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

CHAHEB Nadjette

Titre :

Estimation non paramétrique par noyaux associés multivariée et applications

Membres du Comité d'Examen :

Dr.	**BRAHIMI** Brahim	Président
Dr.	**BERKANE** Hassiba	Encadreur
Dr.	**DHIABI** Samra	Examinateur

2020

Dédicace

Je dédie ce modeste travail

A mon cher père

A ma cher mère

A mes soeurs et fréres

Qui m'ont toujours soutenu et encouragé

A mes camarades de promotion 2019/2020

A tous mes professeurs

$\mathcal{N}$adjette.

REMERCIEMENTS

Tout d'abord, nous tenons à remercier le "BON DIEU" le tout puissant de nous avoir accordé patience, courage et volonté afin de réaliser mener à terme ce modeste travail.

Je tiens à remercier tout particulièrement mon encadreur, Dr BERKANE Hassiba pour ses conseils, sa grande disponibilité et sa générosité. La pertinence de ses questions et de ses remarques ont toujours su me motiver et me diriger.

Je voudrais également remercier tous les membres de jury d'avoir accepté d'évaluer et d'examiner ce travail, merci pour toutes leurs remarques et critiques.

Je remercie chaleureusement toute ma famille, qui m'a soutenu, encouragé et poussé durant toutes mes années d'étude.

J'exprime mes remerciements à tous mes enseignants du département de Mathématiques.

Merci à tous

Table des matières

Table des figures

Liste des tableaux

Introduction

Un des problèmes habituellement rencontrés en statistique est celui de l'estimation de fonctions multivariés. Il s'agit d'un problème bien connu dans de nombreux domaines où l'on a souvent à modéliser des phénomènes complexes. On rencontre ce genre de questions en économie, médecine, sociologie, environnement, marketing, etc. Dans le cas multivarié, ces derniers sont souvent décrits par des vecteurs aléatoires à valeurs réelles et à support connu dans $\aleph_d \subseteq \mathbb{R}_d$ avec $d \in \{1, 2, ...\}$. Cet ensemble peut être formé d'axes univariés continus, discrets (à la fois discrets et continus). On peut se référer à Hilger (1990), Agarwal & Bohner (1999), Bohner & Peterson (2001, 2003) pour d'autres développements. Il faut noter que le support $\aleph_d$ est parfois relatif à des données fonctionnelles. Le lecteur pourra se référer, par exemple, à Cardot et al. (2015), D. Niang & Yao (2013) et Amiri et al. (2014) .

On considère une fonction multivariée f à estimer à partir des n observations du vecteur aléatoire réel ($v.a.r.$) dans $\aleph_d$ ($\subseteq \mathbb{R}_d$). Cette fonction f peut être une fonction de densité, de masse de probabilité (continue et/ou discrète). Puisqu'on ne dispose pas de manière générale de modèles paramétriques, cette mémoire propose l'approche non-paramétrique pour estimer la fonction f par la méthode des noyaux dits associés multivariés dans le cas d'une fonction de densité, l'estimateur à noyau associé multivarié est donné par :

$$\hat{f}_n(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathrm{x},H}(X_i) \qquad \forall x \in \aleph_d$$

où $K_{\mathrm{x},H}(X_i)$ sera le "noyau associé" lequel est intrinsèquement lié à la cible x et à la matrice

des fenêtres H. Les noyaux associés généralisent les "noyaux symétriques" de Rozenblatt (1956) et Parzen (1962). Ces derniers sont très populaires dans la littérature et approprié pour le support continu $\aleph_d = \mathbb{R}_d$.

L'objectif principal de cette mémoire est d'étendre au multivarié les estimateurs à noyaux associés univariés discrets de Senga Kiéssé (2008) et univariés continus de Libengué (2013). En multivarié, les noyaux associés doivent conserver leur aptitude à respecter le support $\aleph_d$ et aussi à atteindre n'importe quelle cible. Dans cet environnement multidimensionnel, on étudiera l'effet de la comparaison de les méthodes d'estimation du matrice de lissage H des fonctions de densité.du support $\aleph_d$.

Dans le premier chapitre, nous allons présenter les principales notions de l'estimation de la densité de probabilité par la méthode du noyau. Ensuite, ses propriétés (biais, variances,...) et les inconvénients du choix des deux paramétres qui constituent un estimateur a noyau, a savoir : le paramètre de lissage h et le noyau symétrique univarié K. Puis, on présente l'idée et la notion de l'estimateur à noyau asymétrique univarié (discrets et continus) dans le cas de variables défnies sur $\mathbb{R}_+$ et le cas de variables défnies sur $\mathbb{N}$. Par la suite, la question du choix de noyau et du paramètre de lissage ainsi que les propriétés des estimateurs conçu dans ce cadre sera présenté.

Dans le deuxième chapitre, on rappelle les principales notions de l'estimation de la densité de probabilité par la méthode du noyau. Ensuite, ses propriétés et les inconvénients du choix des deux paramètres qui constituent un estimateur a noyau symétrique multivariés, a savoir : la matrices de lissage H et le noyau symétrique multiivarié K. Puis, on présente l'idée et la notion de l'estimateur à noyau asymétrique multivarié dans le cas de variables défnies sur le support $\aleph_d \subseteq (\mathbb{R}_d$ ou $\mathbb{N}_d)$. Par la suite, la question du choix de noyau et du matrice de lissage H ainsi que les propriétés des estimateurs.

Le troisième chapitre est construire les résultats de simulation des différentes méthodes de sélection du paramètre ou matrice de lissage. Tous les résultats numériques et graphiques sont effectués à l'aide du logiciel R.

Chapitre 1

Estimation de densité univariée

Dans ce chapitre, nous avons présenté la notion d'une densité de probabilité 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 et expression (le biais, la variance,...). Ensuite, nous allons aborder le problème de choix du noyau et du paramètre de lissage dans ce cas.

1.1 Noyau symétrique

1.1.1 Noyau continu symétrique

Définition 1.1.1 *Soit un échantillon de variables aléatoires (v.a)* $X_1, X_2, ..., X_n$ *; indépendant et identiquement distribuée (i.i.d) de densité de probabilité continue inconnue* f *sur* $\mathbb{R}$ *.*

$$\hat{f}_n(x)=\frac{1}{nh}\sum_{i=0}^{n}K(\frac{x-X_i}{h}) = f_{n,h,K}(x)$$

oú : $x \in \mathbb{R}$ *,* $h > 0$ *où* K *est le noyau associé continu de cible* x *et de fenêtre* h *sur* $\mathbb{R}$*. qui vérifié les conditions suivantes :*

$$K(u) = \frac{1}{2}\mathbf{1}_{(|u|<1)} \ , \quad K(u) \geq 0 \quad et \quad \int_{\mathbb{R}} K(u)du = 1 \tag{1.1}$$

$h = h(n) > 0$ est une paramètre de lissage ou la fenêtre il dépend de n et il vérifie $h(n) \to 0$ lorsque $n \to \infty$.

Définition 1.1.2 *Soit X une v.a de densité $f(x)$ sur $\mathbb{R}$: $(X_1, X_2, ..., X_n)$ un échantillon issu de X , de fonction de répartition $F(x) = \int_{-\infty}^{x} f(t)dt$. On appelle fonction de répartition empirique associée à $X_1, X_2, ..., X_n$, la fonction aléatoire $F_n : \mathbb{R} \to [0; 1]$ définie, pour tout $x \in \mathbb{R}$, par :*

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(X_i \leq x)} \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{(]-\infty;x[)}$$

À partir de la définition d'une densité de probabilité, on aura :

$$\hat{f}_n(x) = \lim_{h \to 0} \frac{F_n(x+h) - F_n(x-h)}{2h}$$

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - X_i}{h}) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) \tag{1.2}$$

où $K_h(x - X_i) = \frac{1}{h}K(\frac{x-X_i}{h})$

L'expression *(1.3)* découle des travaux des pionniers de l'estimation non paramétrique *Rosemblet (1956)*, puis *Parzon (1962)*. Généralement, la fonction K appelée noyau est une fonction positive et bornée satisfaisant les conditions suivantes :

$$\int_{\mathbb{R}} K(u)du = 1 \ , \quad \int_{\mathbb{R}} uK(u)du = 0 \quad \text{et} \quad \int_{\mathbb{R}} u^2K(u)du = \mu_2 < \infty \tag{1.3}$$

Les conditions données dans la formule (1.4) permettent de prouver plusieurs types de convergence *(locale et globale)* de l'estimateur définie dans la formule (1.3), voir Silverman.

Le tableau *(1.1)* donne un récapitulatif des fonctions noyaux continues classiques :
Pour plus de détails sur les types des noyaux, nous pouvons se référer à l'article d'*Epanechnikov (1969)* et le livre de *Tsybakov (2004)*.
Les courbes de ces noyaux sont présentées ci-dessous :

Noyau	Fonction noyau	Domaine de définition	Efficacité
Epanchinkov	$\frac{3}{4}(1-u^2)$	$[-1,1]$	1.000
Cosinus	$\frac{\pi}{4}\cos\left(\frac{\pi u}{2}\right)$	$[-1,1]$	0.999
Biweight	$\frac{15}{16}(1-u^2)^2$	$[-1,1]$	0.994
Triweight	$\frac{35}{32}(1-u^2)^3$	$[-1,1]$	0.987
Triangulaire	$1-\lvert u\rvert$	$[-1,1]$	0.986
Gaussian	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}$	$\mathbb{R}$	0.946
Uniforme	$\frac{1}{2}$	$[-1,1]$	0.930

TAB. 1.1 – Exemples de noyaux continus symétriques univariés.

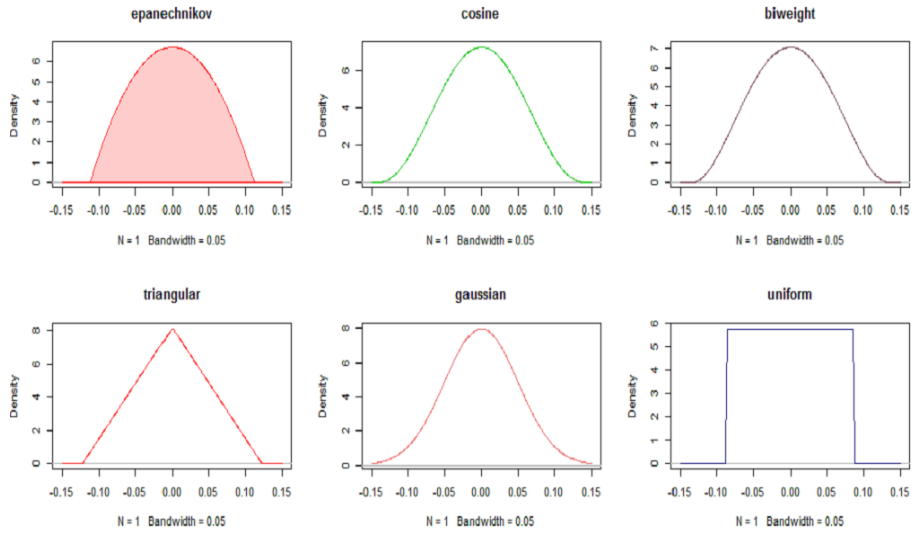

FIG. 1.1 – La Forme des noyaux usuels.

Lemme 1.1.1 *Si le noyau K est positif et $\int_{\mathbb{R}} K(u)du = 1$ alors, l'estimateur $\hat{f}_n$ est de densité de probabilité $\hat{f}_n$ est continu si K est continu.*

Propriétés de l'estimateur à noyau symétrique

Juste après introduction de l'estimateur à noyau de la densité par *Rosenblatt (1956)*, *Parzen (1962)* à étudié ses propriétés fondamentales. Depuis, cet estimateur est devenu un objet classique étudié par les statisticiens. Les expressions de l'espérance, biais et de la variance de l'estimateur à noyau sont données respectivement par :

Esperance de l'estimateur L'espérance mathématique de l'estimateur à noyau est définie par :

$$E\left[\hat{f}_n(x)\right] = E\left[\frac{1}{nh}\sum_{i=0}^{n} K(\frac{x-X_i}{h})\right] = f(x) + \frac{h^2}{2} f"(x)\mu_2(K) + \circ(h^2) \tag{1.4}$$

où $\mu_2 = \int_{\mathbb{R}} u^2 K(u)du$.

Biais de l'estimateur Le biais de l'estimateur $\hat{f}_n$ est :

$$Biais\left[\hat{f}_n(x)\right] = f(x) + \frac{h^2}{2} f"(x)\mu_2(K) + \circ(h^2) - f(x) = \frac{h^2}{2} f"(x)\mu_2(K) + \circ(h^2) \tag{1.5}$$

Variance de l'estimateur Soit x fixé dans $\mathbb{R}$. La variance de l'estimateur $\hat{f}_n$ est définie par :

$$Var\left[\hat{f}_n(x)\right] = Var\left[\frac{1}{nh}\sum_{i=0}^{n} K(\frac{x-X_i}{h})\right] = \frac{f(x)}{nh} R(K) + \circ(\frac{1}{nh}) \tag{1.6}$$

oú : $R(K) = \int_{\mathbb{R}} K^2(y)dy$.

Erreur quadratique moyenne MSE L'erreur quadratique moyenne(en anglais "*Mean Squard Error*") en un point x fixé ; s'exprimer par :

$$MSE\left(\hat{f}_n(x)\right) = E\left\{\left[\hat{f}_n(x) - f(x)\right]^2\right\} = \frac{1}{nh} f(x)R(K) + \frac{h^4}{4}\left[f"(x)\right]^2 \mu_2^2(K) + \circ(h^4 + \frac{1}{nh})$$

$$= Var\left[\hat{f}_n(x)\right] + Biais^2\left[\hat{f}_n(x)\right]$$

oú : $R(K) = \int_{\mathbb{R}} K^2(y)dy$, $\mu_2 = \int_{\mathbb{R}} u^2 K(u)du$.

Erreur quadratique moyenne intégrée (MISE)

Propriété 1.1.1 *L'erreur quadratique moyenne intégrée ("*Mean Intégrated Squard Error*") est la mesure théorique commune la plus utilisée pour évaluer l'erreur entre la fonc-*

tion f et $\hat{f}_n$. Il est également convenable d'évaluer l'erreur globale *sur le support $\mathbb{R}$ de cet estimateur.*

$$MISE(\hat{f}_n) = \int_{\mathbb{R}} MSE(\hat{f}_n)dx = \frac{1}{nh}R(K) + \frac{h^4}{4}R(f"(x))\mu_2^2(K) + \circ(h^4 + \frac{1}{nh})$$

oú : $R(g) = \int_{\mathbb{R}} g^2(y)dy$ pour une fonction g carrée intégrable. Alors $R(K) = \int_{\mathbb{R}} K^2(y)dy$, $\mu_2 = \int_{\mathbb{R}} u^2K(u)du$.

Pour plus de détails voir Silverman [14]

1.1.2 Comportement asymptotique de l'estimateur de noyau

Parzen a élaboré les conditions de plusieurs types de convergence de l'estimateur à noyau ainsi que la convergence de ses propriétés. Une approximation asymptotique de l'espérance de l'estimateur $\hat{f}_n$ est donnée sous les conditions suivantes sur f, h et K. Les principaux résultats obtenus, par l'auteur, sont résumés comme suivant :

1. La fonction de densité f admet la dérivée seconde $f"$ qui peut être une fonction absolument continue.
2. Le paramètre de lissage h est positif $(h > 0)$ et on suppose que h satisfait :

$$\lim_{n\to\infty} h = 0 \text{ et } \lim_{n\to\infty} nh = \infty$$

3. Pour que f soit une densité on suppose que $(K(u) \geq 0)$ bornée, densité $\left(\int_{\mathbb{R}} K(u)du = 1\right)$, symétrique autour de zéro $(K(-x) = K(x))$, soit $\left(\int_{\mathbb{R}} uK(u)du = 0\right)$ et possède un moment d'ordre 2 fini, soit $\left(\int_{\mathbb{R}} u^2K(u)du < +\infty\right)$.
4. Pour que f est une densité bornée dont la dérivée seconde est bornée, alors on a :

$$\left|biais\left(\hat{f}_n(x)\right)\right| \leq c_1h^2 \qquad var\left(\hat{f}_n(x)\right) \leq \frac{c_2}{nh}$$

où c_1 et c_2 est une constantes.

En utilisant les expression (1.3) et (1.4) ; nous permettent de trouver des expressions asymptotiques pour MSE et $MISE$ on note l'approximation asymptotique du MSE par :

$$AMISE(\hat{f}_n) = \frac{1}{nh}R(K) + \frac{h^4}{4}R(f"(x))\mu_2^2(K)$$

oú : $R(g) = \int_{\mathbb{R}} g^2(y)dy$ pour une fonction g carrée intégrable, $\mu_2 = \int_{\mathbb{R}} u^2K(u)du$.

1.1.3 Noyau optimal basé sur le critère de MISE

Choix de noyau

Pour mesurer l'efficacité de chacun des noyaux continus symétriques présenté dans le *tableau 1.1. Epanchinkov (1969)* exhibe un noyau optimal sous ces contraintes en minimisant $MISE$. Le biais est une fonction croissante en h alors que le terme en variance est une fonction décroissante en h si h est grand la variance sera petite et le biais sera fort, donc ; la valeur optimale de h qui minimise le $MISE$ réalise un compromis entre le biais et la variance. Alors pour calculer le h optimal comme suit :

$$\frac{\partial AMISE}{\partial h}(h) = 0 \Longrightarrow h^5 = \frac{R(K)}{nR(f"(x))\mu_2^2(K)} \Longrightarrow h^* = \left[\frac{R(K)}{R(f"(x))\mu_2^2(K)}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

oú : h^* : est h optimale, $R(K) = \int_{\mathbb{R}} K^2(y)dy$.

En substituant h^* dans la formule $AMISE$,on obtient :

$$AMISE_{opt}(h^*) = \frac{5}{4}R^{\frac{4}{5}}(K)\mu_2^{\frac{2}{5}}(K)R^{\frac{1}{5}}(f"(x))n^{-\frac{4}{5}}$$

On pose :

$$M(K) = R^{\frac{4}{5}}(K)\mu_2^{\frac{2}{5}}(K) = \left[R^4(K)\mu_2^2(K)\right]^{\frac{1}{5}}$$

Puisque l'on a aucune information sur $f"(x)$ le paramètre de lissage à été déjà optimiser, alors pour minimiser le $AMISE$, il faut choisir le noyau K qui minimise la valeur de $M(K)$. On objectif est donc de minimiser $M(K)$, Ce qui est équivalent à minimiser $\int_{\mathbb{R}} K^2(u)du$.

Soient deux noyaux continus symétriques fixés K_1 et K_2. Le critère utilisé pour mesurer l'efficacité relative de ces deux noyaux est le suivant :

$$Eff(K_1, K_2) = \frac{M(K_1)}{M(K_2)} = \left[\frac{AMISE(K_1)}{AMISE(K_2)}\right] = \left[\frac{R^4(K_1)\mu_2^2(K_1)}{R^4(K_2)\mu_2^2(K_2)}\right]^{\frac{1}{5}}$$

Ainsi, l'efficacité d'un noyau continu symétrique K par rapport au noyau optimal d'*Epanechnikov (1969)* K_E est donnée par :

$$Eff(K) = \left[\frac{R^4(K_E)\mu_2^2(K_E)}{R^4(K)\mu_2^2(K)}\right]^{\frac{1}{5}} = \frac{3}{5\sqrt{5}} \frac{1}{\sqrt{\int_{\mathbb{R}} u^2K(u)du} \int_{\mathbb{R}} K^2(u)du} \leq 1.$$

Choix de fenêtres

Dans cette partie, Elles comparent plusieurs méthodes pour choisir le paramètre de lissage pour plusieurs distributions différentes. Toutes ces méthodes nous donnent un paramètre de lissage qui est optimale pour la distribution à estimer, on va étudier les méthodes suivantes :

Estimateur Rule Of Thumb (règle de référence) L'estimateur "*Rule of Thumb*" du paramètre de lissage, noté h_r suppose que nous utilisons le noyau gaussien pour estimer une densité f d'une distribution normale centrée la moyenne $\mu = 0$ et de variance σ^2. De se fait :

$$R(f"(x)) = \int_{\mathbb{R}} f"^2(x)dx = \frac{3}{8}\sqrt{\pi}\sigma^{-5}$$

En substituant $R(f")$ et K noyau gaussien par leur formule, on obtient :

$$h_{opt} = \left(\frac{1}{2\sqrt{\pi}}\right)^{\frac{1}{5}} \left(\frac{3}{8\sigma^5}\sqrt{\pi}\right)^{-\frac{1}{5}} n^{-\frac{1}{5}} = 1.06\sigma n^{-\frac{1}{5}}$$

Il suffit donc d'estimer σ à partir des données et de substituer cet estimateur dans la formule ci-dessus. D'aprés *Silverman(1986)* ; cette formule donnera de bon résultat si la population est réellement normalement distribuées, mais ; celle-ci peut donner une distribution trop lissée si la population est plutôt multimodale.

Dans ce cas de meilleurs résultats peuvent être obtenus, si on utilise l'interquartile :

$$IQR = X_{\left[\frac{3}{4}n\right]} - X_{\left[\frac{1}{4}n\right]}$$

Dans le cas où X suit une loi normale, alors h devient :

$$h_{opt} = 1.06 \left(\frac{IQR}{1.34}\right) n^{-\frac{1}{5}} \approx 0.79 IQR n^{-\frac{1}{5}}$$

Enfin ; la fenêtre optimale est :

$$h_{opt} = 1.06 \min\left\{\sigma, \frac{IQR}{1.34}\right\} n^{-\frac{1}{5}},$$

cette correction est insuffisante dans de nombreux cas.si la vraie densité est multimodale.

Méthodes de validation croisée (Cross validation) CV L'idée de bas des méthodes de CV consiste à trouver une fonction de score $CV(h)$ ayant la même structure que le $MISE$ et dont le calcul soit plus simple. On sélectionne alors la fenêtre h minimisant ce critère dont on attend le même comportement asymptotique que h^*.

Validation croisée non biaisée (UCV) Cette méthode a été proposée par *Rudemo(1982)* et *Bowman(1984)*. Elle consiste à choisir le paramètre de lissage qui minimise un estimateur convenable de :

$$UCV(h) = \int_{\mathbb{R}} \left[\hat{f}_n(x) - f(x)\right]^2 dx - \int_{\mathbb{R}} f^2(x)dx = \int_{\mathbb{R}} \hat{f}_n^2(x)dx - 2\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx$$

Puisque $\int_{\mathbb{R}} f^2(x)dx$ ne dépend pas du paramètre de lissage h. On peut choisir alors le paramètre de lissage de façon à ce qu'il minimise un estimateur de :

$$\int_{\mathbb{R}} \hat{f}_n^2(x)dx - 2\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx.$$

Maintenant ; on veut trouver un estimateur de $\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx$. Pour cela remarquons que :

$$\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx = E\left(\hat{f}_n(x)\right).$$

L'estimateur empirique de $\int_{\mathbb{R}} \hat{f}_n(x)f(x)dx$ est alors $\frac{1}{n}\sum_{i=1}^{n} \hat{f}_{n,i}(x_i)$.
Et le critère à minimiser est :

$$UCV(h) = \int_{\mathbb{R}} \hat{f}_n^2(x)dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,i}(x_i)$$

avec $\hat{f}_{n,i}(x_i) = \frac{1}{(n-1)h}\sum_{j=1,\, j\neq i}^{n} K(\frac{x_i - x_j}{h})$, est l'estimateur de la densité construit à partir de l'ensemble de points sauf au point x_i. Nous notons par h_{UCV} l'estimateur de h qui minimise $UCV(h)$.

L'optimalité asymptotique de la validation croisée non biaisée a été obtenue par *Stone*. Cependant, cette méthode présente deux problèmes majeurs. Pour d'autres études voir *Hall, Burman, Scott et Terrel.*

Validation croisée biaisée (BCV) Le critère de validation croisée biaisée, a été introduit par *Scott* et *Terrel (1987)* pour remédier aux problèmes de la méthode "validation croisée biaisée". Il s'agit d'introduire un biais le UCV afin de réduire sa variance. Le paramètre de lissage basé sur la méthode de BCV est la valeur de h qui minimise un estimateur du $AMISE$. il est claire que afin d'estimer l'$AMISE$, il suffit d'estimer $R(f")$. Un estimateur naturel de ce dernier terme est donnée par $R(\hat{f}")$.

Finalement, *Scott* et *Terrel* ont proposé la forme de l'estimateur de $AMISE$ à minimiser

qui se résume comme suit :

Proposition 1.1.1 *(**Scott** et **Terrel**) Soit $X_1, X_2, ..., X_n$ un n-échantillon i.i.d issu d'une variable aléatoire X de fonction de densité f. Pour un noyau K, on obtient :*

$$BCV(h) = \frac{R(K)}{nh} + \frac{h^4}{4n^2}\mu_2^2(K)\sum_i \sum_{j,j\neq i} K_h^{(2)} K_h^{(2)} (X_i - X_j)$$

Des résultats de simulation ont été obtenus pour la méthode de BCV dans le travail de *Park* et *Marron*.

Validation croisée par le maximum de vraisemblance (LCV) Pour un estimateur à noyau f_h de f défini dans l'équation *(1.1)* et de largeur h, la sélection par validation croisée de la vraisemblance est une approche classique. C'est *Habbema*, *Hermans et V andenbroek* en *(1974)* qui ont proposé cette méthode fondée sur un critère non asymptotique du maximum de vraisemblance. Le paramètre de lissage basé sur la méthode de LCV est le paramètre qui maximise la fonction suivante :

$$LCV(h) = \prod_{i=1}^{n} \hat{f}_{n,i}(x_i) = \prod_{i=1}^{n} \frac{1}{(n-1)\,h} \sum_{\substack{j=1 \\ j\neq i}}^{n} K(\frac{x_i - x_j}{h})$$

En utilisant un noyau gaussien on obtient :

$$LCV(h) = \prod_{i=1}^{n} \frac{1}{(n-1)\,h} \sum_{\substack{j=1 \\ j\neq i}}^{n} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_i - x_j)^2}{2h^2}\right)$$

1.2 Noyau asymétrique

Les définitions suivantes présentent les notions du noyau associé, et de l'estimateur à noyau associé pour la fonction de densité f inconnue sur le support $\aleph$.

Définition 1.2.1 *Soit $x \in \aleph$ et $h > 0$. On appelle noyau associé $K_{x,h}$ toute densité de probabilité liée à une variable aléatoire continue ou discrète $\mathcal{K}_{x,h}$ de support $\aleph_{x,h}$. vérifiant*

les quatres conditions suivantes :

$$\aleph_{x,h} \cap \aleph \neq \phi \tag{1.7}$$

$$\underset{x}{\cup} \aleph_{x,h} \supseteq \aleph \tag{1.8}$$

$$\lim_{h\to 0} E\left(\mathcal{K}_{x,h}\right) = x \tag{1.9}$$

$$\lim_{h\to 0} Var\left(\mathcal{K}_{x,h}\right) = 0 \tag{1.10}$$

1.2.1 Noyau associée continu asymétrique

L'estimateur à noyau associée continu asymétrique est approprié pour estimer des densités à support compact et bornées. Soit $X_1, X_2, ..., X_n$ un échantillon de variables aléatoires $(i.i.d)$ de densité de probabilité continue inconnue f à support $\aleph = [a; b]$, avec $(a \in \mathbb{R}$ et $b \in \mathbb{R})$. De manière générale, l'estimateur à noyau continu est de la forme :

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{x,h}(X_i) = f_{n,h,K}(x)$$

où $x \in \aleph$ fixé, $h \in \mathbb{R}; (h > 0)$: est le paramètre de lissage, $K_{x,h}$ est associée à noyau continu asymétrique vérifié : $K(u) \geq 0 \qquad K_{x,h}(\cdot) = \frac{1}{h}K(\frac{x-\cdot}{h}) \qquad$ et $\int_{\mathbb{R}} K(u)du = 1$

Remarque 1.2.1 *Un noyau symétrique est aussi vérifiée la définition du noyau associé asymétrique.*

Dans cette partie, nous donnons les différences propriétés fondamentales de l'estimateur à noyau associé.

Où : $IG(a, b)$,$RIG(a, b)$ sont les lois gaussien inverse et gaussien inverse réciproque respectivement, et $\Gamma(a) = \int_0^{+\infty} t^{a-1}e^{-t}dt$ et $\beta(a, b) = \int_0^1 (1-t)^{b-1}t^{a-1}dt$, $a, b \in \mathbb{R}_+^*$.

Propriété 1.2.1 *Soit $X_1, X_2, ..., X_n$ un échantillon de variables aléatoires (i.i.d) d'une densité de probabilité continue inconnue f de support $\aleph$. Soit $\hat{f}_n$ l'estimateur de f à noyau*

Noyau	$\aleph_{x,h}$	$K_{x,h}(u)$	**Espérance**	**Variance**
$Gamma(a,b)$	$\mathbb{R}_+$	$\frac{1}{\Gamma(a)b^a}u^{a-1}e^{-\frac{u}{a}}$	ab	ab^2
$B\hat{e}ta(a,b)$	$[0;1]$	$\frac{1}{\beta(a,b)}u^{a-1}(1-u)^{b-1}$	$\frac{a}{(a+b)}$	$\frac{a}{\left\{(a+b)^2(a+b+1)\right\}}$
$IG(a,b)$	$\mathbb{R}_+$	$\frac{\sqrt{b}}{\sqrt{2\pi u^3}}\exp\left\{-\frac{b}{2a}\left(\frac{u}{a}-2+\frac{a}{u}\right)\right\}$	a	$\frac{a^3}{b}$
$RIG(a,b)$	$\mathbb{R}_+$	$\frac{\sqrt{b}}{\sqrt{2\pi u}}\exp\left\{-\frac{b}{2a}\left(au-2+\frac{1}{au}\right)\right\}$	$\frac{1}{a}+\frac{1}{b}$	$\frac{1}{ab}+\frac{2}{b^2}$

TAB. 1.2 – Qulques noyaux continus asymétriques.

continu asymétrique $K_{x,h}$ de variable aléatoire $\mathcal{K}_{x,h}$ sur le support $\aleph_{x,h}$ alors : $\forall x \in \aleph$ et $h > 0$ on a :

$$E\left[\hat{f}_n(x)\right] = E\left[f(\mathcal{K}_{x,h})\right].$$

Propriété 1.2.2 *On présente le développement limite de* Taylor-Lagrange *à l'ordre* 2*et au point moyen de la variable aléatoire* $E\left(\mathcal{K}_{x,h}\right) = \mu_{x,h}$ *tel que :*

$$f(\mathcal{K}_{x,h}) = f(\mu_{x,h}) + (\mathcal{K}_{x,h} - \mu_{x,h})\, f'(x) + \frac{(\mathcal{K}_{x,h} - \mu_{x,h})^2}{2!} f"(x) + \circ(h^2). \tag{1.11}$$

En calculant, l'espérance de cette quantité, on obtenue :

$$E\left[f(\mathcal{K}_{x,h})\right] = f\left\{E\left(\mathcal{K}_{x,h}\right)\right\} + \frac{1}{2} Var(\mathcal{K}_{x,h}) f"(x). \tag{1.12}$$

Propriétés de l'estimateur à noyau asymétrique

Biais ponctuel Pour un x fixé, On calcule le biais de l'estimateur à noyau associé continu asymétrique de manière générale.

Propriété 1.2.3 *Soit x fixé de $\aleph$, on a :*

$$Biais\left[\hat{f}_n(x)\right] = E\left[\hat{f}_n(x)\right] - f(x) = \left[f\left\{E\left(\mathcal{K}_{x,h}\right)\right\} - f(x)\right] + \frac{1}{2} Var(\mathcal{K}_{x,h}) f"(x). \tag{1.13}$$

Variance ponctuel Pour un x fixé, On généralise de l'expression de la variance de $\hat{f}_n$.

Propriété 1.2.4 *Soit x fixé de $\aleph$, on a :*

$$Var\left[\hat{f}_n(x)\right] = \frac{1}{n}\int_{\aleph_{x,h}\cap\aleph} K^2_{x,h}(y)f(t)dt - \frac{1}{n}\left[Biais\left[\hat{f}_n(x)\right] + f(x)\right]^2.$$

Erreur quadratique moyenne intégrée (MISE)

Propriété 1.2.5 *En sommant sur l'intersection de des supports le $MISE$ est :*

$$MISE\left(\hat{f}_n(x)\right) = \int_{\aleph_{x,h}\cap\aleph} Biais^2\left[\hat{f}_n(x)\right]dx + \int_{\aleph_{x,h}\cap\aleph} Var\left[\hat{f}_n(x)\right]dx.$$

On suppose dans toute la suite que f admet une dérivée seconde continue sur le support $\aleph$ et que les termes suivants sont finis : $\int_{\aleph}\left[f'(u)\right]^2 du$, $\int_{\aleph}\left[xf"(u)\right]^2 du$ et $\int_{\aleph}\left[x^3 f"(u)\right]^2 du$.

Exemple 1.2.1 *Le noyau associé gamma été introduire par* Chen (2000) *pour estimer des densités à support $\aleph = \mathbb{R}_+$ Il a utilisé la loi gamma pour construire des noyaux associés continus asymétriques. Deux classes de noyaux ont été proposées. La premiére classe des noyaux gamma est :*

$$K_{\mathcal{G}(\rho_h(x);h)}(u) = \frac{u^{\frac{x}{h}}e^{-\frac{u}{h}}}{\Gamma(\rho_h(x))h^{\rho_h(x)}}$$

où $\Gamma(\alpha) = \int_0^\infty e^{(-t)}t^{\alpha-1}dt$, $\alpha > 0$ est la fonction gamma et h est le paramètre de lissage satisfaisant les conditions $h \to 0$ et $nh \to \infty$ quand $n \to \infty$ où :

$$\rho_h(x) = \begin{cases} \dfrac{x}{h} & si \quad x \geq 2h \\ \\ \frac{1}{4}\left(\dfrac{x}{h}\right)^2 + 1 & si \quad x \in [0, 2h[\end{cases}$$

Le premier estimateur à noyau gamma est donnè par :

$$\hat{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathcal{G}(\frac{x}{h}+1;h)}(X_i).$$

L'estimateur à noyau gamma modifié $\tilde{f}_n(x)$ est notée :

$$\tilde{f}_n(x) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathcal{G}(\rho_h(x);h)}(X_i).$$

Le biais asymptotique de $\tilde{f}_n(x)$ est exprimé par la formule suivante :

$$Biais\left[\tilde{f}_n(x)\right] = \begin{cases} \dfrac{h}{2} x f"(x) & si \quad x \geq 2h \\ \\ h\theta_h(x) f'(x) & si \quad x \in [0, 2h[\end{cases}$$

où : $\theta_h(x) = \dfrac{(1-x)\left(\rho_h(x) - \frac{x}{h}\right)}{(1 + h\rho_h(x) - x)}$.
La variance asymptotiques de $\tilde{f}_n(x)$ est donné par :

$$Var\left[\tilde{f}_n(x)\right] = n^{-1} \frac{h^{-1}\Gamma(\frac{2x}{h}+1)}{2^{\frac{2x}{h}+1}\Gamma^2(\frac{x}{h}+1)} f(x) + \circ\left(n^{-1}\right).$$

Le $MISE$, le h optimal (au sens du MISE) ainsi que le MISE associe a ce dernier correspondants aux deux noyaux sont comme suit :
On mesure ainsi le $MISE$:

$$MISE\left[\tilde{f}_n(x)\right] = \frac{1}{4h^2}\int_0^{+\infty} \left[x f"(x)\right]^2 dx + \frac{1}{2n\sqrt{h\pi}}\int_0^{+\infty} x^{-\frac{1}{2}} f(x) dx + \circ(\frac{1}{n\sqrt{h}} + h^2)$$

Paramètre de lissage optimal :

$$h^* = \left[\frac{\frac{1}{2\sqrt{\pi}}\int_0^{+\infty} x^{-\frac{1}{2}} f(x) dx}{4\int_0^{+\infty}\left[x f"(x)\right]^2 dx}\right]^{\frac{2}{5}} n^{-\frac{2}{5}}$$

$MISE$ Optimal :

$$MISE_{opt}(h^*) = \frac{5}{4^{\frac{4}{5}}}\left[\int_0^{+\infty}\left[x f"(x)\right]^2 dx\right]^{\frac{1}{5}}\left[\frac{1}{2\sqrt{\pi}}\int_0^{+\infty} x^{-\frac{1}{2}} f(x) dx\right]^{\frac{4}{5}} n^{-\frac{4}{5}}$$

1.2.2 Choix paramètre de lissage

Méthode Plug-In

Méthode plug-in est à minimiser de deux dominant terms en le $MISE$ de $\hat{f}_n(x)$. Hirukawa(2010) défini le paramètre de lissage plug-in comme suit (pour modifié beta) :

$$h_{MB} = \arg\min_h \left[\frac{b^2}{4} \int_0^1 x^2 \left(1 - x^2\right) \left\{ f_\theta^{(2)}(x) \right\}^2 \upsilon(x)\, dx + \frac{1}{nb^{\frac{1}{2}} 2\sqrt{\pi}} \int_0^1 \frac{f_\theta(x)}{\sqrt{x\,(1-x)}} \upsilon(x)\, dx \right]$$

$$h_{MB} = \arg\min_h \left(MISE \left[\hat{f}_{MB}(x) \right] \right)$$

$$h_{MB} = \arg\min_h \left(MISE \left[\hat{f}_n(x) \right] \right)$$

Pour plus de détails voir : livre*(Masayuki-Hirukawa).*

Méthode de validation croisée

validation croisée (CV) est un autre méthode de choix paramètre de lissage. L'idée de CV est de trouver la valeur de h (minimiser l'erreur carré intégré ISE)

$$ISE\left(\hat{f}_n(x)\right) = \int \left\{ \hat{f}_n(x) - f(x) \right\}^2 dx$$

$$h = \arg\min_h ISE\left[\hat{f}_n(x)\right]$$

Remarque 1.2.2 *La fenêtre optimale dans le cas asymétrique est d'ordre* $\circ(n^{-\frac{2}{5}})$ *inferieur que dans le cas symétrique* $\circ(n^{-\frac{1}{5}})$.

En remplaçant h^* dans l'expression du $MISE$, on va :

$$MISE_{opt}(h^*) = \frac{n^{-\frac{4}{5}}}{4^{\frac{4}{5}}} \left[\int_0^{+\infty} f'(x) + \frac{1}{2} x \left(f"(x)\right)^2 dx \right]^{\frac{1}{5}} \left[\frac{1}{2\sqrt{\pi}} \int_0^{+\infty} x^{-\frac{1}{2}} f(x) dx \right]^{\frac{4}{5}}.$$

Noyau	$K_{x,h}(u)$	$\aleph_{x,h}$
Poisson	$\frac{(x+h)^u}{u!}e^{-(x+h)}$	$\mathbb{N}$
Binomial	$\frac{(x+1)!}{u!(x+1-u)!}\left(\frac{x+h}{x+1}\right)^u\left(\frac{1-h}{x+1}\right)^{x+1-u}$	$\{0,1,...,x+1\}$
Binomial négatif	$\frac{(x+u)!}{u!x!}\left(\frac{x+h}{2x+1+h}\right)^u\left(\frac{x+1}{2x+1+h}\right)^{x+1}$	$\mathbb{N}$
Dirac	$I_{(u=x)}$	$\mathbb{N}$
Triangulaire	$\frac{(a+1)^h-\lvert u-x\rvert^h}{P(a,h)}$	$\{x, x\pm 1, ..., x\pm a\}$

TAB. 1.3 – Quelques noyaux associés discrets en univarié

Dans le but de réduire le biais et par le suite d'erreur entre f et $\hat{f}_n$.

1.2.3 Noyau associée discret asymétrique

Définition 1.2.2 *Soit $X_1, X_2, ..., X_n$ un n-échantillon aléatoire (i.i.d) issu d'une variable aléatoire X de fonction de densité inconnue f sur $\aleph \subset \mathbb{R}$. L'estimateur à noyau associé discret $\hat{f}_n$ de f est défini par :*

$$\hat{f}_n(x)=\frac{1}{n}\sum_{i=1}^{n}K_{x,h}(X_i).$$

avec $x \in \aleph$, $h > 0$ où $K_{x,h}$ est le noyau associé continu ou discret de cible x et de fenêtre h sur $\aleph_{x,h}$.qui vérifié les conditions suivantes : $K_{x,h}(u) \geq 0$, $\sum_{u\in\aleph_{x,h}} K_{x,h}(u) = 1$.

Propriétés de l'estimateur à noyau discret

Dans cette section nous allons introduire la définition quelques propriétés de l'estimateur a noyau discret, qui ont été établis principalement par *Senga Kiessé* .

Propriété 1.2.6 *Soit $X_1, ..., X_n$ un n- échantillon* i.i.d *issu d'une variable aléatoire X de la fonction de mass de probabilité inconnue f sur $\mathbb{N}$, si $\hat{f}_n$ est l'estimateur a noyau asymétrique discret de f, alors, pour $x \in \mathbb{N}$ et $h > 0$ on a :*

$$E\left[\hat{f}_n(x)\right] = E\left[f(\mathcal{K}_{x,h})\right].$$

où : $\mathcal{K}_{x,h}$ est le variable aléatoire de loi $K_{x,h}$ sur $\aleph_{x,h}$. De plus, on a $\hat{f}_n(x) \in [0;1]$ pour $x \in \mathbb{N}$ et

$$E\left[\hat{f}_n(x)\right] = \sum_{u \in \aleph_{x,h} \cap \aleph} K_{x,h}(u) f(u) \rightarrow \ f(x) \ \textit{quand } h \rightarrow 0 \ \textit{lorsque } n \rightarrow \infty$$

- *L'erreur quadratique moyenne (MSE) :*

$$MSE\left(\hat{f}_n(x)\right) = E\left\{\left[\hat{f}_n(x) - f(x)\right]^2\right\} = Var\left[\hat{f}_n(x)\right] + Biais^2\left[\hat{f}_n(x)\right].$$

- L'erreur quadratique moyenne intégrée (MISE) :

$$MISE\left(\hat{f}_n(x)\right) = \sum_{x \in \mathbb{N}} MSE\left[\hat{f}_n(x)\right] = \frac{1}{n}\sum_{x \in \mathbb{N}} f(x)\left[\left[P(K_{x,h} = x)\right]^2 - f(x)\right]$$

$$+ \sum_{x \in \mathbb{N}}\left[f\left(E\left(K_{x,h}\right)\right) - f(x) + \frac{1}{2} Var\left(K_{x,h}\right) f^{(2)}(x)\right]^2 + \circ\left(\frac{1}{n} + h^2\right).$$

Avec $f^{(2)}$ représentent la différence finie d'ordre 2.

Noyaux associés discrets standards

Nous présentons dans cette section la première classe des noyaux associés discrets, dite, classe des noyaux discrets standards ou de premier ordre proposés par *Senga kiessé [2008].*

Exemple 1.2.2 ***Noyau Poissonien :*** *Pour un type de noyau Poissonien $\mathcal{P}(\lambda)$, on considère le noyau discret associé $K_{\mathcal{P}(\lambda)}$ de loi $(\mathcal{P}(x+h))$ sur $\aleph_{x,h} = \mathbb{N}$, avec : $x \in \mathbb{N}$, $u \in \mathbb{N}$ et $h > 0$ est le paramètre de lissage ; tel que :*

$$K_{\mathcal{P}(\lambda)}(u) = \frac{(x+h)^u}{u!} e^{-(x+h)}$$

L'estimateur $\hat{f}_n$ de f à noyau associé est défini par :

$$\hat{f}_n(x)=\frac{1}{n}\sum_{i=1}^{n}K_{\mathcal{P}(\lambda)}(X_i) = \frac{1}{n}\sum_{i=1}^{n}\frac{(x+h)^{X_i}}{X_i!}e^{-(x+h)}$$

Notons que pour une cible $x \in \mathbb{N}$ et pour tout $h > 0$, le noyau associé $K_{P(\lambda)}$ est de support $\mathbb{N}$, et de moyenne égale à la variance égale à $(x + h)$.

- Le biais est donné par :

$$Biais\left[\hat{f}_n(x)\right] = hf'(x) + \frac{1}{2}(x+h)\,f"(x) + \circ(h)$$

- La variance est donnée par :

$$Var\left[\hat{f}_n(x)\right] = \frac{1}{n}f(x)\frac{(x+h)^x}{x!}e^{-(x+h)}$$

Enfin, la valeur du $MISE$ est :

$$MISE\left(\hat{f}_n(x)\right) = \frac{1}{n}\sum_{x\in\mathbb{N}}f(x)\frac{(x+h)^x}{x!}e^{-(x+h)} + \sum_{x\in\mathbb{N}}\left[hf'(x) + \frac{1}{2}(x+h)\,f"(x) + \circ(h)\right]^2$$

Le $MISE$ dépend de la densité inconnue et des ses dérivées première et seconde.

Choix de paramètre de lissage Nous présentons à ce niveau des méthodes de choix de fenêtres pour approcher la valeur optimale de la fenêtre h définie par :

$$h_{opt} = \arg\min_{h\,>0} MISE\left(\hat{f}_n(x)\right).$$

Minimisation de l'erreur quadratique moyenne intégrée Cette méthode consiste à minimiser l'erreur quadratique moyenne intégrée $MISE$ ou asymptotique ($AMISE$).

Nous rappelons que le $MISE$ est donnée par :

$$MISE\left(\hat{f}_n(x)\right) = \sum_{x\in\mathbb{N}} Biais^2\left[\hat{f}_n(x)\right]dx + \sum_{x\in\mathbb{N}} Var\left[\hat{f}_n(x)\right]dx$$

La variance peut être approximée comme suit :

$$Var\left(\hat{f}_n(x)\right) = \frac{1}{n}Var\left(K_{x,h}\left(X\right)\right) = \frac{1}{n}\left[f(u)\sum_{x\in\mathbb{N}}\mathcal{K}_{x,h}^2 - f^2\left(x\right)\right] + \circ\left(\frac{h}{n}\right).$$

Sous la condition $\lim_{h\to 0}\sum_{u\in\aleph_{x,h}} u\ K\left(u\right) = x$, on a l'approximation suivante :

$$\tilde{V}ar\left(\hat{f}_n(x)\right) = \frac{1}{n}f(x)P(\kappa_{x,h} = x)$$

où $\mathcal{K}_{x,h}$ est la variable aléatoire discrète de densité $K_{x,h}$. On approxime le biais de $\hat{f}_n$ en utilisant le développement discret de **Taylor** à l'ordre 2.

$$Biais\left(\hat{f}_n(x)\right) = E\left[\hat{f}_n(x)\right] - f(x) = f\left[E(\mathcal{K}_{x,h})\right] - f(x) + \frac{1}{2}Var\left(\mathcal{K}_{x,h}\right)f"\left(x\right) + \circ\left(h\right)$$

Finalement, le $MISE$ peut être approximer par :

$$AMISE(h) = \frac{1}{n}\sum_{x\in\mathbb{N}} f\left(x\right)P(\mathcal{K}_{x,h} = x) + \sum_{x\in\mathbb{N}}\left[f\left[E(\mathcal{K}_{x,h})\right] - f(x) + \frac{1}{2}Var\left(\mathcal{K}_{x,h}\right)f"\left(x\right) + \circ\left(h\right)\right]^2$$

Le paramètre de lissage h_{AMISE} dans ce cas peut être obtenu de la manière suivante :

$$h_{AMISE} = \arg\min_h AMISE\left(h\right)$$

Le paramètre de lissage h_{AMISE} n'est pas utilisable directement en pratique, car $AMISE(h)$ dépend de la densité discrète inconnue f.

Excès de zéros Dans cette section, le choix de la fenêtre repose sur des données de comptage avec $\aleph = \mathbb{N}$ qui n'est autre que l'excès des zéros dans l'échantillon $X = (X_1, X_2, ..., X_n)$. On peut choisir une fenêtre adaptée $h_0 = h_0(X, K)$ de h satisfaisant :

$$\sum_{i=1}^{n} P(\kappa_{X_i,h_0} = 0) = n_0 \tag{1.14}$$

où n_0 désigne le nombre des zéros dans X ($n_0 = Card(X_i = 0)$). L'équation () s'obtient à partir de l'expression :

$$E\left(\hat{f}_n(x)\right) = \sum_{u \in \aleph_{x,h}} f(u) P(\mathcal{K}_{x,h} = u).$$

Dans laquelle on prend $u = 0$ et $f(0) = 1$ afin d'identifier le nombre de zéros théoriques au nombre de zéros empiriques n_0. La fenêtre h_0 ajuste le nombre de zéros théoriques au nombre de zéros observés.

Type de noyau	h_0
Poisson	$h_0 = \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{X_i}\right)$
Binomial	$n_0 = \sum_{i=1}^{n}\left(\frac{1-h_0}{X_i+1}\right)^{X_i+1}$
Binomial négatif	$n_0 = \sum_{i=1}^{n}\left(\frac{X_i+1}{2X_i+1+h_0}\right)^{X_i+1}$
Dirac	$h_0 = 0$
Triangulaire	*Pas de solution*

TAB. 1.4 – Solutions h0 pour les noyaux associés discrets standards

Chapitre 2

Estimation de densité multivarié

Dans ce chapitre, nous allons présenter l'estimateur de la fonction densité dans le cas multidimensionnel. Nous présentons ses propriétés statistiques, les différents noyaux classiques et associés ainsi que les différentes méthodes pour la sélection de la matrice des fenêtres de lissage.

Nous considérons ainsi les observations (X_{ij}) *i.i.d.* avec $i = 1, ..., n$ et $j = 1, ..., d$, à valeurs réelles de même densité f.

2.1 Noyaux associés continus multivariés

2.1.1 Noyaux symétriques multivariés

Cette section présente la méthode d'estimation de la densité multidimensionnelle par noyaux *continus symétriques*.

Définition 2.1.1 *Une fonction K de support continu $\aleph_d \subseteq \mathbb{R}^d$ est dite noyau symétrique (classique) si elle est une densité de probabilité symétrique $(K(-u) = K(u))$ de vecteur moyen $\left(u_K = \int_{\mathbb{S}_d} uK(u)du = 0\right)$ de matrice de variance-covariance $\mathbf{\Sigma}_K$ $\left(\mathbf{\Sigma}_K = \int_{\mathbb{S}_d} uu^t K(u)du\right)$ de carré intégrable $\left(\int_{\mathbb{S}_d} K^2(u)du < +\infty\right)$. Le noyau symétrique K est positif de masse totale égale à l'unité (*i.e. *pour tout élément* $\mathfrak{u}$ *de* $\aleph_d$ $K(u) \geq 0$ *et* $\int_{\mathbb{S}_d} K(u)du = 1$ *).*

Définition 2.1.2 *Soit* $H = H_n \underset{n\to\infty}{\to} 0_d$ *une matrice des fenêtres symétrique, définie positive de* $d \times d$ *et* K *la fonction noyau vérifiant la (Définition 1). L'estimateur à noyau symétrique multivarié de* f *peut être défini par :*

$$\hat{f}_n(\mathrm{x}, H)=\frac{1}{n\left(\det H\right)^{\frac{1}{2}}}\sum_{i=1}^{n}K(H^{\frac{1}{2}}\left(\mathrm{x}-X_i\right)) = \frac{1}{n}\sum_{i=1}^{n}K_H\left(\mathrm{x}-X_i\right) \tag{2.1}$$

où $K_H\left(\mathrm{x}\right) = \left(\det H\right)^{\frac{1}{2}} K(H^{\frac{1}{2}}\mathrm{x})$*, cette forme a été introduite par* Wand et Jones *(*1994*) pour faciliter une technique de choix de la matrice de lissage.*

Le type d'orientation de la fonction noyau classique K est contrôlé par la paramétrisation de la matrice de lissage. *Wand and Jones* (1994) ont étudié les différentes formes de paramétrisation de la matrice de lissage bivariée H. Les auteurs ont dénombré trois formes principales et trois autres hybrides :

1. La classe des matrices "pleines" symétriques définies positives :

$$H = \begin{pmatrix} h_1 & h_{12} \\ h_{21} & h_2 \end{pmatrix}$$

2. La classe des matrices diagonales :

$$H = diag(h_1, h_2).$$

3. La classe formée du produit d'une constante positive et de la matrice identité $H = h\mathbf{1}_2$.

4. La classe formée du produit d'une constante positive et de la matrice de variance-covariance empirique S,

$$H = \begin{pmatrix} hS_1^2 & hS_{12} \\ hS_{21} & hS_2^2 \end{pmatrix}$$

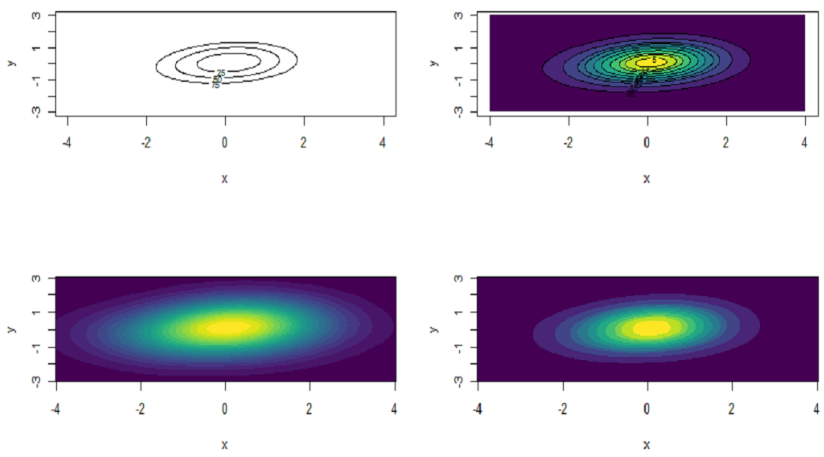

FIG. 2.1 – Les formes de contours.

5. La classe formée d'une constante positive et de la diagonale de S,

$$H = \begin{pmatrix} hS_1^2 & 0 \\ 0 & hS_2^2 \end{pmatrix}$$

6. La classe des matrices obtenues en utilisant le coefficient de corrélation $\rho_{1,2}$ pour déterminer la rotation :

$$H = \begin{pmatrix} h_1^2 & \rho h_1 h_2 \\ \\ \rho h_1 h_2 & h_2^2 \end{pmatrix}$$

Le plus utilisé des noyaux classiques multivariés est la *loi normale multivariée*, donnée par $\mathcal{N}(\mu, \mathbf{\Sigma})$où μ est le vecteur moyenne et $\mathbf{\Sigma}$ est la matrice de variance covariance :

$$K_{\mathcal{N}(\mu,\mathbf{\Sigma})}(\mathrm{u}) = \frac{1}{(2\pi)^{\frac{d}{2}} (\det \mathbf{\Sigma})^{\frac{1}{2}}} e^{\left\{-\frac{1}{2}(\mathrm{u}-\mu)^t \mathbf{\Sigma}^{-1}(\mathrm{u}-\mu)\right\}} \qquad \mathrm{u} \in \mathbb{R}^d$$

En plus de la loi normale multidimensionnelle,

on peut construire plusieurs autres noyaux symétriques multivariés à partir des noyaux univariés. Quelques uns sont rappelés dans la Table (1.1).

Tous ces noyaux univariés vérifient la Définition 1 et ont pour support $\aleph_1 = \mathbb{R}$ ou $[-1, 1]$.

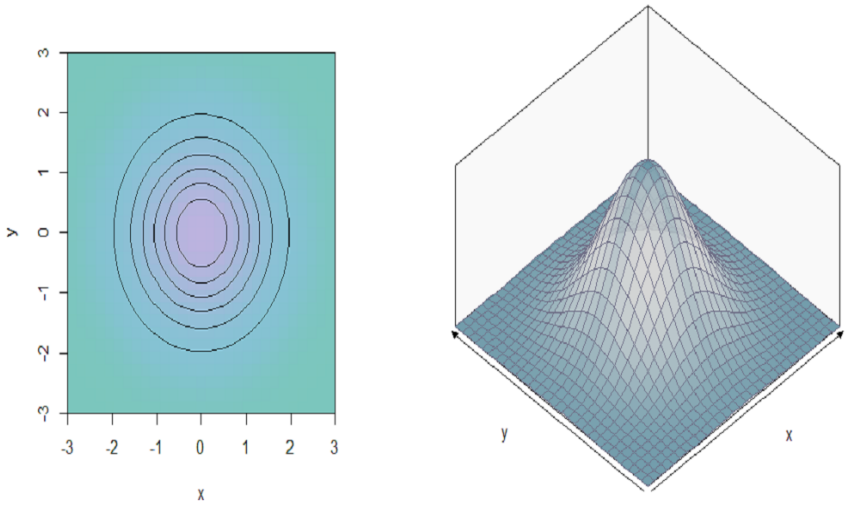

FIG. 2.2 – Bivariate normal.

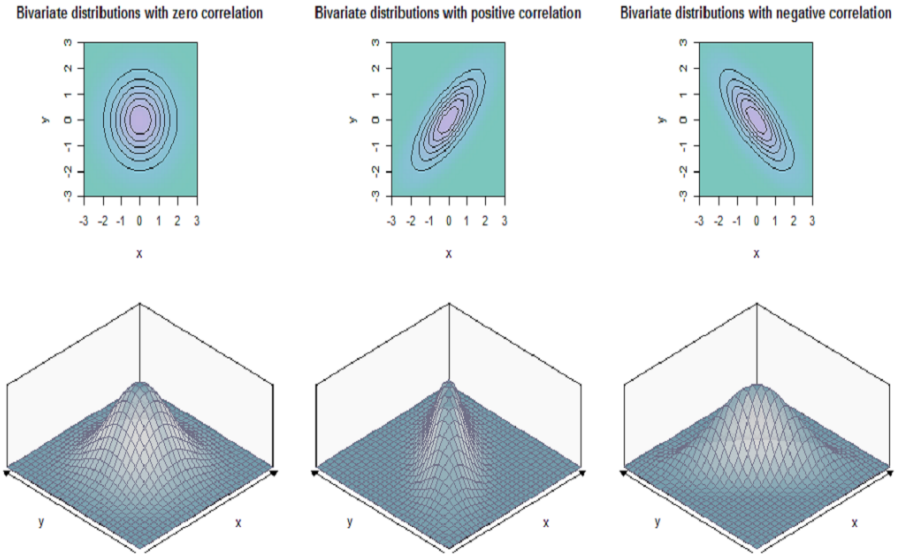

FIG. 2.3 – Normale bivariée.

Propriétés de l'estimateur

Dans cette section nous allons rappeler quelques résultats théoriques qui ont été établis par *Wand and Jones (1995)* et *Somé (2015)*. On commence par :

Espérance mathématique

$$E\left(\hat{f}_n(\mathrm{x})\right) = \int_{\mathbb{S}_d} K_H(\mathrm{t}-\mathrm{x}) f(t) dt = \frac{1}{\det H} \int_{\mathbb{S}_d} K_H\left[H^{-1}(\mathrm{t}-\mathrm{x})\right] f(t) dt$$

En posant :

$$\mathrm{u} = H^{-1}(t-\mathrm{x}) \,(\boldsymbol{ie}\ t = \mathrm{x} - H\mathrm{u}),$$

et en utilisant la formule des changements linéaires de variables, on obtient :

$$E\left(\hat{f}_n(\mathrm{x})\right) = \int_{\mathbb{S}_d} K_H(\mathrm{u}) f(\mathrm{x} - Hu) d\mathrm{u}.$$

Le développement en séries de *Taylor à l'ordre* 2 pour la fonction f au voisinage de x, on donne :

$$f(\mathrm{x} - H\mathrm{u}) = f(\mathrm{x}) - \mathrm{u}^t H^t \nabla f(x) + \frac{1}{2}\mathrm{u}^t H^t \nabla^2 f(x) + \circ(\mathrm{u}^t H^2 \mathrm{u}).$$

Cette quadratique est une matrice 1×1. On rappelle aussi que l'opérateur **trace** satisfait $trace(AB) = trace(BA)$ pour toutes matrices A et B de dimensions respectives $r \times s$ et $s \times r$. De ces résultats de la trace et en injectant dans l'expression de l'espérance, on obtient alors :

$$E\left(\hat{f}_n(\mathrm{x})\right) = \int_{\mathbb{S}_d} K_H(\mathrm{u}) \left\{ f(\mathrm{x}) - \mathrm{u}^t H^t \nabla f(x) + \frac{1}{2}\mathrm{u}^t H^t \nabla^2 f(\mathrm{x}) + \circ(\mathrm{u}^t H^2 \mathrm{u}) \right\} d\mathrm{u}$$

$$= f(\mathrm{x}) + \frac{1}{2} trace\left(\boldsymbol{\Sigma}_K H^t \nabla^2 f(\mathrm{x}) H\right) + \circ\left(trace\left(H^2\right)\right).$$

Biais

Le biais de l'estimateur est :

$$Biais\left(\hat{f}_n(\mathrm{x})\right) = \frac{1}{2} trace\left(\mathbf{\Sigma}_K H\ {}^t\nabla^2 f(\mathrm{x}) H\right) + \circ\left(trace\left(H^2\right)\right)$$

Variance

La variance de l'estimateur est donnée par :

$$Var\left(\hat{f}_n(\mathrm{x})\right) = \frac{1}{n} Var\left(\sum_{i=1}^{n} K\left(\mathrm{x} - X_i\right)\right)$$

$$Var\left(\hat{f}_n(\mathrm{x})\right) = \frac{1}{n\left(\det H\right)^2} \int_{\mathbb{S}_d} K\left[H^{-1}(\mathrm{t} - \mathrm{x})\right] f(t) dt - \frac{1}{n\left(\det H\right)^2} E\left(K\left[H^{-1}(\mathrm{t} - \mathrm{x})\right]\right)$$

Posons aussi $\mathrm{u} = H^{-1}(\mathrm{t} - \mathrm{x})$, on aura :

$$Var\left(\hat{f}_n(\mathrm{x})\right) = \frac{1}{n\left(\det H\right)} \int_{\mathbb{S}_d} K^2(\mathrm{u}) f(\mathrm{x} - H\mathrm{u}) d\mathrm{u} - \frac{1}{n\left(\det H\right)^2} E\left(K\left[H^{-1}(\mathrm{t} - X_1)\right]\right)$$

$$= \frac{1}{n\left(\det H\right)} \int_{\mathbb{S}_d} K^2(\mathrm{u}) f(\mathrm{x}) d\mathrm{u} + R_n + \circ\left(\frac{1}{n}\left(\det H\right)\right)$$

$$= \frac{1}{n\left(\det H\right)} f(\mathrm{x}) \int_{\mathbb{S}_d} K^2(\mathrm{u}) d\mathrm{u} + \circ\left(\frac{1}{n}\left(\det H\right)\right)$$

Critère d'erreur

L'évaluation de la similarité entre l'estimateur $\hat{f}_n$ et la vraie densité f à estimer. La mesure la plus naturelle utilisée est la moyenne intégrée des erreurs quadratiques. On défini d'abord la moyenne des erreurs quadratiques (**MSE**) par :

$$MSE\left(\mathrm{x}\right) = E\left[\left\{\hat{f}_n(\mathrm{x}) - f(\mathrm{x})\right\}^2\right] = \frac{1}{n\left(\det H\right)} f(\mathrm{x}) \int_{\mathbb{S}_d} K^2(\mathrm{u}) d\mathrm{u} + \left[\frac{1}{2} trace\left(\mathbf{\Sigma}_K H^t \nabla^2 f(\mathrm{x}) H\right)\right]^2$$

avec $x \in \aleph_d$; $MISE$ elle est donnée par :

$$MISE\left(\hat{f}_n(\mathrm{x})\right) = \int_{\mathbb{S}_d} MSE(\mathrm{x})d\mathrm{x} = \int_{\mathbb{S}_d} Var\left(\hat{f}_n(\mathrm{x})\right)d\mathrm{x} + \int_{\mathbb{S}_d} Biais^2\left(\hat{f}_n(\mathrm{x})\right)d\mathrm{x}$$

$$= \int_{\mathbb{S}_d}\left[\frac{1}{n\left(\det H\right)}f(\mathrm{x})\int_{\mathbb{S}_d} K^2(\mathrm{u})d\mathrm{u}\right]d\mathrm{x} + \int_{\mathbb{S}_d}\left[\frac{1}{2}trace\left(\mathbf{\Sigma}_K H^t\nabla^2 f(\mathrm{x})H\right)\right]^2 d\mathrm{x}$$

Asymptotique propriétés de l'estimateur de noyau

Une approximation asymptotique d'estimation de densite multivarié est trés simillaire d'estimation de densite univarié, alors on a les conditions suivantes sur f, H et K , sont résumés comme suivant :

$\boldsymbol{C_1}$ La densité f est carré intégrable, admet la dérivée seconde qui doit être une fonction absolument continue.

$\boldsymbol{C_2}$ La matrice de lissage $H = H_n$ est déterministique, définie positive et symétrique matrices, où : quand $n \to \infty$, $vecH \to 0_d$ et $n\left|H\right|^{\frac{1}{2}} \to \infty$

$\boldsymbol{C_3}$ Le noyau K est symétrique possède un moment d'ordre 2 et carré intégrable.

Les expressions approchées du MSE et $MISE$ sont données par :

$$AMISE\left(\hat{f}_n(\mathrm{x})\right) = \frac{1}{n\left(\det H\right)}\int_{\mathbb{S}_d} K^2(\mathrm{u})d\mathrm{u} + \frac{1}{4}\int_{\mathbb{S}_d}\left[trace\left(\mathbf{\Sigma}_K H^t\nabla^2 f(\mathrm{x})H\right)\right]^2 d\mathrm{x}$$

Si on suppose de plus que $\mathbf{\Sigma}_K = \mu_2(K)\mathbf{1}_d$; où $\mu_2(K) > 0$ et $\mathbf{1}_d$ est la matrice unité d'ordre d, alors l'$AMISE$ devient :

$$AMISE\left(\hat{f}_n(\mathrm{x})\right) = \frac{1}{n\left(\det H\right)}\int_{\mathbb{S}_d} K^2(\mathrm{u})d\mathrm{u} + \frac{1}{4}\mu_2^2(K)\int_{\mathbb{S}_d}\left[trace\left(H\ {}^t\nabla^2 f(\mathrm{x})H\right)\right]^2 d\mathrm{x}$$

$$= \frac{1}{n\left(\det H\right)^{\frac{1}{2}}}\int_{\mathbb{S}_d} K^2(\mathrm{u})d\mathrm{u} + \frac{1}{4}\mu_2^2(K)\left(vech^t\ H\right)\psi\left(vech\ H\right)$$

où $vech\ H$ est le vecteur de dimension $\frac{d(d+1)}{2}$.
$vech\,(H) = \left(h_{11}, ..., h_{1d}, h_{22}, ..., h_{2d}, ..., h_{(d-1)(d-1)}, h_{(d-1)d}, h_{dd}\right)^t$ obtenu après élimination des éléments de H situés au dessus de la diagonale et ψ est la matrice carrée d'ordre $\frac{d(d+1)}{2}$ donnée par :

$$\psi = \int_{\mathbb{S}_d} vech\left(2\nabla^2 f(\mathrm{x}) - diag\nabla^2 f(\mathrm{x})\right) vech^t\left(2\nabla^2 f(\mathrm{x}) - diag\nabla^2 f(\mathrm{x})\right) d\mathrm{x}$$

Les éléments de ψ peuvent être obtenus comme suit :
Considérons $p = (p_1, ..., p_d)^t$ où les $p_1, ..., p_d$ sont des entiers positifs et $|p| = \sum_{i=1}^{d} p_i$. Alors la dérivé partielle d'ordre p de f est :

$$f^{(p)}(\mathrm{x}) = \frac{\partial^{|p|}}{\partial x_1^{p_1} ... \partial x_d^{p_d}} f(\mathrm{x})$$

et les dérivées fonctionnelles de ψ sont :

$$\psi_p = \int_{\mathbb{S}_d} f^{(p)}(\mathrm{x}) f(\mathrm{x}) d\mathrm{x} \ = E\left[f^{(p)}(\mathrm{x})\right].$$

Choix de la matrice de lissage

Dans cette section, on rappelle les méthodes pratiques du choix de la matrice des fenêtres H, à savoir la méthode de substitution ("*Plug-in*"), la méthode de validation croisée ("*Cross-validation*") ainsi que l'approche bayésienne avec ses trois variantes (*globale, locale et adaptative*).

Méthode Plug-in

Le critère de la méthode *plug-in* est donné par :

$$PI(H) = \frac{1}{n\,(4\pi)^{\frac{d}{2}}\,(\det H)^{\frac{1}{2}}} + \frac{1}{4}\left(vech^t\ H\right)\psi\,(vech\ H)$$

où $\hat{\psi}$ est une estimation de ψ. En pratique elle est donnée comme suit :

$$\hat{\psi}(H) = \frac{1}{n}\sum_{i=1}^{d} \hat{f}_n^{(p)}(x) = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_H^{(p)}(X_i - X_j)$$

où $\hat{f}_n^{(p)}$ est la dérivée partielle d'ordre p de $\hat{f}$.

Méthode de validation croisée

Des recherches ont été entreprises par *Sain et al (1994)* sur la version multivariée de la méthode validation croisée non biaisée (UCV), mais leurs intérêts s'est porté uniquement sur le noyau produit équivalent à utiliser H diagonale. *Duong and Hazelton* (2005) ont généralisé ces résultats pour une matrice de lissage H symétrique définie positive quelconque.

Le principe de base de cette méthode est de minimiser par rapport à H le critère d'integrate squared error (ISE),

$$ISE = \int_{\mathrm{x}\in\mathbb{S}_d} \left\{\hat{f}_n(\mathrm{x}) - f(\mathrm{x})\right\}^2 d\mathrm{x}$$

La matrice des fenêtres optimale notée H_{LSCV} , est donnée par :

$$H_{LSCV} = \arg\min_{\mathcal{M}} LSCV(H)$$

où $\mathcal{M}$ est l'espace des matrices de lissage symétriques définies positives et

$$LSCV(H) = \int_{\mathbb{S}_d} \hat{f}_n(\mathrm{x})d\mathrm{x} - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{n,-i}(X_i)d\mathrm{x}$$

où $\hat{f}_{n,-i}(X_i) = \dfrac{1}{n-1}\sum_{j\neq i} K_H\left(\mathrm{x} - X_j\right)$ est l'estimateur calculé à partir de l'échantillon privé de l'observation X_i . Cette méthode est la généralisation en multivarié du cas univarié de *Bowman (1984)*.

Approche bayésienne

Dans cette partie, on va faire un rappel sur l'approche bayésienne avec ses trois variantes (globale, locale et adaptative) dans le cas multidimensionnel continu.

Approche bayésienne globale L'approche bayésienne globale a été introduite par *Zhang et al (2006)* en utilisant le noyau *gaussien multivarié*. Dans ce cas, la loi a posteriori est souvent de forme complexe. Cette difficulté est surmontée grâce aux méthodes d'approximation $MCMC$ (Monté Carlo par Chaîne de Markov).

Approche bayésienne locale L'approche bayésienne locale a été introduite par *De Lima and Atuncar (2010)*. Cette méthode est une généralisation au cas multidimensionnel de la méthode décrite par *Gangopadhyay and Cheung (2002)* en utilisant le n*oyau gaussien multivarié*. Elle consiste à estimer H en chaque vecteur cible x et traite donc H comme une quantité aléatoire de loi a priori $\pi(\cdot)$. À partir de la formule de *Bayes*, la loi a posteriori prend la forme suivante :

$$\hat{\pi}(H \mid \mathrm{x}, X_1, ..., X_n) = \pi(H)\hat{f}_n(\mathrm{x}) \left[\int_{\mathcal{M}} \pi(H)\hat{f}_n(\mathrm{x})dH\right]^{-1}$$

L'estimateur de *Bayes* sous la perte quadratique de H en chaque cible x est la moyenne de $\hat{\pi}(H \mid \mathrm{x}, X_1, ..., X_n)$ donnée par :

$$\hat{H}(\mathrm{x}) = \int_{\mathcal{M}} H\hat{\pi}(H \mid \mathrm{x}, X_1, ..., X_n)dH$$

Approche bayésienne adaptative L'approche bayésienne adaptative a été introduite par *Zougab et al (2014)* en utilisant le *noyau gaussien multivarié*. Cette méthode consiste à associer une matrice de lissage H_i pour chaque observation X_i et traiter donc H_i comme une quantité aléatoire de loi a priori $\pi(\cdot)$. L'information apportée par les observations pour

H_i est obtenue par une estimation par validation croisée de $f(X_i)$:

$$\hat{f}\left(X_i \mid \{X_{-i}\}, H_i\right) = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} K_H\left(X_i - X_j\right)$$

où X_{-i} est l'ensemble des observations recueillies en excluant X_i, et K_{H_i} est le noyau symétrique. À partir de la formule de *Bayes* et de l'estimateur $\hat{f}\left(X_i \mid \{X_{-i}\}, H_i\right)$, la loi a posteriori pour chaque H_i prend la forme suivante :

$$\hat{\pi}(H_i \mid X_i) = \hat{f}\left(X_i \mid \{X_{-i}\}, H_i\right)\pi(H_i)\left[\int_{\mathcal{M}} \hat{f}\left(X_i \mid \{X_{-i}\}, H_i\right)\pi(H_i)dH_i\right]^{-1}.$$

2.1.2 Noyaux asymétriques multivariés

Cette section présente la méthode d'estimation de la densité multidimensionnelle par noyaux continus asymétriques. La notion d'un noyau associé multivarié $K_{\mathrm{x},H}$ de vecteur cible x et de matrice des fenêtres de lissage H a été introduite par *Kokonendji and Somé (2015)* dans le cas continu. Les variables à observer $X_1, ..., X_n$ sont des vecteurs aléatoires indépendants et identiquement distribués (*iid*), à valeurs réelles de même densité f. L'estimateur à noyau symétrique multivarié de f peut être défini par :

$$\hat{f}_n(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathrm{x},H}(X_i) \qquad\qquad \forall x \in \aleph_d \subset \mathbb{R}_d$$

où $K_{\mathrm{x},H}(.) = \frac{1}{(\det H)}K(H^{\frac{1}{2}}(\mathrm{x} - .))$.

Définition 2.1.3 *Soient* $\mathrm{x} \in \aleph_d \subseteq \mathbb{R}_d$ *et* H *une matrice de lissage, avec* $\aleph_d$ *est le support de la fonction* f *à estimer.* $K_{\mathrm{x},H}(.)$ *de support* $\aleph_{\mathrm{x},H} \subseteq \mathbb{R}_d$ *est appelée noyau associé multivarié si les conditions suivantes sont satisfaits :*

$$\mathrm{x} \in \aleph_{\mathrm{x},H}, \qquad E\left(\mathcal{K}_{\mathrm{x},h}\right) = \mathrm{x} + a\left(\mathrm{x}, H\right), \qquad Cov\left(\mathcal{K}_{\mathrm{x},h}\right) = B(\mathrm{x}, H)$$

où $a(\mathrm{x}, H) \underset{H\to 0_d}{\to} 0_d$, $B(\mathrm{x}, H) \underset{H\to 0_d}{\to} 0_d$ *(*0_d *est une matrice carré nulle d'ordre d) et* $K_{\mathrm{x},h}$ *est un vecteur de variables aléatoires discrètes de loi* $K_{\mathrm{x},H}$.

Exemple 2.1.1 *Noyau beta bivarié de* Sarmanov(1966) *et* Lee(1996)*, un distribution de* 2 *indépendants beta univariés*

$$g_j(u) = \frac{u^{p_j-1}(1-u)^{q_j-1}}{\mathcal{B}(p_j, q_j)}\mathbf{1}_{[0,1]}(u) \ , \qquad j = 1, 2$$

où $\mathcal{B}(p_j, q_j) = \int_0^1 (1-t)^{q_j-1} t^{p_j-1} dt$ *est fonction usuel de* beta *avec :* $p_j > 0$ *et* $q_j > 0$. *Leurs moyennes et variances sont :*

$$\mu_j = \frac{p_j}{p_j+q_j} = \mu_j(p_j, q_j) \quad et \quad \sigma_j^2 = \frac{p_j q_j}{(p_j+q_j)^2(p_j+q_j+1)} = \sigma_j^2(p_j, q_j)$$

$$\mu = (\mu_1, \mu_2) \quad et \quad \mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix}$$

L'estimateur de cet noyau est :

$$\hat{f}_n(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathcal{BS}_{\beta(\mathrm{x},H)}}(X_i) \ , \qquad \forall \mathrm{x} \in [0,1]\times[0,1]$$

où $K_{\mathcal{BS}_{\beta(\mathrm{x},H)}}$ *est fonction de noyau* beta bivarié.*et* $\forall \mathrm{x} = \binom{x_1}{x_2}$, $H = \begin{pmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{pmatrix}$

$$\beta(\mathrm{x}, H) = \left(\frac{x_1}{h_{11}} + 1, \frac{1-x_1+1}{h_{11}}, \frac{x_2}{h_{22}} + 1, \frac{1-x_2}{h_{22}} + 1, \frac{h_{12}}{(h_{11}h_{22})^{\frac{1}{2}}}\right)^t$$

L'estimateur modifiée de cet noyau est :

$$\tilde{f}_n(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathcal{BS}_{\tilde{\beta}(\mathrm{x},H)}}(X_i) \qquad , \forall \mathrm{x} \in [0,1]\times[0,1]$$

2.2 Noyaux associés discrets multivariés

Dans cette section, nous allons présenter le cas discret multidimensionnel ainsi que quelques exemples des noyaux associés discrets ainsi que le noyaux produit. Ensuite nous introduisons l'estimateur à noyau associé discret multivarié et ses propriétés. Enfin deux méthodes classiques pour le choix de la matrice des fenêtres de lissage seront présentées.

Définition 2.2.1 *(Noyau associé discret multivarié) Soient* $\mathrm{x} \in \aleph_d \subseteq \mathbb{Z}_d$ *et* H *une matrice des fenêtres, avec* $\aleph_d$ *est le support de la fonction* f *à estimer.* $K_{\mathrm{x},H}(.)$ *de support* $\aleph_{\mathrm{x},H} \subseteq \mathbb{Z}_d$ *est appelée noyau associé discret multivarié si :*

$$\mathrm{x} \in \aleph_{\mathrm{x},H} \qquad E\left(\mathcal{K}_{\mathrm{x},h}\right) = \mathrm{x} + a\left(\mathrm{x}, H\right) \qquad Cov\left(\mathcal{K}_{\mathrm{x},h}\right) = B(\mathrm{x}, H)$$

où $a\left(\mathrm{x}, H\right) \underset{H \to 0_d}{\to} 0_d$, $B(\mathrm{x}, H) \underset{H \to 0_d}{\to} 0_d$ *(*0_d *est une matrice carré nulle d'ordre* d *) et* $\mathcal{K}_{\mathrm{x},h}$ *est un vecteur de variables aléatoires discrètes de loi* $K_{\mathrm{x},H}$.

Définition 2.2.2 *(Noyau associé discret multivarié produit) Soient* $\mathbb{S}_1^{[i]}$ *le support des marges univariés de* f, $\mathrm{x} = (x_1, x_2, ..., x_d)^t$ *est le vecteur cible,* h_i , $i = 1, ..., d$ *sont les fenêtres de lissage univariés et* $K_{x_i,h_i}^{[j]}$ *est le* $i^{ème}$ *noyau associé discret univarié de support* $\aleph_{x_i,h_i}$. *Le "noyau associé discret multivarié produit" est définit comme suit :*

$$K_{\mathrm{x},H}(.) = \prod_{i=1}^{d} K_{x_i,h_i}^{[i]}(.) \quad , \qquad \forall x_i \in \aleph_1^{[i]} \subseteq \mathbb{Z}.$$

2.2.1 Estimateur à noyau associé discret multivarié

Soient $X_1, ..., X_n$ des vecteurs aléatoires $i.i.d$ multivariés commune *inconnue* f à estimer sur $\aleph_d$. L'estimateur à noyau associé discret multivarié $\hat{f}_n$ de f est de la forme :

$$\hat{f}_n(\mathrm{x}) = \frac{1}{n} \sum_{i=1}^{n} K_{\mathrm{x},H}(X_i) \quad , \qquad x \in \aleph_d$$

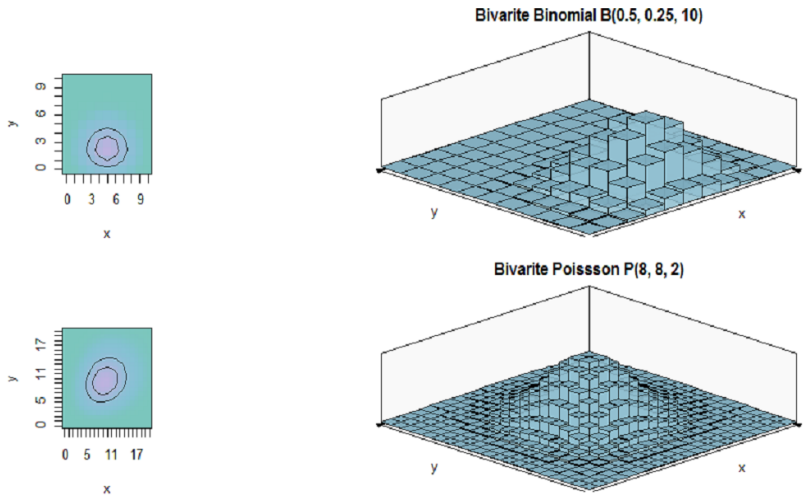

FIG. 2.4 – Bivariate Binomial et Poisson.

Où $K_{\mathrm{x},H}(.)$ est le noyau associé *discret multivarié dépendant* du vecteur cible x et la matrice des fenêtres H symétrique définie positive, qui tend vers la matrice nulle (0_d) quand $n \to \infty$. Estimateur à noyau associé discret *multiple multivarié*

$$\hat{f}_n(\mathrm{x}) = \frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K^{[j]}_{x_j,h_j}(X_{ij}) \qquad , \; x_j \in \aleph_1^{[j]} \subseteq \mathbb{Z}$$

Avec $\aleph_1^{[j]}$ est le support de la marge univarié de f pour $j = 1, ..., n$, $x = (x_1, ..., x_d)^t \in X_{j=1}^{d}\aleph_1^{[j]}$, $X_i = (X_{i1}, ..., X_{id})^t$ pour $i = 1, ..., n$ et $h_1, ..., h_d$ sont les paramètres de lissage unidimensionnels. La fonction $K^{[j]}_{x_j,h_j}$ est le $j^{ème}$ noyau associé discret univarié de support $\aleph_{x_i,h_i} \subseteq \mathbb{Z}$. De *Kokonendji et Somé (2015)*, il est connu que pour les deux formes de l'estimateur $\hat{f}_n$, On a $\hat{f}_n(\mathrm{x}) \in [0, 1]$ pour tout $\mathrm{x} \in \aleph_d$ et $\sum_{\mathrm{x}\in\mathbb{S}_d} \hat{f}_n(\mathrm{x}) = 1$.

2.2.2 Propriétés de l'estimateur

On examine les différentes propriétés statistiques à distance finie et asymptotique de l'estimateur à noyaux associé multiple, en utilisant les noyaux associés discrets du type :

Cas des noyaux standards (Binomial et Poisson).

Biais

On commence par l'espérance mathématique qui est donnée par :

$$E\left[\hat{f}_n(\mathrm{x})\right] = E\left[\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K^{[j]}_{x_j,h_j}(X_{ij})\right] = E\left[f\left(\mathcal{K}^{[1]}_{x_1,h_1},...,\mathcal{K}^{[d]}_{x_d,h_d}\right)\right]$$

Où $\mathcal{K}^{[j]}_{x_j,h_j}$ sont des variables aléatoires indépendantes de loi $K^{[j]}_{x_j,h_j}$ (*binomial ou Poisson*), de moyenne : $\mu_j = x_j + h_j$ et de variance :

$$\sigma_j^2 = \begin{cases} \frac{x_j+h_j-x_jh_j}{x_j+1} & \textit{cas du noyau Binomial} \\ \\ x_j + h_j & \textit{cas du noyau Poisson} \end{cases}$$

En utilisant le développement en séries de *Taylor* à l'ordre 2 et en remplaçant les dérivées partielles par les différences finies, on obtient :

$$f\left(\mathcal{K}^{[1]}_{x_1,h_1},...,\mathcal{K}^{[d]}_{x_d,h_d}\right) = f\left(\mu_1,\mu_2,...,\mu_d\right) + \sum_{j=1}^{d}\left(\mathcal{K}^{[j]}_{x_j,h_j} = \mu_j\right) f_j^{(1)} + \frac{1}{2}\sum_{j=1}^{d}\left(\mathcal{K}^{[j]}_{x_j,h_j} = \mu_j\right)^2 f_{jj}^{(2)}$$

$$+ \sum_{k\neq j}^{d}\left(\mathcal{K}^{[k]}_{x_k,h_k} = \mu_k\right) f_{kj}^{(2)} + \circ\left(\sum_{j=1}^{d} h_j^2\right)$$

Donc

$$E\left[\hat{f}_n(\mathrm{x})\right] = f\left(\mu_1,\mu_2,...,\mu_d\right) + \frac{1}{2}\sum_{j=1}^{d} Var\left(\mathcal{K}^{[j]}_{x_j,h_j}\right) f_{jj}^{(2)} + \circ\left(\sum_{j=1}^{d} h_j^2\right)$$

$$= f\left(x_1,...,x_d\right) + \sum_{j=1}^{d} h_j f_j^{(1)} + \frac{1}{2}\sum_{j=1}^{d} Var\left(\mathcal{K}^{[j]}_{x_j,h_j}\right) f_{jj}^{(2)} + \circ\left(\sum_{j=1}^{d} h_j^2\right)$$

Par conséquent, le biais de l'estimateur est donné comme suit :

$$Biais\left[\hat{f}_n(\mathrm{x})\right] = \sum_{j=1}^{d} h_j f_j^{(1)} + \frac{1}{2}\sum_{j=1}^{d} Var\left(\mathcal{K}_{x_j,h_j}^{[j]}\right) f_{jj}^{(2)} + \circ\left(\sum_{j=1}^{d} h_j^2\right)$$

Où $f_j^{(1)}$, $f_{jj}^{(2)}$ sont *les fonctions discrètes multivariées* avec leurs différences finies partielles correspondantes comme dans *Abdous and Kokonendji (2009)*. Finalement, le biais est :

$$Biais\left[\hat{f}_n(\mathrm{x})\right] = \begin{cases} \sum_{j=1}^{d} h_j f_j^{(1)} + \frac{1}{2}\sum_{j=1}^{d}\left(\frac{x_j+h_j-x_j h_j}{x_j+1}\right) f_{jj}^{(2)} + \circ\left(\sum_{j=1}^{d} h_j^2\right) & \textit{Cas du noyau Binomial} \\ \sum_{j=1}^{d} h_j f_j^{(1)} + \frac{1}{2}\sum_{j=1}^{d}\left(x_j+h_j\right) f_{jj}^{(2)} + \circ\left(\sum_{j=1}^{d} h_j^2\right) & \textit{Cas du noyau Poisson} \end{cases}$$

Variance

La variance de $\hat{f}_n(\mathrm{x})$ se décompose autour du vecteur cible comme suit :

$$Var\left[\hat{f}_n(\mathrm{x})\right] = Var\left[\frac{1}{n}\sum_{i=1}^{n}\prod_{j=1}^{d} K_{x_j,h_j}^{[j]}(X_{ij})\right]$$

$$Var\left[\hat{f}_n(\mathrm{x})\right] = \frac{1}{n}\left[f(\mathrm{x})\left(\prod_{j=1}^{d} P\left(\mathcal{K}_{x_j,h_j}^{[j]} = x_j\right)\right)^2 + \sum_{\mathrm{u}\in\mathbb{S}_{\mathrm{x}}^{d}\setminus\mathrm{x}} f(u)\left(\prod_{j=1}^{d} P\left(\mathcal{K}_{x_j,h_j}^{[j]} = u_j\right)\right)^2\right]$$

$$-\frac{1}{n}\left[\left(\prod_{j=1}^{d} f(\mathrm{x})\, P\left(\mathcal{K}_{x_j,h_j}^{[j]} = x_j\right)\right)^2 + \left(\sum_{\mathrm{y}\in\mathbb{S}_{\mathrm{x}}^{d}\setminus\mathrm{x}} f(u)\left(\prod_{j=1}^{d} P\left(\mathcal{K}_{x_j,h_j}^{[j]} = y_j\right)\right)\right)^2\right]$$

$$Var\left[\hat{f}_n(\mathrm{x})\right] = = \frac{1}{n} f(\mathrm{x})\left(1 - f(\mathrm{x})\right)\left(\prod_{j=1}^{d} P\left(\mathcal{K}_{x_j,h_j}^{[j]} = x_j\right)\right)^2 + \circ\left(\frac{1}{n}\right)$$

Finalement, variance d' estimateur est :

$$Var\left[\hat{f}_n(\mathrm{x})\right] = \begin{cases} \frac{1}{n} f(\mathrm{x})(1-f(\mathrm{x}))\left(\prod_{j=1}^{d} P\left(\mathcal{B}^{[j]}_{x_j,h_j} = x_j\right)\right)^2 + \circ\left(\frac{1}{n}\right) & \text{cas du noyau Binomial} \\ \frac{1}{n} f(\mathrm{x})(1-f(\mathrm{x}))\left(\prod_{j=1}^{d} P\left(\mathcal{P}^{[j]}_{x_j,h_j} = x_j\right)\right)^2 + \circ\left(\frac{1}{n}\right) & \text{cas du noyau Poisson} \end{cases}$$

MSE

L'erreur quadratique moyenne est donnée par :

$$MSE(\mathrm{x}) = \frac{1}{n} f(\mathrm{x})(1-f(\mathrm{x}))\left(\prod_{j=1}^{d} P\left(\mathcal{K}^{[j]}_{x_j,h_j} = x_j\right)\right)^2 + \left[\sum_{j=1}^{d} h_j f_j^{(1)} + \frac{1}{2}\sum_{j=1}^{d} Var\left(\mathcal{K}^{[j]}_{x_j,h_j}\right) f_{jj}^{(2)}\right]^2$$

MISE

L'erreur quadratique moyenne intégrée est donnée par :

$$MISE\left(\hat{f}_n(\mathrm{x})\right) = \sum_{\mathrm{x}\in\mathbb{S}_d} \frac{1}{n} f(\mathrm{x})(1-f(\mathrm{x}))\left(\prod_{j=1}^{d} P\left(\mathcal{K}^{[j]}_{x_j,h_j} = x_j\right)\right)^2$$

$$+ \sum_{\mathrm{x}\in\mathbb{S}_d}\left[\sum_{j=1}^{d} h_j f_j^{(1)} + \frac{1}{2}\sum_{j=1}^{d} Var\left(\mathcal{K}^{[j]}_{x_j,h_j}\right) f_{jj}^{(2)}\right]^2$$

2.2.3 Choix de la matrice de lissage

On propose dans ce travail la méthode least square cross validation ($LSCV$) avec les noyaux associés discrets multivariés. On propose également la méthode d'excès de zéros.

Validation croisée

Cette méthode est la version discrète de la méthode citée dans le chapitre précédent, dont le principe est de minimiser par rapport à H le critère $ISE = \sum_{x\in\mathbb{S}_d}\left[\hat{f}_n(\mathrm{x}) - f(\mathrm{x})\right]^2$. La

matrice des fenêtres optimale notée H_{LSCV}, est donnée par :

$$H_{LSCV} = \arg\min_{\mathcal{M}} LSCV\,(H)$$

Où $\mathcal{M}$ *est l'espace des matrices de lissage symétriques* définies positives et

$$LSCV\,(H) = \sum_{\mathrm{x}\in\mathbb{S}_d}\left(\frac{1}{n}\sum_{i=1}^{n}K_{\mathrm{x},H}(X_i)\right)^2 - \frac{2}{n\,(n-1)}\sum_{i=1}^{n}\sum_{j=i}^{n}K_{\mathrm{x},H}(X_j)$$

Où $\hat{f}_{n,-i}$ est calculé à partir de $\hat{f}_n$ sans l'observation X_i.

Excès de zéros

Le choix de la matrice des fenêtres pour cette méthode repose sur une particularité des données de comptage ($\aleph_d = \mathbb{N}_d$), on va généraliser cette technique à cas unidimensionnel au cas multidimensionnel. Etant donné un noyau associé discret multivarié $K_{\mathrm{x},H}$, on peut choisir H_0 qui satisfait :

$$\sum_{i=1}^{n} P\left(\mathcal{K}^{[1]}_{X_1,H_0} = 0, ..., \mathcal{K}^{[d]}_{X_d,H_0} = 0\right) = n_0$$

Chapitre 3

Simulation et résultats numériques

Le but de cet chapitre est une étude de simulation pour comparer les méthodes d'estimation du matrice de lissage H, pour différents noyaux discrets et continu symétrique et asymétrique. Nous développons cet estimateur dans le cas univarié, ensuite nous le traitons dans le cas multivarié. selon le critère ISE. Une application sera réalisée sur des données pour évaluer les performances des méthodes de choix du matrice de lissage.

3.1 Noyau associé symétrique

3.1.1 Cas univarié

Etude de simulation

Dans cette partie, nous illustrons certains estimateurs à noyaux continus symétriques à savoir le noyau *Gaussien*, le noyau *Cauchy* $C(0,1)$ et le noyau *Uniforme U(-1,1)*. Nous simulons un échantillon de taille $n = (10, 50, 100)$. Pour chaque noyau fixé, la fenêtre optimale est choisie par les méthodes de validation croisée et par Plug-in.

Notons par :

n : la taille de l'échantillon,

h_{ucv} : Le paramètre de lissage optimal obtenu par validation croisée non bias.

h_{bcv} : Le paramètre de lissage optimal obtenu par la technique de validation croisée bias.

h_{pi} : Le paramètre de lissage optimal obtenu par la technique de plug-in.

h_{nrd} : Le paramètre de lissage optimal obtenu par la technique de rule of thumb.

Le critère de comparaison utilisé est le ISE, $MISE$ donné par :

– Pour l'application numérique nous avons considéré les distributions suivantes :

1. Une *loi* Gaussian $\mathcal{N}(0,1)$: $g_1(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$
2. Une *loi* Cauchy $\mathcal{C}(0,1)$: $g_2(x) = [\pi(1+u^2)]^{-1}$
3. Une *loi* Uniforme $\mathcal{U}(-1,1)$: $g_3(x) = \frac{1}{2}\mathbf{1}_{(]-1;1[)}$

Nous obtenons les résultats suivants :

n		Gaussian $\mathcal{N}(0,1)$	Cauchy $\mathcal{C}(0,1)$	Uniforme $\mathcal{U}(-1,1)$
10				
	h_{nrd}	0.5363149	1.159971	0.3595413
	h_{pi}	0.609302	1.286969	0.3467531
	h_{ucv}	0.9548033	0.9996075	0.6195964
	h_{bcv}	0.9996259	0.9995488	0.9996075
50				
	h_{nrd}	0.4757543	0.7216755	0.275477
	h_{pi}	0.4405524	0.7025794	0.2176408
	h_{ucv}	0.4843378	0.8701536	0.2454644
	h_{bcv}	0.6306482	0.8286498	0.4908226
100				
	h_{nrd}	0.4040319	0.5246375	0.251836
	h_{pi}	0.3688296	0.441237	0.1725156
	h_{ucv}	0.3793846	0.1257297	0.1680829
	h_{bcv}	0.4950572	0.507616	0.4898627
$n = 1000$				
	$h.nrd$	0.2520506	0.4002166	0.152942
	$h.pi$	0.2098817	0.1452667	0.100814
	$h.ucv$	0.2057205	0.01039247	0.05294
	$h.bcv$	0.2276954	0.4104829	0.1283487

TAB. 3.1 – Résultats de simulation pour la sélection du paramètre de lissage par : nrd, pi, ucv, bcv.

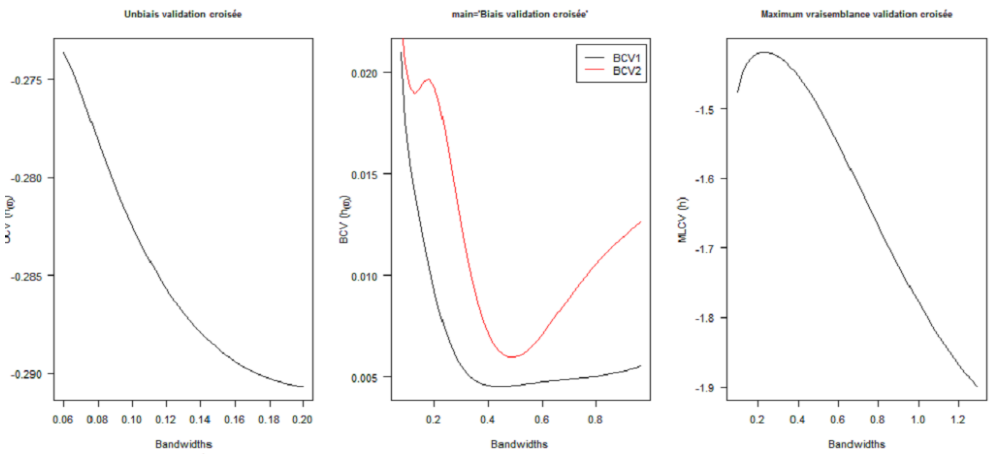

FIG. 3.1 – Les courbes de validation croisée de noyau gaussien.

3.1.2 Cas bivarié

$n = 500$

Le noyau **k :** on a noyau normale multivarié $\mathcal{N}_2\left(\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{\Sigma} = \begin{pmatrix} 1 & -0.75 \\ -0.75 & 2 \end{pmatrix}\right)$

H_{pi} : matrice de lissage obtenu par la technique de plug-in.

H_{lscv} : matrice de lissage obtenu par la technique de validation croisée non bias.

H_{bcv} : matrice de lissage obtenu par la technique validation croisée biais.

La comparaison de estimation kernel density (*KDEs*) .

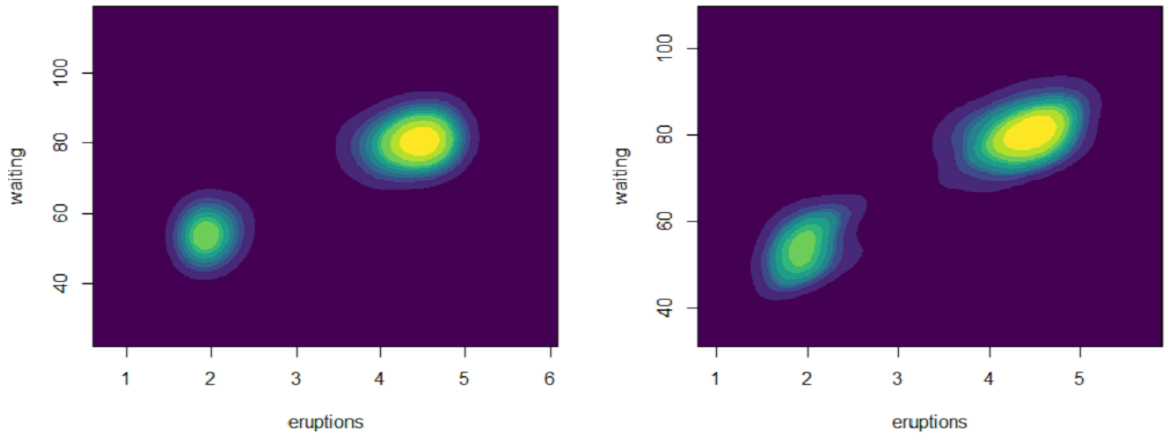

FIG. 3.2 – Comparaison de KDE's.

3.1.3 Application sur des données réelles

On considère les données de qualité de l'air mesuree dans la station Chatelet,Regie autonome des transports parisiens - Departement Developpement, Innovation et Territoires. Accessed 2017-09-27. (Package *ks*)

n		**bivarie**	**trivarie**
10	H_{pi}	$\begin{pmatrix} 1.076893 & 1.540755 \\ 1.540755 & 2.891265 \end{pmatrix}$	$\begin{pmatrix} 0.4043681 & 0.6034478 & 2.799390 \\ 0.6034478 & 1.1484308 & 4.993724 \\ 2.7993901 & 4.9937239 & 132.044996 \end{pmatrix}$
	H_{LSCV}	$\begin{pmatrix} 0.7375502 & 0.1317229 \\ 0.1317229 & 0.9423255 \end{pmatrix}$	$\begin{pmatrix} 0.441934 & 0.459516 & 8.182571 \\ 0.459516 & 1.525472 & 28.064118 \\ 8.182571 & 28.064118 & 571.535510 \end{pmatrix}$
	H_{bcv}	$\begin{pmatrix} 0.89824777 & -0.00716349 \\ -0.00716349 & 3.78591653 \end{pmatrix}$	
30	H_{pi}	$\begin{pmatrix} 23.51231 & 20.85708 \\ 20.85708 & 22.32520 \end{pmatrix}$	$\begin{pmatrix} 16.16799 & 14.02759 & 37.64394 \\ 14.02759 & 14.68434 & 35.30183 \\ 37.64394 & 35.30183 & 319.42706 \end{pmatrix}$
	H_{LSCV}	$\begin{pmatrix} 34.31736 & 45.75647 \\ 45.75647 & 61.00861 \end{pmatrix}$	$\begin{pmatrix} 8.186643 & 10.37092 & 73.27173 \\ 10.370925 & 14.22315 & 115.87869 \\ 73.271733 & 115.87869 & 1485.25626 \end{pmatrix}$
	H_{bcv}	$\begin{pmatrix} 47.9559329 & -0.3123023 \\ -0.3123023 & 33.0569364 \end{pmatrix}$	
1000	H_{pi}	$\begin{pmatrix} 0.26258868 & 0.03375926 \\ 0.03375926 & 0.23619016 \end{pmatrix}$	
	H_{LSCV}	$\begin{pmatrix} 0.2309584 & -0.1278430 \\ -0.1278430 & 0.3090394 \end{pmatrix}$	
	H_{bcv}	$\begin{pmatrix} 7.832413e-01 & -2.529089e-05 \\ -2.529089e-05 & 6.742428e-01 \end{pmatrix}$	

TAB. 3.2 – Exemples de noyaux continus symétriques bivarie et trivarie .

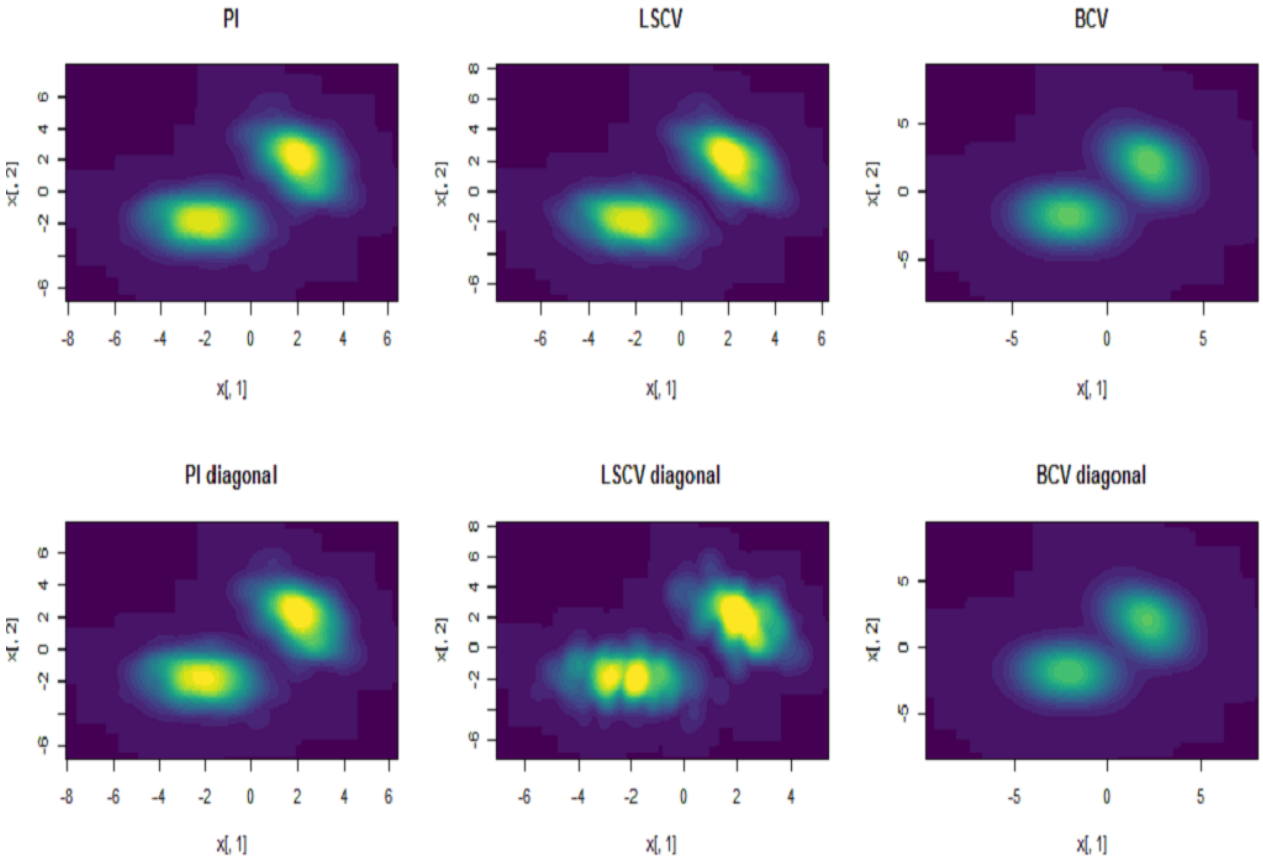

FIG. 3.3 – Comparaison de selection matrice de lissage : $Hbcv$, Hpi, $Hlscv$.

3.2 Noyau associé asymétrique multivarié

Résultats de simulation

Pour beta bivariate (package *bivariate*)

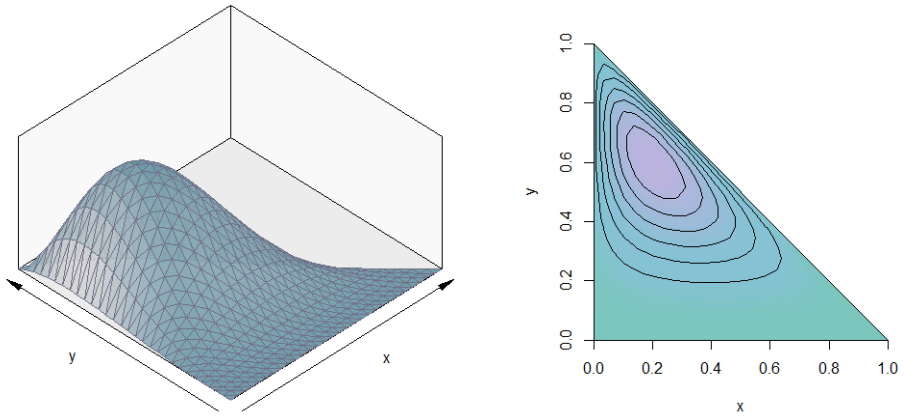

FIG. 3.4 – Beta bivariée (2,4,2)

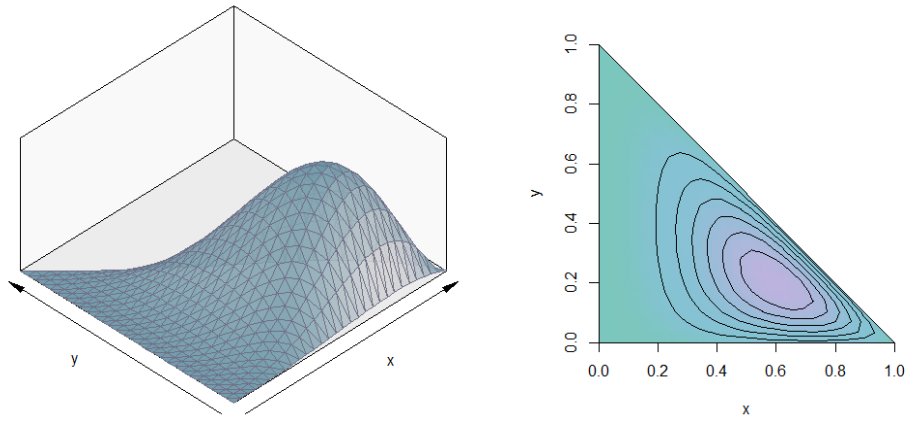

FIG. 3.5 – Beta bivariée (4,2,2)

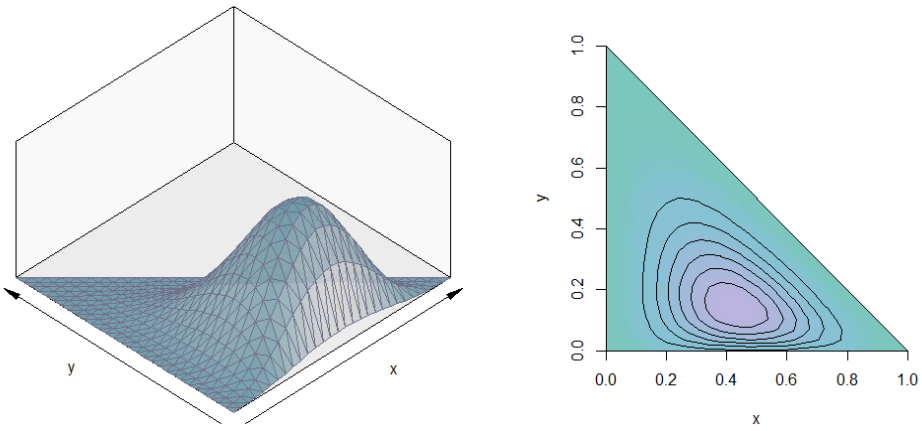

FIG. 3.6 – Beta bivariée (4,2,4)

Conclusion

Ce mémoire porte sur l'estimation non-paramétrique de la densité de probabilité par la méthode du noyau multidimensionnelle à support continu et discret, en utilisant la méthode du noyau associé.

Dans un premier lieu, la notion originale de l'estimateur à noyau d'une densité de probabilité, et qui se base sur des noyaux symétriques, asymétriques, à était mis en évidence et cela en introduisant sa forme, ses propriétés, le choix du noyau ainsi que certaines des procédures de sélection du paramètre de lissage proposées.

Dans un second lieu, On présente d'abord l'état de l'art sur l'estimateur à noyau symétrique et asymétrique de la densité multivarié continue et ses propriétés statistiques. Les différentes techniques de sélection de la matrice de lissage ont été rappelées.

Finalement, l'étude de simulation réalisée dans ce mémoire pour comparer les méthodes d'estimation de la matrice de lissage H, pour différents noyaux discrets et continu symétrique et asymétrique.

Il sera intéressant de compléter ce travail par :

- Réaliser une simulation extensive toute en considérant d'autres noyaux et d'autres lois.
- Poursuivre la réflexion autour du noyau associé mixte (à la fois discrets et continus).
- Concernant les noyaux associés non-classiques et multivariés, il serait intéressant de faire une étude fine de l'algorithme de réduction du biais sur la qualité de l'estimation, et de les comparer aux estimateurs à noyaux standards et à noyaux classiques multivariés.

Bibliographie

[1] A. K. Gangopadhyay and K. N. Cheung. Bayesian approach to the choice of smoothing parameter in kernel density estimation. Journal of Nonparametric Statistics, 14 :655–664, 2002.

[2] A. Spurdle. Package "bivariate". https ://cran.r-project.org/web/packages/bivariate/index.html. 26-02-2020.

[3] A.W. Bowman. An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71 :353–360, 1984.

[4] C.C. Kokonendji and T.S Kiessé. Estimateur à noyau discrèt standard pour une densité de probabilité discrète. Université de Pau et des Pays de l'Adour, France, 36, Juin 2007.

[5] C. C. Kokonendji and S. M. Somé. On multivariate associated kernels for smoothing general density function. arXiv : 1502.01173, 2015.

[6] Cybakov, A. B. Introduction à l'estimation non paramétrique, vol. 41. Springer Science & Business Media, 2003.

[7] Epanechnikov, V. Nonparametric estimation of a multidimensional probability density. Teoriya Veroyatnostei i ee Primeneniya 14, 1 (1969), 156–161.

[8] Hodges, J., and Lehmann, E. The efficiency of some nonparametric competitors of the t-test. Annals Mathematicals statistzcs 27(1) (1956), 324–335.

[9] M. P. Wand and M. C. Jones. Multivariate plug-in bandwidth selection. Computational Statistics, 9 :97–116, 1994.

[10] N. Zougab. Approche Bayéssienne dans l'estimation non paramétrique de la densité de probabilité et de la courbe de régression de la moyenne. Thèse de doctorat en mathématique appliqué, Université A/MIRA, Béjaia, Février 2013.

[11] S. M. Somé. Estimation non-paramétrique par noyaux associés multivariés et applications. Thèse de Doctorat, Université de Bourgogne Franche-Comté, France, 2015.

[12] S. R. Sain, K. A. Baggerly, and D. W. Scott. Cross-validation of multivariate densities. Journal of the American Statistical Association, 89 :807–817, 1994.

[13] Scott, D. W. Multivariate Density Estimation :Theory, Practice, and Visualization. Wiley Interscience, New York, 1992.

[14] Silverman, B. W. Density estimation for statistics and data analysis, vol. 26. CRC press, 1986.

[15] Simonoff, J. S. Smoothing methods in statistics. Springer Science & Business Media, 2012.

[16] T. Duong and M. L. Hazelton. Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics, 32 :485–506, 2005.

[17] T. Duong, M. Wand, J. Chacon, A. Gramacki. Package"ks". https ://cran.r-project.org/web/packages/ks/index.html. 11-02-2020.

[18] T.S. Kiessé. Approche non-paramétrique par noyaux associés discrets des données de dénombrement. Thèse de doctorat en mathématique appliqué, Université de Pau et des Pays de l'Adour, France, Octobre 2008.

[19] Wand, M. P., and Jones, M. C. Kernel smoothing. Crc Press, 1994.

[20] Wansouwé, W. E., Kokonendji, C. C., and Kolyang, D. T. Nonparametric estimation-for probability mass function with Disake. ARIMA Journal, vol. 19, pp. 1-23 (2015)

Abréviations et Notations

$K(u)$	Noyau.
h_{opt}	La fenêtre optimale.
$\mathbf{\Sigma}_K$	Matrice de variance-covariance
$vech(H)$	Demi-vecteur de H, vecteur de $(\frac{1}{2}d(d+1) \times 1)$.
$\pi(x)$, $\pi(\theta\|x)$	*A priori, .A posteriori.*
$f(\theta\|x)$	Loi de probabilité, indexée par le paramètre θ.
S_d	Méthodes de Monte Carlo par chaîne de Markov.
MSE	L'erreur quadratique moyenne.
ISE	L'erreur quadratique intégrée.
$MISE$	L'erreur quadratique moyenne intégrée.
$AMISE$	L'erreur quadratique moyenne intégrée asymptotique.
LCV	Validation croisée de vraisemblance.
BCV	Validation croisée biais.
UCV où $LSCV$	Validation croisée non biais.
PI	Plug-in.

Résumé

Dans ce travail, on a présenté l'estimateur non-paramétrique par noyaux (symétriques et asymétriques) associés multivariés de la fonction de probabilité. On a présenté également ses propriétés statistiques à distance finie (biais, variance, erreur quadratique moyenne et erreur globale) et ses propriétés asymptotiques. La matrice de lissage est estimée par l'approche classique validation croisée ainsi que par méthode plug-in et les approches bayésiennes. Les études de comparaison entres les approches bayésiennes et la validation croisée en utilisant le critère de l'erreur quadratique intégrée, sur des données simulées et réelles montrent les bonnes performances des approches bayésiennes.

Mots clés : Estimation non paramétrique de densité, noyaux associés multivariés, paramètre de lissage , matrice de lissage, validation croisée, plug-in, approche bayésienne.

Abstract

In this work, we presented the nonparametric estimator with multivariate associated kernels (symmetric and asymmetric) for probability function. We also presented its statistical properties (bias, variance, mean square error and the global error) and its asymptotic properties. The matrix of bandwidths is estimated by classical cross validation method, plug-in method and the Bayesian approaches. The comparison studies among Bayesian approaches and cross-validation by using integrated square error on simulated and real count data show the good performances of Bayesian approaches.

Key words : Nonparametric density estimation, multivariate associated kernels, smoothing parameter, matrix of bandwidths, cross validation, plug-in, bayesian approach.

ملخص

في هذا العمل، قدَرنا دالة كتلة الإحتمالات في مجال غير مترابط، بطريقة النواة مُتَعدد المتغيرات. قدمنا لهذا المقدر الخصاص لمسافة محدودة و كذلك التقاربية، تم إختيار مصفوفة النوافذ بالطريقة الكلاسيكية،طريقة التوصيل و الطريقة البيزيانية. قمنا بمقارنة فعالية الطريقة البيزيانية و الطريقة الكلاسيكية ن طريق معطيات مصنوعة و حقيقية. النتائج المتحصل عليها تُبين أن الطرق البيزيانية أحسن من الطريقة الكلاسيكية.

كلمات مفتاحية: دالة كتلة الإحتمالات في مجال غير مترابط، **نواة متعدد المتغيرات، تجانس المعلمة، مصفوفة النوافذ، الطريقة الكلاسيكية، طريقة توصيل، الطريقة البيزيانية.**