Statistique et analyse des données

Alain Berlinet
Luc Devroye

Estimation d'une densité : un point sur la méthode du noyau

Statistique et analyse des données, tome 14, n° 1 (1989), p. 1-32.
<http://www.numdam.org/item?id=SAD_1989__14_1_1_0>

Statistique et Analyse des Données
1989 - Vol. 14 n° 1 - p. 1 - 32

ESTIMATION D'UNE DENSITE :

UN POINT SUR LA METHODE DU NOYAU

Alain BERLINET

Laboratoire de Probabilités et Statistique
U.F.R. de Mathématiques
U.S.T.L.
F-34 060 MONTPELLIER Cedex

Luc DEVROYE

School of Computer Science
Department of Mathematics and Statistics
McGill University
MONTREAL, CANADA H3A 2K6

Résumé

Nous présentons les principales qualités d'un estimateur à noyau basé sur un échantillon $X_1, \ldots, X_n$ de variables indépendantes et de même loi à densité f inconnue. Deux thèmes seront privilégiés dans cette discussion :

(i) connexion entre taille d'échantillon et qualité d'estimation

(ii) propriétés à distance finie des estimateurs.

Le second sera illustré au moyen d'inégalités de déviation. Les principaux ouvrages de référence sont ceux de Bosq et Lecoutre (1987), Tapia et Thompson (1978), Prakasa Rao (1983) ainsi que le survol de Silverman (1986). Les points de vue développés ici ont été présentés de façon plus ou moins nette dans les travaux de Devroye et Györfy (1985) et Devroye (1987a).

Classification AMS : 62 G 05

Classification STMA : 04 180, 04 080

Mots-clés : densité de probabilité, estimateur à noyau, vitesses de convergence, taille d'échantillon, propriétés à distance finie.

Manuscrit reçu le 23. 12. 88, révisé le 24. 3. 89

Plan

Abstract

Our subject is the estimation of an unknown density f from a sample $X_1, \ldots, X_n$ *of iid random variables having density f. Much has been written and said about this topic, so we will limit ourselves to a general discussion focused on the kernel estimate with two main themes :*

(i) the connection between the sample size n and the quality of an estimator

(ii) the non-asymptotic properties of density estimates.

The study of (ii) will be illustrated principally with the aid of inequalities. The subject has been dealt with in a number of books, such as the French text by Bosq and Lecoutre (1987), the overview by Silverman (1986) and the works of Tapia and Thompson (1978) and Prakasa Rao (1983). The viewpoints presented in this note are by-and-large those that are rather clumsily put forward in Devroye and Györfi (1985) and Devroye (1987a).

Keywords and phrases : density estimation, kernel estimate, rates of convergence, sample size, non-asymptotic properties.

Introduction

Dans cet article notre propos sera de présenter les qualités principales du meilleur "généraliste" en matière d'estimation de densité : l'estimateur à noyau. Dans un but didactique, à l'intention des non-spécialistes de l'estimation fonctionnelle, nous avons élagué les considérations mathématiques pour privilégier les idées-forces. Les lecteurs désireux d'entrer dans les techniques de démonstration voudront bien se reporter à la bibliographie. Ce choix nous permettra de présenter au plus grand nombre des résultats extrêmement récents ainsi que des voies de recherche nouvelles, sans désir d'une quelconque exhaustivité.

Le problème que nous traitons ici est l'estimation d'une densité de probabilité inconnue f à partir d'un échantillon $X_1, \ldots, X_n$ de variables aléatoires indépendantes et de même loi à densité f (ce que nous appellerons dans la suite un f- échantillon). Nous aborderons un certain nombre de concepts généraux en estimation fonctionnelle : choix des mesures d'erreur, stabilité relative, problèmes liés à la taille n de l'échantillon, choix du facteur de lissage et de la fonction noyau.

Un bref rappel historique.

C'est à la fin du siècle dernier que Karl Pearson introduisit la célèbre famille de lois de probabilité qui porte son nom et dont chaque élément satisfait à une équation différentielle du type :

$$(1) \qquad f'(x) = \frac{(x-a)\ f(x)}{b_0 + b_1 x + b_2 x^2} \qquad (a, b_0, b_1, b_2) \in \mathbb{R}^4 .$$

La raison pour laquelle cette famille est encore si populaire de nos jours tient probablement dans deux faits : le premier est qu'elle inclut bon nombre de familles paramétriques usuelles (bêta, gamma, exponentielle, normale, de Student, de même que des densités moins connues comme les Pearson IV); le second est qu'elle permet quasiment tous les mariages possibles entre coefficient d'asymétrie (skewness) et d'aplatissement (kurtosis), et donc de refléter fidèlement l'allure d'une distribution empirique lorsque l'on se limite à ces deux paramètres.

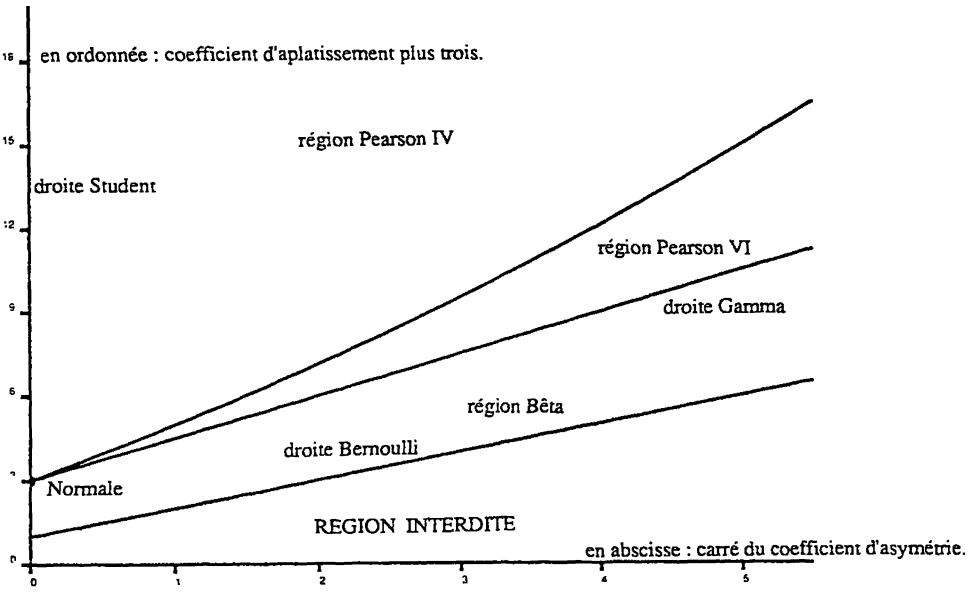

Figure 1. Diagramme de Pearson.

Cette propriété de description est illustrée par la figure 1 : un couple quelconque de réels positifs n'est pas nécessairement la représentation dans le plan d'une loi de probabilité mais, en dehors de la région interdite et de sa frontière (droite Bernoulli) tout point du plan est la représentation d'une loi de Pearson.

Malheureusement le modèle de Pearson présente de graves défauts : outre que des modèles simples très courants en soient exclus (Laplace, log-normale, Weibull, Gumbel, logistique) l'équation (1) nous montre les limitations sévères du modèle en ce qui concerne l'existence de modes. Enfin une méthode basée sur les seuls quatre premiers moments (en fonction desquels s'expriment les constantes a, b_0, b_1 ,b_2) ne peut que se révéler inefficace et non robuste dès que l'on s'écarte quelque peu du modèle. Un exemple d'une telle méthode est celle des moments où l'on substitue les moments empiriques aux moments théoriques dans les expressions des constantes.

A moins d'avoir sur le phénomène aléatoire étudié des informations a priori très précises et indiscutables, le champ d'application d'un modèle paramétrique ne devient

satisfaisant que lorsque l'inflation du nombre de paramètres est telle que les méthodes d'estimation du modèle deviennent tout à fait inefficaces. Pour pallier aux insuffisances et aux défauts des familles paramétriques une démarche élémentaire est de faire appel à l'omniprésent histogramme : l'intervalle d'étude, disons [0 , 1], est partagé en k sous-intervalles de même longueur notés I_1 , ... , I_k et, sur le $j^{ème}$, f(x) est estimé par $k\ N_j/n$ où N_j est le nombre de points tombés dans I_j.

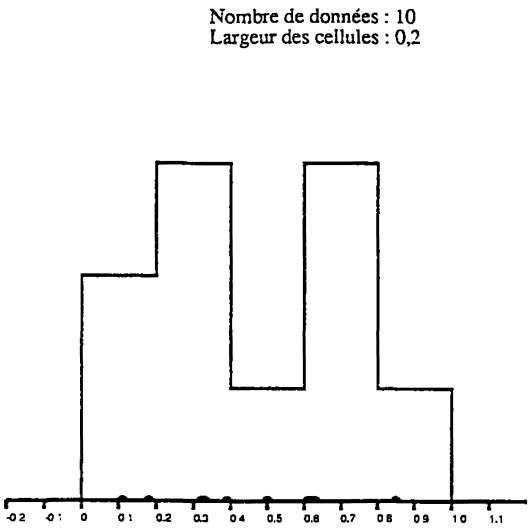

Figure 2. Un histogramme.

L' histogramme convient bien pour des analyses relativement grossières. Néanmoins ses discontinuités n'apparaissent pas très naturelles et, ce qui est plus grave, les points près des bords et les points près du milieu d'un intervalle reçoivent le même poids; ceci explique la variabilité des interprétations statistiques que l'on peut faire d'un histogramme suivant l'origine choisie. En fait, pour des densités raisonnablement lisses, l'histogramme est un estimateur sévèrement limité. Pour remédier à ce problème, Rosenblatt proposa en 1956 de centrer chaque cellule de l'histogramme au point où l'on estime, introduisant ainsi la première forme rudimentaire d'estimateur à noyau.

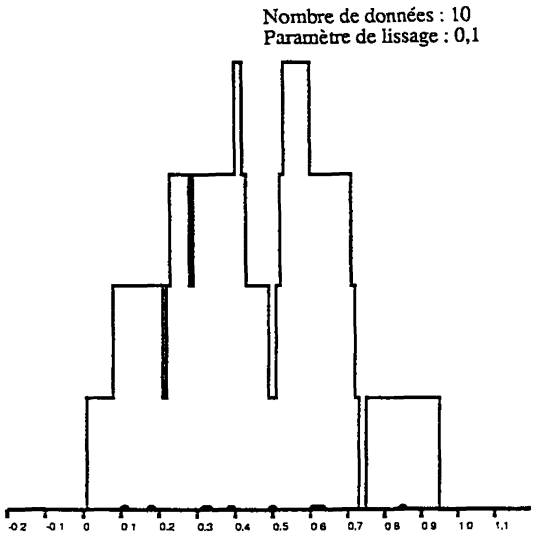

Figure 3. Estimateur à noyau uniforme.

Nous savons maintenant que l'estimateur de Rosenblatt a une meilleure vitesse de convergence que l'histogramme pour la plupart des densités régulières. Il peut s'écrire sous la forme

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

où K, appelé le noyau, est la densité uniforme sur [- 1/2 ; + 1/2] et où $h > 0$ est la largeur de fenêtre ou constante de lissage. Dès que K est une densité, f_n en est également une. On peut obtenir des estimateurs plus esthétiques en utilisant des noyaux K lisses (Rosenblatt, 1956; Parzen, 1962).

Lorsque l'on se limite aux noyaux K positifs les vitesses de convergence varient peu en fonction de K et les critères essentiels de choix du noyau sont la simplicité et la vitesse de calcul d'une part, la régularité de la courbe à obtenir d'autre part. Nous verrons qu'il en est tout autrement lorsque l'on s'autorise à utiliser des noyaux quelconques.

Il est facile de vérifier que pour les noyaux usuels et un ensemble de données fixé la loi de densité f_n converge (étroitement) vers la mesure empirique lorsque h tend vers 0 et

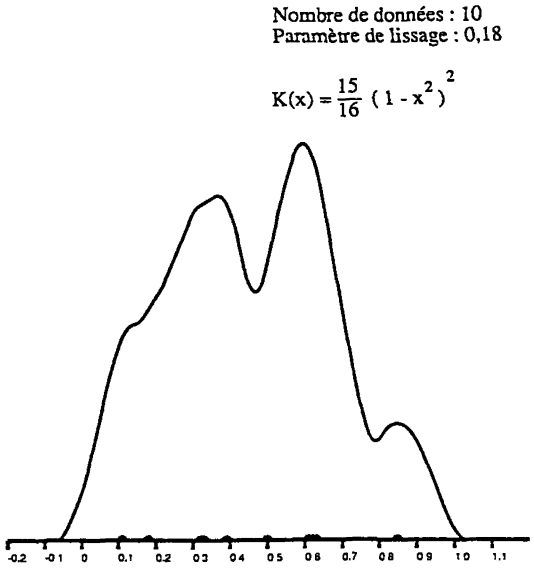

Figure 4. Estimateur à noyau

que f_n tend uniformément vers la fonction nulle lorsque h tend vers l'infini. Le coeur du problème est le choix du facteur de lissage qui, en variant, fait décrire à f_n un ensemble de lois dont les extrêmes sont "proches" de lois discrètes d'un côté, uniformes de l'autres.

A la recherche d'un critère.

Pour comparer des estimateurs de la densité, ou pour tenter de décider d'un choix optimal pour le paramètre de lissage, un critère est nécessaire, qui en fait est déterminé par l'application en vue. Illustrons ceci par deux exemples, l'un de discrimination, l'autre de simulation.

On peut formuler certains problèmes de discrimination de la manière suivante : on dispose de réalisations (X_1 , Y_1), ... , (X_n , Y_n) de couples aléatoires indépendants de même loi que (X , Y) où X est le vecteur des observations et Y élément de { 0 , 1 } est la classe. Pour fixer les idées considérons un exemple où X serait un vecteur (C_1, C_2) dont les composantes aléatoires seraient respectivement les concentrations de deux produits P_1 et P_2 dans le sang d'un individu et Y serait l'indicateur d'une maladie (Y = 1 si l'individu a la maladie, Y = 0 s'il ne l'a pas).

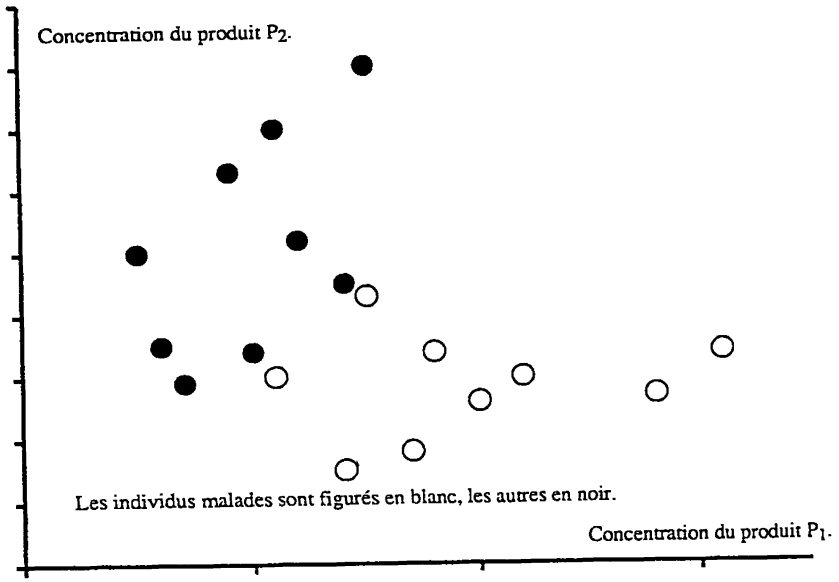

Figure 5. Un exemple de discrimination à deux classes.

La figure 5 donne une représentation d'un groupe de dix-huit individus pour lesquels on a mesuré C_1 et C_2 et dont on s'est assuré par des investigations plus poussées (et plus coûteuses) qu'ils présentaient ou non la maladie.

Sur présentation d'un nouveau couple (X' , Y') où Y' est inconnu et X' est observé, on doit décider si Y' = 0 ou Y' = 1. Ici le critère est évidemment la probabilité d'erreur, probabilité que la valeur attribuée à Y' soit fausse. En fait c'est la probabilité conditionnelle d'erreur étant donné les valeurs initiales. Lorsque, conditionnellement à Y = 0 ou Y = 1, X a les densités respectives f et g alors la probabilité d'erreur est minimisée en décidant que Y' = 1 quand $g(X') > f(X')$ pourvu que $P(Y = 1) = P(Y = 0) = 1/2$, cas le plus simple. Puisque f et g sont inconnues on peut estimer ces densités par f_n et g_n respectivement et conjecturer que Y' = 1 lorsque $g_n(X') > f_n(X')$. Avec cette règle la probabilité d'erreur est une fonction plutôt compliquée des densités et de leurs estimateurs. La décision étant binaire il n'est pas nécessaire que les estimateurs soient très proches des vraies densités pour obtenir une probabilité d'erreur quasiment optimale. Aussi sommes-nous dans une situation où il semble avantageux de travailler directement en termes de probabilité d'erreur plutôt que de s'engager sur la voie consistant à construire d'abord de très bons estimateurs de la densité

En simulation le problème suivant se pose souvent : alors que, pour des raisons de coût ou de faisabilité, on ne dispose que d'un seul f-échantillon $X_1, \ldots, X_n$, on veut mettre en œuvre des procédures recréant le hasard basées sur de nouveaux f-échantillons de taille k (éventuellement supérieur à n). La densité f étant inconnue, ce problème est insoluble.

Il est cependant possible de générer un f_n-échantillon, où f_n est un estimateur de f basé sur les données originelles. Différentes méthodes peuvent être utilisées pour générer un échantillon d'une loi quelconque (Berlinet, 1988), la plus répandue étant la méthode de rejet (Devroye, 1986) : si f_n est majorée par M sur un intervalle [a , b] contenant son support, on génère des variables uniformément distribuées dans le rectangle $[a , b] \times [0 , M]$. Les abscisses des points tombant sous le graphe de f_n constituent un f_n-échantillon.

Les projections des points forment un échantillon de la densité en trait plein.

Les points en blanc doivent être déplacés pour créer un échantillon de la densité en pointillés.

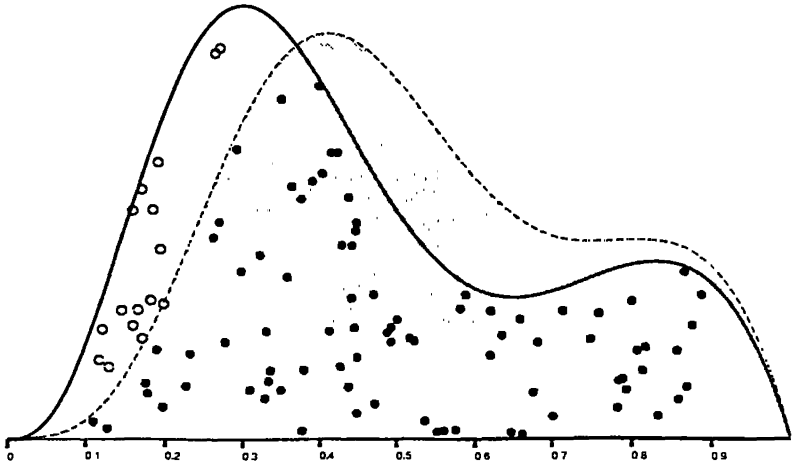

Figure 6. Chirurgie sur un échantillon.

L'erreur commise en simulant à partir de f_n au lieu de f peut être mesurée en termes du nombre de données nouvellement générées qui doivent être déplacées pour transformer le f_n-échantillon de taille k en un f-échantillon de même taille. Ce nombre aléatoire est distribué selon une loi binomiale de taille k et de paramètre

$$\int (f_n - f)_+ = \frac{1}{2}\int |f_n - f|$$

ce qui met en évidence l'importance du critère L_1. Lorsque la distance L_1 entre f_n et f sera par exemple 0,02 il nous faudra remplacer en moyenne 1% des points du f_n-échantillon pour obtenir un f-échantillon.

Le rôle joué en théorie des probabilités par les densités et leurs estimateurs ne doit pas être oublié. Pour la mesure empirique ordinaire μ_n (qui met la masse 1/n à chaque donnée), nous avons une propriété désagréable :

$$\sup |\mu_n(A) - \mu(A)| = 1$$

où μ est la mesure induite par une densité f quelconque et le supremum est pris sur l'ensemble des boréliens A. Nous pouvons définir une mesure empirique lissée basée sur un estimateur f_n de la densité par

$$\mu_n(A) = \int_A f_n \ .$$

On a alors par le théorème de Scheffé (1947) :

$$\sup |\mu_n(A) - \mu(A)| = \sup \left| \int_A f_n - \int_A f \right| = \frac{1}{2}\int |f_n - f| \ .$$

Cette propriété est due au fait que la mesure de densité $(f_n - f)$ est de masse totale nulle et est partagée par toutes les mesures bornées à signe ayant cette propriété. On en déduit donc que le sup tend vers zéro selon un mode stochastique quand la distance L_1 entre f_n et f tend vers zéro selon le même mode. Outre que sa calculabilité ne nécessite aucune hypothèse supplémentaire sur la densité, le critère L_1 est universel : il a une signification physique claire en termes de différences de probabilités et il est possible de l'utiliser pour comparer des estimateurs dans des situations différentes. Le théorème de Scheffé entraîne que la distance L_1 reste invariante sous les transformations continues strictement monotones de l'axe des abscisses. Ceci peut être exploité lorsque l'on visualise sur écran l'erreur L_1 qui est commise dans la queue de la distribution pour une densité ou un estimateur à support infini.

Comme le montre la figure 7, il suffit de transformer la partie intéressante de l'axe réel d'une façon continue monotone en un intervalle borné. Alors que les formes des densités changent sous des transformations non linéaires, les distances L_1 entre elles restent invariantes. Les contributions dues aux queues peuvent par conséquent être rendues visibles.

A gauche : densités sur [0 , 1] après transformation : $z = 1 - 2/y$.

A droite : densités sur [2 , ∞] avant transformation.

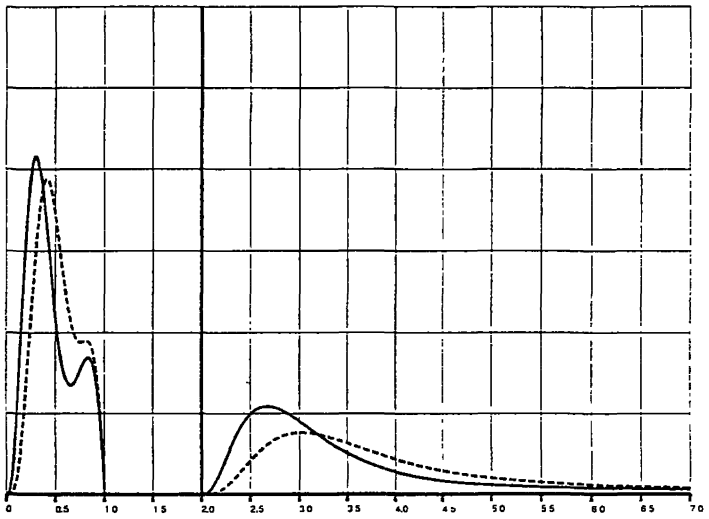

Figure 7. Transformations de densités.

Stabilité relative.

Malheureusement, le critère $\int |f_n - f|$ est une variable aléatoire.

La question se pose d'évaluer sa médiane, ses quantiles, sa moyenne ou tout autre quantité non aléatoire. Il apparaît que sa moyenne conviendra bien puisque la plupart des estimateurs non-paramétriques comme l'estimateur à noyau et l'histogramme sont relativement stables ie $\dfrac{\int |f_n - f|}{E\int |f_n - f|}$ converge vers un en probabilité.

Dans toute la suite l'espérance de la norme L_1 de (f_n - f) sera simplement appelée "erreur". La stabilité relative s'exprime au moyen de l'inégalité simple :

$$\mathrm{Var}\left(\int |f_n - f|\right) \le \frac{4}{n}\left(\int |K|\right)^2$$

vraie pour tout h > 0, toute densité f, tout noyau K et tout n. Pour l'histogramme la variance n'excède pas 4/n (Devroye, 1987b). Dans l'étude des propriétés de l'estimateur à noyau, le caractère non-asymptotique de cette borne est très utile. Néanmoins différentes techniques (Devroye, 1988b) permettent d'obtenir une inégalité exponentielle ayant une uniformité similaire (pour tout n et tout $\varepsilon > 0$) :

$$\sup_{f,\, h>0} P\left\{\left|\int |f_n - f| - E\int |f_n - f|\right| \ge \varepsilon\right\} \le 2\exp\left\{-\frac{n\varepsilon^2}{32\left(\int |K|\right)^2}\right\} .$$

La comparaison des ordres de grandeur dans les situations les plus souvent rencontrées de

$\int |f_n - f|$ ($>> \frac{1}{\sqrt{n}}$) et de $\left|\int |f_n - f| - E\int |f_n - f|\right|$ ($\sim \frac{c}{\sqrt{n}}$) permet de "situer" f_n par rapport à f dans L_1. Elle renforce l'intérêt de l'espérance de la distance L_1 comme critère.

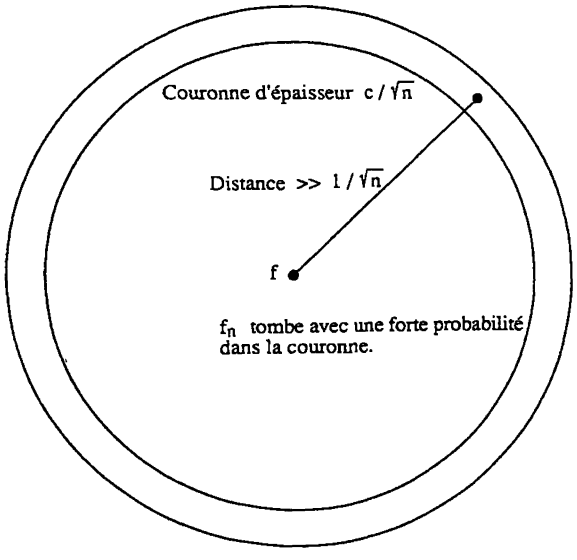

Figure 8.

L'estimateur à noyau.

Parmi les estimateurs non-paramétriques, l'estimateur à noyau est de loin le plus populaire. Ceci est partiellement dû à sa simplicité ($f_n(x)$ est la somme de n variables aléatoires iid) et à sa flexibilité (possibilité de choisir K et h) mais n'aurait pas résisté à l'épreuve du temps s'il ne s'était pas trouvé que f_n converge vers f au sens L_1 pour toute densité f dès que 1/h et nh tendent tous deux vers l'infini. D'autre part, si l'estimateur est convergent, il est convergent dans tous les sens i.e. en probabilité, en moyenne, presque sûrement et presque complètement. En fait

$$P\left\{\int |f_n - f| > \varepsilon\right\} \leq \exp\left(-\frac{1}{3} n \varepsilon^2\right)$$

pour tout $n \geq n_0$, où n_0 dépend de $(h_n)_n$, f, K et ε. Noter le caractère asymptotique de cette inégalité. Nous sommes encore loin d'une réponse à la question cruciale du choix de h. Cette réponse dépend fortement de deux facteurs : la lissité (néologisme commode pour smoothness) et la taille des queues de la densité. Mais commençons avec une borne inférieure universelle valide pour tout n :

$$\inf_{K,h,f} E \int |f_n - f| \geq \frac{1}{\sqrt{528\, n}} .$$

La version asymptotique de cette inégalité a une borne inférieure légèrement meilleure :

$$\frac{0,125 + o(1)}{\sqrt{n}} \quad \text{(Devroye, 1988a).}$$

C'est l'erreur que vous ne pouvez pas éviter, même si vous pouvez choisir préalablement f, h, K ! Notons ici que cette borne est très optimiste : elle ne peut être atteinte à une constante fixe près que lorsque la fonction caractéristique de la densité a un support compact. Des bornes inférieures pour les estimateurs à noyau peuvent être commodément obtenues en utilisant l'inégalité :

$$\int |f_n - f| \geq \sup_t |\phi_n(t) - \phi(t)|$$

où ϕ_n et ϕ sont les fonctions caractéristiques de f_n et f. A la différence de L_2 , où l'identité de Parseval est valide, l'inégalité L_1-L_∞ ci-dessus est en général stricte.

Pour atteindre la vitesse optimale $n^{-1/2}$, il est nécessaire de prendre un noyau K d'intégrale 1 et dont la transformée de Fourier soit égale à 1 dans un voisinage ouvert de l'origine (Devroye et Györfi, 1985). De tels noyaux prennent bien sûr des valeurs négatives. Supposons maintenant que nous nous restreignions aux noyaux qui sont des

densités. Alors nous avons le résultat suivant, plutôt désappointant, vrai pour tout noyau K symétrique :

$$\liminf_{n\to\infty} \inf_{h} n^{2/5} E \int |f_n - f| \geq (c + o(1)) \left(\int \sqrt{f}\right)^{4/5} \left(\int |f^{(2)}|\right)^{1/5} \geq 0{,}86 \ .$$

Ici "c" est une constante positive et il est supposé que f est absolument continue ainsi que f'. Cependant, avec la définition généralisée d'une dérivée seconde donnée dans Devroye et Penrod (1984) ou Devroye et Györfi (1985), les inégalités sont valides pour toutes les densités. L'expression centrale dans la chaîne d'inégalités ci-dessus dépend de f de deux façons, par l'intégrale de la racine de f qui est une mesure de la taille des queues de la loi, et par celle de la valeur absolue de $f^{(2)}$ qui indique le manque de lissité de f. Les densités régulières i.e. pour lesquelles les deux facteurs sont finis peuvent être représentées dans le plan de la figure 9. Les autres ont une queue suffisamment importante pour que leur racine ait une intégrale infinie (comme la densité de Cauchy) ou ont une dérivée seconde généralisée dont la valeur absolue a une intégrale infinie. Par exemple, les densités ayant une simple discontinuité ne se trouvent pas dans le plan de la figure 9.

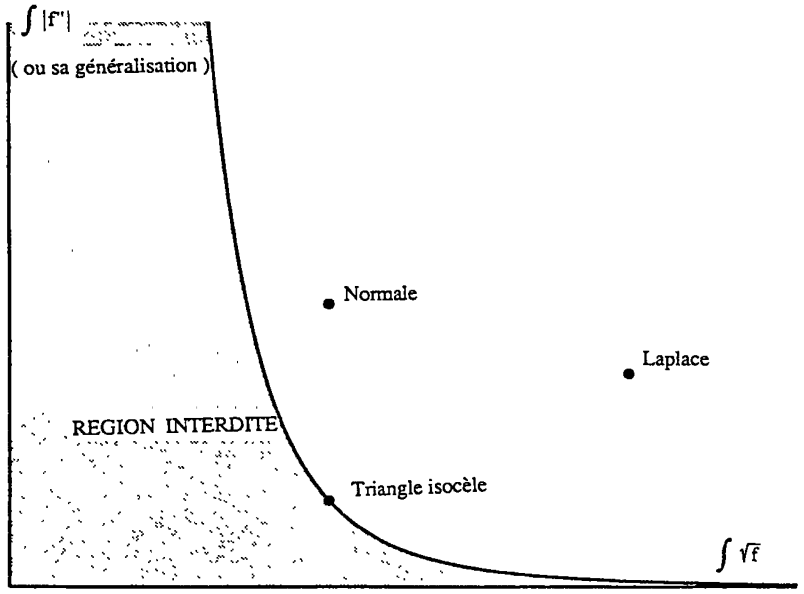

Figure 9. Représentation plane de l'ensemble des densités.

La borne inférieure universelle de 0,86 montre qu'il y a une région interdite dans le plan. Aucune densité ne peut être estimée avec une erreur inférieure à 0,86 $n^{-2/5}$. La raison de l'existence d'une zone interdite se perçoit facilement : la contrainte d'avoir une intégrale de valeur déterminée pour une fonction positive fait que la diminution de la taille des queues impose une augmentation de la courbure et qu' une réduction importante de la courbure oblige à alourdir les queues. Les densités les plus proches de la région interdite sont les plus faciles à estimer. Les seules densités sur la bordure sont les densités triangulaires isocèles. Cependant la densité normale en est très près.

Taille d'échantillon.

Dans la table ci-après, des bornes inférieures sont données pour la taille d'échantillon afin que l'erreur soit inférieure à 0,01 ou 0,001 pour au moins une densité. Rappelons que ceci signifie que dans des nouveaux échantillons de tailles 200 et 2000 tirés selon f_n par la méthode de rejet, nous aurons à remplacer en moyenne un seul point pour rendre l'échantillon conforme. Dans le cas de l'estimateur à noyau, en situation générale, nous avons besoin d'échantillons de taille 157 et 15625 respectivement. Pour la normale les bornes inférieures sont données ligne 2.

Conditions d'estimation	**e < 0,01**	**e < 0,001**
Estimateur à noyau; f, K, h quelconques	n ≥ 157	n ≥ 15 625
Estimateur à noyau; K, h quelconques, f normale	n ≥ 298	n ≥ 40 608
Estimateur à noyau; f, h quelconques, K ≥ 0	n ≥ 68 588	n ≥ 21 689 330
Histogramme, fenêtre quelconque, f "lisse"	n ≥ 681 472	n ≥ 681 472 00

Table 1.

Cependant, si maintenant nous nous restreignons aux noyaux $K \geq 0$, alors nous obtenons les bornes désastreuses de la ligne 3. Les résultats concernant l'histogramme, pour une densité f absolument continue ayant une dérivée f' continue bornée, sont donnés ligne 4. Nous retiendrons de ceci la grande importance du choix du noyau et le fait que l'emploi de l'histogramme doit être réservé à l'analyse exploratoire des données.

Choix du noyau.

Pour tout noyau K dont la valeur absolue peut être majorée par une fonction intégrable unimodale, nous avons la borne inférieure suivante sur la performance (Devroye, 1988a) :

$$\inf_h E \int |f_n - f| \geq (1 + o(1)) \inf_{0 \leq u \leq 1} \max\left(u, \frac{\sqrt{\Phi(2u)} \int \sqrt{f}}{8\sqrt{2\pi n}}\right)$$

où $\Phi(u) = \int_{|\phi(t)| \geq u} dt$, et $\int \sqrt{f}$ est supposé fini.

Cette borne dépend encore de la taille des queues et du manque de régularité de f, mesurés par la vitesse de décroissance vers zéro de la fonction caractéristique ϕ de f. Elle peut être utilisée pour déterminer si un noyau donné K produit des erreurs s'approchant des meilleures possibles.
Par exemple, pour toutes les densités pour lesquelles il existe des constantes positives a, b, c telles que

$$|\phi(t)| \geq a \exp\left(-b|t|^c\right)$$

nous avons, en posant $d = (2c)^{-1}$

$$\liminf_{n \to \infty} \frac{\sqrt{n}}{(\log n)^d} E \int |f_n - f| \geq \frac{1}{8 (2b)^d \sqrt{\pi}} \int \sqrt{f} \ .$$

En particulier, pour la densité normale, l'erreur est au moins $\frac{1 + o(1)}{2^{9/4}\sqrt{n}} \left(\frac{\log n}{\pi}\right)^{1/4}$, ce qui est "proche" de l'optimum. Pour atteindre cette excellente erreur il suffit de

considérer un super noyau (Devroye, 1987a) c'est-à-dire un noyau dont tous les moments sont nuls. Le choix optimal correspondant pour h est de la forme C (log n)$^{-1/2}$. Un super noyau admet une transformée de Fourier qui vaut un dans un voisinage ouvert de l'origine et il prend des valeurs négatives; il peut donc en être de même pour l'estimateur f_n associé. Ceci n'est pas un handicap lorsque le noyau est absolument intégrable : il est alors facile de transformer f_n en une densité tout en réduisant l'erreur.

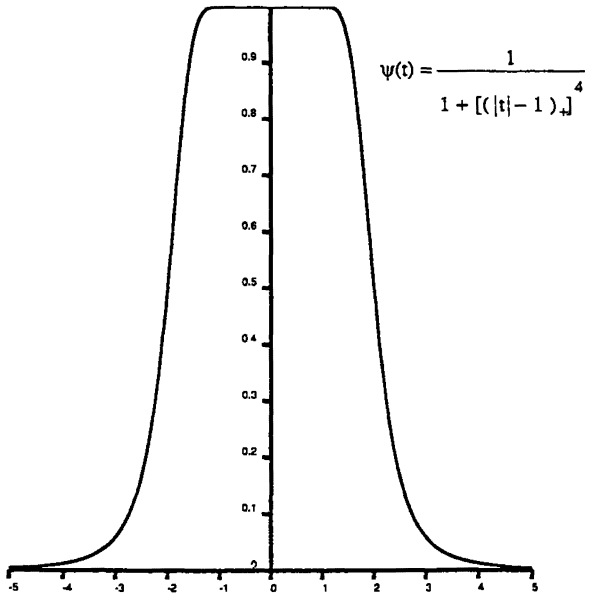

Figure 10. Fonction caractéristique d'un super noyau.

Mentionnons une méthode de construction d'un super noyau. La figure 11 montre la densité de De La Vallée-Poussin. Sa fonction caractéristique est un triangle isocèle. En soustrayant deux fonctions "triangles isocèles" on peut obtenir une fonction "trapèze". Par conséquent une combinaison de deux densités peut conduire facilement à un noyau dont la transformée de Fourier vaut un dans un voisinage ouvert de l'origine, comme requis.

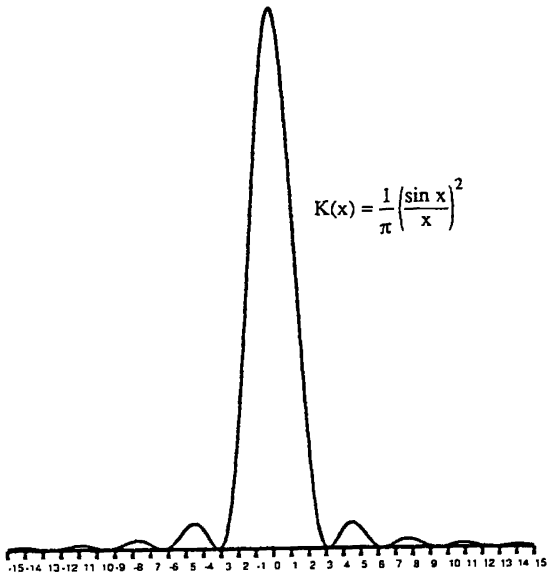

Figure 11. Le noyau de De La Vallée Poussin.

Le problème du choix optimal du noyau dans L_2 a été résolu par Watson et Leadbetter en 1963. Grâce à l'identité de Parseval ils ont obtenu l'égalité suivante :

$$\inf_{K,\,h} E \int (f_n - f)^2 = \frac{1}{2\pi} \int \frac{(1-\phi^2)\phi^2}{1+(n-1)\phi^2}$$

Malheureusement, le noyau K et la fenêtre h permettant d'atteindre le minimum sont des fonctions de la densité inconnue f. En 1975 et 1977, Davis montra qu'il existe un noyau (représenté figure 12), qui s'approche de l'optimum pour les densités de L_2 dont la fonction caractéristique a un module qui décroît vers zéro à l'infini. Ce noyau a une transformée de Fourier rectangulaire mais n'est pas Lebesgue-intégrable.

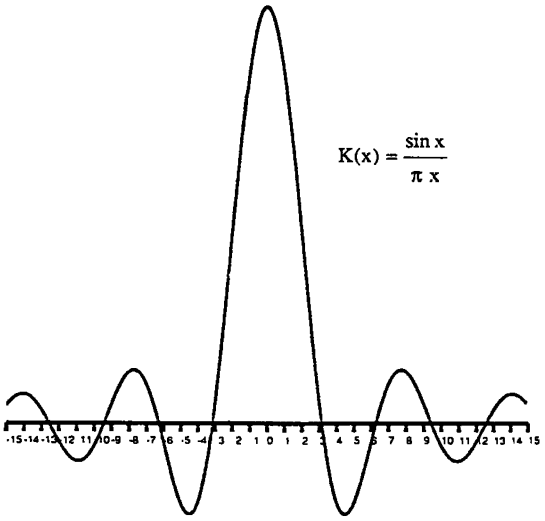

Figure 12. Le noyau de Fourier.

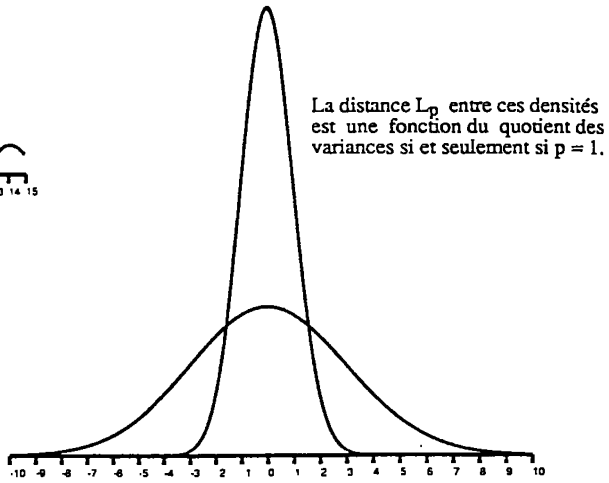

Figure 13. Deux densités de lois normales centrées.

Il peut sembler étrange que la vaste littérature sur les propriétés L_2 des estimateurs de la densité fournisse si peu d'information concernant les erreurs L_1. Ceci est dû au fait que les erreurs L_2 ne sont pas des grandeurs universelles puisqu'elles dépendent de l'échelle choisie. Considérons par exemple les gaussiennes centrées. La distance L_p entre leurs densités est une fonction du rapport des variances si et seulement si $p = 1$. (Signalons que l'information de Kullback de deux telles densités est, elle aussi, une fonction du rapport des variances). De simples changements d'échelle affectent donc l'erreur L_2 . Pour chaque couple (d_1 , d_2) du rectangle $] 0 , 2] \times] 0 , +\infty [$ il existe au moins un couple de densités de L_2 dont la distance L_1 est égale à d_1 et la distance L_2 est égale à d_2. Par voie de conséquence il ne peut exister d'inégalité générale entre les distances L_1 et L_2 de deux densités.

Erreurs minimax.

Comme la performance d'un estimateur dépend de la densité inconnue f, il est important, lorsqu'on se limite à une classe F de densités possibles, de pouvoir donner une garantie relative à l'erreur L_1 qui pourra être commise par un estimateur quelconque f_n sur F. Cette garantie ne peut être que la plus mauvaise performance de f_n pour un élément de F; elle est minorée par l'erreur minimax :

$$\inf_{f_n} \sup_{f \in F} E \int |f_n - f| \ .$$

L'erreur minimax, fonction de n et F, nous indique l'erreur que tout estimateur fait pour au moins une densité de F et présente donc la vue la plus pessimiste sur un estimateur. Pour des considérations plus détaillées sur les bornes minimax on pourra se reporter à Bretagnolle et Huber (1979), Assouad (1983), Birgé (1987a, 1987b) et Devroye (1987a).

Combinaison d'estimateurs.

Beaucoup d'estimateurs sont définis pour des besoins bien précis en dehors desquels ils sont souvent sans utilité : ils sont taillés sur mesure pour des familles particulières de densités que nous appellerons dans la suite "classes cibles" de ces estimateurs. C'est le cas dans les modèles paramétriques où la densité à estimer est décrite par un nombre fini de réels qui sont estimés à partir des données. Les estimateurs non-paramétriques ont un champ d'application très vaste mais bien entendu restent inférieurs aux précédents sur leurs classes cibles.

Une idée naturelle est d'essayer d'imbriquer des modèles de façon à obtenir un estimateur résultant performant sur les classes cibles et conservant à l'extérieur les propriétés et la vitesse de convergence de l'estimateur non-paramétrique. De telles combinaisons sont impératives dans les programmes automatiques d'estimation fonctionnelle. La distance L_1 s'avère être un bon outil de choix de l'estimateur final.

On peut se représenter les classes cibles comme des petites îles dans le vaste océan de l'ensemble de toutes les densités (voir figure 14). A l'intérieur de chaque classe cible, nous pouvons construire un estimateur spécifique; ces estimateurs, appelés $f_{n1}, \ldots, f_{nk}$ sont figurés par des gros points dans les îles de la figure 14. Autour de chacun de ces estimateurs, nous considérons un halo, ou boule d'influence, avec un rayon bien choisi. Un estimateur standard fiable non-paramétrique g_n est aussi construit et est utilisé comme estimateur de la densité à moins qu'il ne tombe dans l'un des halos. Dans ce cas, l'estimateur non-paramétrique est rejeté et remplacé par l'estimateur spécifique de la classe cible. Cette méthode a été proposée par Devroye (1986).

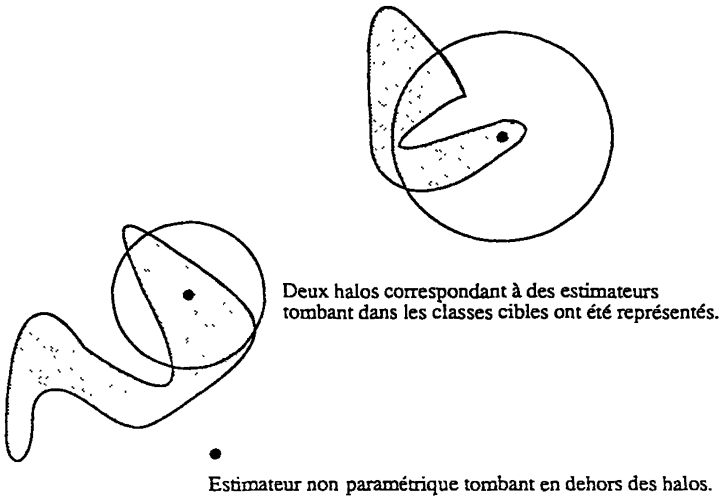

Figure 14. Deux classes cibles dans l'ensemble des densités.

Dans certains exemples simples, comme celui dans lequel l'unique classe cible est celle de toutes les lois normales, il a été démontré que

si f est dans la classe cible 1 $E\int|f_n - f| \sim E\int|f_{n1} - f|$,

si f n'est pas dans la classe cible 1 $E\int|f_n - f| \sim E\int|g_n - f|$.

Il semble donc que nous soyons capables de combiner le meilleur de chacun des points de vue; cependant, beaucoup reste à faire. Pour d'autres approches on pourra se référer à Schuster et Yakowitz (1985) et Olkin et Spiegelman (1987) qui utilisent une méthode de maximum de vraisemblance pour décider entre un choix paramétrique et un choix non-paramétrique. Les équations différentielles associées aux modèles paramétriques (telle que celle de Pearson) pourront également fournir des outils intéressants dans les problèmes de choix de modèles.

Classes cibles.

Les restrictions au niveau des tailles d'échantillons imposées par l'estimateur à noyau même sur les "meilleures densités" sont très lourdes. Il est donc vraisemblable que les principales avancées pratiques en estimation de la densité ne seront pas sur les estimateurs non-paramétriques dans leurs formes générales, mais sur les estimateurs "sur mesure". Les classes cibles sur lesquelles des progrès substantiels pourraient être faits se situent entre les modèles paramétriques et les modèles trop vastes pour que l'on puisse espérer de bonnes vitesses de convergence. Un exemple de telle classe intermédiaire est celle des densités monotones pour lesquels de bons estimateurs ont été développés par Grenander (1956) et Birgé (1987a, 1987b).

Des versions particulières de l'estimateur à noyau se comportent bien pour des classes cibles définies en termes de régularité de f, telle que la classe de toutes les densités analytiques. Pour celles-ci nous savons que la fonction caractéristique ϕ satisfait

$$|\phi(t)| \leq \exp(-a|t|) \quad \text{où} \quad a > 0 \ .$$

Sous une condition additionnelle sur les queues il est possible de choisir K et h de façon à ce que l'erreur de l'estimateur à noyau soit étonnamment petite : $O((\log n / n)^{1/2})$. Il est par ailleurs possible, pour toute classe dont les éléments ont une fonction caractéristique s'annulant en dehors d'un compact $[-T, +T]$, d'exhiber des estimateurs à noyau non biaisés (Ibragimov et Khas'minskii 1982).

Cependant, comment peut-on traiter des classes importantes comme celle de toutes les densités log-concaves (cette classe comprend les lois normale, gamma, de Weibull, bêta, exponentielle, logistique, sécante hyperbolique, gaussienne inverse généralisée, de Gumbel, de Kummer, de Perks, ainsi que la densité de l'arc tangente d'une v.a. de loi Pearson IV) ? D'autres classes non-paramétriques d'usage courant soulèvent la même question. Par exemple celle des mélanges de lois normales ou celle des densités de sommes de k v.a. à densité de support inclus dans [0 , 1].

Problèmes multivariés.

Dans beaucoup de problèmes modernes, la dimension de l'espace euclidien est si grande que la taille des échantillons obtenus en pratique ne permet pas d'obtenir des erreurs acceptables pour des estimateurs universellement applicables comme l'estimateur à noyau. Deux façons de voir, qui d'ailleurs sont loin de s'opposer, semblent s'imposer.

La première consiste à marier méthode du noyau et transformations géométriques globales ou locales (Deheuvels 1977b, Abdous et Berlinet 1986); ces transformations pourront être définies grâce aux méthodes descriptives classiques d'analyse de données. La seconde utilise des modèles flexibles déterminés par un nombre fini mais variable de paramètres interprétables. C'est le cas de l'estimation par directions révélatrices (projection pursuit) pour laquelle la méthode du noyau unidimensionnelle peut se présenter comme un auxiliaire précieux. Il est souhaitable que l'avenir nous apporte dans ce domaine des résultats de fond concernant le comportement des estimateurs. Enfin, le théorème de Rao-Blackwell (Prakasa Rao 1983, Bosq et Lecoutre 1987) invite (en toute dimension) à rechercher des statistiques permettant, par conditionnement, l'amélioration d'estimateurs connus (Berlinet, 1984).

Choix automatique du paramètre de lissage.

Certains chercheurs adoptent le point de vue suivant : ils choisissent d'abord une classe d'estimateurs et ensuite essaient d'en tirer le meilleur parti. Il leur faut accepter les conséquences de cette stratégie : l'erreur est alors bornée inférieurement par

$$V(f,n) = \inf_{f_n \in C} E \int |f_n - f| \ ,$$

où C est la classe des estimateurs considérés. Nous avons déjà mentionné que pour les estimateurs à noyau, une borne inférieure pour V(f,n) peut être déterminée. Parfois il est même possible de calculer des bornes inférieures non triviales pour

$$V(n) = \inf_{f} V(f,n) \ .$$

Il serait intéressant de pouvoir sélectionner à partir des données un f_n de C pour lequel la borne inférieure V(f,n) serait atteinte à une constante multiplicative près. Une telle règle de choix nous permettrait d'obtenir la meilleure vitesse de convergence possible à l'intérieur de C, tout en ignorant ce qu'elle est. Un bon point de départ pour la recherche future est le travail de Stone (1984, 1985) où un problème similaire est successivement résolu dans L_2 pour les estimateurs à noyau K fixé et les histogrammes.

Lorsque K est fixé, choisir le meilleur f_n à partir des données signifie définir une variable aléatoire $H = H(X_1 , \ldots , X_n)$ proche du choix optimal pour h. Nous dirons

qu'une sélection basée sur les données est asymptotiquement optimale pour f lorsque

$$\frac{E\int|f_{nH}-f|}{\inf_{h>0} E\int|f_{nh}-f|} \to 1 \quad \text{quand } n \to \infty$$

où f_{nh} est l'estimateur à noyau avec paramètre de lissage h. Si l'utilisateur n'a à choisir aucun paramètre a priori, nous dirons que la méthode en question est automatique. Plus grande est la classe des densités pour lesquelles une méthode est asymptotiquement optimale, meilleure est cette méthode. Il semble, pourtant, que pour l'estimateur à noyau une méthode asymptotiquement optimale pour toutes les densités n'existe pas; ceci est certainement le cas pour les supernoyaux et est conjecturé pour les noyaux d'intégrale un. Notons encore que si K = 1, tout choix de H conduit à l'optimalité asymptotique, alors que l'estimateur n'est pas consistant.

Dans une première approche, on pourrait analyser le comportement asymptotique de l'erreur sous certaines conditions de régularité pour f, et déterminer les formules optimales pour n et f. Par exemple,pour $K \geq 0$, ceci conduit approximativement à la formule suivante :

$$h^5 = \frac{c\left(\int K^2\right)\left(\int \sqrt{f}\right)^2}{n\left(\int x^2 K\right)^2\left(\int |f^{(2)}|\right)^2} \quad .$$

Les quantités inconnues dans cette expression sont estimées dans un premier temps en remplaçant f par un estimateur g_n , et la valeur estimée pour h est utilisée pour l'estimateur à noyau f_n. En dehors du cercle vicieux (comment choisissons-nous les paramètres dans f_n ?) et la nature asymptotique du processus (la formule donnée pour h n'est valide qu'asymptotiquement, sans aucune garantie pour un n particulier) , nous sommes confrontés au fait qu'il n'est jamais possible 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 pas consistant. Dans L_2 , cette procédure en deux étapes remonte à Woodroofe (1970) et Nadaraya (1974); voir aussi Deheuvels (1977a), Bretagnolle et Huber (1979). Dans L_1, où les composantes de biais et de variation de l'erreur ne peuvent être séparées, l'analyse tient de la gageure; l'optimalité asymptotique sous conditions de régularité fut d'abord

établie par Hall et Wand (1987). Duin (1976), Habbema, Hermans et Vandenbroek (1974) ont proposé une méthode non asymptotique qui consiste à maximiser par rapport à h la validation croisée :

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

où $f_{n,i}$ est l'estimateur à noyau basé sur les (n - 1) observations différentes de X_i . Cette méthode est valide dans certains cas (Chow, Geman et Wu (1983), Hall (1982), Marron (1985), Devroye et Györfi (1987)) mais peut conduire à des estimateurs non consistants lorsque la distribution a des queues qui décroissent à une vitesse au plus exponentielle (Schuster et Gregory (1981)). D'autre part ce critère n'est pas relié à la distance L_1 entre densités.

Une méthode de validation croisée dans L_2 a été proposée par Rudemo (1982) mais elle ne fournit aucune information sur le meilleur facteur de lissage pour L_1. Elle consiste à minimiser un estimateur convenable de

$$\int (f_n - f)^2 - \int f^2 = \int f_n^2 - 2 \int f f_n \ ,$$

par exemple

$$\int f_n^2 - \frac{2}{n} \sum_{i=1}^{n} f_{n,i}(X_i) \ .$$

Son optimalité asymptotique pour toutes les densités bornées et tous les noyaux bornés à support compact a été obtenue par Stone (1984). Pour d'autres études voir Bowman (1984), Hall (1983,1985), Burman (1985), Scott et Terrell (1986), Hall et Marron (1987a). Ici aussi, il n'est pas clair que la procédure soit consistante lorsque la densité n'est pas de carré intégrable. Il ne semble pas non plus facile de l'étendre au critère L_1.

Pour le lissage automatique dans L_1, considérons comme nous l'avons fait pour les estimateurs combinés l'ensemble de toutes les densités et, dans cet ensemble, deux familles d'estimateurs à noyaux basées sur le même échantillon; lorsque h varie de 0 à $+\infty$ chaque estimateur se déplace sur une courbe (figure 15). La famille (f_{nh}) est la famille d'intérêt car elle utilise le noyau K. La famille (g_{nh}) est créée artificiellement ; elle utilise un autre noyau L dont l'ordre (ordre de son premier moment non nul) est supérieur à s, ordre de K. Nous voudrions pouvoir identifier le f_{nh} le plus voisin de f mais, puisque f est inconnue nous utiliserons le f_{nH} correspondant à la valeur H de la fenêtre qui minimise la norme L_1 de ($f_{nh} - g_{nh}$).

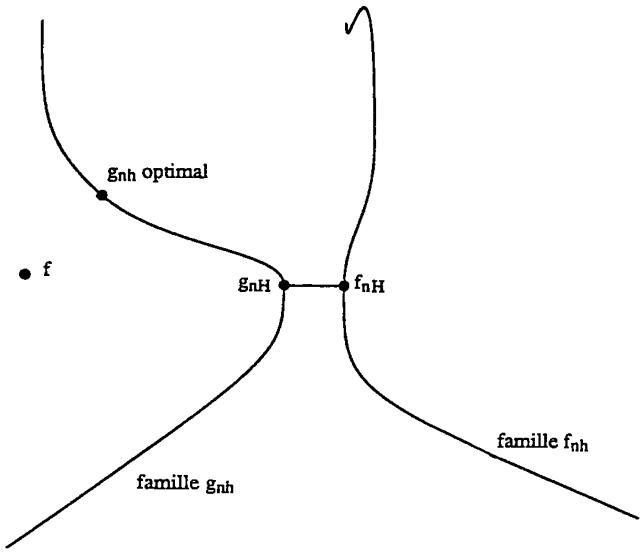

Figure 15. Sur le choix automatique de h.

En trait pointillé : $\frac{75}{16}(1-x^2)_+ - \frac{105}{32}(1-x^4)_+$ [ordre 4]

En trait plein : $\frac{3}{4}(1-x^2)_+$ [ordre 2]

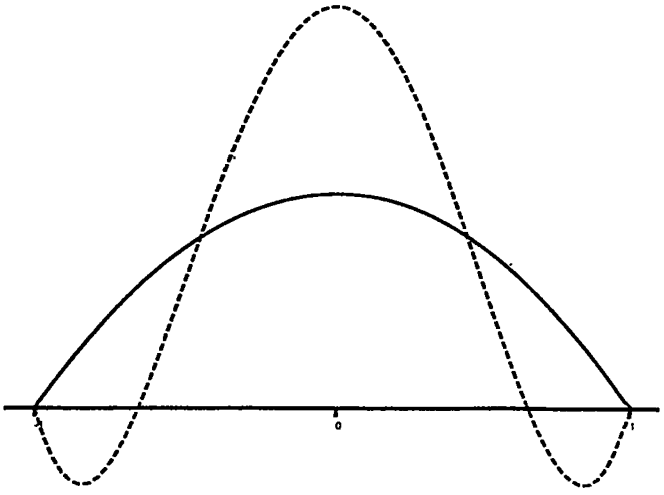

Figure 16. Un noyau d'ordre deux et un noyau d'ordre quatre.

Figure 16, sont représentés un noyau d'ordre 2 et un noyau d'ordre 4. Notons que dans la procédure de minimisation précédente, le facteur de lissage est libre de prendre toute valeur positive. Un fait intéressant est que la méthode fournit toujours un estimateur consistant. Pour beaucoup de densités lisses, on peut obtenir une certaine forme d'optimalité asymptotique :

$$\text{si } \varepsilon > 0\,,\ \int |f^{(s)}| < \infty\,,\ \int \sqrt{f} < \infty \text{ et } \int L^2 < \left(\frac{\varepsilon}{4\varepsilon + 8}\right)^2 \int K^2 \text{ alors}$$

$$\lim_{n \to \infty} \frac{E \int |f_{nH} - f|}{\inf_{h>0} E \int |f_{nh} - f|} \leq 1 + \varepsilon\ .$$

Le fait que L ait un ordre plus élevé que K peut entraîner que g_{nH} soit plus près de f que f_{nH}. Néanmoins, à l'intérieur de la famille (g_{nh}), g_{nH} est d'ordinaire loin d'être asymptotiquement optimal. Le problème du choix de l'ordre d'un noyau a été abordé par Hall et Marron (1987b) qui ont montré l'intérêt d'accorder la lissité de f et l'ordre choisi pour le noyau. Le fait de surévaluer l'ordre peut être néfaste. Pour terminer notons qu'il existe de nombreuses méthodes simples pour obtenir à partir d'un noyau fixé des noyaux d'ordre plus élevé. Par exemple la méthode de "twicing" de Stuetzle et Mittal (1979), basée sur le fait que si l'ordre de K est s, alors celui de (2 K - K * K) est 2 s. D' autre part, les super noyaux introduits ci-dessus ont un ordre infini et n'ont donc pas à être changés quand s varie. Pour plus de détails on pourra se reporter à Devroye (1987c).

Références

B. Abdous et A. Berlinet Convergence uniforme presque sûre d'une classe d'estimateurs à noyau, pour la densité. *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 303,pp. 761-764, 1986.

P. Assouad Deux remarques sur l'estimation. *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 296,pp. 1021-1024, 1983.

A. Berlinet Propriétés locales d'un paramètre. Application à l'estimation. *Comptes Rendus de l'Académie des Sciences de Paris*, vol. 298,pp. 345-348, 1984.

A. Berlinet Simulating to forecast. The Eighth International Symposium on Forecasting, Amsterdam, June 12-15, 1988.

L. Birgé Non-asymptotic minimax risk for Hellinger balls. *Probability and Mathematical Statistics*, Vol. 5, pp. 21-29, 1985.

L. Birgé On estimating a density using Hellinger distance and some other strange facts. *Probability Theory and Related Fields*, vol. 71,pp. 271-291, 1986

L. Birgé On the risk of histograms for estimating decreasing densities. *Annals of Statistics*, vol 15, pp.1013-1022, 1987a.

L. Birgé Estimating a density under order restrictions : non-asymptotic minimax risk. *Annals of Statistics*, vol 15, pp. 995-1012, 1987b.

D. Bosq et J.P. Lecoutre *Théorie de l'Estimation Fonctionnelle*, Economica, Paris, 1987.

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

J. Bretagnolle et C. Huber Estimation des densités : risque minimax. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 47, pp.119-137, 1979.

P. Burman A data dependent approach to density estimation. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, vol. 69, pp.609-628, 1985.

Y.S. Chow, S. Geman et L.D. Wu Consistent cross-validated density estimation. *Annals of Statistics*, vol. 11,pp. 25-38, 1983.

K.B. Davis Mean integrated square error properties of density estimates. *Annals of Statistics*, vol. 5,pp. 530-535, 1977.

P. Deheuvels Estimation non-paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée*, vol. 25,pp. 5-42, 1977a.

P. Deheuvels Estimation non-paramétrique de la densité par histogrammes généralisés (II). *Publications de l'Institut de Statistique des Universités de Paris*, vol. 22,pp. 1-23, 1977b.

L. Devroye The equivalence of weak, strong and complete convergence in L_1 for kernel density estimates. *Annals of Statistics*, vol. 11,pp. 896-904, 1983.

L. Devroye *Non-Uniform Random Variate Generation* Springer-Verlag, 1986.

L. Devroye Nonparametric density estimates with improved performance on given sets of densities. *Statistics (Mathematische Operationsforschung und Statistik)*, 1986, to appear.

L. Devroye *A Course in Density Estimation*. Birkhauser, Boston, 1987a.

L. Devroye An application of the Efron-Stein inequality in density estimation. *Annals of Statistics*, vol. 15,pp. 1317-1320, 1987b.

L. Devroye An L_1 asymptotically optimal kernel estimate. Technical Report, McGill University, Montréal, 1987c.

L. Devroye Asymptotic performance bounds for the kernel estimate. *Annals of Statistics*, vol. 16,pp. 1162-1179, 1988a.

L. Devroye The kernel estimate is relatively stable. *Probability Theory and Related Fields,* vol. 77, pp. 521-536, 1988b.

L. Devroye et L. Györfi *Nonparametric Density Estimation : The L_1 View.* John Wiley, New York, 1985.

L. Devroye et C.S. Penrod Distribution-free lower bounds in density estimation. *Annals of Statistics,* vol. 12,pp. 1250-1262, 1984.

R.P.W. Duin On the choice of smoothing parameters for Parzen estimators of probability density functions. *IEEE Transactions on Computers,* vol. C-25, pp. 1175-1179, 1976.

U. Grenander On the theory of mortality measurement, part II. *Skandinavisk Aktuarietidskrift,* vol. 39, pp. 125-153, 1956.

J.D.F. Habbema, J. Hermans et K. Vandenbroek A stepwise discriminant analysis program using density estimation. in COMPSTAT 1974, ed. G. Bruckmann, pp. 101-110, Physica-Verlag, Wien, 1974.

P. Hall Cross-validation in density estimation. *Biometrika,* vol. 69,pp. 383-390, 1982.

P. Hall Large-sample optimality of least squares cross-validation in density estimation. *Annals of Statistics,* vol. 11,pp. 1156-1174, 1983.

P. Hall Asymptotic theory of minimum integrated square error for multivariate density estimation. *Multivariate Analysis VI,* ed. Krishnaiah, pp. 289-309, North-Holland, Amsterdam, 1985.

P. Hall et J.S. Marron 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, 1987a.

P. Hall et J.S. Marron Choice of kernel order in density estimation. *Annals of Statistics,* vol. 16,pp. 161-173, 1987b.

P. Hall and M.P. Wand Minimizing L_1 distance in nonparametric density estimation. *Journal of Multivariate Analysis*, 1988, to appear.

I. A. Ibragimov et R.Z. Khas'minskii Estimation of distribution density belonging to a class of entire functions.*Theory of Probability and its Applications*, vol. 17, pp. 551-562, 1982.

J.S. Marron An asymptotically efficient solution to the bandwith problem of kernel density estimation. *Annals of Statistics*, vol. 13,pp. 1011-1023, 1985.

E.A. Nadaraya 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, 1974.

I. Olkin et C.H. Spiegelman A semiparametric approach to density estimation. *Journal of the American Statistical Association*, vol. 82,pp. 858-865, 1987.

E. Parzen On the estimation of a probability density function and the mode. *Annals of Mathematical Statistics*, vol. 33,pp. 1065-1076, 1962.

M. Rosenblatt Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, vol. 27,pp. 832-837, 1956.

M. Rudemo Empirical choice of histograms and kernel density estimators. *Scandinavian Journal of Statistics*, vol. 9,pp. 65-78,1982.

H. Scheffé A useful convergence theorem for probability distribution. *Annals of Mathematical Statistics*, vol. 18,pp. 434-438, 1947.

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

E.F. Schuster et S. Yakowitz Parametric / nonparametric mixture density estimation with application to flood frequency analysis. *Water Resources Bulletin,* vol. 21,pp. 797-804, 1985.

D.W. Scott et G.R. Terrell Biased and unbiased cross-validation in density estimation.*Journal of the American Statistical Association,* vol. 82,pp. 1131-1146, 1987.

B.W. Silverman *Density Estimation for Statistics and Data Analysis.* Chapman et Hall, Londres, 1986.

C.J. Stone An asymptotically optimal window selection rule for kernel density estimates. *Annals of Statistics,* vol. 12,pp. 1285-1297, 1984.

C.J. Stone An asymptotically optimal histogram selection rule. *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer,* vol. 2, ed. L. Le Cam et R.A. Olshen, pp. 513-520, Wadsworth, Belmont, 1985.

W. Stuetzle et Y. Mittal Some comments on the asymptotic behavior of robust smoothers. *Proceedings of the Heidelberg Workshop,* ed. T. Gasser et M. Rosenblatt, pp. 191-195, Springer-Verlag, Heidelberg, 1979.

R.A. Tapia et J.R. Thompson *Nonparametric Probability Density Estimation,* Johns Hopkins University Press, Baltimore, 1978.

G.S. Watson et M.R. Leadbetter On the estimation of the probability density. *Annals of Mathematical Statistics,* vol. 34,pp. 480-491, 1963.

M. Woodroofe On choosing a delta sequence. *Annals of Mathematical Statistics,* vol. 41,pp. 1665-1671, 1970.