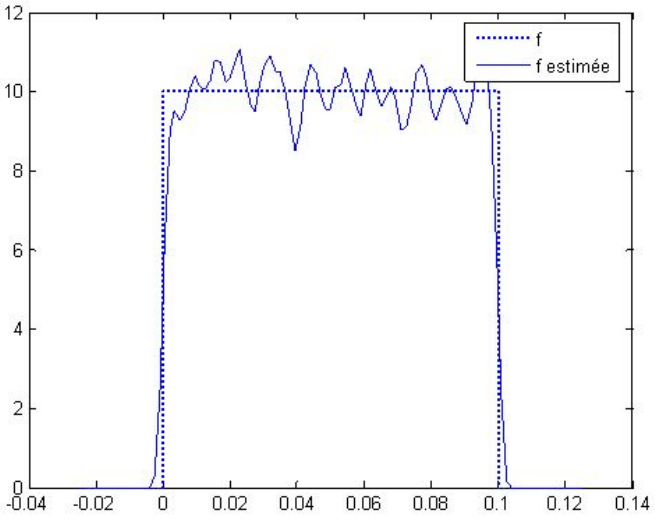

Figure 3. Estimation de la densité de probabilité d'une loi uniforme par la méthode du noyau classique avec ajustement du pas par l'algorithme Plug-in

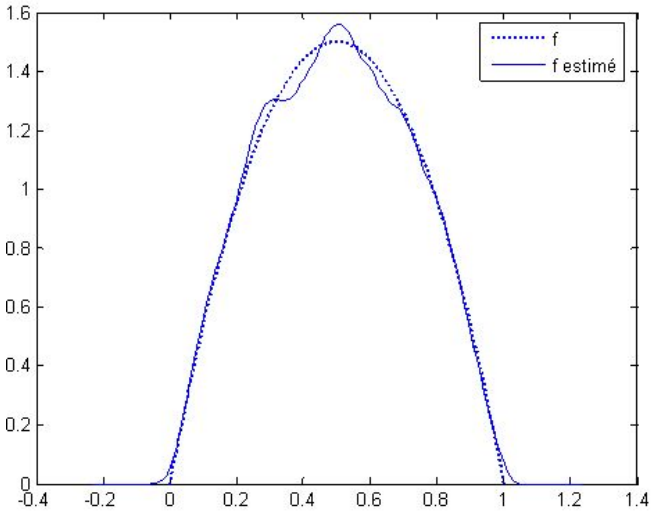

Figure 4. Estimation de la densité de probabilité d'une loi beta par la méthode du noyau classique avec ajustement du pas par l'algorithme Plug-in

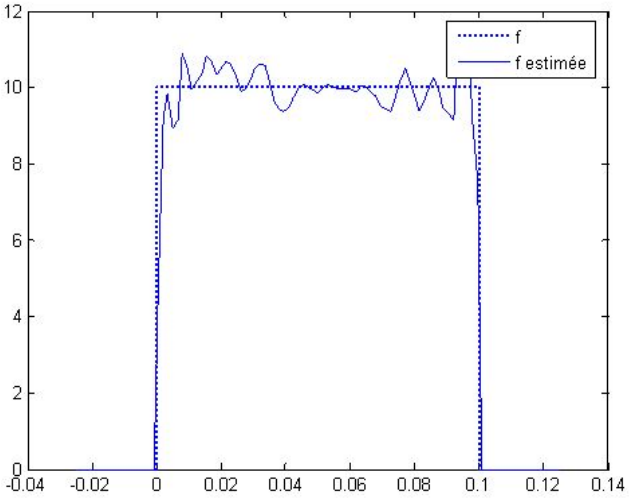

Figure 5. Estimation de la densité de probabilité d'une loi uniforme par la méthode du noyau-difféomorphisme avec ajustement du pas par l'algorithme Plug-in généralisé

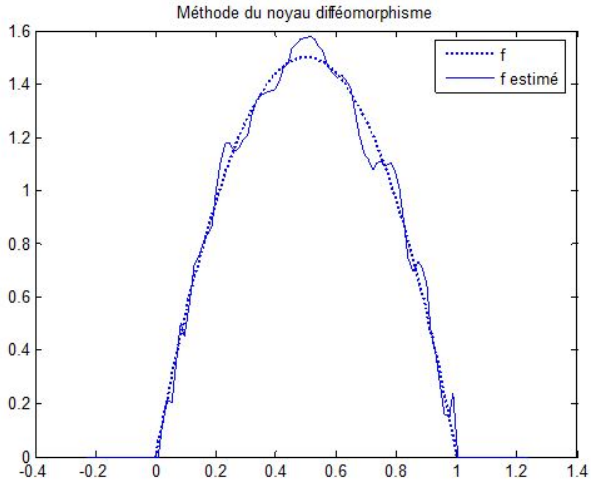

Figure 6. Estimation de la densité de probabilité d'une loi beta par la méthode du noyau-difféomorphisme avec ajustement du pas par l'algorithme Plug-in généralisé

Tableau 1. Évolution de l'EQMI en fonction de la taille d'échantillon pour les estimateurs à noyau et à noyau-difféomorphisme d'une distribution semi bornée (Loi exponentielle)

Taille échantillon		1000	2000	3000	4000	5000
EQMI*10^{-5}	Estimateur à noyau	3.79	2.68	2.00	1.22	0.81
	Estimateur à noyau-difféomorphisme	0.78	0.61	0.55	0.26	0.21

Tableau 2. Évolution de l'EQMI en fonction de la taille d'échantillon pour les estimateurs à noyau et à noyau-difféomorphisme de deux distributions bornées (Loi uniforme et loi beta)

		Distribution beta		Distribution uniforme	
Taille échantillon		1000	5000	1000	5000
EQMI*10^{-5}	Estimateur à noyau	720	170	24540	14200
	Estimateur à noyau-difféomorphisme	1340	280	31390	6650

Tableau 3. Évolution de l'EQMI selon la puissance de $J_\phi(f)$ et la taille d'échantillon pour les estimateurs à noyau et à noyau-difféomorphisme de deux distributions bornées (Loi beta et loi uniforme)

			Distribution beta		Distribution uniforme	
Taille échantillon			1000	5000	1000	5000
EQMI	Estimateur à noyau		0.0072	0.0018	0.2471	0.1432
	Estimateur à noyau-difféomorphisme	Puissance de $J_\phi(f)$				
		0.2	0.134	0.0028	0.3139	0.0665
		0.19	0.009	0.0022	0.2173	0.055
		0.18	0.0068	0.0018	0.1533	0.0372
		0.17	0.0065	0.0017	0.1308	0.0315
		0.16	0.0059	0.0017	0.01210	0.03
		0.15	0.0063	0.0021	0.1212	0.033
		0.14	0.0073	0.0026	0.1232	0.037

Ces perturbations évoquent une convergence de l'algorithme plug-in-généralisé vers une valeur inférieure à celle du pas optimal. Cette divergence pourrait s'expliquer par l'accumulation d'erreurs lors de l'estimation des différentes entités intervenant dans l'estimation de $J_\phi(f)$. En nous inspirant de la méthode du Plug-in rapide publiée antérieurement [18], nous proposons d'essayer d'ajuster de manière empirique le pas optimal en faisant varier les puissances de l'entité $J_\phi(f)$ dans l'expression analytique du pas optimal h_N dont nous rappelons ci-dessous l'expression analytique

$$h_N^* = \left[M_\phi(K)\right]^{0.2}\left[J_\phi(f)\right]^{-0.2} N^{-0.2}$$

Nous présentons dans le tableau 3, les valeurs de l'EQMI estimées en faisant varier la puissance de $J_\phi(f)$ pour les distributions beta et uniforme pour des tailles d'échantillon de 1 000 et de 5 000. Dans les deux cas, nous constatons que les valeurs de l'EQMI sont minimales lorsque les puissances de l'entité $J_\phi(f)$ dans l'expression analytique du pas optimal h_N est de -0.16.

Par conséquent, nous proposons d'ajuster l'expression analytique du paramètre de lissage qui devient :

$$h_N^* = \left[M_\phi(K)\right]^{0.2}\left[J_\phi(f)\right]^{-0.16} N^{-0.2} \quad (24)$$

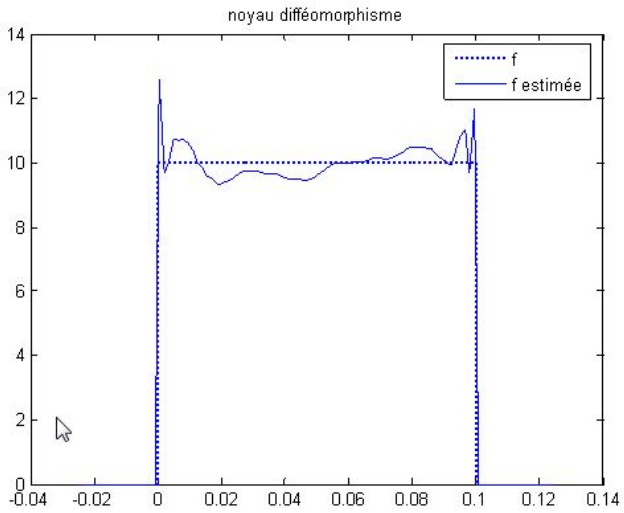

Figure 7. Estimation de la densité de probabilité d'une loi uniforme par la méthode du noyau classique avec ajustement du pas par l'algorithme Plug-in généralisé avec optimisation de la puissance de $J_\phi(f)$

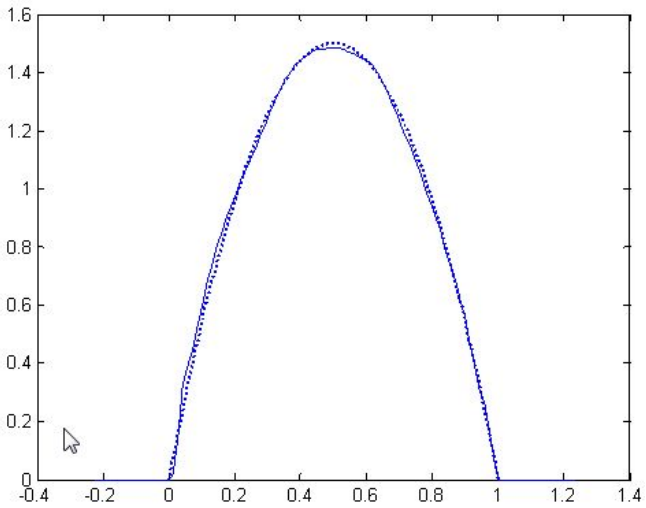

Figure 8. Estimation de la densité de probabilité d'une loi uniforme par la méthode du noyau classique avec ajustement du pas par l'algorithme Plug-in généralisé avec optimisation de la puissance de $J_\phi(f)$

Les figures 7 et 8 visualisent les densités estimées par la méthode du noyau-difféomorphisme avec ajustement du paramètre de lissage par l'algorithme du plug-in généralisé. On remarque une amélioration importante de la qualité de l'estimation confirmée par les valeurs des EQMI reportées dans le tableau 3.

6. Conclusion

Dans cet article, nous nous sommes focalisés sur la méthode du noyau difféomorphisme qui est une généralisation de la méthode du noyau conventionnelle pour une meilleure estimation des distributions à supports bornés ou semi-bornés. Cette méthode présente l'avantage de minimiser de manière importante le phénomène de Gibbs. L'algorithme plug-in pour l'optimisation du paramètre de lissage est également généralisé à la méthode du noyau-difféomorphisme. Il s'ensuit une complexité supplémentaire de l'estimation principalement due aux constatations suivantes :

– L'entité $M_\phi(K)$ n'est plus une constante ne dépendant que du noyau choisi. En effet, elle est liée a la densité inconnue à estimer et doit être évaluée au fil des itérations de l'algorithme plug-in généralisé ;

– L'entité $J_\phi(f)$ est fonction de f et de f' en plus de f'' ;

Comme pour la méthode du noyau classique, l'expression analytique du pas optimal pour la méthode du noyau difféomorphisme résulte d'une étude asymptotique visant à minimiser l'écart quadratique moyen intégré (EQMI). Ainsi, l'algorithme plug-in généralisé permet de converger vers la valeur optimale du paramètre de lissage et par conséquent de garantir une meilleure qualité d'estimation pour les densités bornées ou semi-bornées. Ces résultats sont illustrés par la simulation puis l'estimation de distributions bornées et semi-bornées.

Deux perspectives principales sont envisagées pour la suite de ces travaux. La première concerne l'application de cet estimateur à des données réelles. La seconde s'oriente vers la généralisation de cette méthode vers les densités multivariées.

Bibliographie

Botev Z.I., Grotowski J.F., Kroese D.P. (2010). Kernel density estimation via diffusion. *Annals of statistics*, vol. 38, p. 2916- 2957.

Bowman A.W. (1984). An alternative method of cross validation for smoothing of density estimates. *Biometrika*, vol.7, p. 353- 360.

Deheuvels P., Hominal P. (1980). Estimation automatique de la densité. *Revue de Statistiques Appliquée*, vol. 28, p. 25-55.

Ghorbel F. (2011). *Une approche unifiée des aspects géométriques et statistiques de la reconnaissance des formes planes*. ARTS-PI editions, Tunis, seconde édition.

Hall P. (1982). Comparison of two orthogonal series methods of estimating a density and its derivatives on interval. *J. Multivariate anal*, vol. 12, p. 432-449.

Hall P., Marron J.S. (1987). Estimation of integrated square density derivatives. *J. statistics and probability letters*, vol. 6, p.109-115.

Hall P., Marron J.S. (1987). Extent to which least-squares cross validation minimizes integrated square error in non parametric density estimation. *J. Probability Theory and related fields*, vol. 74, p. 567-581.

Hall P., Marron J.S. (1991). Lower bounds for bandwidth selection in density estimation. *J. Probability Theory and related fields*, vol. 90, p. 149-173.

Hall P., Marron J.S., Byeong U.P. (1992). Smoothed cross validation. *J. Probability Theory and related fields*, vol. 92, p. 1-20.

Hardle W. (1991). *Techniques with implementation in S*. Springer, New York.

Mugadi A.R., Ahmad I.A. (2004). A bandwidth selection for kernel density estimation of functions of randoms variables. *J. Computational statistics and data analysis*, vol. 47, p. 49-62.

Park B.U., Marron J.S. (1990). Comparison of data driver bandwidth selection. *Journal of the American Statistical Association*, vol. 85, p. 66-72.

Parzen E. (1962). On estimation of a probability density function and mode. *Annals of mathematical statistics*, vol. 33, p. 1065-1076.

Rozenblatt R. (1956). Remarks on some non-parametric estimates of a density function. *Annals of mathematical statistics*, vol. 27, p. 832-83.

Saoudi S., Ghorbel F., Hillion A. (1994). Non parametric probability probability density function estimation on a bounded support : applications to shape classification and speech coding. *J. Applied statistic Models and Data Analysis*, vol. 10, p. 215-231.

Saoudi S., Ghorbel F., Hillion A. (1997). Some statisticals properties of the kernel diffeomorphism estimator. *J. Applied statistic Models and Data Analysis*, vol. 10, p. 39-58.

Scott D.W., Terrel G.R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American statistical association*, vol. 82, p. 1131-1146.

Silverman B.W. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London.

Terrel G.R. (1990). The maximal smoothing principle in density estimation. *Journal of the American statistical association*, 85:470–477, 1990.

Troudi M., Alimi A. M., Saoudi S. (2008). Analytical Plug-in Method for Kernel Density Estimator Applied to Genetic Neutrality Study. *Eurasip Journal of Advances in Signal Processing*, vol. 2008, Article ID 739082, doi:10.1155/2008/739082.

Troudi M. (2009). *Optimisation du paramètre de lissage pour l'estimateur à noyau par des algorithmes itératifs : applications à des données réelles*. Thèse en Sciences pour Ingénieurs, Télécom Bretagne, France.