Décision et Signal

version 2 (finale), 11 février 2002

Table des matières

Chapitre 1

Estimation de Densité et Prise de Décision

Alain Berlinet et Gérard Biau[1]

1.1. Introduction

Beaucoup a été dit et écrit dans les cinquante dernières années sur l'estimation d'une densité de probabilité. De nombreux estimateurs ont été définis, étudiés et comparés. Pourtant c'est bien souvent le sacro-saint histogramme que l'on voit utilisé par les praticiens qui n'ont pas reçu de formation spécifique. Il y a d'ailleurs bien peu de logiciels, en dehors de ceux destinés à des spécialistes du calcul scientifique, qui proposent des outils alternatifs efficaces et accessibles. Les lisseurs qui sont parfois proposés, comme l'estimateur à noyau, dépendent de paramètres dont l'interprétabilité n'est pas simple pour le profane. Celui-ci préfère souvent un outil rudimentaire dont il sait interpréter les valeurs à l'aide de fréquences de points tombant dans des intervalles. Notre but dans ce chapitre est de présenter quelques-uns des principaux outils d'estimation de la densité ; nous mettrons l'accent sur les idées de base ayant conduit à leur définition ainsi que sur leurs qualités et défauts éventuels. L'optique "prise de décision" nous a amenés à privilégier un encouragement à la mise en œuvre effective. De manière à ne pas rebuter le lecteur peu familier des techniques mathématiques enseignées en second cycle universitaire, nous renverrons à quelques références clé lorsque des considérations mathématiques plus profondes sont à envisager pour justifier les propriétés mises en avant. Commençons par prendre le mot *densité*

[1]Laboratoire de Probabilités et Statistique, Université Montpellier II, Place Eugène Bataillon, 34095 Montpellier Cedex 5, France, e-mail : berlinet@stat.math.univ-montp2.fr

dans son acception courante pour aborder un problème de Mathématiques expérimentales. Il n'est donc pas question pour le moment de modélisation par variable aléatoire.
Considérons un ensemble donné de points, sur une droite, dans un plan ou bien encore dans un espace de dimension quelconque. Nous voulons avoir une idée sur la façon dont ces points se répartissent localement, sur leur densité, de manière à décider de conjectures à vérifier ultérieurement. Semblent-ils se concentrer au voisinage d'un point, au contraire remplir un certain ensemble, ou bien encore s'éloigner graduellement ? Un premier outil pour décider d'analyses ultérieures est bien sûr l'histogramme. Pour le construire on définit un certain nombre de sous-ensembles dans l'ensemble où l'on fait les observations et on dénombre ensuite les points tombant dans chacun d'eux. Prenons un exemple simple. Fixons un réel a supérieur à 1 et examinons la répartition de la suite des nombres x_n qui sont les parties décimales des nombres a^n, $n \geq 1$. Par exemple, si $a = 1.5$ on a

$$a^2 = 2.25, a^3 = 3.375, a^4 = 5.0625, a^5 = 7.59375, \text{ etc},$$

et donc

$$x_1 = 0.5, x_2 = 0.25, x_3 = 0.375, x_4 = 0.0625, x_5 = 0.59375, \text{ etc.}$$

La répartition de ce type de suites a beaucoup d'importance et d'implications dans la recherche de générateurs de nombres pseudo-aléatoires, et par voie de conséquence en Statistique, Analyse Numérique et Cryptographie.
Considérons les 10000 premières puissances de $a = 1.5$ et dénombrons les termes de la suite $(x_n)_{1 \leq n \leq 10^4}$ associée dans les intervalles de la forme

$$A_i = \left[\frac{i}{50}, \frac{i+1}{50}\right[, \qquad 0 \leq i \leq 49.$$

L'histogramme p associé, représenté sur la figure 1.1, est donc défini par

$$p(x) = \frac{N(x)}{10^4(1/50)} \tag{1.1}$$

où $N(x)$ est le nombre de points parmi les 10000 qui tombent dans l'intervalle contenant x. Cette définition met en relation la fréquence d'un intervalle ($N(x)/10^4$) avec la longueur de cet intervalle (1/50). La somme des aires des rectangles bâtis sur chacun des intervalles est égale à 1 ce qui fait que la définition (1.1) conduit à une fonction constante par morceaux normalisée par son intégrale égale à 1. C'est toujours avec ce sens bien précis que nous utiliserons l'appellation "histogramme". Ce vocable a cependant tendance à recouvrir tout type de diagrammes à base de rectangles (en dimension 2) ou de parallélépipèdes rectangles (en dimension 3) pour lesquels l'expression *diagrammes en*

barres serait préférable.
En notant $\mathbf{1}_A$ la fonction indicatrice d'un ensemble A :

$$\mathbf{1}_A(x) = \begin{cases} 1 & \text{si} \quad x \in A \\ 0 & \text{si} \quad x \notin A \end{cases}$$

on obtient, si $x \in A_i$,

$$N(x) = \sum_{j=1}^{10^4} \mathbf{1}_{A_i}(x_j) = \sum_{i=0}^{49} \sum_{j=1}^{10^4} \mathbf{1}_{A_i}(x)\mathbf{1}_{A_i}(x_j)$$

et donc

$$p(x) = \frac{1}{10^4} \sum_{i=0}^{49} \sum_{j=1}^{10^4} \frac{\mathbf{1}_{A_i}(x)\mathbf{1}_{A_i}(x_j)}{\mathrm{mes}(A_i)} \tag{1.2}$$

où $\mathrm{mes}(A_i)$ est la mesure de l'intervalle A_i au sens usuel, c'est-à-dire sa longueur. Cette dernière écriture de l'histogramme, apparemment plus compliquée, nous sera utile dans la suite.

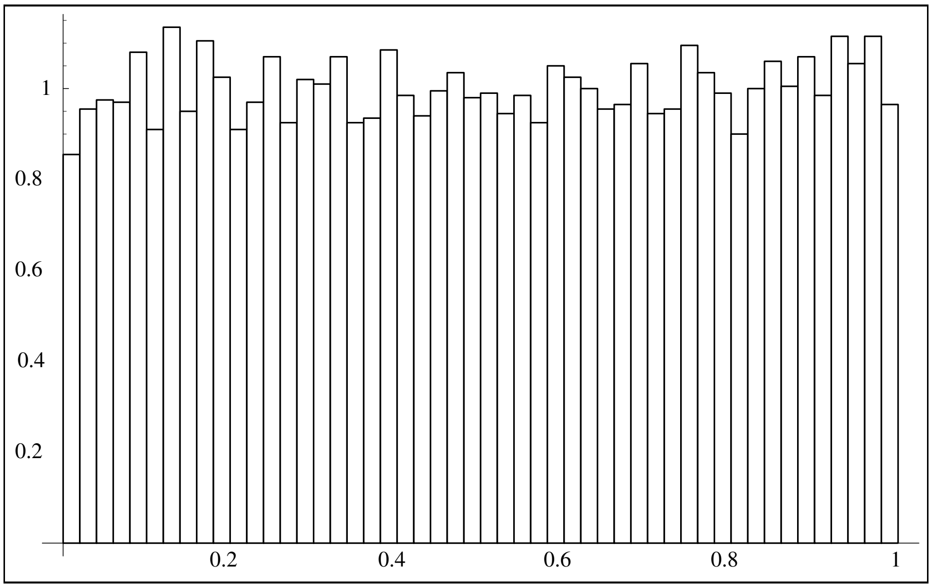

Figure 1.1. *Un histogramme associé à la suite des parties décimales des* 10000 *premières puissances entières de* 1.5.

A l'examen de la figure 1.1 on serait tenté de conjecturer que les nombres x_n ont tendance à occuper uniformément l'intervalle $[0, 1]$. On voit ainsi que l'exploitation des possibilités de calcul et de représentation graphique des ordinateurs permet de prendre des décisions quant à des axes de recherche en Mathématiques. La conjecture précédente est une question ouverte [DRE 01].

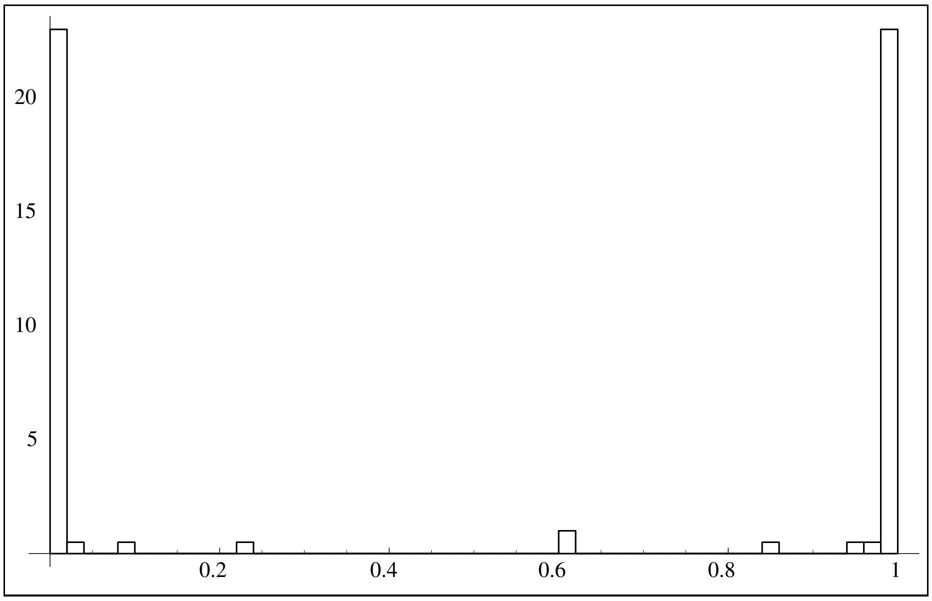

Figure 1.2. Un histogramme associé à la suite des parties décimales des 100 premières puissances entières de $(1 + \sqrt{5})/2$.

La figure 1.2 montre l'histogramme construit à partir des mêmes intervalles pour la suite $(x_n)_{1 \leq n \leq 100}$ obtenue lorsque a est égal au nombre d'or $(1+\sqrt{5})/2$. En dehors des 8 premiers, les termes se retrouvent alternativement dans les intervalles $[0, 0.02]$ et $[0.98, 1]$. Au vu de ce diagramme et de ces résultats numériques on est bien sûr amené à mettre en doute l'équirépartition dans $[0, 1]$ de la suite observée et à formuler l'hypothèse que, pour $n \geq 9$,

si n est pair $x_n \geq 0.98$;

si n est impair $x_n \leq 0.02$.

On pourrait aller un peu plus loin en s'interrogeant sur la convergence (monotone) éventuelle de la sous-suite d'indices pairs vers 1 et de la sous-suite d'indices impairs vers 0.

Si un élément x_n nous était fourni sans que nous ne connaissions n et que nous devions conclure sur la parité de n en nous basant sur l'étude expérimentale des 100 premières valeurs de la suite, nous aurions donc tendance à décider que

n est pair si x_n est proche de 1 ;

n est impair si x_n est proche de 0.

En observant comment se répartissent les nombres x_n en fonction de "l'étiquette de parité" de n, on en vient à construire un processus décisionnel sur cette étiquette. Dans le cas présent, il est facile d'apporter la preuve mathématique des conjectures expérimentales en utilisant les propriétés du nombre d'or.

Lorsque l'on a à prendre une décision au vu d'une fonction p rendant compte d'une densité de points, les caractéristiques importantes sont bien sûr les modes globaux ou locaux $\big($valeurs de x donnant les plus grandes valeurs de $p(x)\big)$ et plus généralement les ensembles de niveau minimal fixé, c'est-à-dire les valeurs de x pour lesquelles on a $p(x) \geq H$ où H est le niveau que l'on s'est fixé. Sont également importants les points où une irrégularité ou une rupture semble se manifester, par exemple un saut dans les valeurs de $p(x)$ ou dans ses taux d'accroissement. Le problème inhérent à la plupart des lisseurs est qu'ils ont plutôt tendance à atténuer les irrégularités : les creux et bosses sont sous-estimés et les ruptures difficilement identifiables, à moins de modifications ad hoc. Ils posent également des problèmes de temps de calcul lorsque les données sont nombreuses ou que les décisions sont à prendre en temps réel, les données arrivant par flux et nécessitant une constante mise à jour. On leur préférera donc souvent des fonctions simples à calculer, dont les valeurs dépendent du nombre de points tombant dans des intervalles autour du point auquel on s'intéresse. Nous les appellerons *densitogrammes* lorsqu'elles sont positives ou nulles et ont une intégrale normalisée à 1. Certains densitogrammes atteignent les meilleures performances possibles. Nous y reviendrons dans la suite.

Le problème décisionnel à traiter se formule souvent en termes d'identification de la source d'une mesure ou d'un signal. On parle comme ci-dessus de *l'étiquette* ou du *label* (anglicisme) d'une variable. L'étiquette est donc une information de nature catégorielle sur une variable X observée. Elle identifie le groupe (ou catégorie) auquel appartient l'individu de la population qui a donné lieu à l'observation X. Le problème général du diagnostic médical se place dans ce cadre. Prenons le cas de l'identification d'une pathologie basée sur la biochimie sanguine. La variable sera alors un vecteur $\boldsymbol{X}$ dont les composantes seront les différents taux évalués lors de l'analyse : glycémie, triglycérides, cholestérol, urée, sodium, potassium, chlore, fer sérique, etc. Si, par exemple, 10 dosages ont eu lieu, $\boldsymbol{X}$ est un vecteur de dimension 10. On dispose d'informations sur les valeurs prises par les composantes de $\boldsymbol{X}$ chez un certain nombre de sujets sains et chez des sujets présentant une pathologie donnée. Ceci a pu être avéré par des examens plus poussés ou d'autres signes cliniques. Un nouveau patient se présente. Il faut décider au vu de ses résultats d'analyse s'il y a lieu de le traiter pour la pathologie en question. Bien entendu le diagnostic final s'appuiera sur bien d'autres paramètres mais limitons-nous ici à la décision binaire, à prendre à partir de $\boldsymbol{X}$ seul,

1 : on retient la possibilité de la pathologie chez le patient
0 : on écarte cette possibilité.

Une telle décision est en général prise selon un processus majoritaire que l'on peut décrire comme ceci. Les sujets étiquetés sains et ceux étiquetés malades ont fourni des observations dont certaines sont "proches" de $\boldsymbol{X}$. Parmi celles-ci on détermine quelle est l'étiquette la plus représentée (plus de sains ou plus de malades) et on l'attribue à $\boldsymbol{X}$. Il ne faut surtout pas se laisser gagner par une impression de simplicité de la démarche. Dans l'exemple présenté ici, la notion de "proximité" que l'on doit envisager est d'une définition très délicate.

Le processus majoritaire peut être mis en œuvre facilement grâce aux fonctions utilisées pour évaluer les "densités" de répartitions de points. Qu'il s'agisse d'histogrammes, de densitogrammes, d'estimateurs à noyau ou de lisseurs un peu plus compliqués à définir, ils s'écrivent sous la forme d'une somme de termes dont chacun ne dépend que d'une des observations effectuées. Soyons plus précis. Notons $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_N$, les N observations dont l'étiquette est connue et $\boldsymbol{X}$ celle pour laquelle il faut prendre une décision d'attribution d'étiquette. Les fonctions que nous venons de mentionner s'écrivent pour la plupart sous la forme

$$\sum_{i=1}^{N} K_N(\boldsymbol{x}, \boldsymbol{X}_i) \tag{1.3}$$

où K_N est une fonction généralement positive qui dépend de N. On obtiendra par exemple l'histogramme (formule 1.2 ci-dessus) en prenant

$$K_N(x,y) = \sum_{j=0}^{49} \frac{\mathbf{1}_{A_j}(x)\mathbf{1}_{A_j}(y)}{N\mathrm{mes}(A_j)}$$

et $N = 10^4$.

On compare alors les contributions des observations à la somme (1.3) en fonction de leur étiquette. Si les étiquettes possibles sont, pour simplifier, $0, 1, 2, \ldots, M$ on calcule

la somme des $K_N(\boldsymbol{X}, \boldsymbol{X}_i)$ pour lesquels l'étiquette de $\boldsymbol{X}_i$ est 0,

la somme des $K_N(\boldsymbol{X}, \boldsymbol{X}_i)$ pour lesquels l'étiquette de $\boldsymbol{X}_i$ est 1,

....

la somme des $K_N(\boldsymbol{X}, \boldsymbol{X}_i)$ pour lesquels l'étiquette de $\boldsymbol{X}_i$ est M.

La *discrimination* entre les étiquettes possibles s'effectue ensuite en attribuant à $\boldsymbol{X}$ l'étiquette correspondant à la plus grande somme.

Pour l'histogramme, ceci revient simplement à chercher l'étiquette majoritaire parmi les observations qui sont tombées dans l'intervalle où se trouve $\boldsymbol{X}$.

C'est bien la "densité" de chacun des groupes possibles autour de l'observation $\boldsymbol{X}$ qui est l'outil essentiel de ce type de décision.

1.2. Estimer une densité

Venons-en maintenant au sens donné en Probabilités au terme *densité.*

1.2.1. *Présentation du problème*

Considérons une variable aléatoire X choisie pour modéliser un phénomène physique stochastique[2] continu. X sera par exemple la concentration d'une substance chimique dans le sang, la hauteur d'eau d'une rivière en crue, la pression atmosphérique en certains points de la surface de la Terre ou encore l'intensité d'un signal reçu par un radio-télescope. La variable X est définie sur un certain espace probabilisé dont on notera $\mathbf{P}$ la probabilité et elle prend ses valeurs dans l'ensemble $\mathbb{R}$ des nombres réels. Considérons d'autre part une fonction p définie sur $\mathbb{R}$ et à valeurs dans l'ensemble $\mathbb{R}^+$ des réels positifs ou nuls à laquelle nous imposerons d'être continue par morceaux et d'intégrale égale à 1.
On dit que X admet p pour *densité (de probabilité)* si la probabilité que X tombe dans un intervalle quelconque $[a, b]$ $(a \leq b)$ s'obtient par intégration de p entre a et b

$$\mathbf{P}(a \leq X \leq b) = \int_a^b p(x)\,\mathrm{d}x.$$

En tout point x où p est continue on a donc

$$p(x) = \lim \frac{1}{b-a} \int_a^b p(t)\,\mathrm{d}t = \lim \frac{\mathbf{P}(a \leq X \leq b)}{\mathrm{mes}\big([a,b]\big)}$$

où la limite est prise pour a et b tendant vers le point fixe x. Lorsque a et b tendent vers x, la probabilité $\mathbf{P}(a \leq X \leq b)$ est de l'ordre de

$$p(x)\ \mathrm{mes}\big([a,b]\big)$$

ce qui fait que le nombre $p(x)$ rend bien compte d'une densité locale de probabilité, en un sens intuitif. Si l'on note F la fonction de répartition de X, c'est-à-dire la fonction définie par

$$F(x) = \mathbf{P}(X \leq x), \qquad x \in \mathbb{R},$$

alors F est dérivable aux points de continuité de p et y admet p pour dérivée.
Plus généralement, on dit qu'un vecteur aléatoire $\boldsymbol{X}$ (*i.e.* une variable aléatoire à valeurs dans $\mathbb{R}^d$) admet une densité si il existe une fonction p (positive,

[2]Stochastique : du grec $\sigma\tau o\chi\acute{\alpha}\zeta\varepsilon\iota\nu$, viser. Qui est de nature aléatoire.

continue par morceaux et d'intégrale 1) telle que, pour tout sous-ensemble A de $\mathbb{R}^d$ de la forme $A = I_1 \times I_2 \cdots \times I_d$ où I_i est un intervalle de $\mathbb{R}$, on ait

$$\mathbf{P}(\boldsymbol{X} \in A) = \int_A p(x_1, \ldots, x_d)\, \mathrm{d}x_1 \ldots \mathrm{d}x_d.$$

La densité p permet de calculer les probabilités d'événements associés à $\boldsymbol{X}$. En tant que fonction, elle nous autorise à accéder au comportement "local" du phénomène aléatoire continu représenté par $\boldsymbol{X}$. Plus la valeur de la densité autour d'un point $\boldsymbol{x}$ sera élevée, plus $\boldsymbol{X}$ aura des chances de se produire (ou de se *réaliser*) au voisinage de $\boldsymbol{x}$.

Exemples

- La variable aléatoire X dont la loi est définie par

$$\mathbf{P}(X \leq x) = 1 - \exp(-\lambda x)$$

pour tout x positif (λ est une constante réelle strictement positive) admet pour densité

$$p(x) = \begin{cases} \lambda \exp(-\lambda x) & \text{si} \quad x \geq 0 \\ 0 & \text{si} \quad x < 0. \end{cases}$$

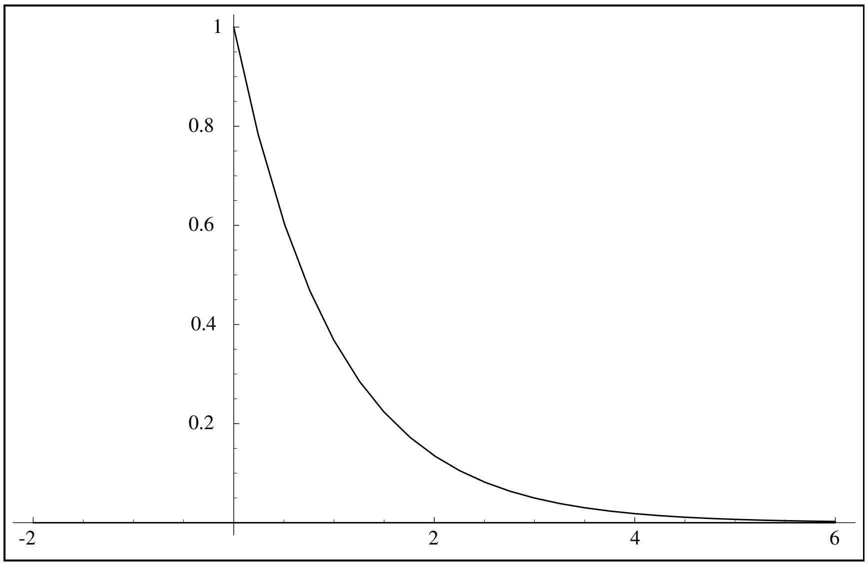

Figure 1.3. *Densité exponentielle de paramètre 1.*

La loi de X est appelée *loi exponentielle* de paramètre λ. Son espérance et son écart-type sont égaux à $1/\lambda$. Elle est couramment utilisée pour représenter des durées de vie, notamment de composants électroniques (figure 1.3).

- Une variable aléatoire réelle X est dite *normale* (ou *gaussienne*) si elle admet une densité du type

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-m)^2}{2\sigma^2}\right), \qquad x \in \mathbb{R},$$

où m est l'espérance de X et $\sigma > 0$ est son écart-type. Plus généralement, un vecteur aléatoire $(X_1, \ldots, X_d)$ est dit normal (ou gaussien) $\mathcal{N}(m, \Sigma)$ d'espérance $\boldsymbol{m} = (m_1, \ldots, m_d)$ et de matrice de covariance Σ si sa densité est

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\sqrt{\text{dét}\Sigma}} \exp\Big(- \frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^t \Sigma^{-1} (\boldsymbol{x}-\boldsymbol{m})\Big).$$

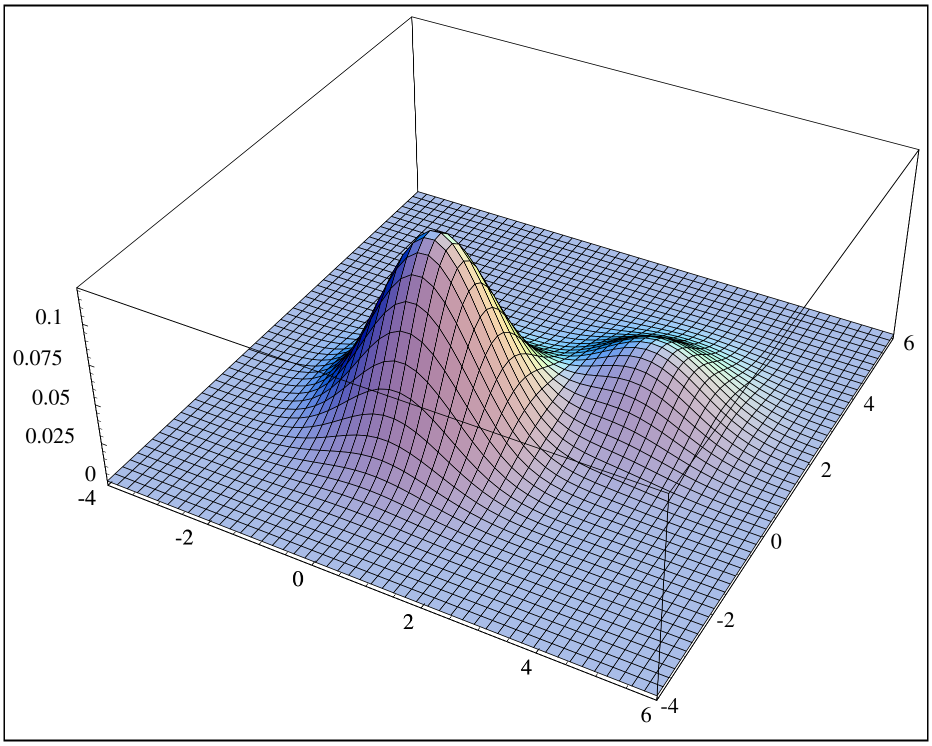

Figure 1.4. *Une densité en dimension $d = 2$, mélange de deux lois normales.*

Dans la pratique, la densité p est inconnue. L'information dont dispose l'expérimentateur se réduit à une collection de valeurs observées $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, réalisations indépendantes du phénomène aléatoire $\boldsymbol{X}$. La question qui se pose est alors la suivante : peut-on, à l'aide des mesures $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, reconstruire ou au moins *estimer*, la densité sous-jacente p ? Un *estimateur* p_N de p est une fonction de $\boldsymbol{x}$ (le point de $\mathbb{R}^d$ où l'on souhaite estimer la valeur de p) et de $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ (ce vecteur aléatoire s'appelle l'*échantillon* : chaque $\boldsymbol{X}_i$ est une variable aléatoire de même loi que $\boldsymbol{X}$ qui, dans le cas présent, s'est réalisée en $\boldsymbol{x}_i$). Le choix de l'estimateur p_N va dépendre de l'application en vue et des caractéristiques que l'expérimentateur exigera de lui : régularité, rapidité de calcul, interprétabilité, flexibilité, stabilité par rapport aux fluctuations d'échantillon ou aux éventuelles valeurs aberrantes, optimalité au regard de certains critères, propriétés de convergence (quand N tend vers l'infini), etc.
La bibliographie sur le sujet est très vaste. Pour une première approche on pourra consulter Delecroix [DEL 83], Silverman [SIL 92], puis Scott [SCO 92]. Pour approfondir on se référera aux livres de Bosq [BOS 87], Devroye et Györfi [DEV 85] et Devroye [DEV 87]. Le récent livre de Devroye, Györfi et Lugosi [DEV 96] développe largement la théorie probabiliste de la discrimination.

1.2.2. *Estimation paramétrique vs estimation non paramétrique*

Afin d'estimer la densité p, une première approche (dite *paramétrique*) consiste à supposer que p appartient à une famille de densités qui peuvent être décrites par un petit nombre (connu) de paramètres réels. Par exemple les densités normales sur $\mathbb{R}$ sont caractérisées par leur espérance et leur variance. Le statisticien qui opte pour une telle approche possède une bonne connaissance a priori du phénomène aléatoire. Il sait, par intuition ou par expérience, que la variable aléatoire X suit une loi normale, tout en ignorant la valeur de son espérance ou de sa variance. Dans un tel contexte, l'estimation de la densité se réduit alors à un problème d'*estimation de paramètres*. C'est ainsi que Karl Pearson introduisit à la fin du XIXème siècle la célèbre famille de densités de probabilité qui porte son nom et dont chaque élément satisfait à une équation différentielle du type

$$p'(x) = \frac{(x-a)p(x)}{b_0 + b_1 x + b_2 x^2}\,, \quad (a, b_0, b_1, b_2) \in \mathbb{R}^4. \tag{1.4}$$

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 coefficients d'asymétrie et d'aplatissement et donc de refléter fidèlement l'allure d'une distribution lorsqu'on se limite à ces deux paramètres (voir Saporta [SAP 90] pour leur définition et leur illustration). Pour les définitions et propriétés des modèles de densités on pourra se reporter au livre de Johnson et Kotz [JOH 69].

Malheureusement, le modèle de Pearson présente de graves défauts : des modèles simples très courants en sont exclus (Laplace, log-normal, Weibull, Gumbel, logistique) et l'équation (1.4) nous montre les limitations sévères du modèle en ce qui concerne l'existence de maxima : la dérivée ne s'annule que si $p(x) = 0$ ou si $x = a$. Pour modéliser une densité à deux ou plusieurs "bosses," il faut donc recourir à un mélange de densités de Pearson, ce qui introduit de nouveaux paramètres. En outre, une méthode d'estimation des paramètres a, b_0, b_1 et b_2 basée sur les quatre premiers moments empiriques (les moments de l'échantillon) sera très sensible aux extrêmes ou à de légères fluctuations de la loi sous-jacente. Un exemple d'une telle méthode est celle qui est dite *des moments,* où l'on substitue simplement les moments empiriques aux moments théoriques dans les expressions des constantes.

Ainsi, à moins d'avoir sur le phénomène aléatoire étudié des informations a priori très précises et indiscutables, le champ d'application d'un modèle paramétrique n'est satisfaisant que lorsque l'inflation du nombre de paramètres est telle que les méthodes d'estimation du modèle deviennent tout à fait inefficaces.

Nous avons évoqué dans le paragraphe précédent l'estimation d'une densité comme mélange de densités appartenant à des modèles connus. Il convient de noter qu'une telle procédure n'entre dans le cadre paramétrique que s'il n'y a pas de problème d'identifiabilité et que le nombre de composantes du mélange est borné. Dans le cas contraire la modélisation par mélange et l'estimation dans ce cadre peuvent se révéler catastrophiques. Pour une discussion des problèmes liés à ce type de modèles et à leurs applications on pourra consulter [TIT 85].

Pour pallier les insuffisances et les défauts des familles paramétriques, une seconde approche (dite *non paramétrique*) propose de "laisser parler les données", sans spécifier au préalable de forme sur p. L'outil d'estimation non paramétrique par excellence (aussi vieux que la statistique elle-même) nous est fourni par l'*histogramme* : une fois les données regroupées en classes de valeurs, les fréquences empiriques sont représentées par des aires rectangulaires dont les bases correspondent aux classes elles-mêmes. 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 tombant près des bords d'une classe et ceux tombant près du milieu ne sont pas différenciés ; ceci explique la variabilité des interprétations statistiques que l'on peut faire d'un histogramme suivant le choix de l'origine et des classes. Pour des densités raisonnablement lisses, l'histogramme apparaît donc comme un estimateur sévèrement limité. Afin de 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.*

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

$$p_N(\boldsymbol{x}) = \frac{1}{N h_N^d} \sum_{i=1}^{N} K\Big(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h_N}\Big),$$

où la fonction K, appelé le *noyau*, est la densité uniforme sur $[-1/2, 1/2]^d$ et où $h_N > 0$ est la *largeur de fenêtre* ou *constante de lissage*. On peut obtenir des estimateurs plus esthétiques (les *estimateurs à noyau*), en utilisant des noyaux K plus réguliers (Rosenblatt [ROS 56] ; Parzen [PAR 62] ; Devroye et Györfi [DEV 85] ; Bosq et Lecoutre [BOS 87] ; Devroye [DEV 87] ; Berlinet et Devroye [BER 89]). Nous y reviendrons en détail dans la suite.
L'apparente opposition entre modèles paramétriques et non paramétriques que pourrait laisser supposer le titre du présent paragraphe est en fait à proscrire. Les vitesses de convergence dans les modèles paramétriques sont certes meilleures mais, en toute occasion, les estimateurs non paramétriques restent des compléments indispensables. Ils pourront tour à tour confirmer une hypothèse paramétrique ou bien se comporter en garde-fous lorsque des données sont contaminées ou lorsque la population étudiée n'a pas l'homogénéité qu'on lui prêtait.

1.2.3. *Estimation de densité et prise de décision*

Bien souvent, le problème de l'estimation de la densité ne constitue pas une fin en soi. Il est au contraire le point de départ nécessaire pour mener à bien d'autres problématiques de nature statistique. Illustrons ceci par un exemple en discrimination.
Comme indiqué en début de chapitre, le but des méthodes de discrimination est de prédire une variable qualitative à M catégories à l'aide d'une fonction d'un vecteur aléatoire numérique continu (Saporta [SAP 90] ; Devroye, Györfi et Lugosi [DEV 96]). La plupart du temps, le problème se pose de la manière suivante. On connaît sur un échantillon de N observations les groupes d'appartenance ainsi que les variables explicatives. A partir de cet échantillon, on souhaite développer une technique permettant d'affecter, ou de *classer*, avec une probabilité d'erreur minimale, de nouvelles observations dans les groupes. Supposons pour cela que les M groupes $G_1, \ldots, G_M$ soient en proportions respectives $\alpha_1, \ldots, \alpha_M$ dans la population totale et que la distribution de probabilité du vecteur des observations $\boldsymbol{X}$ soit donnée pour chaque groupe G_j par une densité de probabilité p_j. La probabilité qu'une nouvelle observation $\boldsymbol{x}$ provienne du groupe G_j est alors donnée par la formule de Bayes (Saporta

[SAP 97]) :

$$\mathbf{P}(G_j/\boldsymbol{x}) = \frac{\alpha_j p_j(\boldsymbol{x})}{\sum\limits_{j=1}^{M} \alpha_j p_j(\boldsymbol{x})}\,.$$

La règle bayesienne consiste alors à affecter l'observation $\boldsymbol{x}$ au groupe qui a la probabilité a posteriori maximale. Le dénominateurs étant les mêmes pour les M groupes, on doit donc chercher le maximum de $\alpha_j p_j(\boldsymbol{x})$. On voit que dans un tel contexte, il sera nécessaire d'estimer de quelle façon s'ordonnent les quantités inconnues $\alpha_j p_j(\boldsymbol{x})$.

1.3. Méthodes d'estimation paramétrique

Dans cette section, nous considérons un échantillon $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ de variables aléatoires indépendantes et de même loi à densité $p_{\boldsymbol{\theta}}$. Ici, $\boldsymbol{\theta}$ représente le paramètre de la densité commune aux variables de l'échantillon. Il appartient à un sous-ensemble de $\mathbb{R}$ ou de $\mathbb{R}^n$, car il peut s'agir d'un vecteur de dimension n dans le cas où la loi dépend de plusieurs constantes inconnues. Par exemple, pour une loi exponentielle le paramètre naturel peut être l'espérance ($\theta = 1/\lambda$) ou son inverse ; pour une loi normale $\mathcal{N}(m, \sigma^2)$, on prendra $\boldsymbol{\theta} = (m, \sigma^2)$. Il faut avoir présent à l'esprit que la paramétrisation résulte toujours d'un choix du statisticien et qu'elle n'a rien d'intrinsèque. Les méthodes les plus courantes d'estimation du paramètre $\boldsymbol{\theta}$ sont la méthode du *maximum de vraisemblance* et celle des *moments* (Tassi [TAS 89] ; Daudin, Robin et Vuillet [DAU 99] ; Montfort [MON 90]). Dans la suite, nous utiliserons les symboles $\mathbf{E}$ et $\mathbf{V}$ pour désigner l'espérance et la variance de variables aléatoires relativement à la probabilité $\mathbf{P}$.

1.3.1. *Maximum de vraisemblance*

Etant donné un échantillon de valeurs observées $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, la méthode du maximum de vraisemblance consiste à prendre comme estimation de $\boldsymbol{\theta}$ une valeur de $\boldsymbol{\theta}$ rendant maximale la densité conjointe des observations, notée $L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta})$ et appelée usuellement *vraisemblance* de l'échantillon. Les variables observées étant supposées indépendantes et de même densité f_θ, nous avons

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}) = \prod_{i=1}^{N} p_{\boldsymbol{\theta}}(\boldsymbol{x}_i)\,.$$

Comme son nom l'indique, la vraisemblance rend compte des chances qu'un échantillon observé $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ a de se réaliser pour une valeur $\boldsymbol{\theta}$ donnée du paramètre. Intuitivement, puisque L est la densité de probabilité du vecteur aléatoire $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$, maximiser L revient donc à déterminer $\boldsymbol{\theta}$ de façon à ce que l'événement qui s'est produit soit le plus "probable". Un estimateur $\hat{\boldsymbol{\theta}}_N$ de $\boldsymbol{\theta}$ réalisant le maximum de la vraisemblance est appelé *estimateur du maximum de vraisemblance*. Il est donc tel que

$$L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \hat{\boldsymbol{\theta}}_N) = \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} L(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N; \boldsymbol{\theta}),$$

où $\boldsymbol{\Theta}$ est le domaine de définition de $\boldsymbol{\theta}$, en général un intervalle de $\mathbb{R}$ ou de $\mathbb{R}^n$ si $\boldsymbol{\theta}$ est un vecteur à n coordonnées.
Etant donnée la forme des densités des lois de probabilité usuelles, on préfère souvent rechercher le maximum de la log-vraisemblance (logarithme de la vraisemblance) plutôt que celui de la vraisemblance (noter que les deux procédures sont équivalentes puisque le logarithme est une fonction strictement croissante).
D'un point de vue théorique, on peut montrer, moyennant quelques hypothèses plus techniques que contraignantes (voir Tassi [TAS 89] ou Monfort [MON 90] pour un traitement détaillé de la question), que l'estimateur du maximum de vraisemblance $\hat{\boldsymbol{\theta}}_N(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ possède de bonnes propriétés, tant à distance finie (le nombre d'observations N est fixé) qu'asymptotiquement (N tendant vers l'infini). En particulier :

1. $(\hat{\boldsymbol{\theta}}_N)$ converge presque sûrement vers $\boldsymbol{\theta}$.
2. Si B_N désigne le biais de l'estimateur $\hat{\boldsymbol{\theta}}_N$, *i.e.*, $B_N = \mathbf{E}(\hat{\boldsymbol{\theta}}_N) - \boldsymbol{\theta}$, on a $\lim_{N\to\infty} B_N = 0$. On dit que $\hat{\boldsymbol{\theta}}_N$ est *asymptotiquement sans biais*.
3. Lorsque $d = 1$ $\hat{\theta}_N$ suit, asymptotiquement, une loi normale. Plus précisément, pour N "grand", nous pouvons écrire

$$\hat{\theta}_N - \theta \approx \mathcal{N}\Big(0, V_N(\theta)\Big).$$

   Le terme $V_N(\theta)$ s'appelle la *borne de Cramer-Rao* : on montre, sous certaines conditions de régularité de p, qu'aucun estimateur sans biais ne peut avoir de variance inférieure à cette borne. Dans la pratique on utilise cette valeur (qui se calcule à partir de la fonction L) comme une approximation de $\mathbf{V}(\hat{\theta}_N)$, variance de l'estimateur pour N fixé.
   Ces résultats se généralisent à d quelconque.
4. Si $\hat{\boldsymbol{\theta}}_N$ est un estimateur du maximum de vraisemblance de $\boldsymbol{\theta}$ et si g est une application bijective, alors $g(\hat{\boldsymbol{\theta}}_N)$ est un estimateur du maximum de vraisemblance de $g(\boldsymbol{\theta})$.

Exemple (Saporta [SAP 90]) Estimation du paramètre de la loi de Weibull standard, de densité

$$p(x; \theta) = \theta x^{\theta-1} \exp(-x^\theta).$$

Sous l'hypothèse d'indépendance des valeurs observées la vraisemblance s'écrit ici

$$L(x_1,\ldots,x_N;\theta)=\theta^N\prod_{i=1}^{N}x_i^{\theta-1}\exp\Big(-\sum_{i=1}^{N}x_i^{\theta}\Big)$$

ou encore

$$\ln L(x_1,\ldots,x_N;\theta)=N\ln\theta+(\theta-1)\sum_{i=1}^{N}\ln x_i-\sum_{i=1}^{N}x_i^{\theta}.$$

Ainsi,

$$\frac{\partial\ln L}{\partial\theta}=\frac{N}{\theta}+\sum_{i=1}^{N}\ln x_i-\sum_{i=1}^{N}x_i^{\theta}\ln x_i.$$

$\hat\theta_N$ est donc solution de l'équation

$$\hat\theta_N=\frac{N}{\sum\limits_{i=1}^{N}(x_i^{\hat\theta_N}-1)\ln x_i}.$$

Cette dernière équation ne peut se résoudre que numériquement par approximations successives et on ne peut donc pas obtenir de forme explicite pour l'estimateur du maximum de vraisemblance. Notons encore que les équations de vraisemblance obtenues comme dans l'exemple par différentiation de la fonction L par rapport à θ peuvent conduire à des optima locaux et sont donc à manipuler avec quelques précautions.

1.3.2. *Méthode des moments*

Cette procédure d'estimation repose d'une part sur le fait que les paramètres d'une loi de probabilité s'expriment souvent en fonction de ses moments $\mathbf{E}(\boldsymbol{X}^k)$ et d'autre part sur la propriété de convergence (presque sûre) des moments empiriques d'un échantillon vers les moments théoriques correspondants. Le modèle paramétrique de Pearson auquel nous avons fait allusion au paragraphe 1.2.2 peut rendre tentante son utilisation. Mais, comme nous l'avons souligné, son manque de robustesse limite son utilisation aux cas où le modèle est très bien identifié et où l'ordre des moments estimés est très faible (souvent deux en pratique).

Soit donc $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_n)$ le paramètre d'intérêt. Dénotons par $m_k(\boldsymbol{\theta})$ le moment théorique d'ordre k de $\boldsymbol{X}$ et par $M_k(N)$ le moment empirique d'ordre k. L'estimateur de $\boldsymbol{\theta}$ par la méthode des moments est obtenu comme la solution

$\tilde{\boldsymbol{\theta}}_N$ du système d'équations suivant :

$$\left\{\begin{array}{ccc} m_1(\boldsymbol{\theta}) & = & M_1(N) \\ & \vdots & \\ m_n(\boldsymbol{\theta}) & = & M_n(N). \end{array}\right.$$

Le choix des moments est guidé par la facilité de résolution du système. On peut prendre des moments tous centrés en $\mathbf{E}(\boldsymbol{X})$, ou tous non centrés, ou un mélange de moments centrés ou non centrés. En outre, il n'y a aucune raison de retenir les n premiers moments, sinon la simplicité de calcul.

Exemple (Tassi [TAS 89]) Considérons une loi bêta de type I de paramètres $p_1 > 0$ et $p_2 > 0$ dont la densité est, sur $[0, 1]$,

$$p(x; p_1, p_2) = \frac{1}{B(p_1, p_2)} x^{p_1 - 1}(1 - x)^{p_2 - 1}, \quad \text{où } B(p_1, p_2) = \frac{\Gamma(p_1)\Gamma(p_2)}{\Gamma(p_1 + p_2)}.$$

Dans la formule précédente, Γ désigne la *fonction Gamma*. Par définition, pour $x > 0$, on a

$$\Gamma(x) = \int_0^\infty \exp(-t) t^{x-1}\, \mathrm{d}x.$$

On trouve, après un calcul facile,

$$\mathbf{E}(X) = \frac{p_1}{p_1 + p_2} \quad \text{et} \quad \mathbf{E}(X^2) = \frac{p_1(p_1 + 1)}{(p_1 + p_2)(p_1 + p_2 + 1)}.$$

La résolution du système fournit alors les estimateurs des moments des paramètres p_1 et p_2 :

$$\tilde{p}_{1,N} = \frac{M_1(N)\big(M_1(N) - M_2(N)\big)}{M_2(N) - M_1(N)^2} \quad \text{et} \quad \tilde{p}_{2,N} = \frac{\big(1 - M_1(N)\big)\big(M_1(N) - M_2(N)\big)}{M_2(N) - M_1(N)^2},$$

où, naturellement,

$$M_1(N) = \frac{1}{N}\sum_{i=1}^{N} x_i \quad \text{et} \quad M_2(N) = \frac{1}{N}\sum_{i=1}^{N} x_i^2.$$

Comme l'estimateur du maximum de vraisemblance, l'estimateur des moments possède d'intéressantes propriétés asymptotiques. Nous renvoyons le lecteur intéressé à Tassi [TAS 89].

1.4. Quelques méthodes d'estimation non paramétrique

1.4.1. *A la recherche d'un critère*

Avant de construire des estimateurs non paramétriques de la densité, d'en mesurer les performances théoriques et, le cas échéant, d'identifier le meilleur, il est nécessaire de spécifier un critère d'erreur qui puisse être éventuellement optimisé. Ce critère est souvent déterminé par l'application en vue.

1.4.1.1. *Dans le cas d'un problème de discrimination*

On dispose de réalisations $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_N, Y_N)$ de couples aléatoires indépendants de même loi que $(\boldsymbol{X}, Y)$ où $\boldsymbol{X}$ est le vecteur des observations et Y un élément de $\{0, 1, \ldots, M\}$ (la catégorie). Pour fixer les idées, revenons sur l'exemple introductif concernant les paramètres sanguins. Supposons que $\boldsymbol{X}$ prenne la forme d'un vecteur (C_1, C_2) dont les composantes aléatoires sont les concentrations de deux substances chimiques P_1 et P_2 dans le sang d'un individu et Y est l'indicateur d'une maladie ($Y = 1$ si l'individu a la maladie, $Y = 0$ s'il ne l'a pas). 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 qu'ils présentaient ou non la maladie.

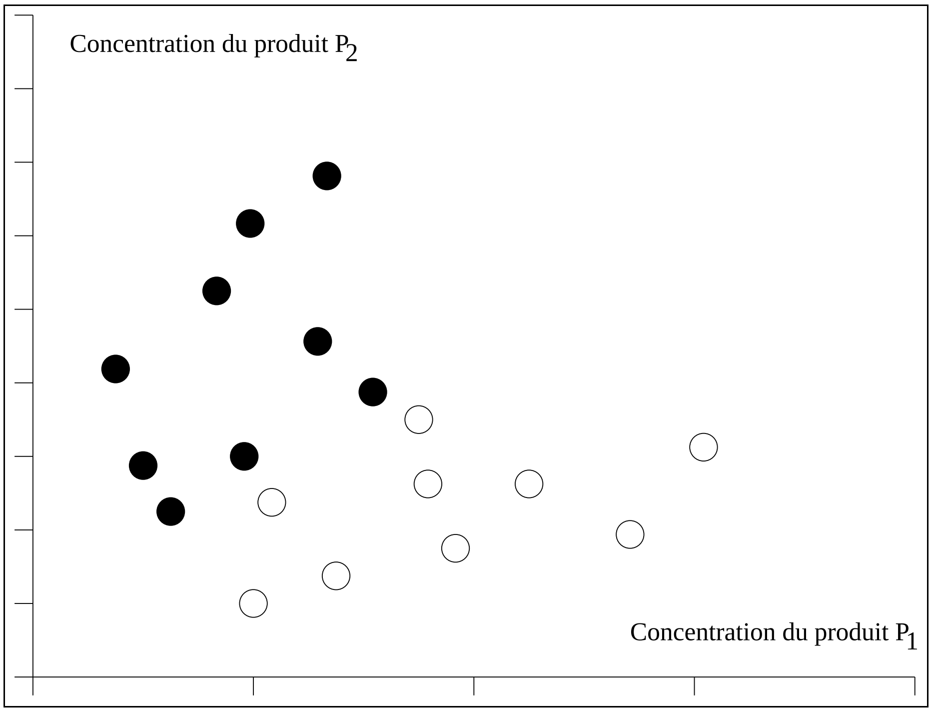

Figure 1.5. *Un exemple de discrimination à deux classes. Les individus malades sont figurés en blanc, les autres en noir.*

Sur présentation d'un nouveau couple $(\boldsymbol{X}', Y')$ où Y' est inconnu et $\boldsymbol{X}'$ 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 l'ensemble d'apprentissage (ensemble des couples (X_i, Y_i), $1 \leq i \leq n$ observés). Lorsque, conditionnellement à $Y = 0$ ou $Y = 1$, $\boldsymbol{X}$ a les densités respectives p_1 et p_2 alors la probabilité d'erreur est minimisée en décidant que $Y' = 1$ quand $p_1(X') > p_2(X')$ pourvu que $\mathbf{P}(Y = 1) = \mathbf{P}(Y = 0) = 1/2$, cas le plus simple. Puisque p_1 et p_2 sont inconnues, on peut estimer ces densités par p_N^1 et p_N^2 respectivement, et conjecturer que $Y' = 1$ lorsque $p_N^1(\boldsymbol{X}') > p_N^2(\boldsymbol{X}')$. 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. Lorsqu'on applique une règle majoritaire, comme on l'a vu auparavant, c'est l'estimation du classement des $\alpha_j p_j(\boldsymbol{x})$ qui est importante, plus que les valeurs elles-mêmes.

1.4.1.2. *Dans le cas général*

Comme on l'a rappelé dans le paragraphe 1.2.3 le rôle essentiel joué par une densité est de permettre le calcul de la probabilité de tout produit d'intervalles (ou pavé) par intégration sur ce pavé. Par conséquent un critère naturel d'erreur nous est fourni par

$$\sup_A \Big| \int_A p_N(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int_A p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \Big|,$$

où le supremum est pris sur l'ensemble des sous-ensembles A de $\mathbb{R}^d$ qui sont des réunions de pavés. C'est la borne supérieure des valeurs absolues des erreurs que l'on peut commettre en utilisant p_N au lieu de p pour calculer la probabilité d'une réunion de pavés. D'après un résultat de Scheffé [SCH 47], on a

$$\sup_A \Big| \int_A p_N(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} - \int_A p(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \Big| = \frac{1}{2} \int_{\mathbb{R}^d} |p_N(\boldsymbol{x}) - p(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x},$$

la dernière intégrale étant la distance L_1 entre p_N et p, la distance naturelle sur l'ensemble des densités. On en déduit donc que le supremum tend vers 0 selon un mode stochastique lorsque la distance L_1 entre p_N et p tend vers 0 selon le même mode. Ce critère L_1 est calculable sans aucune hypothèse additionnelle sur la densité et possède une signification claire en termes de différences de probabilités : si nous savons que $\int_{\mathbb{R}^d} |p_N(\boldsymbol{x}) - p(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x} \leq 0.004$, alors nous savons aussi que les différences entre les probabilités d'ensemble sont au plus de 0.002. Le théorème de Scheffé entraîne également que la distance L_1 reste invariante par transformation bijective de l'espace $\mathbb{R}^d$. Cette propriété peut être exploitée lorsque l'on visualise sur écran l'erreur L_1 qui est commise dans la queue de distribution pour une densité ou un estimateur à support infini. Pour ce faire, 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 être par conséquent rendues visibles. Enfin, l'erreur L_1 commise entre p_N et p peut être aisément visualisée : elle correspond à l'aire comprise entre les courbes représentatives des deux fonctions.

Les raisons évoquées ci-dessus font du critère L_1 un des critères privilégiés des statisticiens. Il convient néanmoins de souligner que bien d'autres mesures d'erreur ont été envisagées, tant du point de vue théorique que pratique. Chacune de ces mesures possède des qualités qui lui sont propres. Citons, pour mémoire

- Les distances L_p :

$$L_p(f,g) = \begin{cases} \left(\displaystyle\int_{\mathbb{R}^d} |f(\boldsymbol{x}) - g(\boldsymbol{x})|^p \,\mathrm{d}\boldsymbol{x} \right)^{1/p} & (0 < p < \infty) \\ \sup |f(\boldsymbol{x}) - g(\boldsymbol{x})| & (p = \infty) \end{cases}$$

- Les distances de Hellinger H_p :

$$H_p(f,g) = \left(\int_{\mathbb{R}^d} \left[f^{1/p}(\boldsymbol{x}) - f^{1/p}(\boldsymbol{x}) \right]^p \mathrm{d}\boldsymbol{x} \right)^{1/p} \quad (p > 0)$$

- L'information (ou divergence) de Kullback-Leibler D :

$$D(f,g) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} f(\boldsymbol{x}) \ln \left(\frac{f(\boldsymbol{x})}{g(\boldsymbol{x})} \right) \mathrm{d}\boldsymbol{x} & \text{si } f \ll g \\ \infty & \text{sinon.} \end{cases}$$

La notation $f \ll g$ utilisée ci-dessus signifie que $f(\boldsymbol{x})$ est nul chaque fois que $g(\boldsymbol{x})$ est nul. Par convention $\ln(0/0)$ est pris égal à 0.

- La divergence du χ^2 :

$$\chi^2(f,g) = \begin{cases} \displaystyle\int_{\mathbb{R}^d} \frac{\big(f(\boldsymbol{x}) - g(\boldsymbol{x})\big)^2}{g(\boldsymbol{x})} \mathrm{d}\boldsymbol{x} & \text{si } f \ll g \\ \infty & \text{sinon.} \end{cases}$$

Par convention $(0/0)$ est pris égal à 0.

Une étude exhaustive de tous ces critères nous entraînerait bien au delà des limites du présent chapitre. Pour plus de détails, nous renvoyons le lecteur intéressé à Devroye [DEV 87].

1.4.2. *Histogrammes et estimateurs basés sur des partitions*

Les moyens technologiques actuels permettent des protocoles de recueil de données mettant le statisticien devant une situation quelque peu paradoxale. D'une part les masses importantes de données l'autorisent à se placer dans des conditions "quasi-asymptotiques". D'autre part le besoin d'estimer des courbes ou des surfaces en temps réel l'oblige à utiliser des estimateurs dont la mise à jour se doit d'être extrêmement rapide. Le temps de calcul de certains estimateurs sophistiqués ayant des vitesses de convergence optimales devient ainsi rédhibitoire. Ces remarques ont motivé l'étude d'estimateurs basés sur des partitions, dont le calcul est très rapide, qui ont un bon comportement à distance

finie et qui atteignent de bonnes vitesses de convergence. Parmi eux on distingue les *densitogrammes* qui ont les propriétés des densités (ils ne prennent pas de valeurs négatives et ont une intégrale égale à un).
Les estimateurs *linéaires par morceaux*, étudiés en particulier par Scott [SCO 87], Jones, Samiuddin, Al-Harbey et Maatouk [JON 98], Beirlant, Berlinet et Györfi [BEI 99], Berlinet, Hobza et Vajda [BER 01a], Berlinet et Vajda [BER 01b] réalisent une bonne synthèse des qualités que nous venons de mentionner mais ne sont pas nécessairement des densitogrammes.
Comme il a été mentionné plus haut, le plus “rudimentaire” (et aussi le plus ancien) des *densitogrammes* est l'estimateur *histogramme.* Donnons-nous une suite de partitions $\mathcal{P}_N$ de $\mathbb{R}^d$ formées de cubes A_{Nj}, $j \in \mathbb{Z}$, de côté h_N, où h_N est un nombre réel strictement positif. L'histogramme construit à partir de l'échantillon $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ et de la partition $\mathcal{P}_N$ est alors défini, pour $\boldsymbol{x}$ dans la classe A_{Nj}, par

$$p_N(\boldsymbol{x}) = \frac{1}{N h_N^d} \sum_{i=1}^{N} \mathbf{1}_{A_{Nj}}(\boldsymbol{X}_i).$$

Cette dernière expression peut encore s'écrire

$$p_N(\boldsymbol{x}) = \frac{\mu_N(A_{Nj})}{h_N^d},$$

où μ_N représente la mesure empirique associée à l'échantillon $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, c'est-à-dire, pour tout sous-ensemble A de $\mathbb{R}^d$,

$$\mu_N(A) = \frac{\text{Cardinal}\Big\{i \in \{1, \ldots, N\} : \boldsymbol{X}_i \in A\Big\}}{N}.$$

Sous les deux conditions $h_N \to 0$ et $N h_N^d \to \infty$ quand N tend vers l'infini, on a, avec probabilité 1,

$$\int_{\mathbb{R}^d} |p_N(\boldsymbol{x}) - p(\boldsymbol{x})| \, \mathrm{d}\boldsymbol{x} \to 0 \quad \text{quand } N \to \infty.$$

Les autres estimateurs linéaires par morceaux sont des fonctions simples de l'histogramme. C'est par exemple le cas du polygone des fréquences, obtenu en joignant le milieu des classes de l'histogramme par des segments de droite. Notons cependant que le polygone des fréquences n'a pas en général une intégrale égale à un.

1.4.3. *Estimateur des plus proches voisins*

L'estimateur des plus proches voisins diffère de l'histogramme par sa méthode de construction. En effet, au lieu de fixer la taille de la cellule et de compter pour chaque cellule le nombre d'observations y appartenant, on se fixe au préalable le nombre d'observations par cellule et on construit les cellules à partir de cette information. Plus précisément, soit k_N une suite de nombres entiers positifs qui tend vers l'infini. L'estimateur des plus proches voisins p_N construit à partir d'un échantillon $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$ est alors défini par

$$p_N(\boldsymbol{x}) = \frac{k_N}{N \text{ mes}\big(B_N(\boldsymbol{x})\big)},$$

où $\text{mes}\big(B_N(\boldsymbol{x})\big)$ est la mesure de la plus petite boule fermée $B_N(\boldsymbol{x})$ de centre $\boldsymbol{x}$ contenant exactement k_N points de l'échantillon. Les études sur l'estimateur des plus proches voisins (voir en particulier Loftsgaarden et Quesenberry [LOF 65] ; Moore et Yackel [MOO 77]) se sont poursuivies récemment dans le cadre des processus. Ces études se situent dans un cadre local, estimation en un point, pour la bonne et simple raison que

$$\int_{\mathbb{R}^d} p_N(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} = \infty.$$

Autrement dit, cet estimateur n'est pas intégrable. En effet, pour $\|\boldsymbol{x}\|$ assez grande, $p_N(\boldsymbol{x})$ est, à une constante multiplicative près, de l'ordre de $1/\|\boldsymbol{x}\|$, qui n'est pas intégrable. En particulier, il n'est pas question d'en étudier la convergence pour la norme L_1. En revanche, on peut montrer, moyennant le fait que $k_N \to \infty$ et $k_N/N \to 0$ quand $N \to \infty$, que

$$\mathbf{E} \int_{\mathbb{R}^d} \big[p_N(\boldsymbol{x}) - p(\boldsymbol{x})\big]^2 \, \mathrm{d}\boldsymbol{x} \to 0,$$

à condition que p soit de carré intégrable et vérifie quelques conditions de régularité. Les méthodes de plus proches voisins sont très utilisées en théorie de la décision en raison de propriétés sur lesquelles nous reviendrons à la fin de ce chapitre.

1.4.4. *Estimateur à noyau*

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 répandu. L'idée naturelle sous-jacente consiste à évaluer la densité p au point $\boldsymbol{x}$ en comptant le nombre d'observations "tombées" dans un certain voisinage de $\boldsymbol{x}$. On peut par exemple choisir un voisinage cubique de $\boldsymbol{x} = (x_1, \ldots, x_d)$ de la forme

$$C\left(\boldsymbol{x}, \frac{h_N}{2}\right) = [x_1 - h_N/2, x_1 + h_N/2] \times \cdots \times [x_d - h_N/2, x_d + h_N/2]$$

où h_N est un nombre réel strictement positif dépendant de N, ce qui conduit à l'estimateur

$$p_N(\boldsymbol{x}) = \frac{\text{Cardinal}\Big\{ i \in \{1,\ldots,N\} : \boldsymbol{X}_i \in \; C\left(\boldsymbol{x}, \frac{h_N}{2}\right) \Big\}}{N h_N^d}.$$

Cette dernière expression peut encore s'écrire

$$p_N(\boldsymbol{x}) = \frac{1}{N h_N^d} \sum_{i=1}^{N} \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]^d} \left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h_N} \right), \tag{1.5}$$

où la fonction $\mathbf{1}_{[-1/2,1/2]^d}$ est la densité de probabilité uniforme sur le cube $[-1/2, 1/2]^d$. En s'inspirant alors de la formule (1.5), et en prenant à la place de la fonction $\mathbf{1}_{[-1/2,1/2]^d}$ une fonction K réelle, bornée, d'intégrale 1 sur $\mathbb{R}^d$, on définit *l'estimateur p_N associé au noyau K* par

$$p_N(\boldsymbol{x}) = \frac{1}{N h_N^d} \sum_{i=1}^{N} K \left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h_N} \right). \tag{1.6}$$

Cette formule est illustrée par la figure 1.6 qui montre que la densité p_N prise comme estimateur est celle du mélange des mesures de densités

$$\frac{1}{h_N^d} K \left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h_N} \right), \qquad 1 \leq i \leq N,$$

chacune étant affectée du même coefficient $1/N$. On a transformé les masses ponctuelles de poids $1/N$ sur chaque donnée X_i en une mesure continue de même masse.

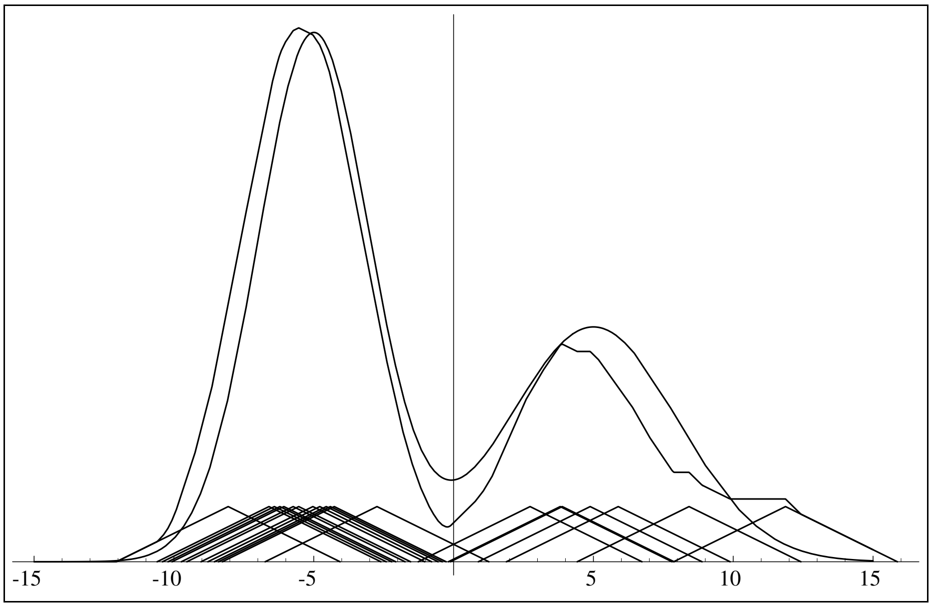

Figure 1.6. *Un estimateur à noyau triangulaire pour un échantillon simulé suivant un mélange de lois gaussiennes. L'estimateur est la somme des contributions triangulaires locales.*

Lorsque le noyau K est choisi positif, l'estimateur p_N a toutes les propriétés d'une densité de probabilité et on parle alors parfois de *la densité de probabilité empirique de noyau* K. 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 (1.6) de $p_N(\boldsymbol{x})$ est extrêmement simple, puisque $p_N(\boldsymbol{x})$ est la somme de N variables aléatoires dont chacune est transformée par le noyau d'une seule des observations. En particulier, lorsque les variables observées sont indépendantes, $p_N(\boldsymbol{x})$ est la somme de N variables aléatoires indépendantes et identiquement distribuées ;
- Ensuite, p_N converge vers p en de nombreux sens, et en particulier au sens L_1 pour toute densité p dès que $1/h_N$ et Nh_N^d tendent tous les deux vers l'infini. 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 [DEV 85]) ;
- 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_N. Le choix de K, qui peut être soumis à certains critères d'optimalité (Berlinet [BER 93]), permet de donner les propriétés que l'on veut à l'estimateur (continuité, dériva-

bilité). Celui de la fenêtre de lissage h_N est abordé dans le paragraphe suivant. Son influence sur le lissage des données est illustrée par les figures 1.7 et 1.8. Lorsque le paramètre h_N est choisi trop petit, des artefacts apparaissent sur l'estimateur. Au contraire lorsqu'il est choisi trop grand les caractéristiques de la vraie densité sous-jacente sont gommées (pics et creux sont amoindris).

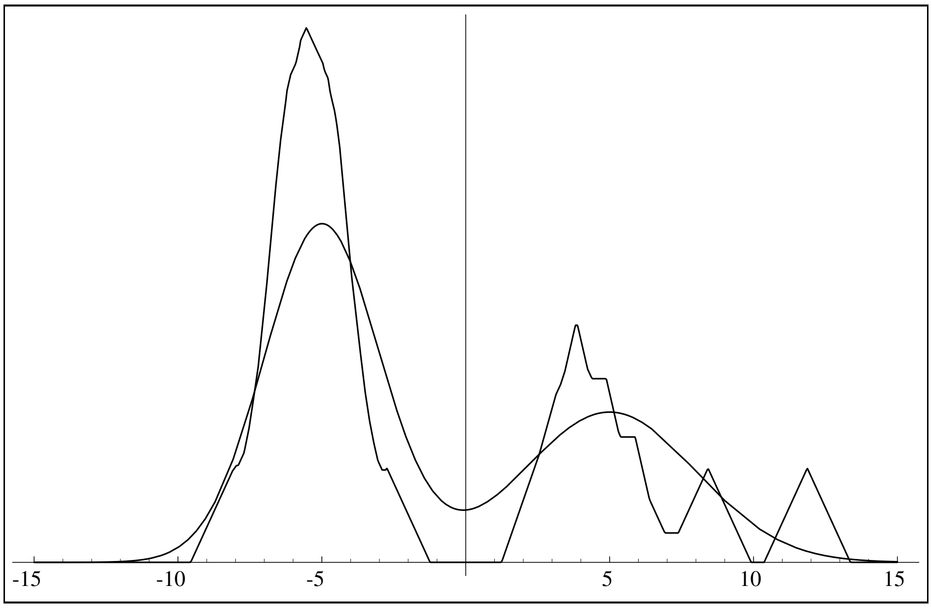

Figure 1.7. *Un exemple de sous-lissage pour un estimateur à noyau triangulaire. La fenêtre h_N est trop petite. En chaque point le biais est faible mais la variance importante.*

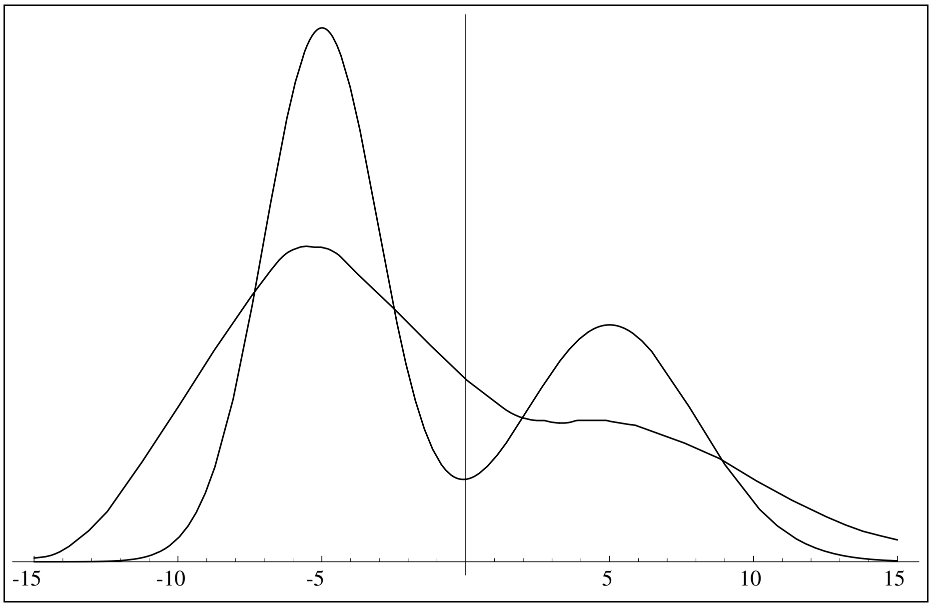

Figure 1.8. *Un exemple de sur-lissage pour un estimateur à noyau triangulaire. La fenêtre h_N est trop grande. En chaque point la variance est faible mais le biais est important.*

1.5. Choix du paramètre de lissage pour l'estimateur à noyau

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_N se révèle crucial aussi bien pour la précision locale que pour la précision globale de l'estimateur p_N. Il est facile de vérifier que, pour les noyaux usuels et un ensemble de données fixé, la loi de densité p_N converge (étroitement) vers la mesure empirique lorsque h_N tend vers 0 et que p_N tend uniformément vers la fonction nulle lorsque h_N tend vers l'infini. En jouant sur la largeur de fenêtre, on peut donc faire décrire à p_N 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 $h(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ 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 [MAR 88], Turlach [TUR 93], Berlinet et Devroye [BER 94], Cao, Cuevas et González-Manteiga [CAO 94], Devroye [DEV 97] ou encore Biau [BIA 99].
Une première approche classique pour déterminer un h_N "optimal" consiste à analyser le comportement asymptotique de l'*Erreur Quadratique Intégrée Moyenne*, que nous noterons dans les formules par son abréviation anglo-saxonne MISE, et qui est définie par

$$\mathrm{MISE}(h_N) = \mathbf{E}\int_{\mathbb{R}^d}\left[p_{h_N}(\boldsymbol{x}) - p(\boldsymbol{x})\right]^2\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}^d}\mathbf{E}\left[p_{h_N}(\boldsymbol{x}) - p(\boldsymbol{x})\right]^2\,\mathrm{d}\boldsymbol{x},$$

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. Moyennant quelques conditions de régularité sur p et K, on déduit (Bosq et Lecoutre [BOS 87]) que le paramètre de lissage h_N^* qui minimise l'Erreur Quadratique Intégrée Moyenne Asymptotique en dimension $d = 1$ s'écrit

$$h_N^* = \alpha(K)\beta(p)N^{-1/5}, \tag{1.7}$$

où

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

où nous choisissons de noter, pour toute fonction g,

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

et où σ_K^2 désigne la variance du noyau K. Notons que h_N^* est une quantité *déterministe* (qui dépend du nombre d'observations N). La largeur de fenêtre optimale h_N^* dépend de la densité inconnue p au travers du paramètre $R(p'')$ 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 $R(p'')$ dans l'expression (1.7) par un estimateur approprié. Les références sur ce sujet ne manquent pas : voir par exemple Hall et Marron [HAL 87a], [HAL 87b], [HAL 87c], [HAL 91b] ou encore Jones et Sheather [JON 91]. Cette approche conduit à un ensemble de méthodes que l'on a coutume de regrouper sous le vocable général de *méthodes plug-in*[3], et qui font l'objet d'une recherche active

[3]Le terme "plug-in" est difficile à traduire (on peut parler de *méthode en deux étapes* ou de *méthode de remplacement*).

(pour des références plus anciennes, consulter Woodroofe [WOO 70] ; Deheuvels [DEH 74] ; Nadaraya [NAD 74] ; Deheuvels [DEH 77] ou encore Deheuvels et Hominal [DEH 80]). Une méthode couramment employée consiste à utiliser un premier estimateur paramétrique pour $R(p'')$ puis à injecter l'estimateur obtenu dans la formule donnant h_N^*. On peut éventuellement itérer (Biau [BIA 99]).
Parallèlement aux algorithmes de sélection de type plug-in, d'autres méthodes se sont également révélées performantes. Nous allons les discuter rapidement. Cette fin de paragraphe un peu technique sur le choix du paramètre de lissage pourra être passée en première lecture.
Habbema, Hermans et Vandenbroek [HAB 74] et Duin [DUI 76] ont proposé une méthode fondée sur un critère non-asymptotique de *maximum de vraisemblance*. Brièvement, il s'agit de maximiser par rapport à h_N la quantité $L(h_N)$ suivante :

$$L(h_N) = \prod_{k=1}^{N} p_{h_N,k}(X_k),$$

où

$$p_{h_N,k}(x) = \frac{1}{(N-1)h_N} \sum_{\substack{1\leq i\leq N \\ i\neq k}} K\left(\frac{x - X_i}{h_N}\right)$$

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 [HAL 82] ; Rudemo [RUD 82] ; Chow, Geman et Wu [CHO 83] ; Devroye et Györfi [DEV 85] ; Marron [MAR 85]) 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 [SCH 81]).
Une méthode dite de *validation croisée* dans L_2 (parfois appelée *validation croisée non-biaisée*) a été proposée par Rudemo [RUD 82] et Bowman [BOW 84]. Le critère consiste ici à choisir le paramètre de lissage qui minimise un estimateur convenable de

$$\int_{\mathbb{R}} \left[p_{h_N}(\boldsymbol{x}) - p(\boldsymbol{x})\right]^2 \mathrm{d}\boldsymbol{x} - \int_{\mathbb{R}} p^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int_{\mathbb{R}} p_{h_N}^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - 2\int_{\mathbb{R}} p_{h_N}(\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x},$$

par exemple

$$\int_{\mathbb{R}} p_{h_N}^2(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} - \frac{2}{N}\sum_{k=1}^{N} p_{h_N,k}(\boldsymbol{X}_k),$$

où

$$p_{h_N,k}(\boldsymbol{x}) = \frac{1}{(N-1)h_N} \sum_{\substack{1\leq i\leq N \\ i\neq k}} K\left(\frac{\boldsymbol{x} - \boldsymbol{X}_i}{h_N}\right).$$

L'optimalité asymptotique de la validation croisée L_2 a été obtenue par Stone [STO 84]. Pour d'autres études, voir également Hall [HAL 82], [HAL 85], Burman [BUR 85], Scott et Terrell [SCO 87]. 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 [HAL 91a]) ; 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 [HAL 87b], [HAL 87c] ; Hall, Marron et Park [HAL 92]). Un critère de validation croisée biaisée, introduit par Scott et Terrell [SCO 87] pour remédier aux défauts de la validation croisée non-biaisée, amène à sélectionner la largeur de fenêtre qui minimise l'estimateur $\widehat{\mathrm{AMISE}}(h_N)$ de l'Erreur Quadratique Intégrée Moyenne Asymptotique défini comme suit :

$$\widehat{\mathrm{AMISE}}(h_N) = \frac{h_N^4}{4}\sigma_K^4\tilde{R}(p'') + \frac{R(K)}{Nh_N},$$

où $\tilde{R}(p'')$ est un estimateur de $R(p'')$.
Des méthodes prometteuses de nature combinatoire ont été récemment explorées. Nous renvoyons le lecteur intéressé à l'ouvrage de Devroye et Lugosi [DEV 01].

1.6. De nouveaux outils

Au confluent de la théorie de l'information et de l'estimation fonctionnelle, les travaux de Barron [BAR 88], Barron, Györfi et van der Meulen [BAR 92], Berlinet, Györfi et van der Meulen [BER 97], Berlinet, Vajda et van der Meulen [BER 98] et Györfi, Liese, Vajda et van der Meulen [GYO 98] ont montré l'intérêt de posséder des estimateurs de la densité convergeant pour des topologies plus fortes que celle définie par la norme L_1, en particulier au sens de l'*information de Kullback-Leibler* ou de la *divergence du* χ^2 (dont les définitions sont données ci-dessus au paragraphe 1.4.1.2). Barron *et al* [BAR 92] ont montré qu'il était possible de construire des estimateurs de la densité convergeant en information de Kullback-Leibler et en information de Kullback-Leibler moyenne, c'est-à-dire tels que $D(p, p_N)$ tende vers 0 (presque sûrement) et tels que $\mathbf{E}D(p, p_N)$ tende vers 0, à condition qu'il existe une densité de probabilité g (dite *de référence*) telle que $D(p, g) < \infty$. Sous la condition plus contraignante $\chi^2(p, g) < \infty$, Györfi *et al* ont étendu en [GYO 98] les résultats de Barron *et al* [BAR 92] au cas de la divergence du χ^2. La famille d'estimateurs p_N introduite par Barron *et al* [BAR 92] est définie à partir des N premiers éléments d'une suite $\{\boldsymbol{X}_i\}_{i\geq 1}$ d'observations indépendantes distribuées suivant la loi de p de la manière suivante :

- Soit $(m_N)_{N\geq 1}$ une suite d'entiers tels que $1 \leq m_N \leq N$, soit $h_N = 1/m_N$ et soit $(a_N)_{N\geq 1}$ une suite de nombres réels dans $]0,1[$;
- Soit g la densité de référence (connue) associée à la loi de probabilité ν ;
- Soit $(\mathcal{P}_N)_{N\geq 1} = \{A_{N1}, A_{N2}, \ldots, A_{Nm_N}\}_{N\geq 1}$ une suite de partitions de $\mathbb{R}^d$ telles que $\nu(A_{Ni}) = h_N$, $i = 1, \ldots, m_N$;

Alors, en notant, pour $\boldsymbol{x} \in \mathbb{R}^d$, $A_N(\boldsymbol{x})$ l'élément de $\mathcal{P}_N$ qui contient $\boldsymbol{x}$ et μ_N la mesure empirique construite à partir de l'échantillon $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N$, on pose

$$p_N(\boldsymbol{x}) = \left[(1-a_N)\frac{\mu_N\big(A_N(\boldsymbol{x})\big)}{h_N} + a_N\right] g(\boldsymbol{x}). \tag{1.8}$$

L'estimateur (1.8) est appelé *histogramme modifié* (*de Barron*). Ses bonnes propriétés de convergence en information de Kullback-Leibler et en divergence du χ^2 font de lui un outil d'avenir dans de nombreux domaines d'application de la statistique (compression de données, réseaux de télécommunications, problèmes de classification, réseaux de neurones, etc).
Les figures 6 à 11 montrent sur données simulées quelques-uns des estimateurs de la densité évoqués ici. La loi sous-jacente est la loi exponentielle de paramètre 1 pour les figures 6, 8, 10 et la loi normale $\mathcal{N}(0,1)$ pour les figures 7, 9 et 11 ; dans chaque cas 200 données ont été simulées. Ces quelques dessins n'ont évidemment pas de valeur statistique sur le comportement des estimateurs. Ils mettent néanmoins en évidence quelques-uns de leurs traits caractéristiques. L'estimateur PLH (Piecewise Linear Histogram) est obtenu en interpolant l'histogramme aux milieux de deux classes consécutives. Il est à noter qu'il atteint la vitesse de convergence L_1 de l'estimateur à noyau.
Nous signalons le site Internet

`http ://siprint.utia.cas.cz/densest/`

où il est possible de faire des essais en ligne sur données simulées.

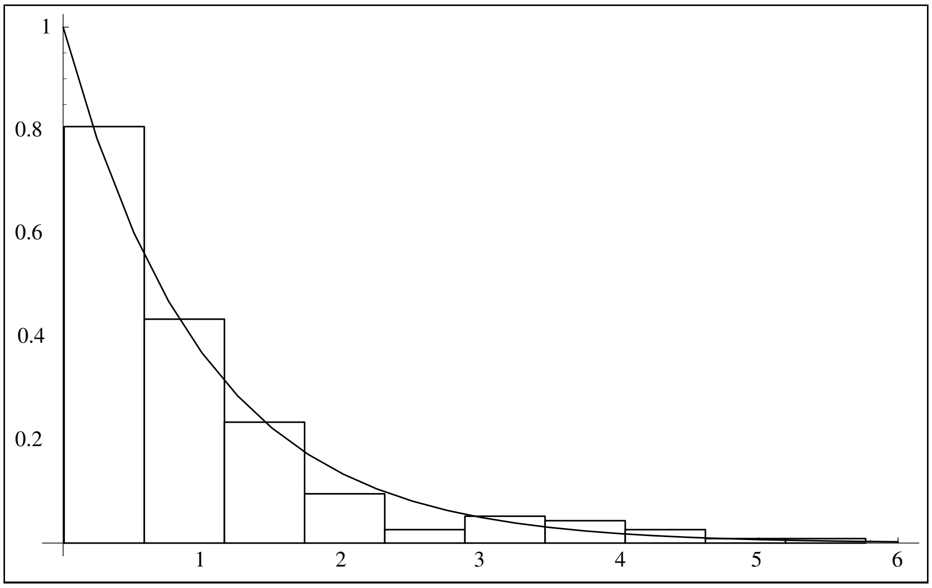

Figure 1.9. *Histogramme et densité exponentielle de paramètre* 1.

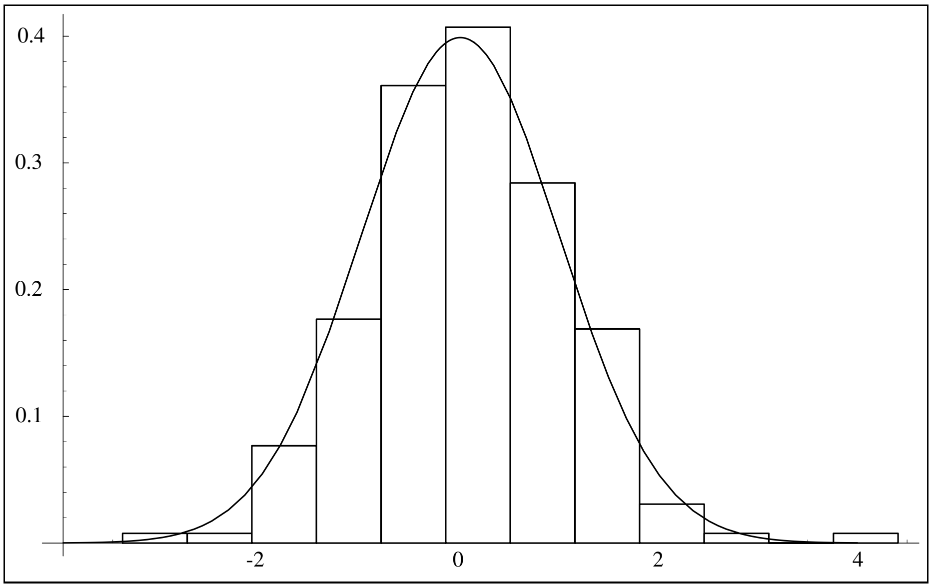

Figure 1.10. *Histogramme et densité normale* $\mathcal{N}(0,1)$.

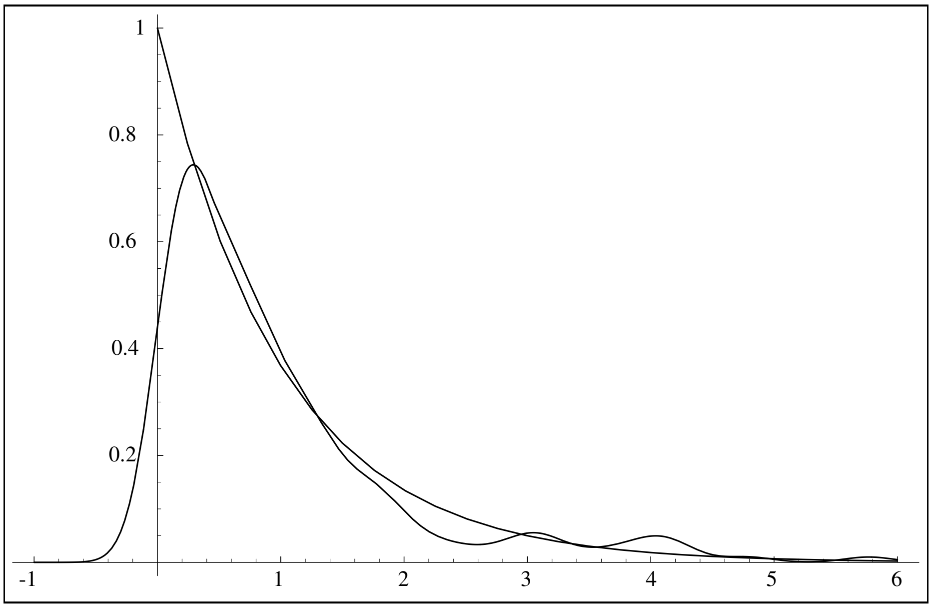

Figure 1.11. *Estimateur à noyau et densité exponentielle de paramètre* 1.

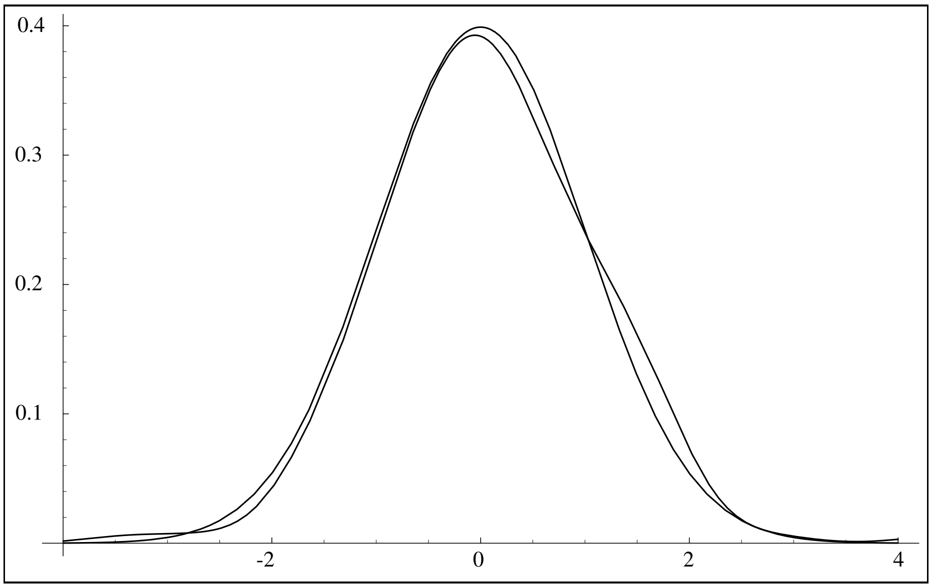

Figure 1.12. *Estimateur à noyau et densité normale* $\mathcal{N}(0, 1)$.

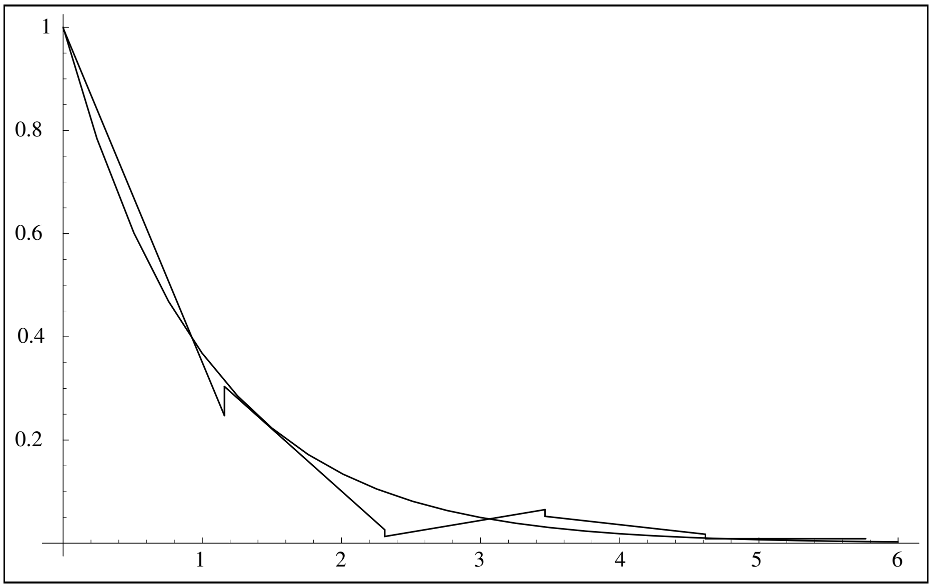

Figure 1.13. *Estimateur PLH et densité exponentielle de paramètre* 1.

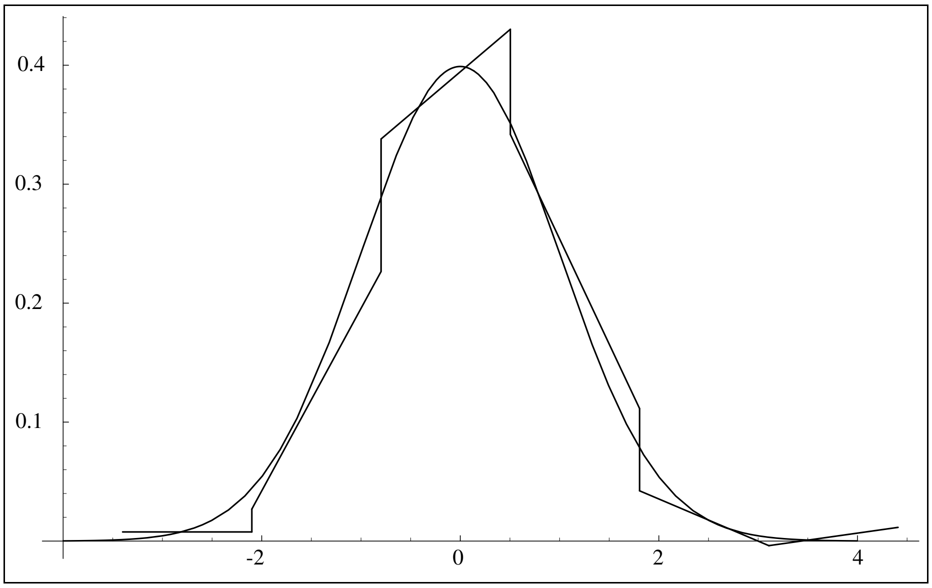

Figure 1.14. *Estimateur PLH et densité normale* $\mathcal{N}(0,1)$.

1.7. Comportement asymptotique des règles de décision

Pour simplifier et alléger les notations, nous considérerons dans la suite un couple de variables aléatoires $(\boldsymbol{X}, Y)$ prenant ses valeurs dans $\mathbb{R}^d \times \{0,1\}$. Autrement dit, l'étiquette Y du vecteur aléatoire $\boldsymbol{X}$ de $\mathbb{R}^d$ peut prendre deux valeurs : 0 et 1. Une règle de décision sera alors une application g définie sur $\mathbb{R}^d$, à valeurs dans $\{0,1\}$, telle que $g(\boldsymbol{X})$ soit une variable aléatoire, étiquette associée à $\boldsymbol{X}$. La probabilité d'erreur associée à cette fonction g est définie par

$$L(g) = \mathbf{P}(g(\boldsymbol{X}) \neq Y).$$

C'est la probabilité de se tromper d'étiquette en se basant sur la règle de décision g. Dans la pratique, cette règle sera bâtie à partir d'observations indépendantes $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_N, Y_N)$ du vecteur $(\boldsymbol{X}, Y)$. Elle permettra ensuite d'attribuer l'étiquette $g(\boldsymbol{x})$ à une nouvelle valeur observée $\boldsymbol{x}$ de $\boldsymbol{X}$. L'intuition nous mène naturellement vers la règle décisionnelle majoritaire évoquée plus haut. Cette démarche est confortée par l'existence d'une règle de décision théorique meilleure que toutes les autres au sens de la probabilité d'erreur (1.7) : la règle de Bayes g^*. Cette règle est définie à partir des probabilités conditionnelles de Y sachant $\boldsymbol{X}$ de la manière suivante :

$$g^*(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad \mathbf{P}(Y=1|\boldsymbol{X}=\boldsymbol{x}) > \mathbf{P}(Y=0|\boldsymbol{X}=\boldsymbol{x}) \\ 0 & \text{sinon.} \end{cases}$$

Elle est optimale relativement à la probabilité d'erreur (Devroye, Györfi et Lugosi [DEV 96]) :

$$L^* = \mathbf{P}(g^*(\boldsymbol{X}) \neq Y) = \min_g \mathbf{P}(g(\boldsymbol{X}) \neq Y),$$

où le minimum est pris sur l'ensemble de toutes les fonctions de décision g possibles. Nous n'avons bien sûr pas accès à cette fonction puisque les probabilités conditionnelles sont inconnues. Les règles majoritaires construites à partir des observations consistent donc à estimer quelle est la plus grande des deux inconnues $\mathbf{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})$ et $\mathbf{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})$. Le champ des investigations et expérimentations est largement ouvert car on peut démontrer qu'il n'existe pas de règle optimale universelle basée sur les données : pour toute règle g, il existe une règle g' qui fera mieux que g pour une certaine distribution du couple $(\boldsymbol{X}, Y)$, au sens où on aura

$$\mathbf{E}\big(L(g')\big) \leq \mathbf{E}\big(L(g)\big).$$

En la matière, il faut donc se méfier des expérimentations numériques sur jeux de données. En revanche, il est important de donner des informations sur le comportement de $L(g)$ par rapport à l'optimum L^*, d'une part lorsque le nombre d'observations est fixé, d'autre part lorsqu'on le fait tendre vers l'infini. On dira qu'une règle de décision est convergente si sa probabilité d'erreur tend vers l'optimum L^* lorsque le nombre de données augmente indéfiniment. Définissons plus précisément cette notion. Supposons donnée une suite $(\boldsymbol{X}_i, Y_i)$, $i \geq 1$, d'observations indépendantes du couple $(\boldsymbol{X}, Y)$. Soit, pour tout N, une règle de décision g_N basée sur les N premières observations de la suite. La probabilité d'erreur associée à g_N est donnée par la probabilité conditionnelle

$$L(g_N) = \mathbf{P}(g_N(\boldsymbol{X}) \neq Y \mid \boldsymbol{X}_1, Y_1, \ldots, \boldsymbol{X}_N, Y_N)$$

et son espérance est donnée par

$$\mathbf{E}L(g_N) = \mathbf{P}(g_N(\boldsymbol{X}) \neq Y).$$

On dit que la suite de règles de décision (g_N) est *convergente*, ou *faiblement convergente*, si

$$\mathbf{E}L(g_N) \to L^* \quad \text{lorsque} \quad N \to \infty.$$

On dit qu'elle est *fortement convergente* si, avec probabilité 1,

$$L(g_N) \to L^* \quad \text{lorsque} \quad N \to \infty.$$

En général, on parle plus brièvement de “règle convergente”. Ces notions de convergence, qui sont souvent équivalentes, par exemple pour la règle des plus proches voisins, sont définies pour une distribution donnée du couple $(\boldsymbol{X}, Y)$. Lorsque la convergence a lieu pour toute distribution de $(\boldsymbol{X}, Y)$ on la qualifie d'*universelle.* Nous allons passer en revue quelques propriétés des principales règles utilisées. Pour plus de détails, nous renvoyons à Devroye *et al* [DEV 96]. Etant donné un estimateur de la densité de $\boldsymbol{X}$ de la forme

$$p_N(\boldsymbol{x}) = \sum_{i=1}^{N} K_N(\boldsymbol{x}, \boldsymbol{X}_i), \tag{1.9}$$

la règle majoritaire prend la forme suivante :

$$g_N(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^N \mathbf{1}_{\{1\}}(Y_i) K_N(\boldsymbol{x}, \boldsymbol{X}_i) > \sum_{i=1}^N \mathbf{1}_{\{0\}}(Y_i) K_N(\boldsymbol{x}, \boldsymbol{X}_i) \\ 0 & \text{sinon.} \end{cases}$$

Si nous autorisons la fonction K_N à dépendre de l'échantillon $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$, nous allons pouvoir englober sous la forme (1.9) les estimateurs des plus proches voisins, les histogrammes automatiques (dont les classes sont construites à partir des données) et d'autres estimateurs de la densité dont les paramètres de lissage pourront dépendre de l'échantillon ou du point d'estimation. Particularisons aux histogrammes et aux estimateurs à noyau.

- **Règle de l'histogramme**

Si on note $A_N(\boldsymbol{x}) = A_N(\boldsymbol{x}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ la classe de la partition qui contient $\boldsymbol{x}$, on peut écrire l'histogramme p_N sous la forme (1.9) avec

$$K_N(\boldsymbol{x}, \boldsymbol{y}) = \frac{\mathbf{1}_{A_N(\boldsymbol{x})}(\boldsymbol{y})}{N\mathrm{mes}\big(A_N(\boldsymbol{x})\big)},$$

et la règle associée prend la forme

$$g_N(\boldsymbol{x}) = \begin{cases} 1 & \text{si } \sum_{i=1}^N \mathbf{1}_{\{1\}}(Y_i)\mathbf{1}_{A_N(\boldsymbol{x})}(\boldsymbol{X}_i) > \sum_{i=1}^N \mathbf{1}_{\{0\}}(Y_i)\mathbf{1}_{A_N(\boldsymbol{x})}(\boldsymbol{X}_i) \\ 0 & \text{sinon.} \end{cases}$$

La décision est prise en fonction de la majorité sur la classe $A_N(\boldsymbol{x})$.

- **Règle du noyau**

Pour un noyau K (que l'on pourrait éventuellement choisir également en fonction des données) et une fenêtre locale $h_N = h_N(\boldsymbol{x}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$, on pose

$$K_N(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{Nh_N^d} K\left(\frac{\boldsymbol{x} - \boldsymbol{y}}{h_N}\right),$$

et la règle associée prend la forme

$$g_N(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^N \mathbf{1}_{\{1\}}(Y_i) K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h_N}\right) > \sum_{i=1}^N \mathbf{1}_{\{0\}}(Y_i) K\left(\frac{\boldsymbol{x}-\boldsymbol{X}_i}{h_N}\right) \\ 0 & \text{sinon.} \end{cases}$$

Pour la règle de l'histogramme à cellules cubiques non aléatoires de côté h_N, on peut montrer la convergence forte universelle dès que les quantités Nh_N^d et $(1/h_N)$ tendent vers l'infini avec N. On peut également prouver une inégalité exponentielle asymptotique pour la probabilité d'erreur (Devroye *et al* [DEV 96]). Pour N assez grand on a

$$\mathbf{P}(L(g_N) - L^* > e) \leq 2\exp\left(-ne^2/32\right).$$

Des résultats du même type peuvent être obtenus pour la règle du noyau.

- **Règle des plus proches voisins**

Ce n'est pas dans une cellule fixée a priori autour de $\boldsymbol{x}$ que la majorité est observée mais dans une cellule centrée en $\boldsymbol{x}$ qui contient un nombre k_N fixé d'observations, k_N pouvant dépendre de $\boldsymbol{x}$ et de $(\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$. L'expression de K_N et la règle associée prennent la même forme que pour l'histogramme :

$$g_N(\boldsymbol{x}) = \begin{cases} 1 & \text{si} \quad \sum_{i=1}^N \mathbf{1}_{\{1\}}(Y_i)\mathbf{1}_{B_N(\boldsymbol{x})}(\boldsymbol{X}_i) > \sum_{i=1}^N \mathbf{1}_{\{0\}}(Y_i)\mathbf{1}_{B_N(\boldsymbol{x})}(\boldsymbol{X}_i) \\ 0 & \text{sinon,} \end{cases}$$

avec cette fois un ensemble $B_N(\boldsymbol{x})$ aléatoire qui est la plus petite boule fermée de centre $\boldsymbol{x}$ qui contient au moins k_N observations. C'est pour cette règle que le premier résultat de convergence universelle a été prouvé par Stone en 1977 [STO 77] dans le cas où la suite (k_N) est déterministe : sous la condition que k_N et N/k_N tendent vers l'infini avec N, la règle des k_N plus proches voisins est faiblement universellement convergente. Sous les mêmes conditions, la convergence forte universelle peut être prouvée lorsque $\boldsymbol{X}$ admet une densité de probabilité [DEV 96]. Dans le cas contraire, on doit définir une stratégie adéquate pour le traitement des ex æquo.
Cette règle et ses variantes ont suscité une abondante littérature. Nous nous contenterons ici de citer une importante inégalité universelle due à Cover et Hart [COV 67] dans le cas de la règle DU plus proche voisin (k_N est fixé égal à 1) :

$$\limsup_{N\to\infty} \mathbf{E}L(g_N) \leq 2L^*.$$

Bien que particulièrement simple, la règle du plus proche voisin fait mieux, pour certaines distributions du couple $(\boldsymbol{X}, Y)$, que toutes les règles des k plus proches voisins ($k > 1$). On pourra se reporter à [DEV 96] pour une discussion sur le nombre de voisins à prendre en compte pratiquement lorsque la probabilité d'erreur optimale L^* est petite.

1.8. Quelques prolongements

Beaucoup d'autres estimateurs de la densité peuvent être envisagés pour répondre à des problèmes de classification. Ils sont associés à différentes méthodes de régularisation de la mesure empirique : polynômes locaux, noyaux d'ordre supérieur, splines, fonctions orthogonales comme les fonctions trigonométriques ou les ondelettes, etc. Lorsque les données sont nombreuses, les problèmes de calculabilité sur ordinateur deviennent prépondérants et certains estimateurs présentés ci-dessus ont le mérite de marier des propriétés d'optimalité à la fois en vitesse de convergence et en temps de calcul. Le domaine commun à l'estimation de densité et à la théorie de la décision est appelé à connaître un important développement dans les prochaines années. Beaucoup de questions se posent en effet en ce qui concerne la comparaison à distance finie d'estimateurs ou de règles de décisions, l'évaluation et la prise en compte de la dépendance des données recueillies et la prise de décision à partir d'enregistrements continus, les étiquettes étant alors associées à des courbes, des surfaces ou des objets plus complexes.

Bibliographie

[BAR 88] A.R. BARRON, The convergence in information of probability density estimators, in *Proceedings of the International Symposium of IEEE on Information Theory*, Kobe, Japan, June 19–24 1988.

[BAR 92] A.R. BARRON, L. GYÖRFI ET E.C. VAN DER MEULEN, Distribution estimation consistent in total variation and in two types of information divergence, *IEEE Transactions on Information Theory*, **Vol. 38**, pp. 1437–1454 1992.

[BEI 99] J. BEIRLANT, A. BERLINET ET L. GYÖRFI, On piecewise linear density estimates, *Statistica Neerlandica*, **Vol. 53**, pp. 287–308 1999.

[BER 89] A. BERLINET ET L. DEVROYE, Estimation d'une densité : un point sur la méthode du noyau, *Statistique et Analyse des Données*, **Vol. 14**, pp. 1–32 1989.

[BER 93] A. BERLINET, Hierarchies of Higher Order Kernels, *Probability Theory and Related Fields*, **Vol. 94**, pp. 489–504 1993.

[BER 94] A. BERLINET ET L. DEVROYE, A comparison of kernel density estimates, *Publications de l'Institut de Statistique de l'Université de Paris*, **Vol. 38**, pp. 3–59 1994.

[BER 97] A. BERLINET, L. GYÖRFI ET E.C. VAN DER MEULEN, Asymptotic normality of relative entropy in multivariate density estimation, *Publications de l'Institut de Statistique de l'Université de Paris*, **Vol. 41**, pp. 3–27 1997.

[BER 01a] A. BERLINET, T. HOBZA ET I. VAJDA, Generalized piecewise linear histograms, *Statistica Neerlandica*, **Vol. 55**, pp. 1–13 2001.

[BER 01b] A. BERLINET ET I. VAJDA, Nonnegative piecewise linear histograms, *Statistics*, à paraître 2001.

[BER 98] A. BERLINET, I. VAJDA ET E.C. VAN DER MEULEN, About the asymptotic accuracy of Barron density estimates, *IEEE Transactions on Information Theory*, **Vol. 44**, pp. 999–1009 1998.

[BIA 99] G. BIAU, Estimateurs à noyau itérés : synthèse bibliographique, *Journal de la Société Française de Statistique*, **Vol. 1**, pp. 41–67 1999.

[BOS 87] D. BOSQ ET J.P. LECOUTRE, *Théorie de l'Estimation Fonctionnelle*, Economica, Paris 1987.

[BUR 85] P. BURMAN, A data dependent approach to density estimation, *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, **Vol. 69**, pp. 609–628 1985.

[BOW 84] A.W. BOWMAN, An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, **Vol. 71**, pp. 353–360 1984.

[CAO 94] R. Cao, A. Cuevas et W. González-Manteiga, A comparative study of several smoothing methods in density estimation, *Computational Statistics and Data Analysis*, **Vol. 17**, pp. 153–176 1994.

[CHO 83] Y.S. Chow, S. Geman et L.D. Wu, Consistent cross-validated density estimation, *The Annals of Statistics*, **Vol. 11**, pp. 25–38 1983.

[COV 67] T. Cover et P. Hart, Nearest neighbor pattern classification, *IEEE Transactions on Information Theory*, **Vol. 13**, pp. 21–27 1967.

[DAU 99] J.J. Daudin, S. Robin et C. Vuillet, *Statistique Inférentielle – Idées, Démarches, Exemples*, Presses Universitaires de Rennes, Rennes 1999.

[DEH 74] P. Deheuvels, Conditions nécessaires et suffisantes de convergence ponctuelle presque sûre et uniforme presque sûre des estimateurs de la densité, *Comptes Rendus Mathématiques de l'Académie des Sciences de Paris*, **Vol. 278**, pp. 1217–1220 1974.

[DEH 77] P. Deheuvels, Estimation non paramétrique de la densité par histogrammes généralisés, *Revue de Statistique Appliquée*, **Vol. 25**, pp. 5–42 1977.

[DEH 80] P. Deheuvels et P. Hominal, Estimation automatique de la densité, *Revue de Statistique Appliquée*, **Vol. 28**, pp. 25–55 1980.

[DEL 83] M. Delecroix, *Histogrammes et estimation de la densité*, Que sais-je 2055, Presses Universitaires de France, Paris 1983.

[DEV 85] L. Devroye et L. Györfi, *Nonparametric Density Estimation : the L_1 View*, Wiley, New York 1985.

[DEV 87] L. Devroye, *A Course in Density Estimation*, Birkhäuser, Boston 1987.

[DEV 96] L. Devroye, L. Györfi et G. Lugosi, *A Probabilistic Theory of Pattern Recognition*, Springer–Verlag, New York 1996.

[DEV 97] L. Devroye, Universal smoothing factor selection in density estimation : theory and practice, *Test*, **Vol. 6**, pp. 223–320 1997.

[DEV 01] L. Devroye et G. Lugosi, *Combinatorial Methods in Density Estimation*, Springer, New York 2001.

[DRE 01] F. Dress et M. Mendès-France, La suite des puissances de 3/2, *La Recherche*, **Vol. 346**, pp. 34–37 2001.

[DUI 76] R.P.W. Duin, On the choice of smoothing parameters for Parzen estimators of probability density functions, *IEEE Transactions on Computers*, **Vol. 25**, pp. 1175–1179 1976.

[GYO 98] L. Györfi, F. Liese, I. Vajda et E.C. van der Meulen, Distribution estimates consistent in χ^2-divergence, *Statistics*, **Vol. 32**, pp. 31–57 1998.

[HAB 74] 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.

[HAL 82] P. Hall, Cross-validation in density estimation, *Biometrika*, **Vol. 69**, pp. 383–390 1982.

[HAL 83] P. Hall, Large-sample optimality of least squares cross-validation in density estimation, *The Annals of Statistics*, **Vol. 11**, pp. 1156–1174 1983.

[HAL 85] P. Hall, 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 1985.

[HAL 87a] P. Hall et J.S. Marron, Estimation of integrated squared density derivatives, *Statistics and Probability Letters*, **Vol. 6**, pp. 109–115 1987.

[HAL 87b] 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 1987.

[HAL 87c] P. Hall et J.S. Marron, On the amount of noise inherent in bandwith selection of a kernel density estimator, *The Annals of Statistics*, **Vol. 15**, pp. 163–181 1987.

[HAL 91a] P. Hall et J.S. Marron, Local minima in cross-validation functions, *Journal of the Royal Statistical Society*, **Vol. B53**, pp. 245–252 1991.

[HAL 91b] P. Hall et J.S. Marron, Lower bounds for bandwith selection in density estimation, *Probability Theory and Related Fields*, **Vol. 90**, pp. 149–173 1991.

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

[JOH 69] N.L. Johnson et S. Kotz, *Distributions in Statistics*, Wiley, New York 1969.

[JON 91] M.C. Jones et S.J. Sheather, Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statistics and Probability Letters*, **Vol. 11**, pp. 511–514 1991.

[JON 98] M.C. Jones, M. Samiuddin, A.H. Al-Harbey et T.A.H. Maatouk, The edge frequency polygon, *Biometrika*, **Vol. 85**, pp. 235–239 1998.

[LIE 87] F. Liese et I. Vajda, *Convex Statistical Distances*, Teubner–Verlag, Leipzig 1987.

[LOF 65] D.O. Loftsgaarden et C.P. Quesenberry, A nonparametric estimate of a multivariate density function, *The Annals of Mathematical Statistics*, **Vol. 36**, pp. 1049–1051 1965.

[MAR 85] J.S. Marron, An asymptotically efficient solution to the bandwith problem of kernel density estimation, *The Annals of Statistics*, **Vol. 13**, pp. 1011–1023 1985.

[MAR 88] J.S. Marron, Automatic smoothing parameter selection : a survey, *Empirical Economics*, **Vol. 13**, pp. 187–208 1988.

[MON 90] A. Monfort, *Cours de Statistique Mathématique*, Economica, Paris 1990.

[MOO 77] D.S. Moore et J.W. Yackel, Large sample properties of nearest neighbour density function estimates, in *Statistical Decision Theory and Related Topics II*, ed. S.S. Gupta et D.S. Moore, Academic Press, New York 1977.

[NAD 74] 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.

[PAR 62] E. Parzen, On the estimation of a probability density function and the mode, *The Annals of Mathematical Statistics*, **Vol. 33**, pp. 1065–1076 1962.

[ROS 56] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, *The Annals of Mathematical Statistics*, **Vol. 27**, pp. 832–837 1956.

[RUD 82] M. Rudemo, Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics*, **Vol. 9**, pp. 65–78 1982.

[SAP 90] G. Saporta, *Probabilités, Analyse des Données et Statistique*, Technip, Paris 1990.

[SAP 97] G. SAPORTA, Méthodes statistiques de discrimination, in *Statistique et méthodes neuronales*, ed. S. Canu, O. Gascuel, S. Thiria et Y. Lechevallier, pp. 20–30 1997.

[SCH 47] H. SCHEFFÉ, A useful convergence theorem for probability distribution, *The Annals of Mathematical Statistics*, **Vol. 18**, pp. 434–438 1947.

[SCH 81] E.F. SCHUSTER ET G.G. GREGORY, 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 1981.

[SCO 87] 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.

[SCO 92] D.W. SCOTT, *Multivariate Density Estimation – Theory, Practice, and Visualization*, Wiley, New York 1992.

[SIL 92] B.W. SILVERMAN, *Density Estimation for Statistics and Data Analysis*, Chapman et Hall, Londres 1986.

[STO 77] C.J. STONE, Consistent nonparametric regression, *The Annals of Statistics*, **Vol. 8**, pp. 1348–1360 1977.

[STO 84] C.J. STONE, An asymptotically optimal window selection rule for kernel density estimates, *The Annals of Statistics*, **Vol. 12**, pp. 1285–1297 1984.

[TUR 93] B.A. TURLACH, Bandwith selection in kernel density estimation : a review, *Rapport technique*, Université Catholique de Louvain 1993.

[TAS 89] P. TASSI, *Méthodes Statistiques, 2ème Edition*, Economica, Paris 1985.

[TIT 85] D.M. TITTERINGTON, A.F.M. SMITH ET U.E. MAKOV, *Statistical Analysis of Finite Mixture Distributions*, Wiley, New York 1985.

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

Index