République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

UNIVERSITÉ MOHAMED KHIDER, BISKRA

FACULTÉ des SCIENCES EXACTES et des SCIENCES de la NATURE et de la VIE

DÉPARTEMENT DE MATHÉMATIQUES

Mémoire présenté en vue de l'obtention du Diplôme **:**

MASTER en Mathématiques

Option : **Statistique**

Par

BOUNAB Bouchra

Titre :

Choix bayésien du paramètre de lissage dans l'estimation à noyau d'une densité conditionnelle

Membres du Comité d'Examen :

Dr.	HASSOUNA Houda	UMKB	Présidente
Dr.	CHERFAOUI Mouloud	UMKB	Encadreur
Dr.	DJABER Ibtissem	UMKB	Examinatrice.

Juin 2018

Dédicace

Je dédie ce humble travail .....

REMERCIEMENTS

.....

Table des matières

Table des figures

Liste des tableaux

Introduction générale

En général, dans les statistiques de toute évidence la densité génère l'échantillon, mais la question qui se pose lorsque étant données un échantillon est-ce-que nous pouvons approximativement recrées leur fonction de densité ?

En pratique, la fonction de densité d'une manière générale et d'une densité conditionnelle en particulier est rarement connue, l'estimation de cette dernière trouve ses applications dans divers domaines, comme la physique, la biologie, la météorologie, etc. La densité est d'une grande importance le fait que son modèle nous permet de répondre à des multiples questions sur les données.

L'une des familles des techniques d'estimation d'une densité de probabilité conditionnelle est l'approche non paramétrique qui consiste en estimation de cette fonction à partir de la seule information disponible dans les données sans aucune hypothèse. Dans la littérature il existe plusieurs techniques d'estimation non paramétrique d'une densité de probabilité mais la technique qui a rencontré le plus de succès est bien que la méthode du noyau vue la simplicité de sa forme, ses modes de convergence multiples et sa flexibilité. Notons que la mise en œuvre de cette technique nécessite le choix d'un noyau K et d'un paramètre de lissage $h = (a, b)$. Quoique le choix du noyau engendre un petit problème mais le vrai problème ça reste dans le choix de paramètre du lissage. Dans la littérature, deux catégories de méthodes pour le choix du h ont été proposées. La première catégorie repose sur la minimisation de l'erreur quadratique moyenne intégrée ($MISE$). L'inconvénient de cette classe est que le paramètre de lissage optimal dépend d'une ou de plusieurs

quantités inconnues. La deuxième catégorie est de type validation croisée, elle est intéressante en pratique car elle se laisse guider seulement par les observations. Cependant ces dernières techniques peuvent produire plusieurs minimums locaux.

Le but de ce travail est de présenter une approche alternative pour le choix du paramètre de lissage h dite "*approche bayésienne*". Contrairement aux méthodes classiques, cette approche est caractérisée par la considération du paramètre de lissage h comme étant une variable aléatoire, en lui associant alors une loi conjointe a priori qui sert à compenser le manque d'informations. Plus précisément, nous proposons l'approche bayésienne globale pour le choix du paramètre de lissage h, dans l'estimation à noyau de la fonction de densité conditionnelle dans deux situation à savoir : lorsque $a = b$ et lorsque $a \neq b$ et pour surmonter le problème de calcul de la loi a posteriori qui est souvent de forme complexe nous faisons appel aux méthodes de Monte Carlo par Chaînes de Markov (MCMC).

Pour répondre à notre objectif, en plus de la présent introduction, nous avons réparti le reste du présent document en trois chapitres une conclusion générale et une liste de références.

Le premier chapitre est consacré à la présentation de l'estimation non paramétrique de la densité de probabilité conditionnelle. En effet, après la présentation de la définition de l'estimateur à noyau de cette fonction de densité et ses propriétés (biais, variances, ...), nous avons abordé le problème du choix du noyau et du paramètre de lissage par les techniques classiques. Le deuxième chapitre concerne le choix du paramètre de lissage par l'approche bayésienne. Nous avons rappelé quelques notions de base sur le formalisme bayésien, ensuite nous avons présenté l'approche bayésienne globale pour l'estimation du paramètre de lissage. Finalement avant de conclure, dans le dernier chapitre, nous avons réalisé une application numérique, sur des données simulées, qui nous permettre d'illustrer l'impact du choix des paramètres de l'estimateur à noyau de la densité $f(y|x)$ sur les performances de ce dernier.

Chapitre 1

Estimation à noyau d'une densité conditionnelle

Introduction

L'une des fonctions de densité la plus usuel dans différents champs des statistiques est bien que la fonction de densité conditionnelle. Dans ce chapitre nous sommes intéressés au problème d'estimation non-paramétrique de cette densité. Plus précisément, nous allons intéresser à l'estimation d'une fonction de densité conditionnelle par la méthode du noyau. En effet, après la présentation de l'estimateur à noyau de cette densité, nous citons quelque unes de ses propriétés (le biais, la variance, la convergence...). Ensuite, nous allons aborder le problème de choix du noyau et du paramètre de lissage dans ce cas.

1.1 Définition de l'estimateur de $f(y|x)$

On cherche l'estimateur de la densité de probabilité conditionnelle de Y sachant que X (on note $Y \mid X$) où X et Y sont des variables aléatoires univariées. On définit la densité

conditionnelle de $Y \mid X$ comme suit :

$$f(y \mid x) = \frac{g(x,y)}{l(x)}, \tag{1.1}$$

où $g(x,y)$ est la densité joint de (X,Y) et $l(x)$ est la densité marginal de X.

Hyndmanet al. (1996) [9] ont considéré la forme modifiée de l'estimateur de Rosenblatt [15] pour définir l'estimateur de la densité conditionnelle, donnée par :

$$\hat{f}(y \mid x) = \frac{\hat{g}(x,y)}{\hat{l}(x)}, \tag{1.2}$$

où $\hat{g}$ et $\hat{l}$ sont les estimateurs respectifs de Parzen-Rosenblatt [11, 15] de g et l, définis respectivement par :

$$\hat{g}(x,y) = \frac{1}{nab}\sum_{j=1}^{n} K\left(\frac{x-x_j}{a}\right) K\left(\frac{y-y_j}{b}\right), \tag{1.3}$$

$$\hat{l}(x) = \frac{1}{na}\sum_{j=1}^{n} K\left(\frac{x-x_j}{a}\right), \tag{1.4}$$

avec K étant un noyau définie sur $\mathbb{R}$, a est un paramètre de lissage dans la direction de x, b est un paramètre de lissage dans la direction de y et $\{(x_1,y_1),...,(x_n,y_n)\}$ est un n-observation issue de la réalisation de la variable aléatoire (X,Y).

En substituant $\hat{g}(x,y)$ et $\hat{l}(x)$ données respectivement dans (1.3) et (1.4) dans 1.2, l'estimateur de la densité conditionnelle s'écrit comme suit :

$$\hat{f}(y \mid x) = \frac{\frac{1}{nab}\sum_{j=1}^{n} K\left(\frac{x-x_j}{a}\right) K\left(\frac{y-y_j}{b}\right)}{\frac{1}{na}\sum_{j=1}^{n} K\left(\frac{x-x_j}{a}\right)} = \frac{\frac{1}{b}\sum_{j=1}^{n} K\left(\frac{x-x_j}{a}\right) K\left(\frac{y-y_j}{b}\right)}{\sum_{j=1}^{n} K\left(\frac{x-x_j}{a}\right)}. \tag{1.5}$$

Si on pose :

$$\omega_j(x) = \frac{K\left(\frac{x - x_j}{a}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{a}\right)},$$

alors, on peut réécrire (1.5) comme suit :

$$\hat{f}(y \mid x) = \frac{1}{b}\sum_{j=1}^{n} \omega_j(x) K\left(\frac{y - y_j}{b}\right). \tag{1.6}$$

Dans le reste du document nous supposons que le noyau $K(.)$ est une fonction réelle, non négative, symétrique et deux fois intégrable, c'est-à-dire :

$$K(u)du = 1, \qquad \int_{\mathbb{R}} uK(u)du = 0 \quad et \quad \sigma_K^2 = \int_{\mathbb{R}} u^2 K(u) < \infty.$$

1.2 Propriétés asymptotiques de l'estimateur

Dans cette section nous allons présenter quelques propriétés de l'estimateur à noyau d'une densité conditionnelle, où nous allons se limiter à la forme du biais et de la variance de l'estimateur, sa convergence ponctuelle en norme L^1 et sa convergence en norme L^2 (ponctuelle et globale).

1.2.1 Biais et variance

En se basant sur la dérivation et le développement de Taylor à l'ordre 2, Hyndman et al. [9] ont obtenu la forme asymptotique du biais et de la variance de l'estimateur et qui sont données respectivement comme suit :

$$\begin{aligned} E\left[\hat{f}(y \mid x)\right] - f(y \mid x) \; &= \frac{a^2\sigma_K^2}{2}\left\{2\frac{l'(x)}{l(x)}\frac{\partial f(y \mid x)}{\partial x} + \frac{\partial^2 f(y \mid x)}{\partial x^2} + \frac{b^2}{a^2}\frac{\partial^2 f(y \mid x)}{\partial y^2}\right\} \\ &\quad + o(a^4) + o\left(b^4\right) + o\left(a^2 b^2\right) + o\left(\frac{1}{na}\right), \end{aligned} \tag{1.7}$$

et

$$Var[\hat{f}(y \mid x)] = \frac{R(K)f(y \mid x)}{nabl(x)}\left[R(K) - bf(y \mid x)\right] + o\left(\frac{1}{n}\right) + o\left(\frac{b}{an}\right) + o\left(\frac{a}{bn}\right), \quad (1.8)$$

où $R(K) = \int K^2(u)du$, a condition que $a \to 0$, $b \to 0$ et $nab \to 0$ lorsque $n \to \infty$.
Avant d'aborder le sujet de convergence de l'estimateur en question rappelons que l'erreur moyenne quadratique intégrée ($MISE$) et l'erreur quadratique intégrée (ISE) correspondants au cas d'une densité conditionnelle sont définies comme suit :
L'erreur moyenne quadratique intégrée :

$$MISE(a, b, \hat{f}, f) = \iint E\{\hat{f}(y \mid x) - f(y \mid x)\}^2 h(x)dxdy. \quad (1.9)$$

L'erreur quadratique intégrée :

$$ISE(a, b, \hat{f}, f) = \iint \{\hat{f}(y \mid x) - f(y \mid x)\}^2 h(x)dxdy. \quad (1.10)$$

1.2.2 Convergence en norme $\mathbf{L}^1$, pour x_0 fixé

Soit $x_0 \in \mathbb{R}$ tel que $f\ (x_0) \neq 0$, nous savons d'une part que la fonction $y \longmapsto f\ (y \mid x_0)$ est une densité. D'autre part ; l'espace naturel pour l'étude des densités est l'espace L^1 [4], de plus nous savons que la convergence ponctuelle presque partout en probabilité (resp. p.s.) implique la convergence L^1 en probabilité (resp. p.s.). Compte tenu de ces arguments et en s'inspirant des travaux de Devroye en 1987 [4], Youndjé [16] a obtenu des conditions suffisantes de convergence en L^1(pour x_0 fixé) du type

$$\int \left|\hat{f}\ (y \mid x_0) - f\ (y \mid x_0)\right| dy \longrightarrow 0,$$

en probabilité, presque sûrement et presque complètement.

1.2.3 Convergence en norme $\mathbf{L}^2$

Sachant que l'erreur quadratique moyenne (MSE) est la somme du carré du biais (1.7) avec la variance (1.8), alors l'erreur quadratique moyenne asymptotique ($AMSE$) est de forme :

$$\begin{aligned} AMSE\left(\hat{f}(x), f(x)\right) &= \frac{a^4\sigma_K^4}{4}\left[2\frac{l'(x)}{l(x)}\frac{\partial f(y \mid x)}{\partial x} + \frac{\partial^2 f(y \mid x)}{\partial x^2} + \frac{b^2}{a^2}\frac{\partial^2 f(y \mid x)}{\partial y^2}\right]^2 \\ &\quad + \frac{R(K)f(y \mid x)}{nabl(x)}\left[R(K) - bf(y \mid x)\right] + o\left(\frac{1}{n}\right) + o\left(\frac{b}{an}\right) \\ &\quad + o\left(\frac{a}{bn}\right) + o(a^6) + o(b^6) + o(a^2b^4) + o(a^4b^2). \end{aligned} \tag{1.11}$$

On constate que cet estimateur est constitué à condition que $a \to 0$, $b \to 0$ et $nab \to \infty$ lorsque $n \to \infty$.

Comme d'autres problèmes de lissage, les petits paramètres de lissage donnent des petits biais et grandes variances, alors que les grands paramètres de lissage donnent des grands biais et petites variances. Les paramètres de lissage qui sont choisis pour minimiser (1.11) donnent en principe un équilibre entre le biais et la variance.

L'erreur quadratique moyenne intégrée asymptotique ($AMISE$) est obtenue en faisant l'intégration par rapport à x et y de l'$AMSE$ pondéré, formé par le produit de (1.11) avec $l(x)$, sa forme est donnée comme suit :

$$MISE \approx \frac{c_1}{nab} - \frac{c_2}{na} + c_3a^4 + c_4b^4 + c_5a^2b^2, \tag{1.12}$$

où les constants c_1, c_2, c_3, c_4 et c_5, qui dépendent du noyau K, la densité conditionnelle

$f(y \mid x)$ et de la densité marginal $l(x)$, sont donnés par :

$$\begin{cases} c_1 &= \int R^2(K)dx, \\ c_2 &= \int\int R(K) f^2(y \mid x)dydx, \\ c_3 &= \int\int \frac{\sigma_K^4 l(x)}{4}\left\{2\frac{l'(x)}{l(x)}\frac{\partial f(y \mid x)}{\partial x} + \frac{\partial^2 f(y \mid x)}{\partial x^2}\right\}^2 dydx, \\ c_4 &= \int\int \frac{\sigma_K^4 l(x)}{4}\left\{\frac{\partial^2 f(y \mid x)}{\partial y^2}\right\}^2 dydx, \\ c_5 &= \int\int \frac{\sigma_K^4 l(x)}{2}\left\{2\frac{l'(x)}{l(x)}\frac{\partial f(y \mid x)}{\partial x} + \frac{\partial^2 f(y \mid x)}{\partial x^2}\right\}\left\{\frac{\partial^2 f(y \mid x)}{\partial y^2}\right\} dydx, \end{cases} \tag{1.13}$$

avec $R(g) = \int g^2(x)dx$.

1.3 Choix du paramètre de lissage

Un noyau approprié aide à surmonter les problèmes des bosses (multi-modes) et de la discontinuité de la densité estimée. Par exemple, si K est une distribution gaussienne, alors la fonction de densité estimée $\hat{f}$ sera lisse et admet des dérivées de toutes ordres. Dans la littérature, il existe plusieurs fonctions qui jouent le rôle d'un noyau, la Table 1.1 résume quelques noyaux les plus usuelles dans la pratique dont leurs formes sont illustrées dans la Figure 1.1.

TABLE 1.1: Noyaux usuels et leurs supports.

Nom	Expression	Domaine
Noyau Uniforme (Rosenblatt)	$K(u) = \frac{1}{2}$	$\lvert u\rvert \leq 1$
Noyau Box (boite)	$K(u) = \frac{1}{2\sqrt{3}}$	$\lvert u\rvert \leq \sqrt{3}$
Noyau Triangulaire	$K(u) = (1 - \lvert u\rvert)$	$\lvert u\rvert \leq 1$
Noyau Cosine	$K(u) = \frac{\pi}{4}\cos(\frac{\pi u}{2})$	$\lvert u\rvert \leq 1$
Noyau Gaussien	$K(u) = \frac{1}{\sqrt{2\pi}}e^{-u^2/2}$	$u \in \mathbb{R}$
Noyau Biweight (Tukey)	$K(u) = \frac{15}{16}(1 - u^2)^2$	$\lvert u\rvert \leq 1$

Noyau Triweight	$K(u) = \frac{35}{32}(1-u^2)^3$	$\|u\| \leq 1$
Noyau Epanechnikov	$K_E(u) = \frac{3}{4\sqrt{5}}\left(1 - \frac{u^2}{5}\right)$	$\|u\| \leq \sqrt{5}$

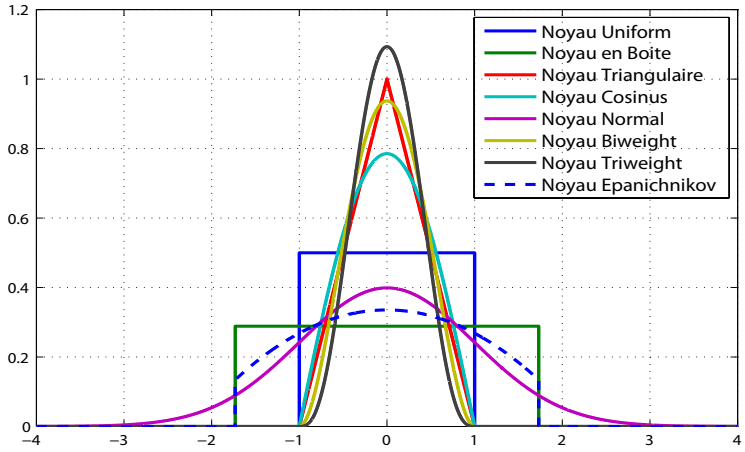

FIGURE 1.1 – La forme des différents noyaux usuels.

1.3.1 Choix optimal

Les largeurs de fenêtres optimales peuvent être obtenues par la différentiation de (1.12) par rapport à a et b et en fixant les dérivés à zéro. En simplifiant ces dérivés, nous obtenons le système d'équations suivant :

$$\begin{cases} -\dfrac{c_1}{n} - \dfrac{c_2 b}{n} + 4c_3 a^5 b + 2c_5 a^3 b^3 = 0 \\ -\dfrac{c_1}{n} + 4c_4 a b^5 + 2c_5 a^3 b^3 = 0 \end{cases}$$

Hyndman et al. (1996) [9] ont montré que la solution du système est approximativement :

$$\begin{cases} a^* &= c_1^{1/6} \left\{ 4\left(\dfrac{c_3^5}{c_4}\right)^{1/4} + 2c_5 \left(\dfrac{c_3}{c_4}\right)^{3/4} \right\}^{-1/6} n^{-1/6}, \\ b^* &= a^* \left(\dfrac{c_3}{c_4}\right)^{1/4} = c_1^{1/6} \left\{ 4\left(\dfrac{c_4^5}{c_3}\right)^{1/4} + 2c_5 \left(\dfrac{c_4}{c_3}\right)^{3/4} \right\}^{-1/6} n^{-1/6}, \end{cases}$$

où c_i $(i = 1, ..., 5)$ sont définie dans (1.13). On remarque que a^* et b^* ne sont pas calculables

car ils dépendent des fonctions inconnues $l(x)$ et $f(y \mid x)$.

1.3.2 La règle de référence

Cette méthode a été proposée dans Bashtannyk et Hyndman (2001) [2]. Elle consiste à supposer que la densité conditionnelle suit une loi normale et de trouver les paramètres de lissage qui minimise le $MISE$. Cette technique est robuste et elle fournie des résultats raisonnables, même pour des densités qui ne sont pas de distribution normale.
Les détails de la construction de cette technique sont comme suit : Les auteurs ont supposé que la densité conditionnelle de Y sachant que $X = x$ suit une loi normale de moyenne $r(x) = u + vx$ et d'écart type $\sigma(x) = p + qx$. D'où, $[Y \mid X = x] \overset{L}{\rightsquigarrow} N\left(u + vx, (p + qx)^2\right)$ et la densité conditionnelle est :

$$f(y \mid x) = \frac{1}{(p+qx)\sqrt{2\pi}} \exp\left\{\frac{-1}{2(p+qx)^2}(y - u - vx)^2\right\}. \tag{1.14}$$

Ils ont également supposé que la densité marginal $l(x)$ est connue où ils ont considéré deux situations, lorsque $l(x)$ est une loi Normale et lorsque $l(x)$ est une loi uniforme sur un intervalle.

Cas de loi uniforme

Pour trouver la règle de référence pour a et b quand $l(x)$ est une distribution uniforme sur $[\alpha, \beta]$, on remplace la densité conditionnelle $f(y \mid x)$ dans (1.14) et la densité marginale dans les constants $c_1, ..., c_5$ voir (1.13), ensuite on fait l'intégration deux fois par rapport à x et y. On trouve ces constants en fonction de p, q, $R(K)$, a et b.
Alors, sous la condition $v \neq 0$, la règle de référence est donnée comme suit :

- **Cas $q \neq 0$:**

$$\begin{cases} a_U &= \left[\dfrac{2^{15/2}\sqrt{\pi}R^2(K)(a-b)^2 q}{3n\sigma_K^2 z w^{3/4}\left(\sqrt{w} + 2v^2 - 3q^2\right)}\right]^{1/6}, \\ b_U &= \dfrac{w^{1/4}}{\sqrt{2}} a_U, \end{cases}$$

où, $z = ((p+qa)^4 - (p-qb)^4)/(p+qa)^4(p-qb)^4$, $w = 19q^4 + 4v^4 + 28q^2v^2$.

- **Cas $q = 0$:**

$$\begin{cases} a_U &= \left[\dfrac{4\sqrt{\pi}R^2(K)(a-b)^2p^5}{3n\sigma_K^2 v^5}\right]^{1/6}, \\ b_U &= va_U, \end{cases}$$

Cas de loi normale

Dans ce cas, il ont supposé que $l(x)$ est normale avec une moyenne constante μ_l et une variance constante σ_l^2. Ils supposent également que la densité (1.14) a une variance constante(i.e. $q = 0$). En refaisant les mêmes étapes que le cas uniforme, ils ont trouvé la règle de référence

$$\begin{cases} a_N &= \left\{\dfrac{16kR^2(K)p^5(288\pi^9\sigma_l^{58}\lambda^2(k))^{1/8}}{n\sigma_K^4 v^{5/2}\gamma^{3/4}(k)[\gamma^{1/2}(k) + v(18\pi\sigma_l^{10}\lambda^2(k))^{1/4}]}\right\}^{1/6}, \\ b_N &= \left\{\dfrac{v^2\gamma(k)}{3\sqrt{2\pi}\sigma_l^5\lambda(k)}\right\}^{1/4} a_N, \end{cases}$$

où, $\lambda(k) = \int_{-\infty}^{k} \phi(t)dt$, $\phi(.)$ est la densité d'une loi normale standard et $\gamma(k) = \sqrt{2\pi}\sigma_l^3(3v^2\sigma_l^2 + 8p^2)\lambda(k) - 16k\sigma_l^2p^2e^{-k^2/2}$. La valeur k contrôle la taille de l'échantillon dans la direction de x, les choix usuels de k sont 2 ou 3.

1.3.3 Validation croisée

Cette méthode a été traité dans le cas où a et b sont supposés égaux. Alors, dans ce cas le problème revient à trouver un seul paramètre de lissage noté h ($i.e$ $a = b = h$).
L'estimateur de la densité conditionnelle sera définis par :

$$\hat{f}(y \mid x) = \frac{1}{h}\sum_{j=1}^{n}\omega_j(x)K\left(\frac{y - y_j}{h}\right), \tag{1.15}$$

avec

$$\omega_j(x) = \frac{K\left(\frac{x - x_j}{h}\right)}{\sum_{j=1}^{n} K\left(\frac{x - x_j}{h}\right)},$$

où K étant un noyau définie sur $\mathbb{R}$, et $h = h(n) \in \mathbb{R}_*^+$ est le paramètre de lissage.

Youndjé [16] a étudié le choix du paramètre de lissage avec la méthode de validation croisée, en se basant sur la minimisation de l'erreur quadratique intégrée pondérée définie par :

$$ISE(\hat{f}, f) = \iint \left\{\hat{f}(y \mid x) - f(y \mid x)\right\}^2 W(x)W'(y)dxdy,$$

où W et W' sont des fonctions de poids positives.

Ensuite, il a démontré que sous certaines hypothèses de régularité on a :

$$ISE(\hat{f}, f) = \frac{c_1}{nh^2} + c_2h^4 + o\left(\frac{1}{nh^2} + h^4\right),$$

où c_1 et c_2 sont deux constants qui dépendent de la fonction de densité inconnue f, alors le paramètre de lissage qui minimise le ISE n'est pas calculable.

Pour trouver un autre paramètre qui minimise le ISE, l'auteur a proposé de décomposer cette quantité de la manière suivante :

$$ISE(\hat{f}, f) = A + B - 2C,$$

où,

$$\begin{aligned} A &= \iint \hat{f}^2(y \mid x)W(\mathbf{x})W'(y)f(x)dxdy, \\ B &= \iint f^2(y \mid x)W(\mathbf{x})W'(y)f(x)dxdy, \\ C &= \iint \hat{f}(y \mid x)f(y \mid x)W(\mathbf{x})W'(y)f(x)dxdy. \end{aligned}$$

Puisque B est indépendant de h, choisir h minimisant ISE revient à choisir h minimisant

$A - 2C$. Le fait que C s'écrit aussi :

$$C = E_{(X,Y)}\left[\hat{f}(y \mid x) f(y \mid x) W(\mathbf{x}) W'(y)\right].$$

Youndjé [16] a approché A et C par :

$$\begin{aligned} \overline{A} &= \frac{1}{n}\sum_{i=1}^{n} \frac{\hat{g}^{-i}(x_i, y_i)}{\hat{l}^{-i}(x_i)} W(X_i) W'(Y_i), \\ \overline{C} &= \frac{1}{n}\sum \int \left(\frac{\hat{g}^{-i}(x_i, y_i)}{\hat{l}^{-i}(x_i)}\right)^2, \end{aligned}$$

où,

$$\begin{aligned} \hat{g}^{-i}(x, y) &= \frac{1}{(n-1)h^2}\sum_{j\neq i}^{n} K\left(\frac{x - x_j}{h}\right) K\left(\frac{y - y_j}{h}\right), \\ \hat{l}^{-i}(x) &= \frac{1}{(n-1)h}\sum_{j\neq i}^{n} K\left(\frac{x - x_j}{h}\right). \end{aligned}$$

Finalement, la règle de sélection du paramètre de lissage par la validation croisée est obtenue en choisissant le h minimisant le critère suivant :

$$CV(h) = \frac{1}{n}\sum_{i=1}^{n} \int \left(\frac{\hat{g}^{-i}(X_i, y)}{\hat{l}^{-i}(X_i)}\right)^2 W'(y) dy W(X_i) - \frac{2}{n}\sum_{i=1}^{n} \frac{\hat{g}^{-i}(X_i, Y_i)}{\hat{l}^{-i}(X_i)} W(X_i) W'(Y_i).$$

Conclusion

Dans le présent chapitre, nous avons mis, dans un premier lieu, en évidence la définition d'un estimateur à noyau d'une densité $f(y \mid x)$, ainsi que les conditions liées à l'existence et la convergence en normes L^1 et L^2 de cet estimateur. En deuxième lieu, nous avons exposé le problème du choix du noyau et du paramètre de lissage par les procédures classiques : choix optimal, règle de référence, validation croisée...

Une alternative du choix du paramètre de lissage proposée dans la littérature ces dernières années est bien que l'approche bayésienne.

Le principe et la procédure de sélection du paramètre de lissage dans le cadre d'estimation à noyau d'une densité conditionnelle par l'inférence bayésien fera l'objet du chapitre suivant.

Chapitre 2

Choix du paramètre de lissage par l'approche bayésienne

Introduction

Dans ce chapitre, nous allons présenter la démarche à suivre pour l'approche bayésienne dans le but d'estimer le paramètre de lissage $h = (a, b)$ dans l'estimation à noyau de la fonction de densité conditionnelle. En effet, dans un premier lieu nous allons rappeler quelques notions de base sur le formalisme bayésien et les estimateurs paramétriques bayésiens ainsi que les méthodes MCMC (Monte Carlo par chaîne de Markov) qui peuvent être employées lors de l'implémentation pratique de l'approche bayésienne.
Dans un seconde lieu, nous allons présenter deux situations du choix du paramètre de lissage par l'inférence bayésienne à savoir : le choix du (a, b) sous l'hypothèse $a = b$ et le choix du (a, b) sous l'hypothèse $a \neq b$.

2.1 Principe Bayésien

Le but de cette section est de rappeler quelques notions sur l'inférence bayésien. L'origine de cette idée remonte à Thomas Bayes en 1761. L'impact du théorème de Bayes provient

de la décision audacieuse de mettre les causes et effets sur le même niveau conceptuel puisque les deux sont aléatoires.

Supposons que $x = (x_1, x_2, ...x_n)$ est le vecteur d'observation et $\theta = (\theta_1, \theta_2, ...\theta_d) \in \Theta$ est le vecteur des paramètres à estimer.

L'idée principale de l'analyse bayésienne repose sur la **loi a posteriori** des paramètres en considérant θ comme variable aléatoire. L'espace des paramètres Θ est muni d'une loi de probabilité π et nous noterons $\theta \sim \pi$. La loi π est appelée **loi a priori** de θ choisie en fonction des connaissances disponibles (**information a priori**) sur θ avant la prise en compte des observations.

2.1.1 Quelques définitions

- On entend par **information a priori** sur le paramètre θ toute information disponible sur θ en dehors de celle apportée par les observations.
- On appelle **loi d'observations** la loi conditionnelle de x sachant θ, où sa densité est notée $f(x \mid \theta)$.
- **La loi a posteriori** : c'est la loi conditionnelle de θ sachant x, où sa densité est notée par $\pi(\theta \mid x)$. En vertu de la formule de Bayes on a :

$$\pi(\theta \mid x) = \frac{f(x \mid \theta)\pi(\theta)}{\int\limits_{\Theta} f(x \mid \theta)\pi(\theta)d\theta}.$$

- **Le modèle statistique paramétrique bayésienne** consiste en la donnée d'une loi a priori π et de la loi des observations.

  En général la distribution a posteriori représente l'actualisation de l'information disponible sur le paramètre θ.

- **La loi du couple** (θ, x) : La densité du couple (θ, x) est donnée par $g(\theta, x) = f(x \mid \theta)\pi(\theta)$.

• **La loi marginale de** x : Sa densité est donnée par $m(x) = \int_{\Theta} f(x \mid \theta)\pi(\theta)\, d\theta$.

2.1.2 Choix de lois a priori : cas priories non informatives

Le point le plus critiquable et le plus critiqué dans l'analyse bayésienne est bien que le choix de la loi a priori. Dans la pratique, il est rare que l'information a priori soit suffisamment précise pour déterminer exactement la loi a priori du paramètre θ. En effet , on peut distinguer deux types : non informative et informative.
Dans le cas d'absence d'information précise sur le paramètre à estimer, les bayésiens proposent d'utiliser des priories non informatives et cela le fait que aucune information n'est disponible, il est impossible de bâtir une distribution a priori sur des considérations subjectives. Dans ce cas, on laisse les données conduire l'inférence qui peut être justifié par les résultats théoriques qui montrent que l'inférence dans les deux paradigmes (bayésien et fréquentiste). Voici quelques considérations sur les priories non informatives.

Densité a priori uniforme : Historiquement, Laplace fut le premier à utiliser des techniques non informatives. Il munit les paramètres d'une loi a priori qui prend en compte son ignorance en donnant la même vraisemblance à chaque valeur du paramètre, donc en utilisant une loi uniforme. Ainsi, la densité a priori d'un paramètre est définit par :

$$\pi(\theta) = c,$$

où c est une constante.

A priori impropre : Lorsque l'espace de définition du paramètre est infini, il n'existe pas de lois de probabilité uniforme sur cet espace. On est alors conduit à choisir des priories impropres. La loi a priori peut être impropre c'est-à-dire $\int_{\Theta} \pi(\theta)\, d\theta = \infty$. Ce choix de type de loi n'a donc plus d'intérêt que calculatoire et s'interprète difficilement.

Lois a priories conjuguée :

En général, les densités $\pi(\theta)$ et $f(x \mid \theta)$ ne sont pas faciles à calculer. Quelque fois $\pi(\theta \mid x)$ ne peut être évaluée de manière explicite. De plus, la complexité augmente lorsque la dimension de l'espace Θ s'accroît. Il convient alors de choisir une distribution a priori qui permette facilement d'exploiter la distribution a posteriori dès le recueil d'une nouvelle information x sur le paramètre à estimer θ.

Définition 2.1.1 *Soit $\mathcal{F}$ une famille de distributions de densité $f(x \mid \theta)$, indexée par θ. Une famille $\Im$ de distributions a priori de densité $\pi(\theta)$ est dite conjuguée par rapport à $\mathcal{F}$, si la distribution a posteriori de densité $\pi(\theta \mid x)$ reste dans la même famille $\Im$ pour tout $\pi \in \Im$ et tout $f \in \mathcal{F}$. Autrement dit, la distribution a posteriori garde la même forme que la distribution a priori.*

Dans ce cas, il n'y a généralement pas besoin de calculer explicitement la loi marginale $m(x)$ parce que les noyaux de fonctions $f(x \mid \theta)\pi(\theta)$, $\pi(\theta \mid x)$ sont les mêmes à un facteur constant près :

$$f(x \mid \theta)\pi(\theta) \propto \pi(\theta \mid x).$$

Lorsque la famille de distributions conjuguées $\Im$ est paramétrée, le passage de la distribution a priori à la distribution a posteriori se réduit à un changement de leurs paramètres. Dans ce cas, la distribution a posteriori est toujours calculable. Une telle considération est fondée sur le principe suivant : l'information apportée par des observations x sur θ est limitée et dont la modification par x ne doit pas conduire à une remise en cause de la forme de $\pi(\theta)$, mais seulement de ses paramètres.

$f(x/\theta)$	$\pi(\theta)$	$\pi(\theta/x)$
$N(\mu, 1/\theta)$	$Gamma(\alpha;\beta)$	$Gamma\left(\alpha+1/2;\beta+(\mu-x)^2/2\right)$
$\mathcal{P}(\theta)$	$Gamma(\alpha;\beta)$	$Gamma(\alpha+x;\beta+1)$
$Gamma(\upsilon;\theta)$	$Gamma(\alpha;\beta)$	$Gamma(\alpha+\upsilon;\beta+x)$
$Beta(n;\theta)$	$Beta(\alpha;\beta)$	$Beta(\alpha+x;\beta+n-x)$

TABLE 2.1 – Quelques densités de lois a priories conjuguées

2.1.3 Estimateur de Bayes

Soit $\hat{\theta}$ un estimateur de θ. On définit une fonction de coût non négative $C(\hat{\theta}-\theta)$ telle que $\hat{\theta}-\theta$ est l'erreur d'estimation pour les observations $x=(x_1,x_2.....x_n)$. Le but est de déterminer l'estimateur $\hat{\theta}$ qui minimise le coût moyen $\mathbf{E}\left(C\left(\hat{\theta}-\theta\right)\right)$ appelé le risque bayésien défini par :

$$\mathbf{E}\left(C\left(\hat{\theta}-\theta\right)\right)=\int C\left(\hat{\theta}-\theta\right)\pi\left(\theta\mid x\right)d\theta.$$

Considérons le cas où $C\left(\hat{\theta}-\theta\right)$ est le coût quadratique (fonction coût la plus utilisée), c'est-à-dire que $C\left(\hat{\theta}-\theta\right)=\left(\hat{\theta}(x)-\theta\right)^2$. L'estimateur qui minimise le risque bayésien en utilisant la fonction de coût quadratique est la moyenne a posteriori donnée simplement par :

$$\hat{\theta}=\mathbf{E}\left(\theta\mid x\right)=\int_{\Theta}\theta\pi\left(\theta\mid x\right)d\theta=\frac{\int_{\Theta}\theta f(x\mid\theta)\pi\left(\theta\right)d\theta}{\int_{\Theta}f\left(x\mid\theta\right)\pi(\theta)d\theta}. \tag{2.1}$$

Pour la construction de l'estimateur de Bayes on peut également utilisé autres fonctions de coût tel que le coût absolu, le coût uniforme,...

Il est a noter que dans le reste du présent document, nous ne utilisons que l'estimateur de Bayes décrit dans (2.1).

Propriétés de l'estimateur de Bayes

Notons que parmi les caractéristiques de l'estimateur de Bayes on a :

- L'estimateur de Bayes est admissible.
- L'estimateur de Bayes est biaisé. Sous certaines hypothèses de régularité le plus souvent satisfaites en pratique, on a les deux propriétés :
  - L'estimateur de Bayes est convergent en probabilité.
  - La loi a posteriori peut être asymptotiquement approximer par une loi normale $\mathbf{N}\left(\mathbf{E}\left(\theta \mid x\right), \mathbf{Var}\left(\theta \mid x\right)\right)$.

Cette dernière propriété est particulièrement utile pour construire des intervalles de confiance a posteriori.

2.1.4 Méthodes de Monté Carlo par Chaîne de Markov (MCMC)

Dans plusieurs situations, le calcul de la densité a posteriori analytiquement est très difficile, voir impossible dans certaines situations, car l'inférence bayésienne se heurte à des problèmes d'intégration. Pour résoudre ce problème, on peut faire recours aux méthodes d'intégration numériques ou encore les méthodes de Monté Carlo par Chaîne de Markov (MCMC). Dans la suite, nous allons rappeler quelques notions intéressantes pour les méthodes MCMC.

Méthodes de Monte Carlo : Techniques d'estimation s'appuyant sur la simulation d'un grand nombre de variables aléatoires.

Chaîne de Markov : Une chaîne de Markov sur un espace D est un processus aléatoire où l'état future de la chaîne étant donnée l'état présent est indépendant de tous les états passés, c'est-à-dire si on considère $X_1,\ X_2, ...,\ X_n$ une suite des variables aléatoire associées aux observations $x_1,\ x_2, ...,\ x_n \in D$, on a :

$$\mathbf{P}\left(X_{n+1} = x_{n+1} | X_n = x_n, ...X_0 = x_0\right) = \mathbf{P}\left(X_{n+1} = x_{n+1} | X_n = x_n\right).$$

Noyau de transition d'une chaîne de Markov : Le noyau de transition, $\mathbb{P}$, associé à une chaîne de Markov est un outil qui détermine la probabilité $P(i, j)$ que la chaîne aille

de l'état i à l'état j avec $i, j \in D$.

Loi stationnaire d'une chaîne de Markov : La loi stationnaire d'une chaîne de Markov est une loi Q telle que $Q = Q\mathbb{P}$, où $\mathbb{P}$ est la matrice (opérateur) des transitions de la chaîne de Markov $P\left(X_{n+1} = x_{n+1} | X_n = x_n\right)$.

Une simulation par chaînes de Markov doit assurer deux convergences :

1. Convergence de la chaîne vers sa loi stationnaire. De plus, cette dernière doit être aussi proche que possible de la loi cible.
2. Convergence des moyennes empiriques vers les espérances correspondantes.

Dans les algorithmes de simulation avec indépendance, la convergence des moyennes empiriques vers les moyennes théoriques est assurée par la Loi des Grands Nombres (LGN), mais cette dernière ne fonctionne plus dans le cadre de simulation avec dépendance. Afin que la simulation via les chaînes de Markov puisse remplacer les méthodes de simulation, indépendantes, la LGN est étendue par le théorème d'ergodicité (voir Robert and Casella [13] et Robert [12]) qui résout le paradoxe des deux types de convergence. Ainsi ce théorème supprime le besoin de produire des suites de variables aléatoires indépendantes.

L'idée des méthodes MCMC

Les méthodes MCMC ont été introduite en 1953 par Metropolis et al. [10] appliquées dans la physique statique. L'idée principale des méthodes MCMC est de construire une chaîne de Markov ergodique dont la distribution stationnaire est la loi a posteriori. En simulant une réalisation suffisamment grande de cette chaîne, on pourra supposer qu'à partir d'un certain rang N_0 les échantillons simulés sont représentatifs de la loi a posteriori (la densité cible).

En supposant que pour $N > N_0$, on échantillonne exactement selon la loi cible. L'inférence sur les paramètres est basée alors sur cet échantillon.

Afin de réduire le biais de l'estimateur bayésien qui est due à l'effet des valeurs initiales, le

nombre d'itérations N_0 est ignoré. Le nombre d'itérations N_0 est appelé phase d'échauffement. L'une des difficultés rencontrée dans l'implémentation des méthodes MCMC est le diagnostic de convergence de la chaîne construite. Plusieurs méthodes ont été proposées dans la littérature pour vérifier la convergence des méthodes MCMC. Nous citons par exemple, le critère de inter-intra chaîne proposé par Gelman and Rubin [6].

Algorithme de Metropolis-Hastings

Nous présentons l'algorithme de Metropolis-Hasting (M-H) proposé par Metropolis et al. [10] en 1953 et généralisé par Hastings [8] en 1970, qu'on peut considérer comme l'algorithme de base d'une grande partie des méthodes MCMC. L'objectif de l'algorithme de M-H est de simuler un échantillon selon une densité cible connue, à une constante multiplicative prés, à partir d'une loi de proposition (instrumental) $q(\theta/\tilde{\theta})$ facilement simulable. Les étapes de l'algorithme M-H sont résumées comme suit :

Algorithme M-H

Étape (1) Initialiser $\theta^{(0)}$

Étape (2) pour $k \in \{1, ..., N\}$

(a) Générer $\tilde{\theta} \sim q(\tilde{\theta}|\theta^{(k-1)})$.

(b) Calculer la probabilité d'acceptation $\rho = \min\left\{1, \frac{\pi(\hat{\theta}|x)\ q(\theta^{(k-1)}|\tilde{\theta})}{\pi\left(\theta^{(k-1)}|x\right)\ q(\tilde{\theta}|\theta^{(k-1)})}\right\}$.

(c) mettre à jour θ :

$$\theta^{(k+1)} = \begin{cases} \tilde{\theta}, & \text{si } u < \rho \text{ et } u \sim U[0,1] \quad (U[0,1] \text{ désigne la loi uniforme dans } [0,1]); \\ \theta^{(k)}, & \text{sinon.} \end{cases}$$

Étape (3) Poser $k = k + 1$ et aller à l'**Étape (2)**.

Étape (4) Calculer l'estimateur de Bayes $\hat{\theta} = \frac{1}{N-N_0} \sum_{k=N_{0+1}}^{N} \theta^{(k)}$.

2.2 Étapes de l'approches bayésienne pour la sélection du paramètre de lissage

Dans cette section nous allons présenter et expliquer les étapes à suivre pour la mise en œuvre de l'inférence bayésienne pour le choix du paramètre de lissage global dans l'estimation à noyau d'une densité de probabilité conditionnelle.

L'estimation bayésienne du paramètre de lissage h, d'une manière générale, conditionné aux données se fait par la densité a posteriori $\pi\,(h|\text{données})$.
Concrètement, nous considérons une séquence $X = (X_1,\ X_2, ...,\ X_n)$ de variables aléatoire réelles, indépendantes et identiquement distribuées par une densité de probabilité inconnu f et de réalisations $x = (x_1,\ x_2, ...,\ x_n)$. L'estimation bayésienne globale du paramètre de lissage est généralement construite par les quatres étapes suivantes :

1. Donner la forme de l'estimateur de la vraisemblance $f(x \mid h)$ de la densité des données sachant le paramètre h ;
2. Choisir la loi a priori sur le paramètre de lissage h ;
3. Calculer la densité a posteriori $\pi\,(h \mid x)$ du paramètre de lissage h sachant les données x en utilisant la formule de Bayes ;
4. Estimer le paramètre de lissage h par la moyenne a posteriori (en utilisant les méthodes de MCMC en cas de nécessité).

   Dans ce qui suit nous allons adapter ces quatres dernières étapes au cas de l'estimation du paramètre de lissage pour densité conditionnelle.

2.2.1 Construction de l'estimateur de la vraisemblance

Considérons une séquence $\{(x_1, y_1); (x_2, y_2); ...; (x_n, y_n)\}$ de réalisation de n variables aléatoire réelles, indépendantes et identiquement distribuées par une densité de probabilité

conjointe inconnu g et de densité conditionnelle f de Y sachant X. La vraisemblance des données $x = (\ x_1,\ x_2, ...,\ x_n)$ sachant $\ y = (y_1, y_2, ..., y_n)$ notée $L\,(x, y)$ est donnée par :

$$L\,(x, y) = \prod_{i=1}^{n} f(y_i \mid x_i).$$

Vue que f est inconnue alors l'estimateur de la vraisemblance sera obtenu en utilisant un estimateur de f.

$$L(x, y \mid h) = \prod_{i=1}^{n} \hat{f}\,(y_i \mid x_i)\,.$$

Dans ce document nous proposons, pour la construction de l'estimateur de la vraisemblance, d'adapter la technique de validation croisée qui consiste a estimer $f(y_i \mid x_i)$ à partir de l'ensemble des points (des observations) sauf le point (x_i, y_i) et cela tout en utilisant la méthode du noyau. Alors, l'estimateur à noyau par la technique de validation croisée de $f(y_i \mid x_i)$ est donnée par :

$$\hat{f}(y_i \mid x_i) = \frac{1}{b} \sum_{j=1, j\neq i}^{n} \omega_j(x_i) K_{y_i,h}\,(y_j)\,,$$

avec :

$$\omega_j(x_i) = \frac{K\left(\dfrac{x_i - x_j}{a}\right)}{\displaystyle\sum_{j=1}^{n} K\left(\frac{x_i - x_j}{a}\right)} \text{ et } K_{y_i,h}\,(y_j) = K\left(\frac{y_i - y_j}{b}\right).$$

Ainsi la vraisemblance est approximée par :

$$L(x, y \mid a, b) = \frac{1}{b^n} \prod_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \omega_j(x_i) K_{y_i,h}\,(y_j)\,. \tag{2.2}$$

Ce dernier estimateur sera utilisé pour le calcul de la densité a posteriori.

2.2.2 Choix de la loi a priori

Rappelons que l'estimation par l'approche bayésienne est caractérisée par le choix d'une loi a priori et notons que le choix n'est pas unique ici. L'utilisation de la loi a priori conjuguée dans ce modèle n'est pas bénéfique, car la vraisemblance de validation croisée donnée par (2.2) s'écrit comme produit de n termes, et le paramètre de lissage h intervient dans chacun des termes.

Malgré que d'une manière générale, en pratique, l'estimateur bayésien n'est pas très sensible au choix de la loi a priori, mais dans notre cas, le choix doit être fait d'une manière minutieuse afin que les méthodes MCMC fonctionnent correctement.

Nous proposons d'utiliser la loi Gamma proposée par Brewer [3] comme loi a priori pour h et nous allons considérer deux situations du choix du paramètre de lissage à savoir : le cas où $a = b = h$ et le cas où $a \neq b$.
La densité $\pi(h)$ est alors donnée dans ce cas a une constante prés par

$$\begin{cases} \pi(h \mid \alpha, \beta) & \propto & h^{\alpha-1} \exp\left(-\frac{h}{\beta}\right) & \text{pour } a = b = h, \\ \pi(a, b \mid \alpha_1, \alpha_2, \beta_1, \beta_2) & \propto & a^{\alpha_1-1} b^{\alpha_2-1} \exp\left(-\frac{a\beta_2+b\beta_1}{\beta_1\beta_2}\right) & \text{pour } a \neq b, \end{cases}$$

où $\alpha,\ \alpha_1,\ \alpha_2,\ \beta,\ \beta_1$ et β_2 sont des hyper-paramètres (paramètres de la loi Gamma).
Dans cette partie, nous n'avons pas discuté le choix des hyper-paramètres parce que, le choix est fait empiriquement de sorte que les méthodes de Monte Carlo vont fonctionner correctement.

2.2.3 Calcul de la loi a posteriori

En utilisant le théorème de Bayes, la loi a posteriori de h prend la forme suivante :

$$\pi(h \mid x, y) = \frac{L(x, y \mid h)\pi(h)}{\pi(x, y)} = \frac{\pi(h) \prod_{i=1}^{n} \hat{f}(y_i \mid x_i)}{\pi(x, y)},$$

où $\pi(x,y) = \int L(x,y \mid h)\pi(h)\,dh$

Alors la densité a posteriori du paramètre de lissage sachant les données $x = (x_1,\ x_2, ...,\ x_n)$ et $y = (y_1, y_2, ..., y_n)$ est donnée à une constante d'intégration prés, sous la forme (2.3) dans le cas $a = b = h$ et sous la forme (2.4) dans le cas $a \neq b$.

$$\begin{aligned}\pi(h/\alpha,\beta,x,y) = \frac{L(x,y/h)\pi(h/\alpha,\beta)}{\pi(x,y)} &\approx \frac{\pi(h/\alpha,\beta)\prod_{i=1}^{n}\hat{f}(y_i/x_i)}{\pi(x,y)},\\ &\propto \frac{\pi(h/\alpha,\beta)}{b^n\pi(x,y)}\prod_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\omega_j(x_i)K_{y_i,h}(y_j), \end{aligned} \tag{2.3}$$

où $\pi(x,y) = \int L(x,y/h)\pi(h/\alpha,\beta)\,dh$ est la loi marginale de l'échantillon.

$$\begin{aligned}\pi(h/\alpha,\beta,x,y) = \pi(a,b/\alpha_1,\beta_1,\alpha_2,\beta_2,x,y) &= \frac{L(x,y/a,b)\pi(a,b/\alpha_1,\beta_1,\alpha_2,\beta_2)}{\pi(x,y)}\\ &\approx \frac{\pi_1(a/\alpha_1,\beta_1)\pi_2(b/\alpha_2,\beta_2)\prod_{i=1}^{n}\hat{f}(y_i/x_i)}{\pi(x,y)},\\ \pi(a,b/x,y) \propto \frac{\pi_1(a/\alpha_1,\beta_1)\pi_2(b/\alpha_2,\beta_2)}{b^n\pi(x,y)}&\prod_{i=1}^{n}\sum_{\substack{j=1\\ j\neq i}}^{n}\omega_j(x_i)k_{y_i,h}(y_j)\end{aligned} \tag{2.4}$$

où : $\pi(x,y) = \int\int L(x,y/a,b)\pi(a,b/\alpha_1,\beta_1,\alpha_2,\beta_2)da\ db$

Vue la complexité des lois a posteriori (2.3) et (2.4), il est impossible d'obtenir directement l'estimateur bayésien de h. Pour cela, on utilise les méthodes MCMC présentées dans la Section **2.1.4** pour estimer le paramètre de lissage h dans les deux situations.

2.2.4 Estimation du paramètre de lissage

Nous utilisons l'algorithme M-H à marche aléatoire pour estimer le paramètre de lissage par la moyenne a posteriori. Cet algorithme est basé sur l'utilisation d'une loi génératrice

de candidats de la forme $Q(.|h^{(k)})$, où h représente le paramètre de lissage. Le candidat $\tilde{h}$ est généré a partir d'une variable aléatoire ε positive de densité $q\left(h|h^{(k)},\gamma_{(k)}\right)$, tels que $h^{(k)}$ est le candidat généré à la $k^{iéme}$ étape, $\gamma_{(k)}$ est un paramètre de réglage choisi de telle sorte à obtenir un taux d'acceptation optimal τ. La variable aléatoire ε est généralement choisie comme gaussienne de moyenne $h^{(k)}$ et de variance $\gamma^2_{(k)}$ (pour plus de détails voir Zougab [17]). Ici, nous la remplaçons par une loi gaussienne tronquée sur $[0,+\infty[$. Alors la densité instrumentale est donnée pour $a=b=h$ sous la forme :

$$q(h \mid h^{(k)},\gamma(k)) = \frac{1}{\gamma(k)\sqrt{2\pi\phi\left(\dfrac{h^{(k)}}{\gamma(k)}\right)}} \exp\left(-\frac{\left(h-h^{(k)}\right)^2}{2\gamma^2(k)}\right),$$

et sous l'hypothèses a et b sont indépendants (i.e. $a \neq b$), elle est donnée sous la forme :

$$q(a,b \mid (a^{(k)},b^{(k)}),\gamma(k)) = \frac{1}{\gamma^2(k)2\pi\sqrt{\phi\left(\dfrac{a^{(k)}}{\gamma(k)}\right)\phi\left(\dfrac{b^{(k)}}{\gamma(k)}\right)}} \exp\left(-\frac{(a-a^{(k)})^2+(b-b^{(k)})^2}{2\gamma^2(k)}\right),$$

où $\phi\,(.)$ est la fonction de répartition de la loi normale standard.

Garthwaite et al. [5] donnent la formule de mise-à-jour du paramètre

$$\gamma_{(k+1)} = \begin{cases} \gamma_{(k)} + \frac{\gamma_{(k)}}{(k)\tau}, & \text{si } \tilde{h} \text{ est accepté,} \\ \gamma_{(k)} - \frac{\gamma_{(k)}}{(k)(1-\tau)}, & \text{si } \tilde{h} \text{ est rejeté,} \end{cases} \tag{2.5}$$

où τ est le taux d'acceptation optimal, qui vont 0.44 dans le cas univarié voir Gelman et al. [7], Roberts and Rosenthal [14] et Garthwaite et al.[5].

Conclusion

Dans ce chapitre, après avoir présenté les notions de base et le principe de l'inférence bayésienne, nous avons mis en évidence la démarche à suivre pour sa mise en œuvre dans l'estimation du paramètre de lissage dans le cadre d'estimation à noyau d'une densité de probabilité en générale et d'une densité conditionnelle en particulier.

A ce stade, il reste à vérifier et à analyser l'apport de cette technique sur les performances de l'estimateur par rapport aux techniques classiques (plug-in et validations croisées). Pour cela, une étude numérique basée sur des échantillons simulés sera présentée dans le chapitre 3.

Chapitre 3

Performances d'un estimateur à noyau d'une densité $f(y|x)$

Introduction

L'objectif du présent Chapitre est de comparer les performances de l'estimateur à noyau d'une densité de probabilité conditionnelle lorsque y a aucune hypothèse sur les paramètres de lissage a et b avec les performances du même estimateur lorsque l'hypothèse d'égalité des deux paramètres de lissage a et b est imposée, c'est-à-dire $a = b = h$. Nous nous somme intéressé principalement à la comparaison de l'ISE des deux estimateurs en question ainsi qu'au temps de calcul.

3.1 Présentation des paramètres de l'application

Dans ce qui suit, afin de ne pas confondre les deux estimateurs nous allons utiliser les notations suivante : f_{ab} dans le cas a et b sont indépendants et f_h dans le cas $a = b = h$. Pour rependre à notre objectif nous avons implémenté un programme sous Matlab dont ses principales étapes sont comme suit :

1. Générer m échantillons (X, Y) de taille n d'une loi cible.

2. Estimer, respectivement, (a^*, b^*) et h^* qui minimise l'ISE moyenne.

3. Calculer f_{ab} et f_h et comparer leurs performances.

Pour implémenter l'étape 2 de cet algorithme sous Matlab, on est contraint à réaliser une discrétisation de la quantité définie dans la formule 1.10. Dans ce sens, pour l'erreur quadratique intégrée, nous proposons la discrétisation de suivante :

$$AISE = \frac{\Delta}{n} \sum_{j=1}^{N} \sum_{i=1}^{n} \left[\hat{f}(y'_j \mid X_i) - f(y'_j \mid X_i) \right]^2, \tag{3.1}$$

où $y' = \{y'_1, ..., y'_N\}$ est un vecteur de points équidistants dans l'espace de Y et $\Delta = y'_{i+1} - y'_i$.

En prenant en considération les m échantillons générer à l'étape 1, on peut définir le $AISE$ moyen comme suit :

$$AISE = \frac{\Delta}{n\ m} \sum_{k=1}^{m} \sum_{j=1}^{N} \sum_{i=1}^{n} \left[\hat{f}(y'_j \mid X_i^{(k)}) - f(y'_j \mid X_i^{(k)}) \right]^2, \tag{3.2}$$

et dans ce cas les paramètre de lissage optimales au sens du $AISE$ moyen seront défini, respectivement, par :

$$(a^*, b^*) = \arg\min_{(a,b)} \frac{\Delta}{n\ m} \sum_{k=1}^{m} \sum_{j=1}^{N} \sum_{i=1}^{n} \left[\hat{f}_{ab}(y'_j \mid X_i^{(k)}) - f(y'_j \mid X_i^{(k)}) \right]^2. \tag{3.3}$$

$$h^* = \arg\min_{h} \frac{\Delta}{n\ m} \sum_{k=1}^{m} \sum_{j=1}^{N} \sum_{i=1}^{n} \left[\hat{f}_h(y'_j \mid X_i^{(k)}) - f(y'_j \mid X_i^{(k)}) \right]^2. \tag{3.4}$$

Pour l'application numérique nous avons repris le premier exemple traités dans [2] et [1], c'est-à-dire nous allons considérer le modèle suivant :

$$Y = 10 + 5X + \varepsilon, \tag{3.5}$$

où X et ε sont deux variables aléatoires indépendantes issues d'une loi normale de para-

mètres $(10, 9)$ et d'une loi normale de paramètres $(0, 100)$, respectivement.

Ensuite, nous avons considéré le modèle suivant :

$$Y = 2\sin(\pi X) + \varepsilon, \tag{3.6}$$

où X et ε sont deux variables aléatoires. Tel que X est uniformément distribuée sur $[0, 2]$ et $\varepsilon_i \mid X_i = W_i N_i + (1 - W_i)M_i$ avec W_i est une variable binaire équiprobable c'est-à-dire $P(W_i = 1) = P(W_i = 0) = 0.05$, N_i suit une loi normale de paramètres $(X_i, 0.09)$ et M_i suit une loi normale de paramètres $(0, 0.09)$.

3.2 Résultats numériques et graphiques

Dans cet exemple, il est facile de démontré que la densité de la variable aléatoire Y sachant X est définie comme suite :

$$f(y \mid x) = \frac{1}{10}\phi\left(\frac{y - 10 - 5x}{10}\right), \tag{3.7}$$

avec, $\phi(.)$ est la densité d'une distribution d'une loi normale centrée réduite.

Pour l'application numérique nous avons considéré ce qui suit :

- Le noyau gaussien et le noyau d'Epanechnikov pour la construction de $\hat{f}(y \mid x)$.
- y' varie entre -10 et 130 avec un pas $140/24$ $(N = 25)$.
- 50 $(m = 50)$ échantillons de différentes tailles n.

Les résultats numériques obtenus sont résumés (rangés) dans les Tables 3.1 et 3.2 et sont présentés dans les Figures 3.1–3.2.

TABLE 3.1: Résultats numériques du cas de noyau Gaussien.

	$a \neq b$			$a = b$		
n	(a^*, b^*)	ISE	*temps* (mn)	h^*	ISE	*temps* (mn)
50	(0.9351, 7.3807)	0.0028	19.2417	2.3412	0.0071	5.2667
100	(0.8152, 6.5415)	0.0022	41.0333	1.9335	0.0055	9.7233
150	(0.7866, 6.0262)	0.0018	60.5250	1.7712	0.0045	16.8017
200	(0.7490, 5.6001)	0.0015	79.3617	1.6063	0.0037	25.1150
250	(0.7151, 5.4174)	0.0013	102.5750	1.5425	0.0034	34.8233
500	(0.6520, 4.7008)	0.0010	220.9018	1.3189	0.0025	72.8457
1000	(0.5824, 4.1568)	0.0007	511.1093	1.1516	0.0018	187.1294

TABLE 3.2: Résultats numériques du cas de noyau d'Epanechnikov.

	$a \neq b$			$a = b$		
n	(a^*, b^*)	ISE	*temps* (mn)	h^*	ISE	*temps* (mn)
50	(0.8877, 7.1378)	0.0029	2.3678	2.1058	0.0077	0.7190
100	(0.8024, 6.1525)	0.0019	5.4912	1.7854	0.0055	1.7917
150	(0.7744, 5.6500)	0.0018	8.0060	1.5944	0.0045	4.5634
200	(0.7299, 5.3290)	0.0015	14.5177	1.5447	0.0040	6.0449
250	(0.7054, 5.0261)	0.0013	18.5651	1.4397	0.0034	7.2891

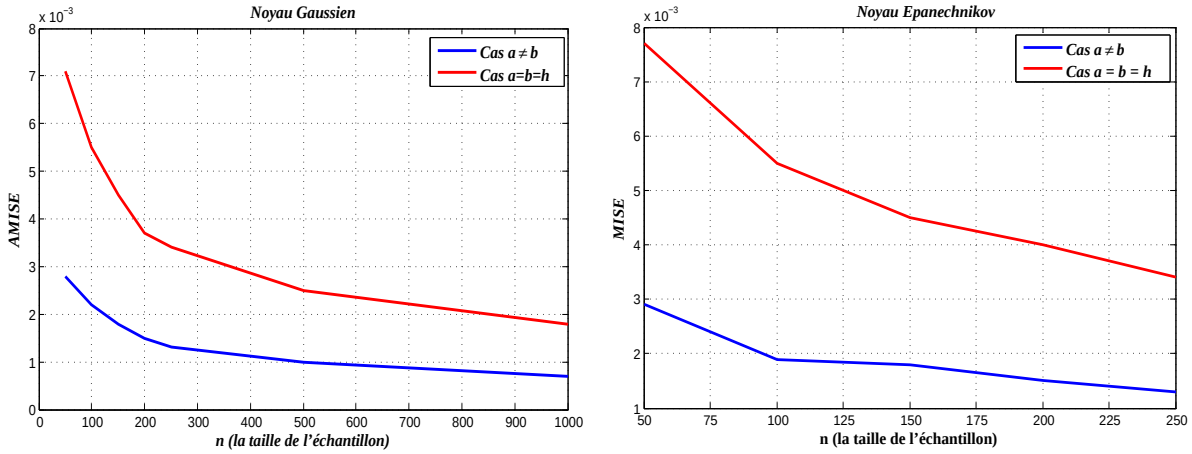

FIGURE 3.1 – Variation du $AISE$ moyen en fonction de la taille de l'échantillon.

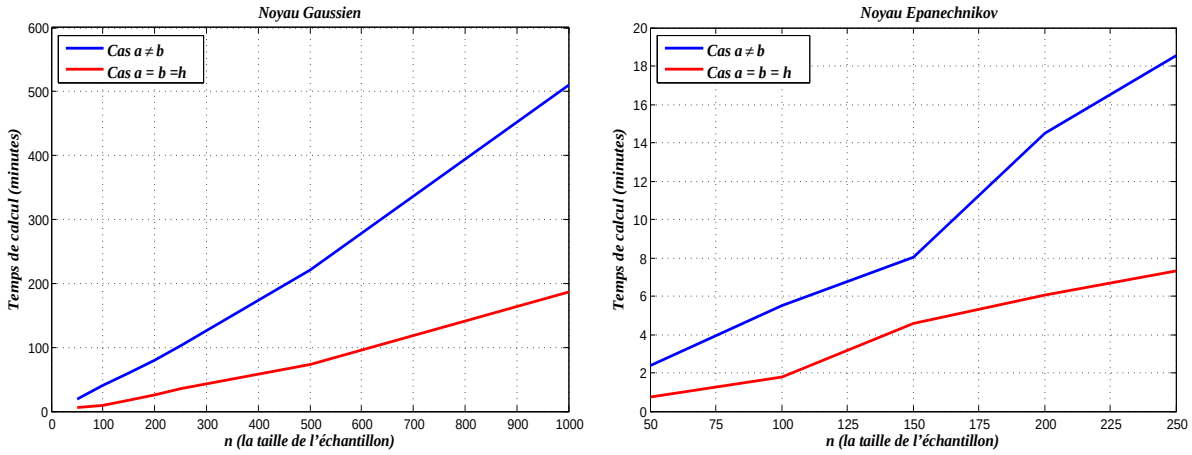

FIGURE 3.2 – Variation du temps de calcul en fonction de la taille de l'échantillon.

3.3 Discussion des résultats

Soit les deux hypothèses suivantes :

Première hypothèse : le paramètre de lissage dans la direction de x est indépendant du paramètre de lissage dans la direction de y, $a \neq b$.

Deuxième hypothèse : le paramètre de lissage dans la direction de x est le même que celui de la direction de y, $a = b = h$.

En tenant compte de ces deux hypothèse est des résultats obtenus dans l'exemples précédent on constate que :

- En fur et à mesure que la taille de l'échantillon augmente les deux paramètres de lissage considérés ainsi que leurs $AISE$ moyens associé décroît et tend vers zéro, ce qui confirme la convergence des deux estimateurs en norme quadratique intégrée lorsque la taille de l'échantillon tend vers l'infinie ($n \to \infty$).
- D'une part, l'estimateur le plus performant au sens du $AISE$ moyen est obtenus dans le cas de la première hypothèse. D'autre part, le temps de calcul est plus considérable dans le cas où la première hypothèse est imposée que dans le cas de la deuxième hypothèse. Cette dernière constatation est bien attendue préalablement à cause de la complexité algorithmique du cas multi-variables plutôt que dans le cas uni-varie. De plus, on constate clairement sur qu'un gain d'une précision d'ordre 10^{-3}, nécessite un excès de cinq heures du temps de calcul lorsque la taille de l'échantillon est fixée à $n = 1000$, ce qui parait très fastidieux.
- Le temps de calcul dépend d'une manière exponentielle de la taille de l'échantillon.
- Le temps de calcul lors de l'utilisation du noyau normale est plus considérable que dans le cas d'utilisation du noyau d'Epanechnikov cela peut être expliqué par la largeur de son support qui nécessite un temps supplémentaire pour le palier. De plus, les $AISE$ moyens engendrés par ces deux noyaux sont pratiquement les mêmes.

En guise de conclusion, le choix de l'hypothèse à imposer dans l'estimation à noyau d'une densité de probabilité conditionnelle dépend des objectifs de l'utilisateurs de l'estimateur en question.

Conclusion

Dans ce chapitre à travers d'une application numérique, basée sur la simulation, nous avons mis en relief le problème du choix du paramètre de lissage dans l'estimation à noyau d'une densité conditionnelle dans le cas d'égalité du paramètre de lissage de la direction de x et le paramètre de lissage de la direction y ($a = b$) et le cas contraire ($a \neq b$).
Les résultats numériquement obtenus dans cette étude indiquent que si l'utilisateur, de l'estimateur en question, s'intéresse à la précision il est préférable d'imposer l'hypothèse $a \neq b$ et d'utiliser le noyau Normal, et si l'utilisateur s'intéresse à réduire le temps de calcul il est préférable d'imposer l'hypothèse inverse c'est-à-dire $a = b = h$ et d'utiliser le noyau d'Epanechnikov.

Conclusion générale

La fonction de densité conditionnelle a un intérêt majeur dans différents domaines statistiques, et elle peut être considérée comme une généralisation à la foi de la densité de probabilité classique et de la régression. Dans ce mémoire nous avons considéré le choix du paramètre de lissage par l'inférence bayésienne dans l'estimation à noyau d'une fonction de densité conditionnelle de Y sachant que $X = x$ avec X et Y sont des variables aléatoires uni-variées.

En premier lieu, nous avons intéressé à l'estimation de la densité conditionnelle par la méthode du noyau. Après avoir défini l'estimateur de la densité conditionnelle, nous avons présenté ses propriétés ainsi que les conditions de convergence de cet estimateur en normes L^1 et L^2. Nous avons abordé également le problème de choix du noyau K et du paramètre de lissage (a, b). Nous sommes focalisés principalement sur le choix du paramètre de lissage le fait que le choix du noyau reste le même que dans l'estimation d'une densité classique, où nous avons distingué deux situations : le cas où les paramètres de lissage a et b sont considérés indépendants $(a \neq b)$ et le cas où a et b sont supposés égaux $(a = b = h)$.

En deuxième lieu, nous avons présenté l'approche bayésienne qui est une alternative pour remédier aux problèmes des techniques classique de sélection du paramètre de lissage. En effet, après avoir abordé le paradigme bayésien nous avons présenté la démarche à suivre pour la mise en œuvre de cette approche pour le choix du paramètre de lissage globale dans l'estimation à noyau d'une densité conditionnelle.

Enfin, nous avons présenté une application numérique basée sur des échantillons artificiels.

Le but de notre application est de comparer les performances de l'estimateur à noyau d'une densité conditionnelle lorsque le paramètre de lissage est sélectionné par l'inférence bayésienne par rapport à ceux conçu via les procédures classiques et cela dans le sens du temps de calcul et de l'erreur moyen associé aux estimateurs considérés.

Les résultats numériques obtenus sur des échantillons de différentes tailles en utilisant le noyau normale et le noyau d'Epanechnikov pour la construction de l'estimateur mettent en relief le problème de choix du paramètre de lissage dans l'estimation d'une densité conditionnelle dans les deux cas : $a \neq b$ et $a = b = h$.

Parmi les perspectives de ce travail, nous pouvons dégager plusieurs axes intéressants, tant sur le plan théorique que pratique :

- Réaliser une simulation extensive afin de confirmer nos résultats.
- Une étude extensive toute en considérant d'autres lois plus complexe (multi-dimensionnelle, multi-modal,...) et d'autres champ d'application.
- Considérer d'autres types de variables aléatoires : cas de variables dépendante, définie sur des supports bornée ou semi-borné...
- Considérer d'autres techniques de calcul bayésienne autres que le MCMC.
- De revoir ce travail pour d'autres choix : de lois à priori, la loi instrumentales,...

Bibliographie

[1] Y. Attayeb, (Juin 2015) *Estimation à noyau d'une fonction de densité de probabilité conditionnelle*, Mémoire de Master, Département de mathématiques, Université de Biskra.

[2] D. M. Bashtannykn and R. J.Hyndman, (2001). *Bandwidth selection for kernel conditional density estimation*. Computational statistics and data analysis **36** : 279–298.

[3] M. J. Brewer, (1998). *A modelling approach for bandwidth selection in kernel density estimation*. In : Proceedings of COMPSTAT Physica Verlag, Heidelberg, 203–208.

[4] L. Devroye, (1987). A Course in Density Estimation. Birkhauser, Boston.

[5] P. H. Garthwaite, Y. Fan, and S. A. Sisson. *Adaptive optimal scaling of metropolishastings algorithms using the robbins-monro process*. Working paper, (2010).

[6] A. Gelman and D. B. Rubin, (1992). *Inference from iterative simulation using multiple sequences. Statistical Science*, **7**, 457–511.

[7] A. Gelman, G. O. Roberts, and W. R. Gilks, (1996). *Efficient metropolis jumping rules*. Bayesian Statistics, **5**, 599–608.

[8] W. K. Hastings, (1970). Monte carlo sampling methods using Markov chains and their applications. Biometrika, 57(1), 97–109.

[9] R. J. Hyndman, D.J. Bashtannyk, and G. K. Grunwald (1996) *Estimating and visualizing conditional densities*. Journal of Computational and Graphical Statistics **5** : 315–336.

[10] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, (1953). *Equations of state calculations by fast computing machines*. Journal of Chemical physics, **21** 1087–1091.

[11] E. Parzen (1962) *On estimation of a probability density function and mode*. Ann. Math. Statist. **33** : 1065-1076.

[12] C. P. Robert, *Le choix bayésien(Principes et pratique)*. Springer-Verlag France, Paris (2006).

[13] C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer-Verlag, 2004.

[14] G. O. Roberts and J. S. Rosenthal, (2009). *Examples of adaptive MCMC*. Journal of Computational and Graphical Statistics, **18(2)**, 349–367.

[15] M. Rosenblatt (1956) *Remarks in some nonparametric estimates of a density function*. Ann. Math. Statist. **27** : 832–837.

[16] É. Youndjé (2011) *Contribution à l'estimation non-paramétrique par la méthode du noyau*. Mémoire d'Habilitation à Diriger des Recherches Spécialité : mathématiques appliquées, statistique. Universités de Rouen et du Havre, INSA de Rouen.

[17] N. Zougab (2013) *Approche Bayésienne dans l'estimation non paramétrique de la densité de probabilité et la courbe de régression de la moyenne*. Thèse Doctorat, Université de Béjaia, Algérie.

Résumé

Ce travail, porte sur la sélection du paramètre de lissage lors de l'estimation d'une densité de probabilité conditionnelle. Plus précisément, nous sommes intéressés à la construction d'une procédure de sélection de ce paramètre par l'inférence bayésienne.

Nous avons également réalisé une application numérique, sur des données simulées, qui nous a permet de mettre en évidence l'impact du choix des paramètres de l'estimateur à noyau, de la densité $f(y|x)$, sur ses performances.

Abstract

This work deals with the selection of the smoothing parameter when estimating a conditional probability density. More precisely, we are interested in the construction of a procedure for the selection of this parameter by Bayesian inference.

We also realized a numerical application, on simulated data, which enabled us to highlight the impact of the choice of the parameters of the kernel estimator, of a conditional density, on its performances.