Estimateurs à noyau itérés : synthèse bibliographique

Gérard BIAU

Département de Mathématiques
Laboratoire de Probabilités et Statistique
Université Montpellier II
Place Eugène Bataillon
34095 Montpellier cedex 5
France

e-mail : biau@ensam.inra.fr

Résumé

Un estimateur à noyau d'une densité de probabilité dépend d'un paramètre de lissage réel appelé largeur de fenêtre ou plus simplement fenêtre. Le choix de ce paramètre est crucial, aussi bien pour la précision locale que pour la précision globale de l'estimateur. A ce jour, de nombreuses méthodes (dites de sélection) ont été proposées afin de choisir la meilleure fenêtre possible. Dans cet article, nous effectuons un point bibliographique sur l'un de ces algorithmes de sélection : le plug-in itéré.

Mots-clés **— Densité de probabilité, estimateur à noyau, paramètre de lissage, plug-in, méthodes itératives.**

Classification AMS **: 62G05.**

Abstract

A kernel estimate of a probability density depends on a smoothing parameter called bandwith. The choice of this parameter is crucial, as well for the local as for the global precision of the estimate. Numerous methods (called selection methods) were proposed to choose the best bandwith. In this article, we review one of these selection algorithms: the iterated plug-in.

Keywords and phrases **— Probability density, kernel estimate, smoothing parameter, plug-in, iterative methods.**

AMS Classification**: 62G05.**

1 Introduction

Dans cet article, nous nous intéressons au problème de l'estimation d'une densité de probabilité inconnue f à partir d'un échantillon $X_1, \dots, X_n$ de variables aléatoires indépendantes et de même loi à densité f. Il s'agit d'un problème fondamental de la statistique non paramétrique qui a connu, durant ces quarante dernières années, des développements théoriques et pratiques à la fois rapides et nombreux. L'idée la plus naturelle consiste à évaluer la densité f au point x en comptant le nombre d'observations "tombées" dans un certain voisinage de x. Sur $\mathbb{R}^d$ ($d \geq 1$), on peut par exemple choisir un voisinage cubique de $x = (x_1, \dots, x_d)$ de la forme $[x_1 - h/2, x_1 + h/2] \times \dots \times [x_d - h/2, x_d + h/2] = [x - h/2, x + h/2]^d$ où h est un nombre réel strictement positif dépendant de n, ce qui conduit à l'estimateur

$$f_h(x) = \frac{\#\left\{i \in \{1, \dots, n\} : X_i \in \ [x - h/2, x + h/2]^d\right\}}{nh^d},$$

le symbole $\#A$ désignant le cardinal de l'ensemble fini A. Cette dernière expression peut encore s'écrire

$$f_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} 1\!\!1_{[-\frac{1}{2},\frac{1}{2}]^d}\left(\frac{x - X_i}{h}\right), \tag{1}$$

où la fonction $1\!\!1_{[-1/2,1/2]^d}$ est la densité de probabilité uniforme sur $[-1/2, 1/2]^d$. En s'inspirant alors de la formule (1), et en définissant K comme étant une fonction réelle, bornée d'intégrale 1 sur $\mathbb{R}^d$, on définit *l'estimateur f_h associé au noyau K* par

$$f_h(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{2}$$

Lorsque le noyau K est choisi positif, l'estimateur f_h est une densité de probabilité et on parle alors parfois de *la densité de probabilité empirique de noyau K*. Parmi les multiples estimateurs non paramétriques de la densité aujourd'hui à la disposition des utilisateurs, l'estimateur à noyau est, de loin, le plus populaire (Akaike [1], Rosenblatt [52], Parzen [51], Silverman [64], Bosq et Lecoutre [5], Devroye [15], Berlinet et Devroye [2], Scott [57], Simonoff [65]). Le succès rencontré par l'estimateur à noyau auprès de la communauté des utilisateurs peut essentiellement s'expliquer en trois points :

- D'abord, l'expression théorique (2) de $f_h(x)$ est extrêmement simple, puisque $f_h(x)$ est la somme de n variables aléatoires indépendantes et identiquement distribuées ;

- Ensuite, f_h converge vers f en de nombreux sens, et en particulier au sens L_1 pour toute densité f dès que $1/h$ et nh^d tendent tous les deux vers l'infini (la dépendance de h en n sera toujours sous-entendue). D'autre part, si l'estimateur est convergent, il est convergent dans tous les modes, *i.e.* en probabilité, en moyenne, presque sûrement et presque complètement (Devroye et Györfi [19]) ;

- Enfin, l'estimateur à noyau 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 simplicité et la vitesse de calcul d'une part, la régularité de la courbe à obtenir d'autre part. En revanche, le choix du paramètre de lissage h se révèle crucial aussi bien pour la précision locale que pour la précision globale de l'estimateur f_h. Il est facile de vérifier que, pour les noyaux usuels et un ensemble de données fixé, la loi de densité f_h converge (étroitement) vers la mesure empirique lorsque h tend vers 0 et que f_h tend uniformément vers la fonction nulle lorsque h tend vers l'infini. En jouant sur la largeur de fenêtre, on peut donc faire décrire à f_h un ensemble de lois dont les extrêmes seront "proches" de lois discrètes d'un côté, uniformes de l'autre.

Lorsque le noyau K est fixé, sélectionner le meilleur h à partir des données signifie définir une variable aléatoire $h(X_1, \ldots, X_n)$ approchant au mieux la largeur de fenêtre optimale (déterministe) *au sens d'un certain critère à définir*. Si l'utilisateur n'a à choisir aucun paramètre a priori, nous dirons qu'une telle méthode de sélection est *automatique*. A ce jour, de nombreuses méthodes de sélection automatique ont été proposées, testées et comparées. Pour une revue exhaustive et mise à jour de ces méthodes, nous renvoyons à Marron [44], à Turlach [68], à Berlinet et Devroye [3], à Cao, Cuevas et González-Manteiga [8] ou encore à Devroye [18].

Dans cette note, nous proposons de faire le point sur une procédure de sélection connue sous le nom de *plug-in itéré*, intéressante tant par ses performances que par les concepts qu'elle véhicule. Depuis environ vingt ans, le plug-in itéré a fait l'objet d'une littérature peu abondante, parfois appliquée, souvent technique, en tous cas trop disparate dans le temps et dans la forme pour être accessible au plus grand nombre. Il nous a donc semblé intéressant d'effectuer une synthèse bibliographique sur le sujet. Dans un but didactique, nous avons volontairement éludé les considérations mathématiques trop techniques, afin de privilégier les idées forces et leur enchaînement. Les lecteurs désireux d'approfondir les techniques de démonstration se reporteront avec profit à la bibliographie.

La suite du présent article se divise en deux parties suivies d'une conclusion. La première partie (Section 2) présente brièvement les techniques dites de *plug-in* et expose les fondements du plug-in itéré. La seconde partie (Section 3) est consacrée à une analyse des techniques de sélection par plug-in itéré les plus récentes.

2 Du plug-in au plug-in itéré

2.1 Le plug-in ou procédure en deux étapes

La décision d'un choix optimal pour la constante de lissage suppose la spécification d'un critère d'erreur qui puisse être éventuellement optimisé. Bien sûr, l'optimalité n'est pas un concept absolu : elle est intimement liée au choix du critère, qui peut faire intervenir à la fois la densité inconnue f et l'estimateur f_h (donc h et le noyau K). Dans le cas qui nous intéresse, on cherche à minimiser l'*Erreur Quadratique Intégrée Moyenne*, que nous noterons dans les formules par son abréviation anglo-saxonne MISE, et qui est définie par

$$\mathrm{MISE}(h) = \mathbf{E} \int_{\mathbb{R}^d} \left[f_h(x) - f(x)\right]^2 \,\mathrm{d}x = \int_{\mathbb{R}^d} \mathbf{E}\left[f_h(x) - f(x)\right]^2 \,\mathrm{d}x,$$

où la permutation entre l'intégrale et l'espérance est justifiée par une application immédiate du théorème de Fubini. L'écriture précédente montre que l'Erreur Quadratique Intégrée Moyenne possède deux interprétations équivalentes : c'est à la fois une mesure de l'erreur globale moyenne et une mesure de l'erreur moyenne ponctuelle accumulée.

Dans la suite de cet article, nous nous restreignons, pour simplifier, au cas univarié ($d = 1$). Les résultats présentés se transposent avec plus ou moins de travail au cas multivarié ($d > 1$). Concernant l'Erreur Quadratique Intégrée Moyenne, on dispose en premier lieu du résultat suivant, dont on trouvera la preuve par exemple dans Scott [56] :

Théorème 2.1 *Si f a une dérivée seconde absolument continue, si $f''' \in L_2$ et si le noyau $K \in L_2$ est une densité de probabilité continue, symétrique de variance $\sigma_K^2 > 0$, alors, sous les conditions $h \to 0$ et $nh \to \infty$, on a le développement asymptotique :*

$$\mathrm{MISE}(h) = \frac{h^4}{4}\sigma_K^4 R(f'') + \frac{R(K)}{nh} + \mathrm{O}\Big(h^5 + \frac{1}{n}\Big), \tag{3}$$

où nous choisissons de noter, pour toute fonction g,

$$R(g) = \int_{\mathbb{R}} g^2(x)\,\mathrm{d}x.$$

Il est instructif de souligner que le premier terme du membre de droite du développement (3) est un terme de biais, alors que le second terme est un terme de variance. Visiblement, ces deux termes varient en sens inverse par rapport à h : une largeur de fenêtre trop importante entraînera une augmentation du biais et une diminution de la variance (phénomène de surlissage), alors qu'une largeur de fenêtre trop petite provoquera une inflation de la variance et une diminution du biais (phénomène de sous-lissage). De l'expression (3), on déduit sans peine que le paramètre de lissage h^* qui minimise l'Erreur Quadratique Intégrée Moyenne Asymptotique s'écrit

$$h^* = \alpha(K)\beta(f)n^{-1/5}, \tag{4}$$

où

$$\alpha(K) = \left[\frac{R(K)}{\sigma_K^4}\right]^{1/5} \quad \text{et} \quad \beta(f) = \left[\frac{1}{R(f'')}\right]^{1/5},$$

sous l'hypothèse (que nous supposerons toujours vérifiée dans la suite) $R(f'') \neq 0$ (la finitude de $R(f'')$ étant assurée par les hypothèses du Théorème 2.1, voir Rosenblatt [53] pour plus de détails). Notons que h^* est une quantité *déterministe* (qui dépend du nombre d'observations n). Dans la suite, nous désignerons par h_{MISE} le paramètre de lissage (inconnu, déterministe et dépendant de n) qui minimise l'Erreur Quadratique Intégrée Moyenne donnée par la formule (3).

La discussion précédente montre que le paramètre de lissage h^*, optimal au sens du critère de l'Erreur Quadratique Intégrée Moyenne Asymptotique, devra réaliser un compromis entre les valeurs de la variance et celles du biais. Outre sa nature asymptotique, la largeur de fenêtre optimale h^* dépend de la densité inconnue f au travers du paramètre $R(f'')$ 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 $R(f'')$ dans l'expression (4) par un estimateur approprié. Les références sur ce sujet ne manquent pas : voir par exemple Hall et Marron [31], [32], [33], [35] ou encore Jones et Sheather [40]. Cette approche conduit à un ensemble de méthodes que l'on a coutume de regrouper sous le vocable général de *méthodes plug-in*[1], et qui font l'objet d'une recherche active (pour des références plus anciennes, consulter Woodroofe [70], Deheuvels [12], Nadaraya [47], Deheuvels [13] ou encore Deheuvels et Hominal [14]). Par exemple, Sheather et Jones [63] suggèrent l'estimateur $\hat{R}_a(f'')$ suivant :

$$\hat{R}_a(f'') = \frac{1}{n^2 a^5} \sum_{i,j=1}^{n} L^{(iv)}\left(\frac{X_i - X_j}{a}\right),$$

où $L^{(iv)}$ désigne la dérivée quatrième d'un noyau suffisamment lisse L et où a est un nouveau paramètre de lissage (parfois appelé *paramètre pilote*). Cet estimateur s'obtient en écrivant $\hat{R}_a(f'') = R(f''_a)$ et en remarquant que, sous des conditions de régularité suffisantes,

$$\int_{\mathbb{R}} (g'')^2(x)\,\mathrm{d}x = \int_{\mathbb{R}} g^{(iv)}(x)\,g(x)\,\mathrm{d}x$$

(voir la troisième partie, ainsi que Deheuvels et Hominal [14] et Hall et Marron [31]). Dans ce cas, quelques considérations théoriques montrent que le paramètre de lissage optimal a^* s'écrit

$$a^* = C_0 n^{-1/7} \quad \text{avec} \quad C_0 = \left[\frac{2L^{(iv)}(0)}{\sigma_L^2 R(f''')}\right]^{1/7}.$$

A nouveau, la constante C_0 dépend du paramètre inconnu $R(f''')$ et il peut donc sembler que l'on ne réalise de la sorte aucune avancée significative. Néanmoins, pour sortir de ce

1. Le terme "plug-in" est difficile à traduire (on peut parler de *méthode en deux étapes* ou de *méthode de remplacement*). A défaut de traduction adéquate, nous continuerons à utiliser cette expression anglo-saxonne.

cercle vicieux, on peut toujours admettre que la quantité $R(f''')$ (qui dépend cette fois de la dérivée troisième de f) devrait être robuste par rapport à une erreur de spécification de la densité f. Dans cet esprit, Cao, Cuevas et González-Manteiga [8] suggèrent d'estimer $R(f''')$ en remplaçant simplement f par un modèle de référence paramétrique gaussien.

Une étude détaillée des différentes techniques (pour ne pas dire tours de main) de type plug-in nous entraînerait bien au-delà des limites du présent article. De l'exemple précédent, on retiendra surtout que les méthodes en deux étapes peuvent se décliner de manière pratiquement infinie. En dehors du cercle vicieux (comment choisissons-nous les paramètres de seconde étape ?) et de la nature asymptotique du processus (encore une fois, la formule donnée pour h^* n'est valable qu'asymptotiquement, sans aucune garantie pour un n particulier), nous sommes confrontés au fait qu'il est difficile de vérifier a priori que les conditions de convergence sont satisfaites. Lorsqu'elles ne le sont pas, il peut arriver que le h^* estimé se conduise mal et que l'estimateur de la densité résultant ne soit même pas convergent.

2.2 Le plug-in itéré : genèse d'une méthode

Dans un article en date de 1977, Scott, Tapia et Thompson [59] proposent de faire le point sur l'état des connaissances relatives à l'estimateur à noyau de la densité. Après avoir rappelé quelques-unes des principales propriétés de l'estimateur f_h, les trois auteurs s'interrogent sur le choix d'une valeur optimale pour la largeur de fenêtre h. En adoptant le critère de l'Erreur Quadratique Intégrée Moyenne (et sans réellement motiver leur approche), Scott, Tapia et Thompson choisissent d'estimer le paramètre $R(f'')$ de l'équation (4) à l'aide de l'estimateur naturel $\hat{R}_h(f'')$ défini comme suit :

$$\hat{R}_h(f'') = R(f_h''),$$

où f_h'' désigne la dérivée seconde de l'estimateur à noyau f_h. Avec un noyau K deux fois dérivable, il est facile de voir que

$$f_h''(x) = \frac{1}{nh^3} \sum_{i=1}^{n} K''\left(\frac{x - X_i}{h}\right),$$

ce qui conduit, en choisissant par exemple le classique noyau gaussien

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{u^2}{2}\right),$$

à l'estimateur $\hat{R}_h(f'')$ suivant :

$$\hat{R}_h(f'') = \frac{3}{8\sqrt{\pi}n^2h^9} \sum_{i,j=1}^{n} \left[h^4 - (X_i - X_j)^2h^2 + \frac{1}{12}(X_i - X_j)^4\right] \exp\left[-\frac{(X_i - X_j)^2}{4h^2}\right].$$

Il est très important de noter que la largeur de fenêtre h contrôlant l'estimateur $\hat{R}_h(f'')$ de $R(f'')$ a été volontairement choisie *identique* à la largeur de fenêtre intervenant dans l'estimation f_h de f. En conjecturant que la quantité $R(f'')$ devrait être robuste par rapport à une erreur de spécification sur f, Scott, Tapia et Thompson [59] proposent finalement d'injecter l'estimateur $\hat{R}_h(f'')$ dans l'expression (4). Cette approche amène naturellement à considérer l'équation numérique suivante en h :

$$h = \alpha(K)\beta(f_h)n^{-1/5}, \tag{5}$$

où, bien entendu,

$$\beta(f_h) = \left[\frac{1}{R(f_h'')}\right]^{1/5}.$$

Toute solution à l'équation (5) constitue un candidat potentiel à l'estimation de la largeur de fenêtre asymptotique optimale h^*. Cette équation, que l'on peut par exemple tenter de résoudre à l'aide d'un algorithme de type Newton, admet toujours la solution triviale 0. Lorsque l'équation admet plusieurs solutions, les auteurs proposent de choisir la plus grande (principe de surlissage) : nous la noterons alors h_∞ (la dépendance de h_∞ en les observations $X_1, \ldots, X_n$ étant sous-entendue, pour alléger les notations). Dans le cas contraire (absence de solution non-triviale), nous dirons que l'algorithme de sélection automatique ainsi défini est *dégénéré*.

Formulée différemment, la méthode de sélection suggérée par Scott, Tapia et Thompson revient à examiner les éventuels points fixes du système dynamique discret Φ défini sur $\mathbb{R}^+$ de la façon suivante :

$$h_{i+1} = \Phi(h_i), \tag{6}$$

où

$$\Phi(h_i) = \alpha(K)\beta(f_{h_i})n^{-1/5}.$$

La Figure 1 présente deux exemples de telles fonctions Φ. Sur chaque graphique nous avons fait figurer la première bissectrice, qui permet de détecter les éventuels points fixes du système dynamique. A notre connaissance, il s'agit de la première fois qu'une problématique de type dynamique (la détermination des points fixes d'un système) est utilisée pour résoudre un problème purement statistique (la détermination d'un paramètre de lissage optimal). Vu sous l'angle dynamique, la méthode de sélection pionnière de Scott, Tapia et Thompson consiste en fin de compte à répéter l'opération de plug-in (*cf.* le paragraphe précédent) jusqu'à une éventuelle convergence du paramètre de lissage : on parle alors de *plug-in itéré*.

Le Tableau 1 reporte une partie des simulations effectuées par Scott, Tapia et Thompson (il est à noter que les résultats ne sont pas discutés par les auteurs). Ces simulations ont été conduites à partir de quatre densités distinctes : une densité gaussienne $\mathcal{N}(0,1)$, un mélange à 50% de deux densités gaussiennes $\mathcal{N}(-1.5,1)$ et $\mathcal{N}(1.5,1)$, une densité de

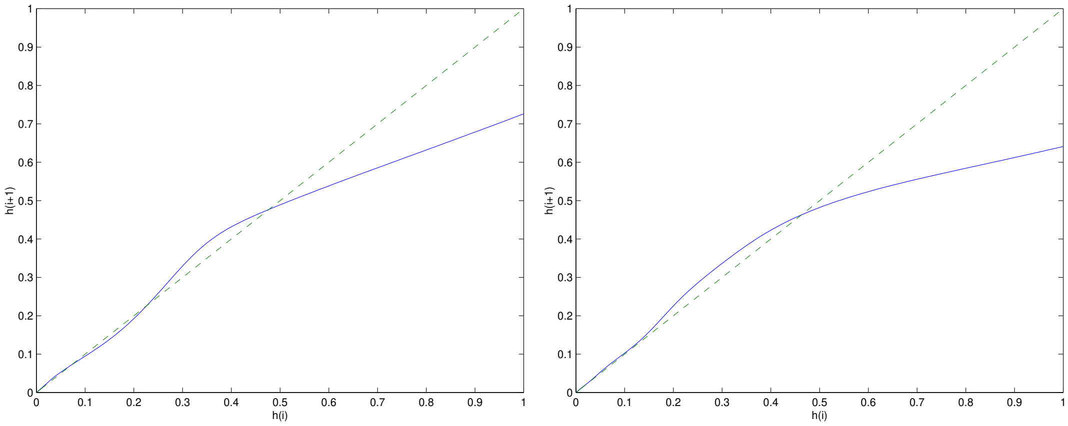

FIG. 1 – **En traits pleins** : deux exemples de fonctions Φ $\left(cf.\right.$ équation (6)$\left.\right)$ obtenues à partir d'un noyau gaussien et d'un échantillon de données gaussiennes. A gauche : $\mathbf{n = 20}$; à droite : $\mathbf{n = 100}$. **En pointillés** : la première bissectrice.

Student centrée t_5 et une densité de Fisher-Snedecor $F_{10,10}$. Afin d'étudier l'influence de la taille d'échantillon sur l'algorithme de sélection, deux tailles ($n = 25$ et $n = 100$) ont été considérées. Pour chaque type de densité et pour chaque taille d'échantillon, 25 répétitions d'expérience ont été conduites. Nous avons reporté les résultats relatifs à la largeur de fenêtre h_∞ (moyenne et écart-type de h_∞ sur les 25 répétitions, valeurs extrêmes et nombre de cas dégénérés).

Densité	n	Dégé.	Moyenne (écart-type)	Extr.	Eff.	h^*
$\mathcal{N}(0,1)$	25	1	0.54 (0.17)	0.20-0.80	0.67	0.56
$\frac{1}{2}\mathcal{N}(-\frac{3}{2},1)+\frac{1}{2}\mathcal{N}(\frac{3}{2},1)$	25	2	0.77 (0.41)	0.09-1.35	0.41	0.66
t_5	25	0	0.59 (0.19)	0.25-0.96	0.91	0.41
$F_{10,10}$	25	2	0.25 (0.09)	0.02-0.42	0.97	0.20
$\mathcal{N}(0,1)$	100	0	0.35 (0.10)	0.09-0.51	0.70	0.42
$\frac{1}{2}\mathcal{N}(-\frac{3}{2},1)+\frac{1}{2}\mathcal{N}(\frac{3}{2},1)$	100	0	0.43 (0.17)	0.12-0.76	0.68	0.50
t_5	100	0	0.37 (0.09)	0.13-0.54	0.81	0.31
$F_{10,10}$	100	1	0.15 (0.04)	0.05-0.20	0.77	0.15

TAB. 1 – Résultats des simulations effectuées par Scott, Tapia et Thompson relatifs au paramètre de lissage $\mathbf{h}_\infty$. Le noyau choisi est gaussien. $\mathbf{n}$ est la taille d'échantillon ; "Dégé." représente le nombre de solutions dégénérées obtenues sur les $\mathbf{25}$ répétitions ; "Extr." donne la plus petite et la plus grande des valeurs obtenues pour $\mathbf{h}_\infty$; "Eff." est une estimation de l'efficacité moyenne de l'estimateur ($cf.$ texte) ; $\mathbf{h}^*$ est le paramètre de lissage asymptotique optimal pour le modèle de densité testé $\left(cf.\right.$ équation (4)$\left.\right)$.

Ce premier tableau montre que les performances moyennes de l'estimateur h_∞ se dégradent lorsque la taille d'échantillon augmente. Cette observation est confirmée par une analyse de l'avant-dernière colonne du tableau, qui donne une estimation de l'efficacité moyenne de l'estimateur, définie comme l'espérance du rapport $\mathrm{MISE}(h_{\mathrm{MISE}})/\mathrm{MISE}(h_\infty)$. En outre, si l'estimateur semble biaisé (pour le mélange gaussien et la distribution de Student notamment), il apparaît en revanche relativement stable (écart-type faible). Nous reviendrons sur ces constatations à la fin du paragraphe.

L'étude originelle de Scott, Tapia et Thompson [59] a été reprise et complétée quatre ans plus tard par Scott et Factor [58]. Dans ce nouvel article, les deux auteurs proposent, entre autres choses, de comparer l'algorithme de sélection de Scott, Tapia et Thompson avec un autre algorithme de sélection automatique plus ancien proposé par Habbema, Hermans et Vandenbroek [26] et Duin [22], fondé sur un critère non-asymptotique de *maximum de vraisemblance*. Brièvement, il s'agit de maximiser par rapport à h la quantité $L(h)$ suivante :

$$L(h) = \prod_{k=1}^{n} f_{h,k}(X_k),$$

où

$$f_{h,k}(x) = \frac{1}{(n-1)h} \sum_{\substack{i=1 \\ i\neq k}}^{n} K\left(\frac{x - X_i}{h}\right)$$

est l'estimateur à noyau basé sur les $n-1$ observations différentes de X_k. Cette méthode de sélection n'est valide que dans certains cas (Hall [27], Rudemo [54], Chow, Geman et Wu [11], Devroye et Györfi [19], Marron [42]) et peut conduire à des estimateurs non-convergents lorsque la distribution a des queues qui décroissent à une vitesse au plus exponentielle (Schuster et Gregory [55]).

Des simulations analogues à celles conduites auparavant par Scott, Tapia et Thompson ont été mises en place par Scott et Factor pour l'estimateur h_∞ et pour l'estimateur du maximum de vraisemblance, disons h_{V}. Afin de souligner les éventuels dangers d'un modèle paramétrique insuffisamment spécifié, les deux auteurs proposent également d'estimer l'Erreur Quadratique Intégrée Moyenne à partir de l'ajustement d'un modèle paramétrique gaussien $\mathcal{N}(\overline{x}, \hat{\sigma}^2)$.

Comme on peut raisonnablement s'y attendre, l'ajustement d'un modèle paramétrique faux conduit à une Erreur Quadratique Intégrée Moyenne substantiellement plus importante que dans le cas non paramétrique. Cela étant, il semble que les performances du plug-in itéré et du maximum de vraisemblance soient proches, avec néanmoins un avantage accru pour l'estimateur h_{V} lorsque la taille d'échantillon devient grande. Dans ce dernier cas, la qualité de l'estimateur h_∞ se dégrade nettement, confirmant par là-même les enseignements tirés quatre ans plus tôt des travaux de Scott, Tapia et Thompson. Dans la dernière partie de leur article, Scott et Factor observent les variations du comportement des estimateurs h_∞ et h_{V} relativement à des données extrêmes. A la lumière de plusieurs simulations, les auteurs concluent que l'estimateur h_∞ est pratiquement insensible à l'addition d'une donnée extrême dans l'échantillon. En revanche, cela ne semble pas être le cas pour l'estimateur

h_{V}, dont le comportement peut vite devenir erratique (avec, en particulier, un accroissement catastrophique du biais). Enfin, les auteurs font remarquer que, d'un point de vue informatique, le calcul de h_∞ est bien plus rapide que celui de h_{V}.

Que faut-il retenir des travaux fondateurs de Scott, Tapia et Thompson [59] et Scott et Factor [58] ? Essentiellement deux points. D'abord, l'algorithme de sélection automatique par plug-in itéré est incontestablement pourvu d'avantages : stabilité, insensibilité aux données aberrantes et rapidité de calcul. Cependant, la méthode souffre manifestement, dans sa forme originelle, d'un défaut grave de convergence.

Ce défaut s'explique facilement. D'une part, l'algorithme repose sur le pari que la quantité $R(f'')$ est probablement robuste par rapport à une erreur de spécification sur f. Cette supposition tient de la gageure. D'autre part, quelques considérations théoriques $\big($*cf.* par exemple Scott [57]$\big)$ montrent que le paramètre de lissage optimal h^{**} qui minimise l'Erreur Quadratique Intégrée Moyenne Asymptotique entre f_h'' et f'' est en $n^{-4/5}$, ce qui implique donc $\big($*cf.* équation (4)$\big)$ que

$$\lim_{n\to\infty} \frac{h^*}{h^{**}} = +\infty. \tag{7}$$

Ainsi, lorsque la taille d'échantillon devient trop importante, la qualité de l'estimateur $R(f_h'')$ utilisé pour résoudre l'équation (5) se dégrade et l'algorithme a donc toutes les chances de mal se conduire.

3 Le plug-in itéré moderne

3.1 Une version convergente du plug-in itéré

Il faut attendre 1990 et un article de Park et Marron [49] pour voir apparaître une version convergente du plug-in itéré. Les travaux de Park et Marron s'appuient sur des résultats relatifs à l'estimation du paramètre $R(f'')$ obtenus trois ans plus tôt par Hall et Marron [31], et prolongent les idées de Sheather [61], [62] sur la mise au point d'une méthode de sélection pour le paramètre de lissage local (voir la troisième remarque du paragraphe suivant). Park et Marron proposent d'estimer le $R(f'')$ de l'expression (4) du paramètre de lissage optimal h^* à l'aide d'un estimateur $\tilde{R}_a(f'')$, a désignant une nouvelle constante de lissage réelle strictement positive. Une première idée consisterait à choisir l'estimateur $\tilde{R}_a(f'')$ égal à l'estimateur $\hat{R}_a(f'')$ $\big(= R(f_a'')\big)$. Cependant, quelques lignes de calcul élémentaire montrent que

$$\hat{R}_a(f'') = \int_{\mathbb{R}} \big[f_a''(x)\big]^2\,\mathrm{d}x = \frac{1}{na^5}R(K'') + \frac{1}{n^2a^5}\sum_{\substack{i,j=1\\ i\neq j}}^{n} K''*K''\bigg(\frac{X_i-X_j}{a}\bigg),$$

où le symbole $*$ dénote le produit de convolution (rappelons que le noyau K est supposé symétrique, voir le Théorème 2.1). En accord avec les travaux de Hall et Marron [31], qui conjecturent que le terme non-stochastique $R(K'')/(na^5)$ s'assimile à un terme de biais inutile, Park et Marron proposent de travailler avec l'estimateur $\tilde{R}_a(f'')$ défini par

$$\tilde{R}_a(f'') = \frac{n}{n-1}\Big[\hat{R}_a(f'') - \frac{1}{na^5}R(K'')\Big],$$

c'est-à-dire

$$\tilde{R}_a(f'') = \frac{1}{n(n-1)a^5}\sum_{\substack{i,j=1\\ i\neq j}}^{n} K'' * K''\Big(\frac{X_i - X_j}{a}\Big). \tag{8}$$

Il faut immédiatement souligner que, dans l'expression (8), le paramètre a fait référence à une largeur de fenêtre *éventuellement distincte* de h. Il s'agit de *la différence fondamentale* entre l'approche de Park et Marron et celle de Scott, Tapia et Thompson [59]. En introduisant cette nouvelle largeur de fenêtre, les auteurs ambitionnent de résoudre le problème lié à la convergence (7).

Un résultat dû à Hall et Marron [31] affirme alors, sous des conditions de régularité suffisantes (voir le Théorème 3.1 plus bas pour une liste exhaustive des hypothèses requises), que le paramètre de lissage optimal a^* qui minimise l'*Erreur Quadratique Moyenne* $\mathbf{E}\big[\tilde{R}_a(f'') - R(f'')\big]^2$ entre $\tilde{R}_a(f'')$ et $R(f'')$ admet la représentation asymptotique suivante :

$$a^* = C_1(K)C_2(f)n^{-2/13}, \tag{9}$$

où

$$C_1(K) = \left[\frac{18R(K^{(iv)} * K)}{\sigma^4_{K*K}}\right]^{1/13} \quad \text{et} \quad C_2(f) = \left[\frac{R(f)}{R^2(f''')}\right]^{1/13}.$$

L'équation (9) peut alors être combinée avec l'équation (4) pour fournir une expression reliant les deux largeurs de fenêtre a^* et h^*. Il vient ainsi :

$$a^* = C_3(K)C_4(f)(h^*)^{10/13}, \tag{10}$$

avec

$$C_3(K) = \left[\frac{18R(K^{(iv)} * K)\sigma^8_K}{\sigma^4_{K*K}R^2(K)}\right]^{1/13} \quad \text{et} \quad C_4(f) = \left[\frac{R(f)R^2(f'')}{R^2(f''')}\right]^{1/13}. \tag{11}$$

A ce niveau, Park et Marron soulignent que *la dépendance en f de la constante $C_4(f)$ est moins cruciale qu'auparavant.* Le calcul effectif de la constante $C_4(f)$ que nous avons effectué pour un certain nombre de modèles vient conforter cet argument (*cf.* le Tableau 2). Forts de cette observation, les auteurs choisissent alors d'estimer $C_4(f)$ en remplaçant f par un modèle paramétrique g_λ défini comme suit :

$$g_\lambda(x) = \frac{1}{\lambda}g_1\Big(\frac{x}{\lambda}\Big). \tag{12}$$

Densité	$C_4(f)$
$\mathcal{N}(0,1)$	0.788
Cauchy	0.637
Gumbel	0.719
$\frac{1}{2}\mathcal{N}(-1/2,1)+\frac{1}{2}\mathcal{N}(1/2,1)$	0.727
$\frac{1}{2}\mathcal{N}(0,1)+\frac{1}{2}\mathcal{N}(0,0.5)$	0.739
$\frac{1}{2}\mathcal{N}(0,1)+1/\left(2\pi(1+x^2)\right)$	0.725

TAB. 2 – Valeurs de la constante $\mathbf{C_4(f)}$ pour certaines densités.

Dans cette dernière expression, λ représente une mesure de l'échelle de la densité f, par exemple son écart-type ou son écart interquartile, et g_1 désigne une densité de référence connue d'échelle 1, typiquement la densité gaussienne $\mathcal{N}(0,1)$. Puisque

$$C_4(g_\lambda) = \lambda^{3/13} C_4(g_1),$$

l'équation (10) conduit naturellement à envisager la largeur de fenêtre $a_\lambda(h)$ définie par

$$a_\lambda(h) = C_3(K) C_4(g_1) \lambda^{3/13} h^{10/13}.$$

Park et Marron suggèrent finalement de choisir le paramètre de lissage h_∞ comme la plus grande des racines de l'équation

$$h = \alpha(K)\tilde{\gamma}(f) n^{-1/5},$$

où le terme $\alpha(K)$ est identique au terme de l'équation (5) initialement envisagée par Scott, Tapia et Thompson [59], mais où le terme $\beta(f_h)$ a été remplacé par

$$\tilde{\gamma}(f) = \left[\frac{1}{\tilde{R}_{a_{\hat{\lambda}(h)}}(f'')}\right]^{1/5},$$

$\hat{\lambda}$ désignant un bon estimateur (*i.e.* convergent avec une vitesse en $\sqrt{n}$) de λ. Moyennant quelques hypothèses de régularité sur la densité f et sur le noyau K, il est alors possible de donner des résultats concernant la loi asymptotique et la vitesse de convergence du rapport des largeurs de fenêtre $h_\infty / h_{\text{MISE}}$ (rappelons que les deux paramètres h_∞ et h_{MISE} dépendent de n, mais que seul h_∞ revêt un caractère stochastique). Avant d'énoncer le théorème fondamental, il nous faut introduire la définition technique suivante :

Définition 3.1 *Soient l un nombre entier et η un nombre réel dans* $]0,1]$. *On dit que f a un* degré de lissage d'ordre (l,η) *si f est au moins 2 fois dérivable et si il existe une constante réelle $M > 0$ telle que*

$$|f^{(2+l)}(x) - f^{(2+l)}(y)| \leq M|x-y|^\eta \quad \textit{pour tous } x, y.$$

Le résultat fondamental, initialement dû à Park et Marron, et amélioré par Park [48], est alors le suivant :

Théorème 3.1 *Sous les cinq hypothèses :*

(i) *Le noyau $K \in L_2$ est une densité de probabilité symétrique ; K est quatre fois différentiable et ses dérivées sont lipschitziennes ;*

(ii) *f admet une dérivée seconde absolument continue ;*

(iii) *f admet une dérivée quatrième lipschitzienne, $f^{(iv)} \in L_2$;*

(iv) *f admet un degré de lissage d'ordre (l, η) ;*

(v) *Il existe deux constantes $0 < \underline{B} < \overline{B}$ telles que*

$$h \in [\underline{B}n^{-1/5}, \overline{B}n^{-1/5}] \, ;$$

on a, en posant $\nu = l + \eta$,

(a) *Si $0 < \nu \leq 1$,*

$$\frac{h_\infty}{h_{\text{MISE}}} = 1 + \mathrm{O}_{\mathrm{P}}(n^{-4\nu/13}) \, ;$$

(b) *Si $\nu > 1$,*

$$n^{4/13}\left(\frac{h_\infty}{h_{\text{MISE}}} - 1\right) \xrightarrow{\mathcal{L}} \mathcal{N}(\mu_\infty, \sigma_\infty^2), \tag{13}$$

où

$$\mu_\infty = \frac{1}{5}\big[C_3(K)C_4(g_\lambda)\big]^2 R(f''')R^{4/13}(K)\sigma_K^{10/13}R^{-17/13}(f'')$$

et

$$\sigma_\infty^2 = \frac{2}{25}\frac{\sigma_K^{72/13}R(K'' * K'')R(f)R^{-18/13}(K)R^{-8/13}(f'')}{\big[C_3(K)C_4(g_\lambda)\big]^9}$$

(voir le Théorème 2.1 pour la définition de σ_K et les équations (11) et (12) pour la définition de $C_3(K)$, $C_4(g_\lambda)$, λ et g_λ).

En particulier,

$$\frac{h_\infty}{h_{\text{MISE}}} = 1 + \mathrm{O}_{\mathrm{P}}(n^{-4/13}).$$

Ce théorème appelle plusieurs remarques et commentaires :

1. Les conditions (i) et (ii) sont relatives au développement asymptotique de l'Erreur Quadratique Intégrée Moyenne (*cf.* le Théorème 2.1).

2. Les conditions (i) et (iii) sont essentielles pour l'existence de la largeur de fenêtre optimale a^*.

3. Selon Park et Marron, la condition (v), a priori restrictive, ne pose aucun problème majeur dans la mesure où l'intervalle $[\underline{B}n^{-1/5}, \overline{B}n^{-1/5}]$ contient toutes les largeurs de fenêtre raisonnables, pourvu que $\underline{B}$ soit choisi assez petit et $\overline{B}$ assez grand. Cette justification est hautement insuffisante et ne trouve de sens que dans un cadre asymptotique. En effet, imposer la condition (v) revient en fait, à distance finie, à tronquer la quantité aléatoire h_∞ dans l'intervalle $[\underline{B}n^{-1/5}, \overline{B}n^{-1/5}]$, et il faudrait alors se poser beaucoup plus sérieusement la question importante du choix des deux constantes $\underline{B}$ et $\overline{B}$.

4. Des résultats similaires de convergence et de normalité asymptotique sont aussi disponibles pour les fenêtres h_{nb} (Hall et Marron [32]) et h_{b} (Scott et Terrell [60]), voir plus bas pour les définitions. On retiendra que, pour une densité f suffisamment lisse, la vitesse de convergence de ces deux estimateurs est en $n^{-1/10}$, ce qui est significativement plus lent que la vitesse en $n^{-4/13}$ de l'estimateur h_∞.

5. L'égalité (10) montre que
$$\lim_{n\to\infty} \frac{h^*}{a^*} = +\infty.$$
Cela signifie que, asymptotiquement, la largeur de fenêtre optimale h^* pour l'estimateur à noyau de f n'est pas du même ordre que la largeur de fenêtre optimale a^* pour l'estimateur à noyau $\tilde{R}_a(f'')$ de $R(f'')$. On retrouve ici la cause de l'échec de l'algorithme de sélection proposé treize ans auparavant par Scott, Tapia et Thompson [59]. L'idée essentielle introduite par Park et Marron pour remédier à cette difficulté consiste à estimer la quantité $R(f'')$ avec un paramètre de lissage $a_{\hat{\lambda}(h)}$ du même ordre asymptotique que a^*. Au prix de cet effort supplémentaire (très peu coûteux en temps de calcul), l'estimateur h_∞ ainsi obtenu est convergent, avec de surcroît une vitesse de convergence supérieure à celle de ses principaux compétiteurs.

Park et Marron proposent de comparer leur version moderne du plug-in itéré avec deux autres algorithmes de sélection : *la validation croisée* L_2 et *la validation croisée biaisée.* Brièvement :

- Le critère de validation croisée L_2 (parfois appelée *validation croisée non-biaisée*) a été proposé par Rudemo [54] et Bowman [6]. Ce critère consiste à choisir le paramètre de lissage h_{nb} qui minimise un estimateur convenable de
$$\int_{\mathbb{R}} \left[f_h(x) - f(x)\right]^2 \mathrm{d}x - \int_{\mathbb{R}} f^2(x)\,\mathrm{d}x = \int_{\mathbb{R}} f_h^2(x)\,\mathrm{d}x - 2\int_{\mathbb{R}} f_h(x)f(x)\,\mathrm{d}x,$$
par exemple
$$\int_{\mathbb{R}} f_h^2(x)\,\mathrm{d}x - \frac{2}{n}\sum_{k=1}^{n} f_{h,k}(X_k),$$

où

$$f_{h,k}(x) = \frac{1}{(n-1)h} \sum_{\substack{i=1 \\ i \neq k}}^{n} K\left(\frac{x - X_i}{h}\right).$$

L'optimalité asymptotique de la validation croisée L_2 a été obtenue par Stone [66]. Pour d'autres études, voir également Hall [28], [29], Burman [7], Scott et Terrell [60]. Cette méthode de sélection possède deux défauts importants : d'une part, la fonctionnelle cible à minimiser a souvent tendance à présenter plusieurs minima locaux (Hall et Marron [34]) ; d'autre part, le résultat de l'estimation peut se révéler extrêmement variable d'un échantillon à l'autre (le sous-lissage est fréquent ; voir Hall et Marron [32], [33], Marron [43], Hall, Marron et Park [36]).

- Le critère de validation croisée biaisée, introduit par Scott et Terrell [60] pour remédier aux défauts de la validation croisée non-biaisée, amène à sélectionner la largeur de fenêtre h_{b} qui minimise l'estimateur $\widehat{\mathrm{AMISE}}(h)$ de l'Erreur Quadratique Intégrée Moyenne Asymptotique (*cf.* équation (3)) défini comme suit :

$$\widehat{\mathrm{AMISE}}(h) = \frac{h^4}{4}\sigma_K^4 \tilde{R}(f'') + \frac{R(K)}{nh},$$

  où $\tilde{R}(f'')$ est un estimateur de $R(f'')$. Naturellement, Park et Marron proposent de travailler avec l'estimateur $\tilde{R}_h(f'')$ calqué sur (8).

Les résultats des simulations effectuées par Park et Marron sont présentés au Tableau 3. Ces résultats ont été obtenus pour les estimateurs h_∞, h_{nb} et h_{b} à partir de données simulées suivant une densité gaussienne $\mathcal{N}(0,1)$, un mélange à 50% de deux densités gaussiennes $\mathcal{N}(-1,4/9)$ et $\mathcal{N}(1,4/9)$, un mélange $0.75\mathcal{N}(0,1)+0.25\mathcal{N}(0,0.04)$ et un mélange $\frac{1}{2}\mathcal{N}(0,1)+\frac{1}{2}\mathcal{N}(0,0.09)$. Deux tailles d'échantillon ont été testées ($n = 100$, $n = 400$) et 500 répétitions d'expérience effectuées. En outre, la cinquième colonne du tableau donne un intervalle de confiance empirique à 95% pour l'efficacité moyenne de l'estimateur f_{h_∞} (voir page 9 pour la définition de l'efficacité moyenne et Marron [45] pour la construction de cet intervalle de confiance). On trouvera également dans le tableau les résultats obtenus pour un autre candidat à la largeur de fenêtre optimale, l'estimateur h_{os}, dit *paramètre de surlissage* (oversmoothing parameter en anglais), et défini par

$$h_{\mathrm{os}} = 7^{1/2}\left[\frac{2R(K)}{45\sigma_K^2}\right]^{1/5} \hat{\sigma}^{-1/5},$$

où $\hat{\sigma}$ représente l'écart-type empirique de l'échantillon. On montre en effet (Terrell et Scott [67]) que, sous les conditions du Théorème 2.1, l'ensemble des largeurs de fenêtre optimales

$$\left\{h^*(f), f \text{ vérifie les conditions du Théorème 2.1}\right\}$$

admet une borne supérieure. h_{os} est alors construit comme un estimateur de cette borne supérieure. Intuitivement, il est clair que la largeur de fenêtre h_{os} devrait donner de bons

Densité	Méthode	n	Moyenne (écart-type)	I	h^*
$\mathcal{N}(0,1)$	h_∞	100	0.479 (0.066)	0.928-0.942	0.445
-	h_{nb}	-	0.439 (0.118)	0.799-0.837	-
-	h_{b}	-	0.508 (0.056)	0.914-0.932	-
-	h_{os}	-	0.441 (0.031)	0.986-0.989	-
-	h_∞	400	0.344 (0.026)	0.978-0.983	0.330
-	h_{nb}	-	0.322 (0.069)	0.865-0.894	-
-	h_{b}	-	0.348 (0.026)	0.976-0.981	-
-	h_{os}	-	0.334 (0.012)	0.996-0.997	-
$\frac{1}{2}\mathcal{N}(-1,4/9)+\frac{1}{2}\mathcal{N}(1,4/9)$	h_∞	100	0.508 (0.083)	0.828-0.861	0.385
-	h_{nb}	-	0.410 (0.130)	0.810-0.846	-
-	h_{b}	-	0.787 (0.116)	0.474-0.536	-
-	h_{os}	-	0.526 (0.027)	0.827-0.860	-
-	h_∞	400	0.317 (0.031)	0.922-0.938	0.272
-	h_{nb}	-	0.271 (0.058)	0.894-0.915	-
-	h_{b}	-	0.375 (0.140)	0.641-0.697	-
-	h_{os}	-	0.399 (0.010)	0.702-0.751	-
$0.75\mathcal{N}(0,1)+0.25\mathcal{N}(0,0.04)$	h_∞	100	0.263 (0.069)	0.778-0.818	0.185
-	h_{nb}	-	0.204 (0.077)	0.810-0.846	-
-	h_{b}	-	0.441 (0.093)	0.437-0.499	-
-	h_{os}	-	0.382 (0.033)	0.528-0.589	-
-	h_∞	400	0.156 (0.021)	0.878-0.903	0.126
-	h_{nb}	-	0.129 (0.029)	0.904-0.923	-
-	h_{b}	-	0.195 (0.066)	0.602-0.660	-
-	h_{os}	-	0.290 (0.012)	0.335-0.393	-
$\frac{1}{2}\mathcal{N}(0,1)+\frac{1}{2}\mathcal{N}(0,0.09)$	h_∞	100	0.212 (0.038)	0.896-0.917	0.186
-	h_{nb}	-	0.191 (0.052)	0.837-0.868	-
-	h_{b}	-	0.290 (0.109)	0.551-0.612	-
-	h_{os}	-	0.324 (0.033)	0.518-0.579	-
-	h_∞	400	0.145 (0.013)	0.964-0.972	0.134
-	h_{nb}	-	0.133 (0.029)	0.890-0.912	-
-	h_{b}	-	0.148 (0.014)	0.953-0.963	-
-	h_{os}	-	0.245 (0.013)	0.413-0.474	-

TAB. 3 – Résultats des simulations effectuées par Park et Marron relatifs aux paramètres de lissage $\mathbf{h}_\infty$, $\mathbf{h}_{\mathbf{nb}}$, $\mathbf{h}_{\mathbf{b}}$ et $\mathbf{h}_{\mathbf{os}}$. $\mathbf{n}$ est la taille d'échantillon ; $\mathbf{I}$ est un intervalle de confiance empirique à $\mathbf{95}$ % pour l'efficacité moyenne de l'estimateur (*cf.* texte) ; $\mathbf{h}^*$ est le paramètre de lissage asymptotique optimal relatif au modèle de densité testé (*cf.* équation (4)). Le noyau utilisé n'est pas précisé.

résultats lorsque la densité à estimer présente un faible niveau de complexité structurelle (unimodalité, par exemple). En revanche, sa qualité devrait se dégrader pour des densités cibles plus complexes, la fenêtre étant alors trop large pour permettre à l'estimateur à noyau de rendre compte des variations de la densité (phénomène de surlissage).

En moyenne, le paramètre de lissage h_{nb} approche mieux que h_∞ la valeur asymptotique optimale h^*. Park et Marron expliquent ce résultat par la présence du terme de biais perturbateur μ_∞ de la formule (13), égal à 0 dans le cas de la validation croisée. Néanmoins, l'estimateur h_∞ est plus stable (variance plus faible) et ses résultats relatifs à l'Erreur Quadratique Intégrée Moyenne sont souvent les meilleurs, comme pouvait le laisser supposer sa bonne vitesse de convergence. On notera que les performances relatives de h_∞ s'améliorent bien avec la taille d'échantillon, en accord avec le Théorème 3.1 et la Remarque 4. Si, comme cela est prévisible, h_{os} est bien le meilleur pour la densité gaussienne, il est en revanche de qualité médiocre pour les mélanges bimodaux. Sur la base du Théorème 3.1 et des simulations précédentes, Park et Marron concluent finalement leur article en louant la supériorité de l'algorithme de sélection par plug-in itéré.

Afin de compléter l'étude de Park et Marron, nous présentons au Tableau 4 les résultats de quelques simulations supplémentaires. Si les résultats obtenus pour la densité gaussienne $\mathcal{N}(0,1)$ et la densité de Gumbel confirment les conclusions de Park et Marron, les résultats obtenus pour la densité de Cauchy sont en revanche fort décevants. Pour expliquer ce phénomène, il suffit de constater que, pour la densité de Cauchy, le paramètre de lissage asymptotique optimal h^* donné par la formule (4) est manifestement trop petit, voire aberrant (les courbes d'estimation sont très irrégulières), même pour une taille d'échantillon "grande" (*cf.* Figure 2, à droite). Cet exemple nous rappelle que la formule (4) ne revêt de sens que dans un cadre asymptotique, et que l'utilisation de cette formule à distance finie doit être entourée de beaucoup de précautions. Dans cet esprit, un des rapporteurs souligne le fait que l'expression (4) du paramètre de lissage optimal ne dépend de la densité à estimer qu'au travers de la quantité $R(f'')$, toutes les autres caractéristiques de la densité cible f ayant été "balayées" dans le $\mathrm{O}(h^5+1/n)$ de la formule (3). Ainsi, pour deux densités de probabilité f_1 et f_2 telles que $R(f_1'') = R(f_2'')$, la formule (4) fournit deux largeurs de fenêtre optimales identiques, mais dont les précisions asymptotiques (transparentes pour l'utilisateur) peuvent être fort différentes. Choisissons par exemple pour f_1 une densité de Cauchy et pour f_2 une densité gaussienne $\mathcal{N}(0, (2^{4/5} \times \pi^{1/10})/2 \approx 0.976)$. Il est facile de voir que, dans ce cas, $R(f_1'') = R(f_2'') = 3/(4\pi)$ et que le paramètre de lissage optimal commun aux deux modèles de densités pour $n = 1000$ a pour valeur $h^* = 0.260$. La Figure 2 montre que, si ce paramètre de lissage est correct pour la densité gaussienne (à gauche), il est en revanche visiblement trop petit pour la densité de Cauchy (à droite).

3.2 Quelques compléments

1. L'algorithme de sélection par plug-in itéré développé par Park et Marron [49] s'inscrit dans la continuité des travaux de Hall et Marron [31] relatifs à l'estimation du para-

Densité	Méthode	n	Moyenne (écart-type)	Eff.	h^*
$\mathcal{N}(0,1)$	h_∞	100	0.458 (0.070)	0.870	0.422
-	$h_{\rm nb}$	-	0.440 (0.123)	0.767	-
-	$h_{\rm b}$	-	0.504 (0.058)	0.919	-
-	h_∞	400	0.339 (0.034)	0.914	0.320
-	$h_{\rm nb}$	-	0.319 (0.078)	0.816	-
-	$h_{\rm b}$	-	0.351 (0.027)	0.962	-
Gumbel	h_∞	100	0.476 (0.074)	0.900	0.408
-	$h_{\rm nb}$	-	0.454 (0.129)	0.782	-
-	$h_{\rm b}$	-	0.540 (0.072)	0.897	-
-	h_∞	400	0.345 (0.029)	0.941	0.309
-	$h_{\rm nb}$	-	0.321 (0.074)	0.852	-
-	$h_{\rm b}$	-	0.354 (0.029)	0.965	-
Cauchy	h_∞	100	1.225 (2.987)	0.665	0.412
-	$h_{\rm nb}$	-	0.505 (0.157)	0.863	-
-	$h_{\rm b}$	-	0.741 (0.201)	0.786	-
-	h_∞	400	0.732 (0.823)	0.599	0.312
-	$h_{\rm nb}$	-	0.345 (0.080)	0.871	-
-	$h_{\rm b}$	-	0.406 (0.045)	0.917	-

TAB. 4 – Résultats de simulations complémentaires relatifs aux paramètres de lissage $\mathbf{h}_\infty$, $\mathbf{h}_{\mathbf{nb}}$ et $\mathbf{h}_{\mathbf{b}}$. Le noyau choisi est gaussien. $\mathbf{n}$ est la taille d'échantillon ; "Eff." est une estimation de l'efficacité moyenne de l'estimateur (*cf.* texte) ; $\mathbf{h}^*$ est le paramètre de lissage asymptotique optimal relatif au modèle testé (*cf.* équation (4)).

mètre $R(f'')$. Rappelons qu'à l'estimateur naturel $\hat{R}_a(f'') = R(f''_a)$, Hall et Marron préfèrent l'estimateur $\tilde{R}_a(f'')$ donné par

$$\tilde{R}_a(f'') = \frac{1}{n(n-1)a^5} \sum_{\substack{i,j=1 \\ i\neq j}}^{n} K'' * K''\Big(\frac{X_i - X_j}{a}\Big),$$

qui fait disparaître le terme de biais "artificiel" (déterministe) $R(K'')/(na^5)$ contenu dans l'estimateur $\hat{R}_a(f'')$. En 1991, dans un article parallèle à celui de Hall et Marron, Jones et Sheather [40] préconisent au contraire l'utilisation de l'estimateur naturel $\hat{R}_a(f'')$, en faisant observer que le terme de biais "artificiel" est (au moins pour la plupart des noyaux usuels) positif et peut donc servir à annihiler le terme de biais (négatif) de l'Erreur Quadratique Moyenne entre $\hat{R}_a(f'')$ et $R(f'')$. Forts de cette constatation, Sheather et Jones [63] proposent de tester l'algorithme de sélection de Park et Marron [49] avec l'estimateur $\hat{R}_a(f'')$. Afin de faire disparaître quelques effets indésirables du terme de biais, les deux auteurs sont contraints de mettre en place une procédure de type plug-in en trois étapes. Au prix de cet effort supplémentaire,

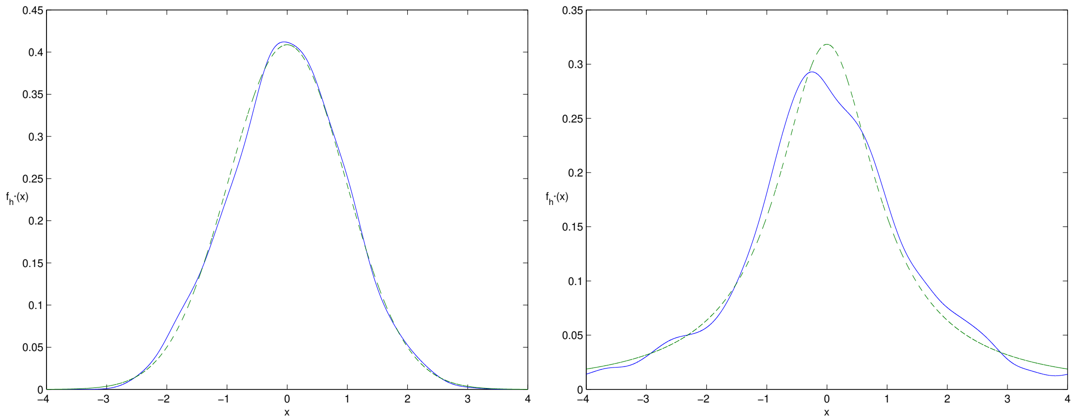

FIG. 2 – **A gauche** : Estimation à noyau à partir de **1000** données réparties suivant une densité gaussienne $\mathcal{N}(\mathbf{0}, \mathbf{0.976})$ (en pointillés). **A droite** : Estimation à noyau à partir de **1000** données réparties suivant une densité de Cauchy (en pointillés). Dans les deux cas, l'estimation a été obtenue à partir du paramètre de lissage asymptotique optimal $\mathbf{h}^* = \mathbf{0.260}$ donné par la formule (4).

leur estimateur h'_∞ vérifie, pour une densité f suffisamment lisse :

$$\frac{h'_\infty}{h_{\text{MISE}}} = 1 + \mathrm{O}_{\mathrm{P}}(n^{-5/14}).$$

Malgré les complications supplémentaires introduites par la procédure en trois étapes, le gain de vitesse par rapport à h_∞ n'est pas extraordinaire. Sous certaines hypothèses, Hall et Marron [35] ont montré que la vitesse maximale que pouvait atteindre un estimateur $\hat{h}$ de la fenêtre optimale était en $n^{-1/2}$, autrement dit, dans le meilleur des cas :

$$\frac{\hat{h}}{h_{\text{MISE}}} = 1 + \mathrm{O}_{\mathrm{P}}(n^{-1/2}).$$

A ce jour, plusieurs algorithmes de sélection convergeant avec une vitesse en $n^{-5/14}$ ont été proposés (Hall, Marron et Park [36], Engel, Herrmann et Gasser [23]) et certaines méthodes en $n^{-1/2}$ commencent à être disponibles (Chiu [9], [10], Hall *et al* [27], Jones, Marron et Park [39], Marron [46], Kim, Park et Marron [41]). Pour des compléments théoriques concernant ces méthodes, on pourra se reporter à Fan et Marron [24] ainsi qu'à Park et Marron [50].

2. La formule (10) donne le lien entre le paramètre de lissage (asymptotique) optimal a^* pour l'estimateur $\tilde{R}_a(f'')$ et le paramètre de lissage (asymptotique) optimal h^* pour l'estimateur f_h. Pour mémoire :

$$a^* = C_3(K)C_4(f)(h^*)^{10/13},$$

où

$$C_3(K) = \left[\frac{18R(K^{(iv)} * K)\sigma_K^8}{\sigma_{K*K}^4 R^2(K)}\right]^{1/13} \quad \text{et} \quad C_4(f) = \left[\frac{R(f)R^2(f'')}{R^2(f''')}\right]^{1/13}.$$

Il est facile de voir que cette dernière égalité peut se réécrire

$$a^* = C_3(K,p)C_4(f,p)(h^{*p})n^{-q}, \tag{14}$$

où p est un paramètre réel, $q = p/5 - 2/13$,

$$C_3(K,p) = \left[\frac{18R(K^{(iv)} * K)\sigma_K^{52p/5}}{\sigma_{K*K}^4 R^{13p/5}(K)}\right]^{1/13} \quad \text{et} \quad C_4(f,p) = \left[\frac{R(f)R^{13p/5}(f'')}{R^2(f''')}\right]^{1/13}.$$

Lorsque $p = 10/13$ (et donc $q = 0$), on retrouve $C_3(K,p) = C_3(K)$ et $C_4(f,p) = C_4(f)$. Park [48] propose de laisser flotter le paramètre réel p et d'introduire le modèle paramétrique g_λ $\big($*cf.* (12)$\big)$ dans l'expression (14) plutôt que dans l'expression (10) de a^*. Cela le conduit donc à envisager la fenêtre $a_\lambda(h,p)$ définie par

$$a_\lambda(h,p) = C_3(K,p)C_4(g_\lambda,p)h^p n^{-q},$$

qui fournit l'estimateur $h_\infty(p)$ par résolution classique de l'équation de plug-in itéré. Park montre que l'estimateur $h_\infty(p)$ possède les mêmes propriétés que l'estimateur h_∞ et qu'il vérifie, en particulier, le Théorème 3.1. En outre, Park observe que l'Erreur Quadratique Moyenne Asymptotique (AMSE en anglais) entre $h_\infty(p)$ et h_{MISE} peut s'écrire

$$\text{AMSE}(p) = C(f,K)[Q_p^4 + \frac{4}{9}Q_p^{-9}],$$

où $C(f,K)$ est une constante ne dépendant que du noyau K et de la densité cible f, et où

$$Q_p^{13} = \left[\frac{R(g_1)}{R(f_1)}\right]\left[\frac{R(f_1''')}{R(g_1''')}\right]^2\left[\frac{R(g_1'')}{R(f_1'')}\right]^{13p/5},$$

avec

$$f_1(x) = \lambda f(\lambda x).$$

Sans résoudre le problème, Park conjecture qu'un choix judicieux du paramètre réel p pourrait permettre de rendre l'Erreur Quadratique Moyenne Asymptotique insensible au choix de la densité de référence g_1.

3. Comme nous l'avons souligné au début de la première partie, l'Erreur Quadratique Intégrée Moyenne représente une mesure *globale* de la qualité de l'estimateur f_h. Cependant, dans de nombreux contextes (estimation de la médiane par exemple), on s'intéresse plus volontiers aux performances *locales* de l'estimateur à noyau. Dans ce dernier cas, afin de déterminer la largeur de fenêtre optimale en un point x_0 d'intérêt,

on cherche souvent à minimiser l'Erreur Quadratique Moyenne (MSE) entre $f_h(x_0)$ et $f(x_0)$, définie comme suit :

$$\mathrm{MSE}(x_0, h) = \mathbf{E}\big[f_h(x_0) - f(x_0)\big]^2.$$

Moyennant quelques conditions de régularité sur le noyau K et sur la densité sous-jacente f, Parzen [51] a montré que la largeur de fenêtre optimale $h^*(x_0)$ qui minimise l'Erreur Quadratique Moyenne Asymptotique entre $f_h(x_0)$ et $f(x_0)$ s'écrit

$$h^*(x_0) = \alpha(K)\beta\big[f(x_0), f''(x_0)\big]n^{-1/5}, \tag{15}$$

où

$$\alpha(K) = \left[\frac{R(K)}{\sigma_K^4}\right]^{1/5} \quad \text{et} \quad \beta\big[f(x_0), f''(x_0)\big] = \left[\frac{f(x_0)}{\big[f''(x_0)\big]^2}\right]^{1/5}.$$

Dans deux articles, Sheather [61], [62] s'intéresse au problème de l'estimation du paramètre $h^*(x_0)$ par plug-in itéré. L'auteur choisit d'estimer la dérivée seconde de f au point x_0 à l'aide de l'estimateur à noyau naturel $f_a''(x_0)$ et montre que le paramètre de lissage optimal $a^*(x_0)$ qui minimise l'Erreur Quadratique Moyenne Asymptotique entre $f_a''(x_0)$ et $f''(x_0)$ s'écrit

$$a^*(x_0) = \delta(K)\gamma\big[f(x_0), f^{(iv)}(x_0)\big]n^{-1/9}, \tag{16}$$

avec

$$\delta(K) = \left[\frac{5R(K'')}{\sigma_K^4}\right]^{1/9} \quad \text{et} \quad \gamma\big[f(x_0), f^{(iv)}(x_0)\big] = \left[\frac{f(x_0)}{\big[f^{(iv)}(x_0)\big]^2}\right]^{1/9}.$$

En remplaçant la densité inconnue f par le modèle de référence paramétrique g_λ $\big($équation (12)$\big)$ et en regroupant les deux équations (15) et (16), il vient

$$a_\lambda\big[h(x_0)\big] = \mu(K)\tau\big[g_1''(x_0), g_1^{(iv)}(x_0)\big]\lambda^{4/9}\big[h(x_0)\big]^{5/9},$$

où l'on a noté

$$\mu(K) = \left[\frac{5R(K'')}{R(K)}\right]^{1/9} \quad \text{et} \quad \tau\big[g_1''(x_0), g_1^{(iv)}(x_0)\big] = \left[\frac{g_1''(x_0)}{g_1^{(iv)}(x_0)}\right]^{2/9}.$$

Après plusieurs essais, Sheather propose finalement de choisir le paramètre de lissage $h_\infty(x_0)$ solution de l'équation

$$h(x_0) = \alpha(K)\beta\big[f_{h(x_0)}(x_0), f''_{a_{\hat\lambda}[h(x_0)]}(x_0)\big]n^{-1/5},$$

$\hat\lambda$ désignant un bon estimateur de λ. Les simulations reportées par Sheather [62] donnent des résultats excellents. On remarquera la similitude entre cette approche et l'approche développée par Park et Marron [49] dans le cas global. Le lecteur désireux d'en savoir plus sur les méthodes de type plug-in dans le cas local consultera avec profit les articles de Hall [30] et Gijbels et Mammen [25].

4 Conclusion et perspectives

Parmi les multiples études comparatives disponibles dans la littérature, nombreuses sont celles qui mettent en évidence la compétitivité - sinon la supériorité - des méthodes de type plug-in par rapport à leurs principales concurrentes. De par leurs bonnes performances, leur rapidité et leur facilité de mise en oeuvre, les algorithmes plug-in ont progressivement accru leur popularité au sein de la communauté des utilisateurs.

La sélection du paramètre de lissage par plug-in itéré, à laquelle cette revue bibliographique a été consacrée, ne doit pas être vue comme une nouvelle méthode de sélection venant grossir le flot déjà important des méthodes existantes. Il nous semble au contraire plus juste d'envisager le plug-in itéré comme une amélioration, pour ne pas dire un aboutissement naturel, des algorithmes de type plug-in. Le plug-in itéré possède les atouts des méthodes plug-in et y ajoute de bonnes propriétés de convergence.

Cependant, trois points importants doivent être impérativement soulignés. D'abord, et nous avons eu l'occasion de le rappeler à plusieurs reprises, la sélection de la largeur de fenêtre par plug-in (ou plug-in itéré) n'est valable qu'asymptotiquement. A taille d'échantillon fixée, l'analyse s'avère délicate. Ensuite, les formules qui sont au coeur des méthodes plug-in imposent des restrictions sur la densité inconnue f souvent difficiles à vérifier dans la pratique. Par exemple, l'expression classique (4) de la largeur de fenêtre asymptotique optimale h^* n'est plus valide pour une densité cible uniforme ou exponentielle. Enfin, il est essentiel de comprendre qu'il n'existe pas de méthode de sélection du paramètre de lissage qui soit intrinsèquement "meilleure" que toutes les autres. L'expérience montre que chaque algorithme de sélection possède, en quelque sorte, ses densités de prédilection : un premier algorithme fonctionnera par exemple correctement pour des densités unimodales, alors qu'un second fournira de bien meilleurs résultats pour des densités à queues lourdes, etc. Le lecteur intéressé par ces considérations pourra consulter les illustrations finales des études comparatives de Berlinet et Devroye [3] et Devroye [18], à l'issue desquelles on peut lire : "Our experiments show why density estimation is fascinating - every method seems to "like" certain types of densities". Ces deux études montrent que les méthodes plug-in donnent de bons résultats pour des densités unimodales suffisamment lisses. Eu égard au plug-in itéré, il serait donc intéressant - dans un travail futur - d'effectuer des simulations extensives et d'observer la robustesse de la méthode par rapport aux propriétés de la densité cible. En tout état de cause, nous ne saurions trop recommander aux éventuels utilisateurs de toujours tester plusieurs algorithmes de sélection avant d'opérer un choix définitif.

Comme la performance d'une méthode de sélection conduisant à une largeur de fenêtre $H = H(X_1, \ldots, X_n)$ dépend de la densité inconnue f, il est important de pouvoir donner une garantie relative à l'erreur commise par l'estimateur à noyau f_H lorsqu'on se limite à une classe de densités possibles. Dans cet esprit, on cherche souvent à mettre en évidence une classe de densités (notée F) suffisamment grande (ou *riche*) et une constante C telles que

$$\sup_{f\in F}\limsup_{n\to\infty}\frac{\mathbf{E}\int|f_H-f|}{\inf_{h>0}\mathbf{E}\int|f_h-f|}\leq C. \tag{17}$$

Pour de plus amples informations sur cette problématique, dite *de type minimax*, le lecteur consultera avec profit Devroye [15] ainsi que les articles plus récents de Berlinet et Devroye [3] et Devroye [16], [17]. Hall et Wand [38] ont montré la validité de la formule (17) lorsque le paramètre de lissage est sélectionné par une méthode de type plug-in, à condition de choisir le supremum dans une classe de densités univariées suffisamment lisses et à queues légères. Dans le cadre de l'estimation L_2, un des résultats les plus remarquables a été récemment obtenu par Wegkamp [69], qui propose un algorithme de sélection inspiré des travaux de Stone [66], Devroye et Lugosi [20], [21] et Birgé et Massart [4] fournissant une fenêtre H telle que

$$\sup_{||f||_\infty\leq C}\limsup_{n\to\infty}\frac{\mathbf{E}\int|f_H-f|^2}{\inf_{h>0}\mathbf{E}\int|f_h-f|^2}=1. \tag{18}$$

A notre connaissance, la recherche de propriétés minimax telles que (17) ou (18) pour l'estimateur à noyau f_{h_∞} présenté dans cet article n'a pas encore été amorcée, et devrait donner lieu à de nombreux problèmes intéressants. D'une manière générale, on peut noter que la plupart des méthodes de sélection automatique du paramètre de lissage restreignent leur champ d'application à une classe plus ou moins grande de densités. Dès lors, toute comparaison objective et équitable entre les différents algorithmes de sélection devient difficile, voire impossible. De nombreux auteurs insistent aujourd'hui sur la nécessité de s'orienter vers des méthodes de sélection "plus universelles", c'est-à-dire vers des méthodes de sélection moins contraignantes quant aux hypothèses faites sur la densité cible f. En particulier, la mise au point d'une méthode de sélection fournissant une largeur de fenêtre $H=H(X_1,\ldots,X_n)$ telle que

$$\limsup_{n\to\infty}\frac{\mathbf{E}\int|f_H-f|}{\inf_{h>0}\mathbf{E}\int|f_h-f|}\leq 1$$

pour toute densité f est encore un problème ouvert.

Remerciements. Qu'il me soit permis de remercier deux rapporteurs anonymes : leurs commentaires, leurs critiques et leurs suggestions constructives m'ont permis d'améliorer substantiellement la qualité de ce travail.

Références

[1] Akaike, H. (1954) An approximation to the density function, *Annals of the Institute of Statistical Mathematics*, **Vol. 6**, pp. 127–132.

[2] Berlinet, A. et Devroye, L. (1989) Estimation d'une densité : un point sur la méthode du noyau, *Statistique et Analyse des Données*, **Vol. 14**, pp. 1–32.

[3] Berlinet, A. et Devroye, L. (1994) A comparison of kernel density estimates, *Publications de l'Institut de Statistique de l'Université de Paris*, **Vol. 38**, pp. 3–59.

[4] Birgé, L. et Massart, P. (1997) From model selection to adaptative estimation, in *Festschrift for LeCam*, pp. 55–87, Springer–Verlag, New York.

[5] Bosq, D. et Lecoutre, J.P. (1987) *Théorie de l'Estimation Fonctionnelle*, Economica, Paris.

[6] Bowman, A.W. (1984) An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **Vol. 71**, pp. 353–360.

[7] Burman, P. (1985) A data dependent approach to density estimation, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **Vol. 69**, pp. 609–628.

[8] Cao, R., Cuevas, A. et González-Manteiga, W. (1994) A comparative study of several smoothing methods in density estimation, *Computational Statistics and Data Analysis*, **Vol. 17**, pp. 153–176.

[9] Chiu, S.T. (1991) Bandwith selection for kernel density estimation, *The Annals of Statistics*, **Vol. 19**, pp. 1883–1905.

[10] Chiu, S.T. (1992) An automatic bandwith selector for kernel density estimation, *Biometrika*, **Vol. 79**, pp. 771–782.

[11] Chow, Y.S., Geman, S. et Wu, L.D. (1983) Consistent cross-validated density estimation, *The Annals of Statistics*, **Vol. 11**, pp. 25–38.

[12] Deheuvels, P. (1974) 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*, **Vol. 278**, pp. 1217–1220.

[13] Deheuvels, P. (1977) Estimation non paramétrique de la densité par histogrammes généralisés, *Revue de Statistique Appliquée*, **Vol. 25**, pp. 5–42.

[14] Deheuvels, P. et Hominal, P. (1980) Estimation automatique de la densité, *Revue de Statistique Appliquée*, **Vol. 28**, pp. 25–55.

[15] Devroye, L. (1987) *A Course in Density Estimation*, Birkhäuser, Boston.

[16] Devroye, L. (1989) A universal lower bound for the kernel estimate, *Statistics and Probability Letters*, **Vol. 8**, pp. 419–423.

[17] Devroye, L. (1994) On good deterministic smoothing sequences for kernel density estimates, *The Annals of Statistics*, **Vol. 22**, pp. 886–889.

[18] Devroye, L. (1997) Universal smoothing factor selection in density estimation: theory and practice, *Test*, **Vol. 6**, pp. 223–320.

[19] Devroye, L. et Györfi, L. (1985) *Nonparametric Density Estimation: the L_1 View*, Wiley, New York.

[20] Devroye, L. et Lugosi, G. (1996) A universally acceptable smoothing factor for kernel density estimates, *The Annals of Statistics*, **Vol. 24**, pp. 2499–2512.

[21] Devroye, L. et Lugosi, G. (1997) Nonasymptotic universal smoothing factors, kernel complexity, and Yatracos classes, *The Annals of Statistics*, **Vol. 25**, pp. 2626–2637.

[22] Duin, R.P.W. (1976) On the choice of smoothing parameters for Parzen estimators of probability density functions, *IEEE Transactions on Computers*, **Vol. 25**, pp. 1175–1179.

[23] Engel, J., Herrmann, E. et Gasser, T. (1994) An iterative bandwith selector for kernel estimation of densities and their derivatives, *Journal of Nonparametric Statistics*, **Vol. 4**, pp. 21–34.

[24] Fan, J. et Marron, J.S. (1992) Best possible constant for bandwith selection, *The Annals of Statistics*, **Vol. 20**, pp. 2057–2070.

[25] Gijbels, I. et Mammen, E. (1998) Local adaptivity of kernel estimates with plug-in local bandwith selectors, *Board of the Foundation of the Scandinavian Journal of Statistics*, **Vol. 25**, pp. 503–520.

[26] Habbema, J.D.F., Hermans, J. et Vandenbroek, K. (1974) A stepwise discriminant analysis program using density estimation, in *Compstat 1974*, ed. G. Bruckmann, pp. 101–110, Physica–Verlag, Wien.

[27] Hall, P. (1982) Cross-validation in density estimation, *Biometrika*, **Vol. 69**, pp. 383–390.

[28] Hall, P. (1983) Large-sample optimality of least squares cross-validation in density estimation, *The Annals of Statistics*, **Vol. 11**, pp. 1156–1174.

[29] Hall, P. (1985) Asymptotic theory of minimum integrated square error for multivariate density estimation, in *Multivariate Analysis VI*, ed. P.R. Krishnaiah, pp. 289–309, North-Holland, Amsterdam.

[30] Hall, P. (1993) On plug-in rules for local smoothing of density estimators, *The Annals of Statistics*, **Vol. 21**, pp. 694–710.

[31] Hall, P. et Marron, J.S. (1987) Estimation of integrated squared density derivatives, *Statistics and Probability Letters*, **Vol. 6**, pp. 109–115.

[32] Hall, P. et Marron, J.S. (1987) Extent to which least squares cross-validation minimises integrated square error in nonparametric density estimation, *Probability Theory and Related Fields*, **Vol. 74**, pp. 567–581.

[33] Hall, P. et Marron, J.S. (1987) On the amount of noise inherent in bandwith selection of a kernel density estimator, *The Annals of Statistics*, **Vol. 15**, pp. 163–181.

[34] Hall, P. et Marron, J.S. (1991) Local minima in cross-validation functions, *Journal of the Royal Statistical Society*, **Vol. B53**, pp. 245–252.

[35] Hall, P. et Marron, J.S. (1991) Lower bounds for bandwith selection in density estimation, *Probability Theory and Related Fields*, **Vol. 90**, pp. 149–173.

[36] Hall, P., Marron, J.S. et Park, B.U. (1992) Smoothed cross-validation, *Probability Theory and Related Fields*, **Vol. 92**, pp. 1–20.

[37] Hall, P., Sheather, S.J., Jones, M.C. et Marron, J.S. (1991) On optimal data-based bandwith selection in kernel density estimation, *Biometrika*, **Vol. 78**, pp. 263–269.

[38] Hall, P. et Wand, M.P. (1988) Minimizing L_1 distance in nonparametric density estimation, *Journal of Multivariate Analysis*, **Vol. 26**, pp. 59–88.

[39] Jones, M.C., Marron, J.S. et Park, B.U. (1991) A simple root n bandwith selector, *The Annals of Statistics*, **Vol. 19**, pp. 1919–1932.

[40] Jones, M.C. et Sheather, S.J. (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statistics and Probability Letters*, **Vol. 11**, pp. 511–514.

[41] Kim, W.C., Park, B.U. et Marron, J.S. (1994) Asymptotically best bandwith selectors in kernel density estimation, *Statistics and Probability Letters*, **Vol. 19**, pp. 119–127.

[42] Marron, J.S. (1985) An asymptotically efficient solution to the bandwith problem of kernel density estimation, *The Annals of Statistics*, **Vol. 13**, pp. 1011–1023.

[43] Marron, J.S. (1987) A comparison of cross-validation techniques in density estimation, *The Annals of Statistics*, **Vol. 15**, pp. 152–162.

[44] Marron, J.S. (1988) Automatic smoothing parameter selection: a survey, *Empirical Economics*, **Vol. 13**, pp. 187–208.

[45] Marron, J.S. (1989) Comments on a data-based bandwith selector, *Computational Statistics and Data Analysis*, **Vol. 8**, pp. 155–170.

[46] Marron, J.S. (1992) Bootstrap bandwith selection, in *Exploring the Limits of Bootstrap*, ed. R. le Page et L. Billard, pp. 249–262, Wiley, New York.

[47] Nadaraya, E.A. (1974) On the integral mean square error of some nonparametric estimates for the density function, *Theory of Probability and its Applications*, **Vol. 19**, pp. 133–141.

[48] Park, B.U. (1989) On the plug-in bandwith selectors in kernel density estimation, *Journal of the Korean Statistical Society*, **Vol. 18**, pp. 107–117.

[49] Park, B.U. et Marron, J.S. (1990) Comparison of data-driven bandwith selectors, *Journal of the American Statistical Association*, **Vol. 85**, pp. 66–72.

[50] Park, B.U. et Marron, J.S. (1992) On the use of pilot estimators in bandwith selection, *Journal of Nonparametric Statistics*, **Vol. 1**, pp. 231–240.

[51] Parzen, E. (1962) On the estimation of a probability density function and the mode, *The Annals of Mathematical Statistics*, **Vol. 33**, pp. 1065–1076.

[52] Rosenblatt, M. (1956) Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, **Vol. 27**, pp. 832–837.

[53] Rosenblatt, M. (1971) Curve estimates, *The Annals of Mathematical Statistics*, **Vol. 42**, pp. 1815–1842.

[54] Rudemo, M. (1982) Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics*, **Vol. 9**, pp. 65–78.

[55] Schuster, E.F. et Gregory, G.G. (1981) On the nonconsistency of maximum likelihood nonparametric density estimators, in *Computer Science and Statistics*: *Proceedings of the* 13^{th} *Symposium on the Interface*, ed. W.F. Eddy, pp. 295–298, Springer–Verlag, New York.

[56] Scott, D.W. (1985) Averaged shifted histograms: effective nonparametric density estimators in several dimensions, *The Annals of Statistics*, **Vol. 13**, pp. 1024–1040.

[57] Scott, D.W. (1992) *Multivariate Density Estimation*: *Theory, Practice, and Visualisation*, Wiley, New York.

[58] Scott, D.W. et Factor, L.E. (1981) Monte Carlo study of three data-based nonparametric probability density estimators, *Journal of the American Statistical Association*, **Vol. 76**, pp. 9–15.

[59] Scott, D.W., Tapia, R.A. et Thompson, J.R. (1977) Kernel density estimation revisited, *Nonlinear Analysis, Theory, Methods and Applications*, **Vol. 1**, pp. 339–372.

[60] Scott, D.W. et Terrell, G.R. (1987) Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association*, **Vol. 82**, pp. 1131–1146.

[61] Sheather, S.J. (1983) A data-based algorithm for choosing the window width when estimating the density at a point, *Computational Statistics and Data Analysis*, **Vol. 1**, pp. 229–238.

[62] Sheather, S.J. (1986) An improved data-based algorithm for choosing the window width when estimating the density at a point, *Computational Statistics and Data Analysis*, **Vol. 4**, pp. 61–65.

[63] Sheather, S.J. et Jones, M.C. (1991) A reliable data-based bandwith selection method for kernel density estimation, *Journal of the Royal Statistical Society*, **Vol. B53**, pp. 683–690.

[64] Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

[65] Simonoff, J.S. (1996) *Smoothing Methods in Statistics*, Springer–Verlag, New York.

[66] Stone, C.J. (1984) An asymptotically optimal window selection rule for kernel density estimates, *The Annals of Statistics*, **Vol. 12**, pp. 1285–1297.

[67] Terrell, G.R. et Scott, D.W. (1985) Oversmoothed nonparametric density estimates, *Journal of the American Statistical Association*, **Vol. 80**, pp. 209–214.

[68] Turlach, B.A. (1993) Bandwith selection in kernel density estimation: a review, *Rapport technique*, Université Catholique de Louvain.

[69] Wegkamp, M.H. (1999) Quasi-universal bandwith selection for kernel density estimators, *The Canadian Journal of Statistics*, **Vol. 27**, pp. 409–420.

[70] Woodroofe, M. (1970) On choosing a delta sequence, *The Annals of Mathematical Statistics*, **Vol. 41**, pp. 1665–1671.