République Algérienne Démocratique et Populaire.

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique.

Université de Béjaïa.

Faculté des Sciences et des Sciences de l'Ingénieur

MÉMOIRE DE MAGISTER

En Mathématiques Appliquées

Option

Modélisation Mathématique et Techniques de Décision

Thème

Sur l'estimation non-paramétrique de la densité de probabilité dans le cas multidimensionnel.

Présenté par :

Said BEDDEK

Devant le jury composé de :

D. AISSANI	**Président**	**Professeur**	**U. de Béjaïa**
S. ADJABI	**Rapporteur**	**M. C. A**	**U. de Béjaïa**
M. O. BIBI	**Examinateur**	**Professeur**	**U. de Bejaïa**
Z. MOHDEB	**Examinateur**	**Professeur**	**U. de Constantine**
K. Cherfi-LAGHA	**Invitée**	**M. C. B**	**U. de Béjaïa**

Béjaïa, Avril 2011

Table des matières

Table des figures

Liste des tableaux

Introduction générale

Les problèmes d'estimation fonctionnelle concernent essentiellement les modèles statistiques non-paramétriques. En effet, un tel modèle, d'après sa définition même, ne peut être indexé que par un paramètre évoluant dans un espace vectoriel topologique à une infinité de dimensions, c'est-à-dire un espace fonctionnel.

Un exemple classique de modèle non-paramétrique pour une variable aléatoire réelle est obtenu en postulant que la loi de probabilité de cette variable est absolument continue et possède une densité uniformément continue. On peut choisir cette densité comme paramètre fondamental du modèle et l'on voit aisément que l'ensemble de ses valeurs est une partie convexe de l'espace vectoriel constitué par les fonctions réelles uniformément continues et intégrables sur $\Re^d$.

La statistique non-paramétrique a connu un développement considérable depuis une cinquantaine d'années. Ce qui a donc impliqué une intense activité dans le domaine de la recherche sur l'estimation fonctionnelle, en se fixant pour objectif quelque thèmes principaux, en particulier :

1. Méthode de construction des estimateurs.
2. Propriétés statistiques des estimateurs (convergence et vitesse de convergence).
3. Etude de l'optimalité des estimateurs.
4. Estimation adaptive.

Un problème spécial a plus retenu l'attention des chercheurs, il s'agit de l'estimation de la densité de probabilité et ses applications. Le nombre de publications qui lui furent consacrées est impressionnant, il doit s'élever à des milliers et ce flot ne semble guère tari.

Parmi les livres de référence sur ce sujet, on peut citer entre autres : Le livre de **Silverman** ([85], 1986) qui a été cité plus de deux milles fois, **Bosq et Lecoutre** ([10], 1987), **Scott** ([79], 1992), **Simonoff**([86], 1996), **Wand et Jones**([99], 1995) et celui de **Tsybakov** ([94], 2004).

L'estimation de la densité de probabilité a investi plusieurs domaines d'application, notamment :

– L'archéologie : **Baxter, Beardah et Westwood** ([4], 2000).

- Les banques : **Tortosa et Ausina** ([93], 2002).
- La météologie : **Ferreyra and al.** ([36], 2001).
- L'économie : **Dinardo, Fortin et Lemieux** ([28], 1996).
- La génétique : **Segal et Wiemels** ([83], 2002).
- L'hydrologie : **Kim et Heo** ([59], 2002).
- La physiologie : **Paulsen et Heggelund** ([71], 1996).
- Le traitement d'image : **Aurélie bugeau et Patrick perez** ([15], 2007).

Le premier objectif de ce mémoire est d'introduire l'estimateur à noyau de la densité de probabilité multidimensionnelle et d'étudier ses propriétés statistiques. Le deuxième objectif est de faire le point sur les différentes procédures de sélection de la matrice de lissage H. Le dernier objectif de ce travail est d'appliquer et comparer les différentes méthodes de sélection de la matrice de lissage sur plusieurs distributions connues.

Pour cela nous avons structuré notre travail en quatre parties principales. Le premier chapitre est une introduction générale à l'estimation fonctionnelle dans le cas multidimensionnel, dans lequel nous avons fixé le cadre général et présenté les notions et les concepts élémentaires de cette théorie. En particulier, nous nous sommes intéressé à l'estimation de la densité de probabilité multivariée par la classe d'estimateurs à noyau généralisé qui regroupe la plupart des estimateurs usuels connus dans la littérature (l'histogramme, estimateur par des séries orthogonales, estimateur à noyau de convolution...). Le deuxième chapitre est consacré à l'estimation de la densité de probabilité multivariée par la méthode du noyau de convolution ou tout simplement méthode du noyau, qui est un cas particulier très important de la méthode du noyau généralisé. La méthode d'estimation à noyau est la plus populaire parmi les différentes méthodes d'estimation de la densité de probabilité, car elle présente de bonne propriétés statistiques et elle est simple à interpréter et à implémenter. Les différentes méthodes de sélection de la matrice de lissage pour l'estimateur à noyau sont exposées dans le troisième chapitre.
Enfin, dans le quatrième et dernier chapitre , nous présentons les résultats des simulations conduites à partir de plusieurs densités cibles connues, afin de :

- Comparer les différents algorithmes de sélection de la matrice de lissage ;
- Étudier la performance de ces algorithmes ;
- Étudier l'influence de la taille de l'échantillon sur ces différents algorithmes.

Ce mémoire se termine par une conclusion générale et quelques propositions d'axes de recherche sur ce problème d'estimation de la densité de probabilité multidimensionnelle.

1

Généralités sur l'estimation fonctionnelle

1.1 Introduction

L'objectif de ce chapitre est de fixer le cadre général de la théorie de l'estimation fonctionnelle ou non-paramétrique, d'introduire les notions et les concepts élémentaires de cette théorie et de présenter ainsi quelques résultats classiques connus dans la littérature sur ce sujet.

1.2 Cadre général de l'estimation fonctionnelle

Pour fixer le cadre général de l'estimation fonctionnelle, reprenons la présentation de **Bosq et Lecoutre** ([10], 1987) :

Considérons un modèle statistique $(E; A; P)$, où E est l'espace des observations, A une tribu sur E et P une famille de mesures de probabilités sur $(E; A)$. On écrit généralement la famille P sous la forme :

$$P \quad = \quad \{ \quad P_\vartheta \quad , \quad \vartheta \in \Theta \quad \} .$$

Le problème posé est celui de l'estimation du paramètre $g(P_\vartheta)$, où :
– g est une fonction définie sur P et à valeurs dans un espace Θ'.
– P_ϑ est une loi inconnue dans P.
– Θ' est un espace vectoriel de dimension infinie, autrement dit un espace fonctionnel.

1.3 Modèle non-paramétrique

Bosq et Lecoutre ([10], 1987) ont admis qu'il n'y a pas de définition précise d'un modéle non-paramétrique car la frontière entre paramétrique et non-paramétrique est assez floue. Cependant, pour fixer les idées, nous adoptons dans notre travail la définition suivante, définition implicitement admise dans une grande partie de la littérature :

Définition 1.3.1. *Le modèle d'estimation du paramètre* $g(P_\vartheta)$ *est dit non-paramétrique lorsque les hypothèses sur* $(E; A; P)$ *ne permettent pas d'écrire* $g(P_\vartheta)$ *en fonction d'un nombre fini de paramètres réels.*

Remarque 1.3.1.

1. *Compte tenu de la nature de l'espace* Θ' *et de la définition ci-dessus (1.3.1), nous allons parler dans la suite indifféremment d'estimation fonctionnelle ou d'estimation non-paramétrique.*

2. *Généralement, on considère le modèle statistique*

$$(E; A; P) = \left(\left[\Re^d\right]^n ; \beta^n ; [\mu^n]_{\mu\in P_0}\right),$$

où :
– β *est la tribu borélienne de* $\Re^d$,
– P_0 *est une famille de probabilité sur* $\Re^d$,
– $\mu \in P_0$ *est une loi de probabilité sur* $\Re^d$ *et* μ^n *est la probabilité produit sur* $\left[\Re^d\right]^n$,

et on pose : $X = (X_1; ...; X_n)$ *, où les :* X_i ; $1 \leq i \leq n$ *; sont des variables aléatoires à valeurs dans* $\Re^d$ *muni de sa tribu borélienne* β*. On suppose en outre que les :* X_i ; $1 \leq i \leq n$ *; sont indépendantes et de même loi de probabilité* $\mu \in P_0$*. Autrement dit :* $X_1; ...; X_n$ *; est un échantillon de taille* n *et de loi* μ *.*

Définition 1.3.2. *Un estimateur* $\widehat{g}_n(P_\vartheta)$ *est une fonction de l'échantillon. Autrement dit :*

$$\hat{g}_n = L(X_1; ...; X_n),$$

pour une certaine fonction $L : \left[\Re^d\right]^n \longmapsto \Re$ *.*

Remarque 1.3.2.
Selon la définition (1.3.2), un estimateur n'est rien d'autre qu'une statistique. La seule différence est que cette statistique sert spécifiquement à obtenir de l'information sur le paramètre inconnu $g(P_\vartheta)$ *.*

1.4 Exemples de modèles non-paramétriques

Nous allons maintenant donner quelques exemples de modèles non-paramètriques, les plus connus dans la littérature.

Fonction de répartition

C'est la fonction définie par :

$$F_\mu\left(x_1;...;x_d\right)=\mu\left\{\prod_{i=1}^{d}\ (-\infty\ ;\ x_i\ ]\right\},$$

où : $(x_1;...;x_d)\in\Re^d$.

Le paramètre g est l'application de l'ensemble $P\ =\ \{\mu^n\ ;\ \mu\in P_0\}$ dans $D_b\left(\Re^d\right)$ définie par :

$$\mu^n\longmapsto F_\mu,$$

avec : $D_b\left(\Re^d\right)$ est l'espace des fonctions réelles sur $\Re^d$, bornées, continues par morceaux et n'admettant que des discontinuités de première espèce.

Fonction caractéristique

Pour tout entier i tel que : $1\ \leq\ i\ \leq\ n$; on définit la fonction caractéristique de la variable aléatoire X_i par :

$$\hat{\mu}\left(t\right)\ =\ E_\mu[e^{(j<t;X_i>)}].$$

Où :

– j est le nombre complexe imaginaire vérifiant $j^2=-1$.

– $t=(\ t_1;...;t_d\ )\ \in\ \Re^d$.

– $X_i=\left(\ X_i^1;...;X_i^d\ \right)$.

– $<.;.>$ est le produit scalaire usuel sur $\Re^d$.

Le paramètre g est l'application de l'ensemble $P\ =\ \{\mu^n\ ;\ \mu\in P_0\}$ dans $C_b\left(\Re^d\right)$ définie par :

$$\mu^n\longmapsto\hat{\mu};$$

avec $C_b\left(\Re^d\right)$ est l'espace des fonctions réelles ou complexes définies sur $\Re^d$ continues et bornées.

Paramètres de régression

Posons $X_i=(X_i^1;X_i^2)\,;1\leq i\leq n\,;$ où X_i^1 prend ses valeurs dans $\Re^{d_1}$ et X_i^2 prend ses valeurs dans $\Re^{d_2}$ avec $d_1+d_2=d$.

Soit pour tout $i\,;\ (1\leq i\leq n)\,;\ \left\{\mu^x_{X_i^2};x\in\Re^{d_1}\right\}$ une famille de versions spécifiées des lois de X_i^2 sachant que $X_i^1=x$.

Toute fonction r de la forme $x\longmapsto r\left(\mu^x_{X_i^2}\right)$ est un paramètre de régression.

Le plus important est la fonction de régression (ou espérance conditionnelle), notée :

$$m_{\psi}(x) = E\left[\psi\left(X_i^2\right)/X_i^1 = x\right]$$

Où $\psi(.)$ est une fonction mesurable à valeurs dans $\Re^{d_2}$.

Fonction quantile

Pour $d = 1$, la fonction quantile est définie par :

$$F_{\mu}^{-1}(p) = inf\left\{t \in \Re : F_{\mu}(t) \geq p\right\}, \quad 0 < p < 1.$$

F_{μ}^{-1} est un paramètre à valeurs dans l'espace des fonctions réelles définies sur] 0 ; 1 [, non décroissantes et continues à gauche.

La densité de probabilité et ses dérivées

Si la loi de probabilité μ est absolument continue par rapport à la mesure de **Lebesgue** λ de $\Re^d$ (on dit aussi que P_0 est dominée par λ), alors d'après le théorème de **Radon-Nikodym**, il existe une application mesurable f_{μ} de $\Re^d$ dans $\Re^+$, intégrable-Lebesgue, telle que pour toute partie mesurable B de $\Re^d$:

$$\mu(B) = \int_B f_{\mu}(x)\, dx.$$

f_{μ}, unique à une équivalence prés (pour l'égalité λ-presque partout sur $\Re^d$), est appelée densité de la variable aléatoire X_i pour tout i $(1 \leq i \leq n)$.

Enfin, si f_{μ} est différentiable, on définit de nouveaux paramètres fonctionnels : les dérivées partielles de la densité.

Fonction du hasard

En combinant les exemples précédents, on peut construire de nouveaux paramètres, les paramètres composés, notamment la fonction du hasard définie en tout point x de $\Re^d$, par :

$$\lambda(x) = \frac{f_{\mu}(x)}{1 - F_{\mu}(x)},$$

avec $F_{\mu}(x) < 1$.

Dans le cas $d = 1$ et lorsque la variable X représente une durée de vie, la fonction λ correspond à une probabilité instantanée de décès à l'instant x.

1.5 Estimation fonctionnelle par la méthode du noyau généralisé

Dans ce paragraphe, nous allons présenter et étudier une classe particulière d'estimateurs, introduite pour la première fois par **A. Földes** et **P. Révész** ([38], 1974) puis indépendamment par **J. Bleuez** et **D. Bosq** ([5], 1976), appelés estimateurs à noyau généralisé. Cette classe contient de nombreux estimateurs usuels (L'histogramme, méthode du noyau de convolution, méthode par projection orthogonale...). Il s'agit d'estimateurs de la forme :

$$\widehat{g}_n(t) = \frac{1}{n} \sum_{i=1}^{n} K_{r_n}(X_i; t) \qquad t \in E,\ n \in N^*, \tag{1.1}$$

Où : $(E; A; \mu)$ est un espace mesuré σ-fini, $(K_r, r \in I)$ est une famille de fonctions numériques définies sur $E \times E$ (on dit aussi que (K_r) est une famille du noyaux généralisés sur E) , I désignant une partie non-bornée de $\Re_+$ et (r_n) une suite qui tend vers l'infini dans I.

1.5.1 Existence de l'estimateur à noyau généralisé

Soit $(E; A; \mu)$ un espace mesuré où μ est une mesure σ-finie. Soit P_0 l'ensemble des probabilités sur $(E; A)$ qui admettent, par rapport à μ, une densité de carrée μ-intégrable. On note par $\ell^k(\mu)$ l'espace des fonctions de k^{ime} puissance μ-intégrable $(1 \leq k < +\infty)$ et par $L^k(\mu)$ l'espace quotient de $\ell^k(\mu)$ par la relation d'égalité μ-presque partout.

On définie une application φ de P_0 dans $L^2(\mu)$ par la formule :

$$\varphi(p) = \frac{dp}{d\mu} \ ; \ p \in P_0, \tag{1.2}$$

Enfin on se donne $P \subseteq P_0$ et une application Ψ de $\varphi(P)$ dans $L^2(\mu)$.

Dans toute cette partie on supposera que $L^2(\mu)$ est séparable et que les opérateurs de $L^2(\mu)$ sont munis de leur relation d'ordre usuelle.

Définition 1.5.1. *Soit T un estimateur d'ordre n de $\Psi o \varphi$. Il sera dit sans biais (pour la version T^*) si T^* est une version régulière de T telle que :*

1. $\forall p \in P$, $T^*(., ..., .; t) \in \ell^1(p^{\otimes n})$ *pour μ presque tous t.*
2. $\forall p \in P$, $\int T^*(x_1, ..., x_n; .)\, dp^{\otimes n}(x_1, ..., x_n)$ *est une version régulière de $\Psi o \varphi(p)$.*

Remarque 1.5.1.

Rappelons que T^*, version de T, est dite régulière si $T^*(., ..., .; .)$ est A^{n+1}-mesurable. Nous supposerons que les estimateurs considérés possèdent une telle version.

Lemme 1.5.1. *Si T est un estimateur d'ordre n dont la norme est de carrée $p^{\otimes n}$-intégrable, pour tout $p \in P$, alors T est sans biais au sens de la définition précédente (1.5.1) si et seulement si il l'est au sens usuel de l'intégrable de bochner.*

Démonstration

Une démonstration complète de ce lemme à été donnée par **D. Bosq** dans ([8]; 1977a) (voir Lemme 2).

Proposition 1.5.1. *Une condition nécessaire et suffisante pour que $\Psi o\varphi$ possède un estimateur sans biais d'ordre 1 est que Ψ soit un opérateur intégral dont le noyau K soit tel que $K(x,.) \in \ell^2(\mu)$ pour tout $x \in E$.*

Démonstration

C'est une conséquence directe de la définition précédente (1.5.1). (voir **J. Bleuez** et **D. Bosq** dans ([6], 1978)).

Proposition 1.5.2. *Si la tribu des événements symétriques est complète pour P et si $K(.,.) \in \ell^2(p \otimes \mu)$ pour tout $p \in P$, alors, pour tout entier $n \geq 1$, la formule :*

$$T_n^*(x_1, ..., x_n; t) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, t); \qquad (x_1, ..., x_n; t) \in E^{n+1},$$

définit une version régulière de l'unique estimateur sans biais d'opérateur de covariance minimum pour $\Psi o\varphi$.

Démonstration

La proposition se déduit aisément de la version hilbertienne du théorème de **Lehmann-Scheffé**, voir **D. Bosq** dans ([7]; 1976) (Proposition 5).

1.5.2 Application à l'estimation de la fonction de répartition

Soit $a = \left(a^1, ..., a^d\right)$ un élément de $\Re^d$. Sur $\left(\Re^d, \beta_{\Re^d}, \lambda\right)$ (λ : mesure de Lebesgue sur $\Re^d$) considérons le noyau K défini par :

$$K(x, a) = I_{(-\infty, a^1] \times ... \times \left(-\infty, a^d\right]}(x);$$

où I est la fonction indicatrice et $x = (x^1, ..., x^d) \in \Re^d$.

On a :

$$\int_{\Re^d} K(x,a) f(x) dx = F(a);$$

où F désigne la fonction de répartition associée à f.

Les deux propositions (1.5.1) et (1.5.2) montrent alors que la fonction de répartition empirique :

$$F_n(a) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, a);$$

pour $x_i = \left(x_i^1, ..., x_i^d\right) \in \Re^d$, est sans biais et d'opérateur de covariance minimum comme estimateur de la fonction de répartition dans la famille P_0.

1.5.3 Application à l'estimation de la densité de probabilité

On commence par remarquer que, d'après la proposition (1.5.1), l'existence d'un estimateur sans biais d'ordre 1 de la densité signifie que la restriction de l'identité de $L^2(\mu)$ à $\varphi(P)$ est un opérateur intégral, or ceci n'est pas toujours vérifié. En général, la restriction de l'identité de $L^2(\mu)$ à $\varphi(P)$ est limite d'une suite d'opérateurs intégraux.

Dans la suite, on supposera que les densités de $\varphi(P)$ sont définies par une version privilégiée et l'on notera D le sous-ensemble de $\ell^2(\mu)$ formé par les versions considérées. D'autre part, V désignera l'espace vectoriel engendré par D. On a la définition suivante :

Définition 1.5.2. *Un estimateur d'ordre n à valeurs dans $\ell^2(\mu)$ sera toujours défini par une application numérique A^{n+1}-mesurable et un estimateur T_n de la densité sera dit sans biais si :*

$$E_p[T_n(X_1, ..., X_n; t)] = f_p(t); \qquad p \in P,\ t \in E;$$

où f_p est la version privilégiée de $\frac{dp}{d\mu}$ et où E_p désigne l'espérance prise par rapport à p.

Enfin on dira que T_n estimateur d'ordre n de la densité, vérifie la condition (1.3) si

$$T(., x_2, ..., x_n; t) \in V \quad pour \quad tout \quad (x_2, ..., x_n, t) \in E^n. \tag{1.3}$$

Maintenant on va introduire les résultats élémentaires suivants dus à **J. Bleuez** et **D. Bosq** ([6] ; 1978) :

Proposition 1.5.3. *Une condition nécessaire et suffisante pour qu'il existe un estimateur sans biais de la densité , d'ordre 1 et localement μ^2intégrable, est que μ soit une mesure discrète.*

Proposition 1.5.4. *Soit un entier $n \geq 1$. Si P est convexe et si μ n'est pas discrète, il n'existe aucun estimateur d'ordre n de la densité qui soit sans biais et de norme de carré intégrable.*

Théorème 1.5.1. *Une condition nécessaire et suffisante pour qu'il existe un estimateur des densités de D, d'ordre 1, sans biais et vérifiant la condition (1.3), est que V, muni du produit scalaire usuel de $\ell^2(\mu)$, soit un espace préhilbertien séparé à noyau reproduisant mesurable.*

Corollaire 1.5.1. *Si D est convexe, si $I_E \in V$ (I est la fonction indicatrice) et si V est un espace préhilbertien séparé à noyau reproduisant mesurable de noyau K, alors la formule :*

$$T_n(X_1, ..., X_n; t) = \frac{1}{n} \sum_{i=1}^{n} K(X_i, t); \qquad t \in E;$$

définie l'unique estimateur sans biais symétrique vérifiant la condition (1.3) et tel que :

$$T(., ..., .; t) \in \ell^1\left(p^{\otimes n}\right) \quad \texttt{pour tout} \quad p \in P \quad \texttt{et tout} \quad t \in E.$$

Enfin, on va conclure ce paragraphe par un résultat plus général donné par **Terrell** et **Scott** ([91], 1992) :

Théorème 1.5.2. *Tout estimateur de la densité de probabilité multivariée f, qui est une fonctionnelle de la fonction de répartition empirique $\hat{F}_n$ continue et Gâteaux différentiable, peut être écrit sous la forme :*

$$\hat{f}(t) = \frac{1}{n} \sum_{i=1}^{n} K(x_i, t, \hat{F}_n);$$

où K est la dérivée de Gâteaux de $\hat{f}$ suivant les variation de x_i.

Ce théorème repris par plusieurs auteurs est devenu un résultat fondamental pour l'estimation de la densité de probabilité multivariée, il introduit un certain ordre dans ce domaine et généralise tous les résultats obtenus jusqu'à présent. Une démonstration complète de ce théorème a été donnée par **Terrell** et **Scott** ([91], 1992) et par **Scott** ([79], 1992).

Rappel :

Soit f une fonction définie sur un ensemble S de $\Re^d$ à valeurs dans $\Re$ (on dit que f est un champ scalaire), x_0 un point intérieur de S et y un point arbitraire dans $\Re^d$. Pour

un scalaire $h \in \Re$ tel que $h \neq 0$ mais suffisamment petit pour garantir que $(x_0 + hy) \in S$, formons le quotient :

$$\frac{f(x_0 + hy) - f(x_0)}{h}. \tag{1.4}$$

Le numérateur de ce quotient nous dit comment varie la fonction lorsque nous bougeons de x_0 à $(x_0 + hy)$.

Définition 1.5.3. *La dérivée de f en x_0 par rapport à y, notée $f'(x_0; y)$, est définie par l'égalité*

$$f'(x_0; y) = lim_{h \to 0} \frac{f(x_0 + hy) - f(x_0)}{h},$$

lorsque la limite définie dans le membre de droite de l'égalité ci-dessus existe.

Définition 1.5.4. *Si y est un vecteur unitaire (c'est à dire $\|y\| = 1$), la dérivée $f'(x_0; y)$ est appelée dérivée directionnelle de f en x_0 suivant la direction y. En particulier, si $y = e_k$ (le k-ième vecteur unitaire des coordonnées), la dérivée directionnelle $f'(x_0; e_k)$ est appelée la dérivée partielle de f par rapport à e_k et également notée $D_k f(x_0)$. Ainsi, nous avons*

$$D_k f(x_0) = f'(x_0; e_k).$$

La fonction f est dite Gâteaux-dérivable en x_0 si et seulement si f est dérivable en x_0 suivant toutes les directions y, et que l'application $y \to f'(x_0; y)$ est linéaire. Cette dernière application linéaire est alors appelée ***dérivée de Gâteaux*** *de f en x_0.*

Cas particulier :

- **Estimation de la densité par la méthode du noyau de convolution**

En posant dans (1.1) :

$$K_{r_n}(X_i, t) = K_H(t - X_i), \quad \texttt{pour} \quad t \in \Re^d;$$

où :
- $K_H(t) = |H|^{-1/2} K\left(H^{-1/2} t\right)$,
- $r_n = |H|^{-1/2}$,
- $K(.)$: est une fonction numérique à d variables, appelée noyau,
- H : est une matrice d'ordre d, symétrique et définie positive, appelée matrice des paramètres de lissage ou matrice des fenêtres.

Sous cette forme l'estimateur de f apparaît comme la densité obtenue en régularisant la mesure empirique $\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$ (δ_a est la mesure de dirac au point a) par convolution avec $|H|^{-1/2} K\left(H^{-1/2}.\right)$. Ceci suggère de définir une classe d'estimateurs

en choisissant une famille convenable de fonctions K que nous appellerons (noyaux de convolution) ou plus simplement (noyaux), en anglais (Kernels). Ces estimateurs ferons l'objet d'une étude détaillée dans le prochain chapitre.

- **Estimation de la densité par projection**

Il est souvent possible d'approcher un paramètre fonctionnel f par un paramètre de dimension finie. Si ce nouveau paramètre est facile à estimer, on obtient un estimateur de f. Le cas le plus simple est celui où f prend ses valeurs dans un espace de Hilbert, on peut alors l'approcher par sa projection orthogonale sur un sous-espace de dimension finie convenablement choisi.

Ainsi, la densité est estimable par projection.

Supposons que $L^2(\mu)$ est de dimension infinie et séparable. On considère un sous-espace vectoriel H_n de $L^2(\mu)$, de dimension d_n finie et on désigne par π_n le projecteur orthogonal de H_n. Pour estimer f, on se propose de construire un estimateur sans biais de $\pi_n f$. En choisissant d'une façon homogène les représentants des éléments de H_n, on peut le considérer comme un espace vectoriel de fonctions. Soit alors K_{d_n} le noyau reproduisant de H_n et posons dans (1.1) :

$$K_{r_n} = K_{d_n},$$

on aura ainsi, un estimateur f_n de la densité à valeurs dans H_n, sans biais et symétrique par rapport aux observations :

$$f_n(X_1, ..., X_n; t) = f_n(t) = \frac{1}{n}\sum_{i=1}^{n} K_{d_n}(X_i, t).$$

Nous dirons que f_n est l'estimateur de la densité par projection sur le sous-espace H_n.

Définition 1.5.5. *soit maintenant $h_1, ..., h_{d_n}$ une base orthonormale de H_n, comme :*

$$K_{d_n}(x,t) = \sum_{i=1}^{d_n} h_i(x)h_i(t), \qquad (x,t) \in E \times E,$$

alors f_n s'écrit sous la forme :

$$f_n(t) = \sum_{i=1}^{d_n} a_{in}h_i(t), \qquad t \in E, \tag{1.5}$$

où a_{in} est un estimateur symétrique sans biais du coefficient de Fourier :

$$a_i = \int f h_i d\mu;\ i = 1, ..., d_n,$$

et :

$$a_{in} = \frac{1}{n}\sum_{j=1}^{n} h_i(x_j).$$

Remarque 1.5.2. *L'expression (1.5) explique pourquoi f_n est souvent désigné comme l'estimateur de la densité par la méthode des fonctions orthogonales.*

C'est **Cencov** ([18], 1962) puis **Tiago de oliveira** ([92], 1963) qui ont introduit la méthode des fonctions orthogonales pour estimer la densité. Pour plus de détails sur ce sujet, on poura consulter **Tarter** et **Lock** ([89], 1993) et **Saadi** et **Adjabi** ([74], 2009).

- **Estimation de la densité par l'histogramme**

Supposons maintenant que (E, A) est un espace métrique séparable avec A sa tribu borélienne et λ une mesure σ-finie, diffuse et ayant une valeur finie pour toute boule bornée. Nous désignons par F l'ensemble des densités uniformément continues sur (E, A, λ). On se propose d'estimer f, élément de F, à partir d'un échantillon $(X_1, ..., X_n)$ de loi μ et de densité f. Pour cela, nous avons besoin de l'hypothèse de régularité suivante, il existe une suite :

$$P_n = \{\pi_{nq}; q \in N_n\}, \quad N_n \subset N,$$

de partition équilibrées de E en boréliens. Pour estimer le paramètre $f = \frac{d\mu}{d\lambda}$ en un point x de π_{nq}, il est naturel de choisir $f_n(x) = \frac{\mu_n(\pi_{nq})}{\lambda(\pi_{nq})}$ où μ_n est la mesure empirique, ce qui conduit à la définition suivante :

Définition 1.5.6. *L'estimateur f_n associé à la partition P_n appelé histogramme des fréquence est défini par :*

$$f_n(x) = \frac{v_n(\pi_{nq})}{n\lambda(\pi_{nq})}, \qquad x \in \pi_{nq}, q \in N_n,$$

où $v_n(B)$ représente le nombre de points de l'échantillon qui appartiennent au borélien B.

La fonction f_n est P_n-simple, i.e. constante sur chacun des éléments de la partition P_n et n'est donc pas en général un estimateur strict de f. Elle vérifie évidemment :

$$\int f_n d\lambda = 1.$$

Pour assurer la convergence uniforme de f_n vers f, il est nécessaire que les partitions deviennent de plus en plus fines, condition que nous supposons réalisée.

- Cas particulier $E = \Re^d$:

On construit une suite de partitions de $\Re^d$ en pavés rectangulaires :

$$\pi_{nq} = [(q_1 - 1)h_n, q_1 h_n[\times ... \times [(q_d - 1)h_n, q_d h_n[, \quad q = (q_1, ..., q_d) \in Z^d;$$

où h_n est un nombre réel positif dépendant de n. L'histogramme s'écrit alors :

$$f_n(x) = \frac{v_n(\pi_{nq})}{nh_n^d}, \quad x \in \pi_{nq}, \quad q \in Z^d.$$

L'histogramme associé à une partition $\left(\pi_{n_q}, q \in Z\right)$ de $\Re^d$ correspond au choix dans (1.1) de :

$$K_{r_n}(x, y) = \sum_{q \in Z} I_{\pi_{n_q}}(x)\ I_{\pi_{n_q}}(y); \quad x, y \in \Re^d;$$

avec $r_n = 1/h_n$. On pourra se référer à **Scott** ([79], 1992), pour plus de détails.

1.6 Conclusion

Nous avons présenté dans ce chapitre des généralités sur l'estimation fonctionnelle, et nous avons abordé les différentes méthodes d'estimation de la densité de probabilité dans le cas multivarié. Dans le chapitre suivant, nous allons introduire l'estimateur le plus populaire pour estimer la densité de probabilité, à savoir l'estimateur à noyau de convolution dans le cas multivarié et nous donnerons les principales propriétés de cet estimateur.

2

Estimation de la densité de probabilité multivariée par la méthode du noyau

2.1 Introduction

Pour estimer la densité de probabilité multivariée, nous allons nous intéresser dans ce chapitre à une certaine classe d'estimateurs appelés : estimateurs à noyau de convolution ou tout simplement estimateurs à noyau (Multivariate kernel density estimator). C'est un cas particulier très important de la famille d'estimateurs à noyau généralisé introduite dans le premier chapitre .

L'estimation de la densité de probabilité par la méthode du noyau est devenue aujourd'hui, très répandue, elle est la plus utilisée par rapport aux autres méthodes (l'histogramme, méthode des séries orthogonales, estimateurs spline,...), car l'estimateur à noyau possède de bonnes propriétés, il est simple à interpréter et à implémenter. L'idée consiste à évaluer la densité f au point $x \in \Re^d$ ($d \in N^*$), en comptant le nombre d'observations tombées dans un certain voisinage de x dans $\Re^d$. Les principes de base de cette théorie ont été introduits par **Fix** et **Hodges** ([37], 1951) et **Akaike** ([3], 1954).

Ainsi, dans le cas univariée, si on dispose d'un n-échantillon $X_1, ..., X_n$, issu d'une variable aléatoire unidimensionnelle X admettant f comme densité de probabilité, alors l'estimateur à noyau $\hat{f}_n(x)$ de la densité f au point $x \in \Re$ fixé est de la forme suivante :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{j=1}^{n} K\left(\frac{x - X_j}{h}\right) \tag{2.1}$$

où :

- h est un paramètre fonction de n, appelé paramètre de lissage.
- K est une fonction définie sur $\Re$ appelé noyau (Kernel). Généralement, K est une densité de probabilité.

Le premier estimateur univarié de cette forme, avec : $K \equiv U_{[-1;1]}$ (loi uniforme sur $[-1, 1]$) a été proposé par **Fix** et **Hodges** ([37], 1951). La forme générale (2.1) a été étudiée par **Rosenblatt** ([72],1956) et **Parzen** ([70],1962).

Dans le cas multivarié, les premiers travaux remontent à **Maniya** ([63],1961) et **Nadaraya** ([68],1964), qui ont proposé des estimateurs à noyau pour la fonction densité bi-dimensionnelle (bivariée).

Dans son article publié en 1966, **Théophilos cacoullos** ([16],1966) a repris l'idée de **Nadaraya** ([68], 1964) et a proposé une extension d-dimensionnelle ($d \geq 2$) de l'estimateur à noyau avec un seul paramètre de lissage h. Dans ce cas, h est un scalaire strictement positif. **Epanechnikov** ([33],1969) a traité le cas où h est un vecteur de paramètres de lissage, c'est à dire : $h = (h_1, ..., h_d)^T$ avec $h_i > 0$ pour $i = 1, ..., d$.

Deheuvels ([26],1977b) est le premier à avoir introduit la forme générale de l'estimateur à noyau. L'estimateur dépend non pas d'un vecteur mais d'une matrice H de paramètres de lissage, symétrique et définie positive. **Singh** ([87],1976) a développé ces idées pour estimer les dérivées partielles de la densité multivariée. Enfin, **Wand** et **Jones** ([99],1995) ont présenté une monographie détaillée sur ce sujet.

2.2 Estimateur à noyau

Soit $X = (X_1, ..., X_d)$ un vecteur aléatoire d-dimensionnel, de fonction densité d-variée $f(x_1, ..., x_d)$. Soit $X^{(1)}, ..., X^{(i)}, ..., X^{(n)}$ un n-échantillon aléatoire issu de X, c'est-à-dire : $X^{(i)} = (X_1^{(i)}, ..., X_d^{(i)})$, $i = 1, ..., n$.

Définition 2.2.1. *Dans sa forme générale, l'estimateur à noyau de la densité de probabilité d-dimensionnelle f s'écrit :*

$$\hat{f}(x; H) = n^{-1} \sum_{i=1}^{n} K_H(x - X^{(i)}), \tag{2.2}$$

avec

$$K_H(x) = |H|^{\frac{-1}{2}} K(H^{\frac{-1}{2}} x), \tag{2.3}$$

où

- *H est une matrice carrée d'ordre d, symétrique et définie positive, appelée matrice des paramètres de lissage ou matrice des fenêtres (bandwidth matrix).*
- *$K(.)$ appelée fonction noyau d-dimensionnelle, est une application de $\Re^d$ dans $\Re$, bornée et vérifiant $\int_{\Re^d} K(x)dx = 1$.*

2.3 Construction du noyau multidimensionnel

Il y a deux façons pour construire le noyau d-dimensionnel K à partir d'un noyau univarié et symétrique w.

Définition 2.3.1. *On appelle noyau produit (product Kernel), le noyau K^p défini par :*

$$K^p(x) = \prod_{i=1}^{d} w(x_i). \tag{2.4}$$

Définition 2.3.2. *On appelle noyau sphérique (spherically or radially symmetric Kernel), le noyau K^s défini par :*

$$K^s(x) = C_{w.d}\ w\{(x^T x)^{\frac{1}{2}}\}, \tag{2.5}$$

avec

$$C_{w.d}^{-1} = \int w\{(x^T x)^{\frac{1}{2}}\}dx. \tag{2.6}$$

Exemple 2.3.1. *Pour le noyau Gaussien :*

$$w(u) = \frac{1}{\sqrt{2\pi}}\ e^{\frac{-u^2}{2}}, \quad u \in \mathbb{R}.$$

le noyau produit est défini par :

$$K^p(x) = \prod_{i=1}^{d} w(x_i) = (2\pi)^{-\frac{d}{2}} exp(-\frac{1}{2}x^T x).$$

Le noyau sphérique est défini par :

$$K^s(x) = C_{w,d}\ w\{(x^T x)^{\frac{1}{2}}\} = \frac{w\{(x^T x)^{\frac{1}{2}}\}}{\int w\{(x^T x)^{\frac{1}{2}}\}dx} = (2\pi)^{-\frac{d}{2}}\ exp(-\frac{1}{2}x^T x).$$

Nous remarquons que dans les deux cas, on a :

$$\begin{aligned} K_H(x - X^{(i)}) &= |H|^{-\frac{1}{2}}\, K\{H^{-\frac{1}{2}}(x - X^{(i)})\} \\ &= (2\pi)^{-\frac{d}{2}}\, |H|^{-\frac{1}{2}}\, exp\{-\frac{1}{2}(x - X^{(i)})^T H^{-1}(x - X^{(i)})\}. \end{aligned}$$

C'est-à-dire $K_H(x - X^{(i)})$ est la densité de la loi normale multidimensionnelle $\mathcal{N}(X^{(i)}, H)$ de moyenne $X^{(i)}$ et de matrice variance-covariance H.
Le noyau normal multidimensionnel est le seul noyau dont la version produit coïncide avec la version sphérique.

Noyaux usuels

Les noyaux les plus utilisés dans l'estimation de la densité de probabilité sont donnés dans le tableau suivant :

Noyau	w(u)
Rectangulaire	$\frac{1}{2}I_{[-1,1]}(u)$
Triangulaire	$(1-\|u\|)I_{[-1,1]}(u)$
Gaussien	$\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}, \qquad u \in R$
Epanechnikov	$\frac{3}{4\sqrt{5}}(1-\frac{u^2}{5})I_{[-\sqrt{5},\sqrt{5}]}(u)$
Biweight	$\frac{15}{16}(1-u^2)^2 I_{[-1,1]}(u)$
Triweight	$\frac{35}{32}(1-u^2)^3 I_{[-1,1]}(u)$
Cosine	$\frac{\pi}{4}\cos(\frac{\pi u}{2})I_{[-1,1]}(u)$
Gamma	$w_{\lambda,r}(u) = \frac{\lambda^r}{\Gamma(r)}e^{-\lambda u}u^{r-1}I_{[0,+\infty[}(u)$
Beta	$\frac{u^{\alpha-1}(1-u)^{\beta-1}}{\beta(\alpha,\beta)}I_{]0,1[}(u)$

TABLE 2.1 – Noyaux usuels

Remarque 2.3.1. *Pour l'estimation de la densité de probabilité multivariée, le noyau gaussien standard est le plus populaire lorsque le support de la densité est $\Re^d$. Dans ce cas, l'estimateur possède de bonnes propriétés asymptotiques (voir* ***Silverman*** *([85],1986),* ***Scott*** *([79],1992) et* ***Wand*** *et* ***Jones*** *([99],1995)). Cependant, cet estimateur n'est pas adapté, lorsque certaine variables sont bornées. Un problème de biais se pose aux bornes, ce qui entraîne la divergence de l'estimateur. Le problème de biais aux bornes a été largement étudié dans le cas univarié (voir par exemple* ***Schuster*** *([78],1985),* ***Müller*** *([67],1991) et* ***Cheng*** *et* ***al.*** *([20],1997)). Ce problème devient plus sévère dans le cas multivarié, car il s'ajoute au problème de dimension du support. Une première solution à ce problème dans le cas multivarié a été donnée par* ***Bouzermani*** *et* ***Rombouts*** *([11],2007). Ils proposent d'utiliser l'estimateur :*

$$\hat{f}(x_1,...,x_d) = \frac{1}{n}\sum_{i=1}^{n}\prod_{l=1}^{d}K^l(h_l, X_l^{(i)})(x_l),$$

où $(h_1,...,h_d)^T$ est le vecteur des paramètres de lissage et K^l est un noyau pour la variable x_l. Ces deux auteurs ont étudié deux cas de variables à support bornée :

1. *Dans le cas où le support de la variable est non-négative, On utilise K^l comme étant l'un des trois noyaux suivants :*
   - ***Le noyau*** K_L *(**local linear kernel**) :*

$$K_L(h,t)(x) = \frac{a_2(x,h) - a_1(x,h)y}{a_0(x,h)a_2(x,h) - a_1^2(x,h)}K(y),$$

où $y = \frac{x-t}{h}$, *K est un noyau symétrique univarié à support compact [-1,1] et*

$$a_l(x,h) = \int_{-1}^{x/h} t^l K(y)dy.$$

- ***Le noyau gamma*** K_G *:*

$$K_G(h,t)(x) = \frac{t^{\frac{x}{h}}\ exp\{-\frac{t}{h}\}}{h^{\frac{x}{h}+1}\Gamma(\frac{x}{h}+1)} = \Gamma(\frac{1}{h}, \frac{x}{h}+1). \tag{2.7}$$

- ***Le noyau gamma*** K_{NG} *:*

$$K_{NG}(h,t)(x) = \frac{t^{\rho(x)-1}exp\{-\frac{t}{h}\}}{h^{\rho(x)}\Gamma(\rho(x))} = \Gamma(\frac{1}{h}, \rho(x)), \tag{2.8}$$

où

$$\rho(x) = \begin{cases} \frac{x}{h} & si\ x \geq 2h \\ \frac{1}{4}(\frac{x}{h})^2 + 1 & si\ x \in [0, 2h). \end{cases}$$

2. *Si la variable est à support compact (pour simplifier nous considérons le support* $[0,1]$ *), on utilise le noyau Beta :*

$$K(h,t)(x) = B\left(\frac{x}{h}+1; \frac{1-x}{h+1}\right), \tag{2.9}$$

ou bien le noyau Beta modifié :

$$K_{NB}(h,t)(x) = \begin{cases} B(\rho(x)\ ;\ \frac{1-x}{h}) & si\ x \in [0, 2h) \\ B(\frac{x}{h}\ ;\ \frac{1-x}{h}) & si\ x \in [2h, 1-2h) \\ B(\frac{x}{h}\ ;\ \rho(1-x)) & si\ x \in [1-2h, 1]. \end{cases}$$

$B(\alpha, b)$ *est la fonction densité de la loi Beta de paramètres* α *et* β, h *est le paramètre de lissage et*

$$\rho(x) = 2h^2 + 2.5 - \sqrt{4h^4 + 6h^2 + 2.25 - x^2 - \frac{x}{h}}.$$

2.4 Choix de la paramétrisation

Soit F l'ensemble des matrices carrées d'ordre d, symétriques et définies positives. En général, la matrice $H \in F$ possède $\frac{1}{2}d(d+1)$ éléments indépendants et qui, même pour des dimensions modérées de l'espace, reste un nombre assez important de paramètres de lissage à choisir. Cependant, des simplifications peuvent être obtenues en imposant des restrictions à la matrice $H \in F$. Ce qui va nous conduire à considérer les différents cas suivants :

• 1^{er} **cas** :

Le plus simple est de prendre $H \in S$, tel que :

$$S = \{H \in F / H = h^2 I, h > 0\}.$$

C'est à dire, on choisit le même paramètre de lissage h pour chaque axe de coordonnées dans l'espace. On vient de définir l'estimateur de **Cacoullos** ([16],1966) :

$$\hat{f}_n(x;h) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x_1 - X_1^{(i)}}{h}, ..., \frac{x_d - X_d^{(i)}}{h}\right),\ x = (x_1, ..., x_d)^T \in \Re^d.$$

• 2^{eme} **cas** :
On peut aussi considérer $H \in D$, tel que :

$$D = \{\ H \in F\ /\ H = diag(h_1^2, ..., h_d^2)\ ;\ h_i > 0\ (i = \overline{1, d})\ \}.$$

C'est à dire, on définit un paramètre de lissage différent pour chaque axe de coordonnées. Nous venons ainsi de définir l'estimateur d'**Epanechnikov** ([14],1969) :

$$\hat{f}(x;h_1, ..., h_d) = n^{-1}(\prod_{l=1}^{d} h_l)^{-1} \sum_{i=1}^{n} K\left(\frac{x_1 - X_1^{(i)}}{h_1}, ..., \frac{x_d - X_d^{(i)}}{h_d}\right),\ x = (x_1, ..., x_d)^T \in \Re^d.$$

• 3^{eme} **cas** :
Il y a des situations où on est amené à lisser dans des directions différentes à celles définies par les axes des coordonnées. Dans ce cas, on prend $H \in A$ tel que : $A = F - D$.

Remarque 2.4.1.

- *Il est clair que : $S \subset D \subset F$ et $A \subset F$.*

- *Pour illustrer ces idées, situons nous dans le contexte bivarié ($d = 2$) :*

  La figure 2.1 représente les contours-plots pour le noyau gaussien bivarié avec différentes paramètrisations, on a respectivement de gauche à droite :
  – (a) : $H \in S$
  – (b) : $H \in D - S$
  – (c) : $H \in A$
  Nous remarquons, tout d'abord, que dans chaque cas les contours du noyau sont elliptiques. Cependant, lorsque $H \in S$, ces ellipses deviennent des cercles et pour $H \in D - S$, ces ellipses sont telles que les directions de leurs axes correspondent à celles prises par les axes des coordonnées. Par contre, dans le cas où $H \in A$, des ellipses à orientations arbitraires ont été obtenues.
  Il y'a un grand intérêt pratique à comparer la performance de l'estimateur suivant les différents types de paramétrisation. Évidemment, il y'aurait une perte d'efficacité

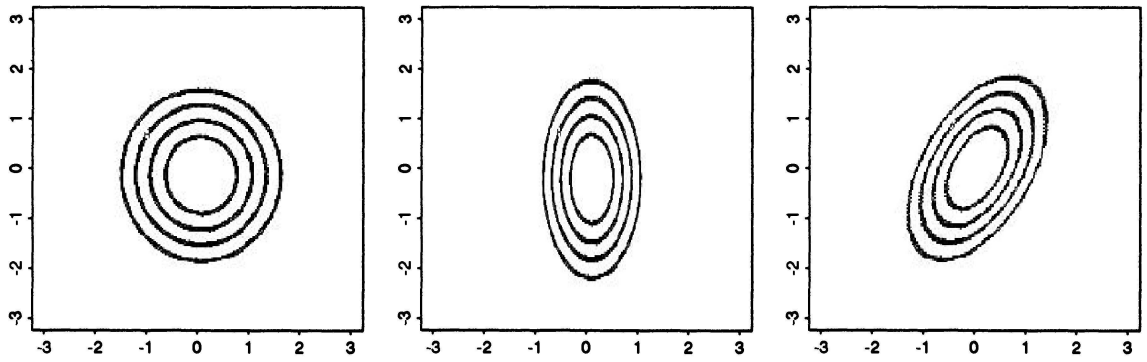

FIGURE 2.1 – Les contours-plots du noyau gaussien

lorsqu'on impose des restrictions sur la forme de la matrice H. Cette perte d'efficacité se traduit par une erreur d'estimation qui dépend de la forme particulière de la densité et de son orientation par rapport aux axes des coordonnées. Pour bien comprendre ceci, considérons deux densités f_1 et f_2 dont les contours-plots sont données par la figure (2.2) suivante :

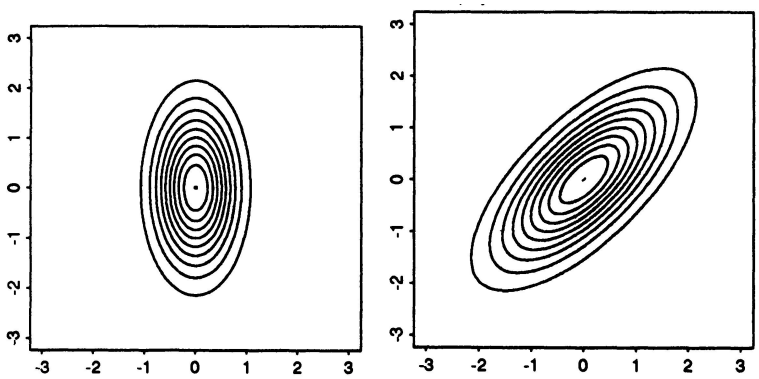

FIGURE 2.2 – Contours-plots des densités f_1 et f_2.

La première f_1 est la densité de la loi normale bivariée non corrélée $N(0, 0, \frac{1}{4}, 1, 0)$ et la seconde f_2 est celle de la loi normale bivariée corrélée $N(0, 0, 1, 1, \frac{9}{10})$.
En utilisant le noyau Gaussien bivarié, il est clair que la meilleure paramétrisation pour l'estimateur dans le premier cas est $H \in D - S$ (car les contours sont elliptiques et ils sont orientés suivant les axes des coordonnées). Pour le deuxième cas, la paramétrisation appropriée est $H \in A$ (car l'orientation des ellipses est arbitraire).

– *Une approche plus pratique pour construire la matrice de lissage H a été proposée par* ***Wand*** *et* ***Jones*** *([99],1995). Cet idée due à* ***Fukunaga****([40],1972) (voir aussi* ***Silverman****([85],1986)) consiste au début à réaliser une transformation linéaire de données afin d'avoir une matrice variance-covariance de l'échantillon unitaire (En*

anglais, on dit "sphering the data"). Puis, on applique l'estimateur de ***Cacoullos*** *([16],1966) aux données ainsi obtenues ("sphered data") et en faisant une transformation inverse ("backtranforming"), on obtient l'estimateur de la densité pour l'échantillon original ("the original data"). La matrice de lissage H s'écrit alors, sous la forme :*

$$H = h^2 S,$$

où S est la matrice variance-covariance de l'échantillon.
Wand *et* ***Jones*** *([97],1993) ont montré que dans le cas bivarié, cette méthode est la plus appropriée si la densité f à estimer est celle d'une loi normale $N(M, \Sigma)$. Mais en général, elle est mal indiquée et elle peut même être nuisible pour les densités non-normales. Ceci est du au fait que les éléments de la matrice variance-covariance ne peuvent pas, en général, exprimer les courbures de la densité f et son orientation par rapport aux axes des coordonnées.*

– *Deux autres méthodes ont été considérées par* ***Wand*** *et* ***Jones*** *([97],1993) dans le cas bivarié. La première consiste à prendre :*

$$H = h^2 S',$$

avec $S' = diag\ S$.
La deuxième méthode consiste à utiliser le coefficient de corrélation pour déterminer la rotation par rapport aux axes des coordonnées. La matrice de lissage s'écrit alors :

$$H = \begin{pmatrix} h_1^2 & \rho_{12} h_1 h_2 \\ \rho_{12} h_1 h_2 & h_1^2 \end{pmatrix},$$

où : $h_1 > 0$, $h_2 > 0$, et $\rho_{12} = \frac{S_{12}}{(S_{11} S_{21})^{\frac{1}{2}}}$ est le coefficient de corrélation.
Ces trois méthodes de paramétrisation, dites hybrides, ne sont pas en général appropriées.

2.5 Qualité et performance de l'estimateur

La qualité de l'estimateur à noyau, défini par la formule (2.2), dépend de sa proximité de la densité cible. On mesure cette proximité par le MISE (l'erreur quadratique moyenne intégrée) ou le AMISE (le MISE asymptotique).
Commençons tout d'abord, par l'introduction du théorème suivant qui est une version multivariée du théorème de Taylor :

Théorème 2.5.1. *Soit g une fonction d-dimensionnelle et $\{\alpha_n = (\alpha_1^n, ..., \alpha_d^n)^T, n \in N\}$ une suite de vecteurs de dimension d où chaque composante α_i^n $(i = \overline{1, d})$ tend vers 0 lorsque n tend vers l'infini.*
Soit $D_g(x)$ le vecteur des dérivées partielles d'ordre 1 de g et $\chi_g(x)$ la matrice Hessienne de g, c'est à dire la matice carrée d'ordre d où l'élèment (i, j) est égal à :

$$\frac{\partial^2}{\partial x_i \partial x_j} g(x).$$

Alors, si chaque élément de $\chi_g(x)$ est continu dans un voisinage de x, on a :

$$g(x+\alpha_n) = g(x) + \alpha_n^T D_g(x) + \frac{1}{2}\alpha_n^T \chi_g(x)\alpha_n + o(\alpha_n^T \alpha_n).$$

C'est le développement de Taylor à l'ordre 2 pour la fonction g au point x.

• Pour ce qui va suivre, nous aurons besoin de faire les hypothèses supplémentaires suivantes sur f, H et K :

1. Chaque élément de la matrice Hessienne de f, qu'on notera $\chi_f(x)$, est borné, continu et de carré-intégrable pour tout $x \in \Re^d$.
2. $H = H(n)$ est une suite de matrices de lissage où chaque élément tend vers zéro quand $n \longrightarrow \infty$ et $\lim_{n\longrightarrow\infty} n^{-1}|H|^{-\frac{1}{2}} = 0$.
3. K est le noyau d-varié, vérifiant :

$$\int K(z)dz = 1, \quad \int zK(z)dz = 0, \quad \int zz^T K(z)dz = \mu_2(K) \times I,$$

où $\mu_2(K) = \int z_i^2 K(z)dz$ est fini, indépendant de i.

Notons que la condition (3) est satisfaite par tous les noyaux sphériques symétriques et les noyaux produits construits à partir d'un noyau symétrique univarié à variance finie.

Nous aurons besoin aussi des définitions et résultas suivants :

Définition 2.5.1. *soit A une matrice carrée. La trace de A, notée $tr(A)$, est la somme des éléments diagonaux de A et on a*

$$tr(AB) = tr(BA), \tag{2.10}$$

pour toute matrice carrèe B de même ordre que A.

Définition 2.5.2. *(**Henderson** et **Searle** ([53],1979)*
Soit A une matrice carrée d'ordre d . Le vecteur de A (the vector of A), noté par vecA, est le $(d^2 \times 1)$ vecteur obtenu en mettant les colonnes de A l'une sous l'autre dans l'ordre, de gauche à droite. Le demi-vecteur de A (the vector-half of A), noté par vechA, est le $(\frac{1}{2}d(d+1)\times 1)$ vecteur obtenu de vecA par l'élimination des éléments de A situés au-dessus de la diagonale.

Exemple 2.5.1. *Si* $A = \begin{pmatrix} 1 & 4 \\ 7 & 3 \end{pmatrix}$, *alors* $vecA = \begin{pmatrix} 1 \\ 7 \\ 4 \\ 3 \end{pmatrix}$ *et* $vechA = \begin{pmatrix} 1 \\ 7 \\ 3 \end{pmatrix}$.

Il est clair que si A est symétrique, alors $vechA$ est composé des éléments distincts de A, est donc $vecA$ est composé des éléments de $vechA$ avec quelques répétitions.

Lemme 2.5.1. *(**Magnus** et **Neudecker**([62], 1988), p.49)*
Il existe une unique $(d^2 \times \frac{1}{2}d(d+1))$ *matrice* D_d *composée de zéros et de* **1** *telle que, pour toute* $(d \times d)$ *matrice symétrique* A*, on a :*

$$D_d vechA = vecA. \tag{2.11}$$

D_d est appelée matrice de duplication d'ordre d (the duplication matrix of order d).

Exemple 2.5.2. *La matrice de duplication d'ordre 2, est donnée par :* $D_2 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

- Un résultat utile et valable pour toutes les matrices carrées A d'ordre d, est donné par la formule :

$$D_d^T vecA = vech(A + A^T - dgA), \tag{2.12}$$

où dgA est la même que la matrice A, mais avec tous les éléments non-diagonaux égaux à zéro.

- Un autre résultat important est :

$$tr(A^T B) = (vec^T A)(vecB). \tag{2.13}$$

- Enfin, pour des changements linéaires de variables quand on intègre sur $\Re^d$, on a le résultat :

$$\int g(Ax)dx = |A| \int g(y)dy, \tag{2.14}$$

où A représente une matrice carrée d'ordre d inversible.

2.5.1 Espérance, biais et variance de l'estimateur

Espérance mathématique de l'estimateur

- L'espérance mathématique de $\hat{f}(x;H)$ est :

$$\begin{aligned} E\hat{f}(x;H) &= \int K_H(x-y)f(y)dy \\ &= \int |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}}(x-y))f(y)dy. \end{aligned}$$

En posant : $z = H^{-\frac{1}{2}}(x-y)$, c'est à dire : $y = x - H^{\frac{1}{2}}z$, et et en utilisant la formule des changements linéaires de variables (2.14), on obtient :

$$E\hat{f}(x;H) = \int K(z)f(x - H^{\frac{1}{2}}z)dz.$$

Le développement de **Taylor** à l'ordre 2 pour la fonction f (théorème 2.5.1) nous donne :

$$\begin{aligned} E\hat{f}(x;H) &= \int K(z)\{f(x) - (H^{\frac{1}{2}}z)^T D_f(x) + \frac{1}{2}(H^{\frac{1}{2}}z)^T \chi_f(x)(H^{\frac{1}{2}}z)\}dz + o\{tr(H)\} \\ &= f(x) - \int z^T H^{\frac{1}{2}} D_f(x) K(z)dz + \frac{1}{2}\int z^T H^{\frac{1}{2}} \chi_f(x) H^{\frac{1}{2}} z K(z)dz + o\{tr(H)\} \\ &= f(x) + \frac{1}{2} tr\{H^{\frac{1}{2}} \chi_f(x) H^{\frac{1}{2}} \int zz^T K(z)dz\} + o\{tr(H)\}. \end{aligned}$$

Donc

$$E\hat{f}(x;H) = f(x) + \frac{1}{2}\mu_2(K) tr\{H\chi_f(x)\} + o\{tr(H)\}.$$

Biais de l'estimateur

Le biais de l'estimateur est :

$$\begin{aligned} biais\{\hat{f}(x;H)\} &= E\hat{f}(x;H) - f(x) \\ &= \frac{1}{2}\mu_2(K) tr\{H\chi_f(x)\} + o\{tr(H)\}. \end{aligned}$$

Variance de l'estimateur

La variance de $\hat{f}(x;H)$ est donnée par :

$$\begin{aligned} Var\hat{f}(x;H) &= n^{-1} Var(K_H(x-y)) \\ &= n^{-1}[E\{K_H(x-y)\}^2 - \{EK_H(x-y)\}^2] \\ &= n^{-1}[\int K_H(x-y)^2 f(y)dy - \{\int K_H(x-y)f(y)dy\}^2] \\ &= n^{-1}[|H|^{-1}\int K(H^{-\frac{1}{2}}(x-y))^2 f(y)dy - \{\int |H|^{-\frac{1}{2}} K(H^{-\frac{1}{2}}(x-y))f(y)dy\}^2]. \end{aligned}$$

Posons : $z = H^{-\frac{1}{2}}(x-y)$, c'est à dire : $y = x - H^{\frac{1}{2}}z$ et d'après le formule (2.14) sur les changements linéaires de variables, on aura :

$$Var\hat{f}(x;H) = n^{-1}[|H|^{-\frac{1}{2}}\int K(z)^2 f(x - H^{\frac{1}{2}}z)dz - \{\int K(z)f(x - H^{\frac{1}{2}}z)dz\}^2].$$

Le développement de **Taylor** à l'ordre 1 pour la fonction f (théorème 2.5.1), nous donne :

$$Var\hat{f}(x;H) = n^{-1}|H|^{-\frac{1}{2}} R(K) f(x) + o(n^{-1}|H|^{-\frac{1}{2}}).$$

2.5.2 Erreur quadratique moyenne intégrée asymptotique

Comme dans le cas univarié, nous pouvons également obtenir une approximation asymptotique simple de l'erreur quadratique moyenne intégrée $MISE$ pour l'estimateur à noyau de la densité multivariée.
Ainsi, D'après les hypothèses (1), (2) et (3) sur f, H et K, on aura :

$$AMISE\{\hat{f}(.;H)\} = n^{-1}\,|H|^{-\frac{1}{2}}\,R(K) + \frac{1}{4}\mu_2(K)^2 \int tr^2\{H\chi_f(x)\}dx, \tag{2.15}$$

Le deuxième terme peut être développé en utilisant les relation (2.10), (2.11) et (2.13) comme suite :

$$\begin{aligned}\int tr^2(H\chi_f(x))dx &= \int (vec^T H)\{vec\chi_f(x)\}\{vec^T\chi_f(x)\}(vecH)dx \\ &= \int (vech^T H)D_d^T\{vec\chi_f(x)\}\{(vec^T\chi_f(x)\}D_d(vechH)dx \\ &= (vech^T H)\Psi_4(vechH)\end{aligned}$$

où, d'après (2.12), Ψ_4 est la matrice carrée d'ordre $\{\frac{1}{2}d(d+1)\}$ définie par :

$$\Psi_4 = \int vech\{2\chi_f(x) - dg\chi_f(x)\} \times vech^T\{2\chi_f(x) - dg\chi_f(x)\}dx.$$

Ainsi, on obtient :

$$AMISE\{\hat{f}(.;H)\} = n^{-1}\,|H|^{-1}\,R(K) + \frac{1}{4}\mu_2(K)^2(vech^T H)\Psi_4(vechH).$$

A première vue la matrice Ψ_4 semble vraiment compliquée. Cependant, une formule simple pour les éléments de Ψ_4 peut être obtenue en faisant une intégration par partie.
Pour une fonction d-dimensionnelle g et un vecteur $R = (r_1, ..., r_d)$ d'entiers non négatifs, nous utiliserons la notation :

$$g^{(R)}(x) = \frac{\partial^{|R|}}{\partial x_1^{r_1}, .., \partial x_d^{r_d}} g(x),$$

en supposant, bien sûr, que cette dérivée existe. On note par $|R|$ la somme des éléments de R, c'est-à-dire : $|R| = \sum_{i=1}^{d} r_i$.
Nous pouvons montrer que :

$$\int f^{(R)}(x)f^{(R')}(x)dx = (-1)^{|R|}\int f^{(R+R')}(x)f(x)dx, \tag{2.16}$$

Si $|(R + R')|$ est pair et 0 sinon.
Il découle de ceci que chaque élément de Ψ_4 peut être écrit sous la forme :

$$\psi_R = \int f^{(R)}(x)f(x)dx,$$

où $|R|$ est pair.

Exemple 2.5.3. *Considérons le cas où $d = 2$. D'après la formule (2.15), On a :*

$$\Psi_4 = \begin{pmatrix} \psi_{4,0} & 2\psi_{3,1} & \psi_{2,2} \\ 2\psi_{3,1} & 4\psi_{2,2} & 2\psi_{1,3} \\ \psi_{2,2} & 2\psi_{1,3} & \psi_{0,4} \end{pmatrix}$$

Remarque 2.5.1. *Contrairement au cas univarié, les expressions explicites de la matrice de lissage H optimale qui minimise le $AMISE$, ne sont pas disponibles en général et cette quantité peut seulement être obtenue numériquement (voir Wand ([96] ,1992a)).*
Des formules très simples pour le $AMISE$ sont possibles dans le cas où $H \in S$ et $H \in D$. Ainsi, nous pouvons montrer que si :

$$H = diag(h_1^2, ..., h_d^2),$$

alors

$$AMISE\left\{\hat{f}(.;H)\right\} = n^{-1}R(K)(\prod_{j=1}^{d} h_j)^{-1} + \frac{1}{4}\mu_2(K)^2(h_1^2, ..., h_d^2)^T\Psi_D(h_1^2, ..., h_d^2),$$

où : Ψ_D est la matrice carrée d'ordre d ayant l'élément (i, j) égale à $\psi_{2e_i+2e_j}$, avec e_i est le vecteur d-dimensionnel ayant 1 pour la i^{eme} composante et 0 sinon.

Et si : $H = h^2I$, on obtient :

$$AMISE\left\{\hat{f}(.;H)\right\} = n^{-1}h^{-d}R(K) + \frac{1}{4}h^4\mu_2(K)^2\int\{\nabla^2 f(x)\}^2dx,$$

où $\nabla^2 f(x) = \sum_{i=1}^{d}(\frac{\partial^2}{\partial x_i^2}f(x))$.
Pour ce cas, la matrice de lissage optimale qui minimise le $AMISE$ à une formule explicite donnée par :

$$h_{AMISE} = \left[\frac{dR(K)}{n\mu_2(K)^2\int\{\nabla^2 f(x)\}^2dx}\right]^{\frac{1}{(d+4)}}.$$

Le minimum de $AMISE$ correspondant est alors :

$$\inf_{h>0} AMISE\{\hat{f}(., h)\} = \frac{d+4}{4d}(\mu_2(K)^{2d}\{dR(K)\}^4\left[\int\{\nabla^2 f(x)\}^2\}dx]^d n^{-4}\right])^{\frac{1}{(d+4)}}.$$

Notons que d'après cette dernière formule, la vitesse de convergence de $\inf_{h>0} AMISE\{\hat{f}(., h)\}$ est de l'ordre de $n^{-\frac{4}{d+4}}$, un taux qui devient plus lent à mesure que la dimension de l'espace augmente. Cette lenteur dans la convergence, due essentiellement au problème de la dimension de l'espace, peut sévèrement compromettre l'implémentation pratique des estimateurs à noyau de la densité et rendre ainsi leur utilisation inappropriée dans des dimensions plus élevées. Cependant, cette méthode reste un outil très pratique pour l'analyse de données dans des dimensions modérées de l'espace (voir : **Scott** et **Wand** ([81],1991) et **Scott** ([79],1992).

Pour qu'une meilleure compréhension de l'exécution pratique de la méthode d'estimation à noyau soit obtenue, nous allons évaluer le $AMISE$ pour des formes particulières de densités multivariées. Il est utile de pouvoir faire ceci pour une large classe de densités. La comparaison entre les différentes formes de paramétrisation qui sera présentée ultérieurement nécessite de tels calculs. Le problème principal dans le calcul de $AMISE$, c'est qu'il implique l'évaluation d'intégrales multiples qui sont, pour beaucoup de densités, difficiles à obtenir. Ainsi, dans la pratique, on fait appel à une large classe de densités pour lesquelles les intégrales ψ_R ont une forme simple à calculer. La famille de densités mélange de lois normales multivariées satisfait cette condition.

Rappelons que la densité de la loi normale d-variée $N(0,\Sigma)$ est donnée par :

$$\Phi_\Sigma(x) = (2\pi)^{-\frac{d}{2}} \, |\Sigma|^{-\frac{1}{2}} \, exp(-\frac{1}{2} x^t \Sigma^{-1} x).$$

Définition 2.5.3. *Une variable aléatoire d-dimensionnel absolument continu X suit une loi mélange gaussiene multivariée, si sa fonction densité s'écrit sous la forme :*

$$f(x) = \sum_{l=1}^{k} \; w_l \; \phi_{\Sigma_l}(x - \mu_l),$$

où $w = (w_1, ..., w_k)^T$ avec $w_l \geq 0$ $(l = \overline{1,k})$ et $\sum_{l=1}^{k} w_l = 1$, μ_l $(l = \overline{1,k})$ est un vecteur de dimension d et Σ_l $(l = \overline{1,k})$ est la matrice variance-covariance, c'est à dire une matrice carrée d'ordre d, symétrique et définie positive.

Théorème 2.5.2. *(**Wand** et **Jones** ([99],1995))*
Pour deux distributions normales multivariées $N(\mu,\Sigma)$ et $N(\mu',\Sigma')$, on a :

$$\int \Phi_\Sigma(x-\mu)\Phi_{\Sigma'}(x-\mu')dx = \Phi_{\Sigma+\Sigma'}(\mu-\mu').$$

Démonstration

On a :

$$\Phi_\Sigma(x-\mu)\Phi_{\Sigma'}(x-\mu') = \Phi_{\Sigma+\Sigma'}(\mu-\mu')\Phi_{\Sigma(\Sigma+\Sigma')^{-1}\Sigma'}(x-\mu^*),$$

où

$$\mu^* = \Sigma'(\Sigma+\Sigma')^{-1}\mu + \Sigma(\Sigma+\Sigma')\mu'.$$

Ce qui donne le résultat.

Théorème 2.5.3. *(**Wand** et **Jones** ([99],1995))*
Si K est le noyau gaussien d-varié Φ_I et f est la densité mélange gaussienne d-variée , alors :

$$MISE(H) = n^{-1}(4\pi)^{-\frac{d}{2}} \, |H|^{-\frac{1}{2}} + w^T\{(1-n^{-1})\Omega_2 - 2\Omega_1 + \Omega_0\}w,$$

où Ω_a est la matrice carrée d'ordre k ayant l'élément (l,l') égal à :

$$\Phi_{aH+\Sigma_l+\Sigma_{l'}} = (\mu_l - \mu_{l'}).$$

Démonstration

On a :

$$MISE\{\hat{f}(.;H)\} = \int Var\{\hat{f}(x;H)\}dx + \int \{E\hat{f}(x;H) - f(x)\}^2 dx$$

où

$$Var\{\hat{f}(x;H)\} = n^{-1}[E\Phi_H(x-X)^2 - \{E\Phi_H(x-X)\}^2].$$

D'après le Théorème (2.5.2) et la symétrie de Φ_H, on obtient :

$$\begin{aligned}\int E\Phi_H(x-X)^2 dx &= \int\int \Phi_H(x-y)^2 dx f(y)dy \\ &= \int \Phi_H(z)^2 dz = \Phi_{2H}(0) = (2\pi)^{-\frac{d}{2}} |2H|^{-\frac{1}{2}}.\end{aligned}$$

De plus,

$$\begin{aligned}E\Phi_H(x-X) &= \sum_{l=1}^{k} w_l \int \Phi_H(y-x)\Phi_{\Sigma_l}(y-\mu_l)dy \\ &= \sum_{l=1}^{k} w_l \Phi_{H+\Sigma_l}(y-\mu_l).\end{aligned}$$

Par conséquent,

$$\begin{aligned}\int \{E\Phi_H(x-X)^2 dx\} &= \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \int \Phi_{H+\Sigma_l}(x-\mu_l)\Phi_{H+\Sigma_{l'}}(x-\mu_{l'})dx \\ &= \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \Phi_{2H+\Sigma_l+\Sigma_{l'}}(\mu_l - \mu_{l'}).\end{aligned}$$

En suivant le même raisonnement pour le deuxième terme et en introduisant la notation Ω_a, On obtient le résultat.

Théorème 2.5.4. *(**Wand** et **Jones**([97],1993))*
Pour chaque deux distributions normales multivariées $N(\mu,\Sigma)$ et $N(\mu',\Sigma')$, on a :

$$(-1)^{\sum_{i=1}^{d} r_i} \int \Phi_{\Sigma}^{(R)}(x-\mu)\Phi_{\Sigma'}^{(R')}(x-\mu')dx = \Phi_{\Sigma+\Sigma'}^{(R+R')}(\mu-\mu'),$$

où $R = (r_1,...,r_d)$ et $R' = (r'_1,...,r'_d)$ sont deux vecteurs dont les composantes sont des entiers non-négatifs.

Démonstration

Posons, pour deux fonctions réelles d-variées f et g :

$$(f * g)(x) = (2\pi)^{-\frac{d}{2}} \int f(\mu)g(x-\mu)d\mu$$

et

$$FT_f(t) = (2\pi)^{-\frac{d}{2}} \int f(x) e^{-it^T x} dx.$$

Ainsi, en utilisant les résultats élémentaires sur la transformée de Fourier pour les fonctions multivariables (voir **Rudin**([26], 1973)) et en notant $c^R = (c_1^{r_1}, ..., c_d^{r_d})$ pour un vecteur complexe d-dimensionnel c, on obtient :

$$\begin{aligned} FT_{\Phi_\Sigma^{(R)}(.-\mu)*\Phi_{\Sigma'}^{(R')}(.-\mu')}(t) &= FT_{\Phi_\Sigma^{(R)}(.-\mu)}(t) FT_{\Phi_{\Sigma'}^{(R')}(.-\mu')}(t) \\ &= (it)^{R+R'} e^{-it^T(\mu+\mu')} \Phi_{\Sigma^{-1}}(t) \Phi_{\Sigma'^{-1}}(t) \left|\Sigma\right|^{-\frac{1}{2}} \left|\Sigma'\right|^{-\frac{1}{2}} \\ &= (it)^{R+R'} e^{-it^T(\mu+\mu')} (2\pi)^{-\frac{d}{2}} \Phi_{(\Sigma+\Sigma')^{-1}}(t) \left|\Sigma+\Sigma'\right|^{-\frac{1}{2}} \\ &= (2\pi)^{-\frac{d}{2}} FT_{\Phi_{\Sigma+\Sigma'}^{(R+R')}}(.-\mu-\mu')(t). \end{aligned}$$

D'où, d'après le théorème d'inversion de Fourier :

$$\Phi_\Sigma^{(R)}(.-\mu) * \Phi_{\Sigma'}^{(R')}(.-\mu')(x) = \Phi_{(\Sigma+\Sigma')}^{(R+R')}(x-\mu-\mu').$$

En remplaçant μ par $-\mu$ et en posant $x = 0$, on obtient le résultat.

Théorème 2.5.5. *(**Wand** et **Jones**([97],1993))*
Si K est le noyau Gaussien bivarié Φ_I et f est la densité mélange Gaussienne bivariée, alors :

$$\begin{aligned} AMISE(H) &= (4\pi n)^{-1} \left|H\right|^{-\frac{1}{2}} + \frac{1}{4} w^T \{\Lambda_{(4,0)} h_1^4 + 4\Lambda_{(3,1)} h_1^2 h_{12} + 2\Lambda_{(2,2)} (h_1^2 h_2^2 + 2h_{12}^2) \\ &\quad + 4\ \Lambda_{(1,3)} h_2^2 h_{12} + \Lambda_{(0,4)} h_2^4\} w, \end{aligned}$$

avec $H = \begin{pmatrix} h_1^2 & h_{12} \\ h_{12} & h_2^2 \end{pmatrix}$ *;* $h_1, h_2 > 0$ *;* $|h_{12}| < h_1 h_2$ *et* Λ_R *est la matrice carrée d'ordre* k *ayant l'élément* (l, l') *égal à :* $\Phi_{(\Sigma_l + \Sigma_l')}^{(R)}(\mu_l - \mu_l')$.

Démonstration
C'est le résultat immédiat du Théorème (2.5.4) et de l'introduction de la notation Λ_R.

Théorème 2.5.6. ***(Wand ([96],1992a)***
Pour chaque deux distributions normales multivarirées $N(\mu, \Sigma)$ et $N(\mu', \Sigma')$, on a :

$$\chi_{\Phi_\Sigma(.-\mu)(x)} = \Phi_\Sigma(x-\mu)(\Sigma^{-1}(x-\mu)(x-\mu)^T - I_d)\Sigma^{-1}.$$

Démonstration
Elle découle directement de la définition de la matrice Hessienne.

Lemme 2.5.2. *(**Seber**([82],1977)*

$$cov(X^T AX, (X-c)^T B(X-c)) = 2tr[A\Sigma B\{\Sigma + 2(\mu-c)\mu^T\}],$$

où

- *X est la variable aléatoire d'une loi normale multivariée $N(\mu, \Sigma)$.*
- *A et B deux matrices carrées d'ordre d symétriques et constantes.*
- *c est un vecteur constant d-dimensionnel.*

Théorème 2.5.7. ***(Wand**([96],1992a)*
Si f est la densité mélange gaussienne d-variée, alors :

$$AMISE\{\hat{f}(.;H)\} = n^{-1}R(K)\,|H|^{-\frac{1}{2}} + \frac{1}{4}\mu_2(K)^2 w^T \Theta w,$$

avec Θ est la matrice carrée d'ordre k ayant l'élément (l, l') égal à :

$$\Phi_{(\Sigma_l+\Sigma_{l'})}(\mu_l - \mu_{l'})\{2tr(HA_{ll'}HB_{ll'}) + tr^2(HC_{ll'})\};$$

et où
$A_{ll'} = (\Sigma_l + \Sigma_{l'})^{-1}$;
$B_{ll'} = A_{ll'}\{I - 2(\mu_l - \mu_{l'})(\mu_l - \mu_{l'})^T A_{ll'}\}$;
$C_{ll'} = A_{ll'}\{I - (\mu_l - \mu_{l'})(\mu_l - \mu_{l'})^T A_{ll'}\}$.

Démonstration
D'après la formule (2.15) de $AMISE$ et en utilisant le Théorème (2.5.7), on obtient

$$\begin{aligned}\int tr^2(H\chi_f(x)) &= \sum_{l=1}^{k}\sum_{l'=1}^{k} w_l w_{l'} \Phi_{\Sigma_l+\Sigma_{l'}}(\mu_l - \mu_{l'}) \times E(tr[H\Sigma_l^{-1}\{(y-\mu_l)(y-\mu_l)^T\Sigma_l^{-1} - I_d\}] \\ &\quad \times tr[H\Sigma_{l'}^{-1}\{(y-\mu_{l'})(y-\mu_{l'})^T\Sigma_{l'}^{-1} - I_d\}] \qquad (2.17)\end{aligned}$$

où :
Y est une variable aléatoire de loi

$$\mathcal{N}(\mu^*_{ll'}, \Sigma_l(\Sigma_l + \Sigma_{l'})^{-1}\Sigma_{l'})$$

et

$$\mu^*_{ll'} = \Sigma_{l'}(\Sigma_l + \Sigma_{l'})^{-1}\mu_l + \Sigma_l(\Sigma_l + \Sigma_{l'})^{-1}\mu_{l'}.$$

Comme $E(UV) = cov(U, V) + E(U)E(V)$ pour deux variables aléatoires U et V, alors l'expression ci-dessus peut s'écrire :

$$\begin{aligned}cov\{(y-\mu_l)^T\Sigma_l^{-1}H\Sigma_l^{-1}(y-\mu_l), (y-\mu_{l'})^T\Sigma_{l'}^{-1}H\Sigma_{l'}^{-1}(y-\mu_{l'})\} \\ +tr(H\Sigma_l^{-1}[E\{(y-\mu_l)(y-\mu_l)^T\}\Sigma_l^{-1} - I_d]) \\ \times tr(H\Sigma_{l'}^{-1}[E\{(y-\mu_{l'})(y-\mu_{l'})^T\}\Sigma_{l'}^{-1} - I_d]).\end{aligned}$$

Sachant que : $\mu^*_{ll'} - \mu_l = \Sigma_l(\Sigma_l + \Sigma_{l'})^{-1}(\mu_{l'} - \mu_l)$.
Le lemme (2.5.1) nous permet d'écrire le terme de covariance sous la forme :

$$2tr[H(\Sigma_l + \Sigma_{l'})^{-1}H(\Sigma_l + \Sigma_{l'})^{-1}\{I_d - 2(\mu_l - \mu_{l'})(\mu_l - \mu_{l'})^T(\Sigma_l + \Sigma_{l'})\}].$$

En utilisant le fait que :

$$E\{(y-\mu_l)(y-\mu_l)^T\} = \Sigma_l(\Sigma_l+\Sigma_{l'})^{-1}\Sigma_{l'} + (\mu^*_{ll'}-\mu_l)(\mu^*_{ll'}-\mu_l)^T,$$

nous pouvons montrer que chaque facteur dans le second terme est égal à :

$$-tr[H(\Sigma_l+\Sigma_{l'})^{-1}\{I_d-(\mu_l-\mu_{l'})(\mu_l-\mu_{l'})^T(\Sigma_l+\Sigma_{l'})^{-1}\}].$$

Combinons ces deux derniers résultats, remplaçons les dans (2.17) et en utilisant les définitions de $A_{ll'}, B_{ll'}$ et $C_{ll'}$ on obtient le résultat.

2.6 Noyau Optimal

Comme nous l'avons déjà vu, l'estimateur à noyau multivarié dépend de deux paramètres : le noyau K et la matrice H des paramètres de lissage.
Dans cette section, nous allons étudier l'influence de la forme du noyau sur la qualité de l'estimateur $\hat{f}(.,H)$. Ceci revient à faire une comparaison qui va nous permettre de choisir le noyau optimal pour la construction d'un estimateur efficace d'une densité f donnée.
En pratique, on choisit le noyau optimal K_0 qui minimise le $AMISE\{\hat{f}(.,H)\}$. Cette tâche n'est pas facile car la formule de $AMISE$ dépend non seulement de K, mais aussi de la matrice H.
L'idée développée par **Wand** et **Jones** ([99],1995) consiste à réécrire la formule de $AMISE$ comme le produit de deux facteurs indépendants : l'un s'écrit en fonction de K seulement et l'autre en fonction du H seulement.
Considérons, pour un noyau K donné, la famille N_K tel que :

$$N_K = \{K_\delta; \delta > 0\} \quad \text{et} \quad K_\delta(.) = K(./\delta)/\delta.$$

Rappelons que :

$$AMISE\{\hat{f}(.,H)\} = n^{-1}\,|H|^{-\frac{1}{2}}\,R(K) + \frac{1}{4}\mu_2(K)^2(vech^TH)\Psi_4(vechH).$$

Pour séparer les deux paramètres, on doit choisir δ de telle sorte que :

$$R(K_\delta) = \mu_2(K_\delta)^2.$$

Ceci est satisfait pour δ égale à :

$$\delta_0 = \{R(K)/\mu_2(K)^2\}^{\frac{1}{d+4}}.$$

On obtient :

$$AMISE\{\hat{f}(.,H)\} = C_d(K_{\delta_0})\{n^{-1}\,|H|^{-\frac{1}{2}} + \frac{1}{4}(vech^TH)\Psi_4(vechH)\},$$

où $C_d(K) = \{R(K)^4 \mu_2(K)^{2d}\}^{\frac{1}{d+4}}$.

En effet,

$$C_d(K_{\delta_0}) = R(K_{\delta_0}) = \mu_2(K_{\delta_0})^2.$$

Notons que $C_d(K)$ est invariant dans N_K, c'est à dire que : $C_d(K_{\delta_1}) = C_d(K_{\delta_2})$ pour tout $\delta_1, \delta_2 > 0$.

Définition 2.6.1. *On appelle noyau canonique pour la famille N_K, le noyau $K^c = K_{\delta_0}$ qui permet de séparer K et H dans la formule de $AMISE$.*

Exemple 2.6.1. *Pour d=1, soit $K = \Phi$ le noyau gaussien standard. Alors le noyau canonique pour la classe $\{\Phi_\delta, \delta > 0\}$ est :*

$$\Phi^c(x) = \Phi_{(4\pi)^{-\frac{1}{10}}}(x) \quad et \quad C_1(\Phi) = (4\pi)^{-\frac{2}{5}}.$$

Si $\hat{f}^c(x;h) = n^{-1}\sum_{i=1}^n \Phi_h^c(x - X_i)$ est l'estimateur à noyau de f, alors :

$$AMISE\{\hat{f}^c(.;h)\} = (4\pi)^{-\frac{2}{5}}\{(nh)^{-1} + \frac{1}{4}h^4 R(f'')\}.$$

L'efficacité du noyau sphérique K^s par rapport au noyau produit K^p se mesure par l'expression :

$$E_{s,p} = \{C_d(K^s)/C_d(K^p)\}^{\frac{(d+4)}{4}} = \{R(K^s)\mu_2(K^s)^{\frac{d}{2}}\}/\{R(K^p)\mu_2(K^p)^{\frac{d}{2}}\}.$$

Ainsi, si cette quantité est inférieure à 1 alors le noyau sphérique K^s est plus efficace que le noyau produit K^p et si elle est supérieure à 1 alors c'est le noyau produit K^p qui est efficace par rapport au noyau sphérique K^s.

Le tableau (2.2) donne les valeurs de $E_{s,p}$ pour certains noyaux. Pour le noyau gaussien, la version produit coïncide avec la version sphérique. Dans les autres cas, nous remarquons qu'il y'a une légère supériorité pour la version sphérique et celle-ci augmente lorsque la dimension de l'espace augmente.

Noyau	d=2	d=3	d=4
Uniforme	0.955	0.888	0.811
Epanechnicov	0.982	0.953	0.916
Biweight	0.983	0.953	0.915
Triweight	0.984	0.956	0.919
Gaussien	1	1	1

TABLE 2.2 – $E_{s,p}$ pour différents noyaux et différentes valeurs de d.

Pour minimiser le $AMISE$ par rapport au noyau K, il suffit de minimiser la quantité $C_d(k)$.
Ainsi, pour le noyau produit K^p, la quantité à minimiser est :

$$C_d(K^p) = \{R(w)^4 \mu_2(w)^2\}^{\frac{d}{d+4}}.$$

Le noyau optimal dans ce cas est le noyau produit d'**Epanechnikov** ([33],1969) :

$$K_0 = K_*^p(x) = (\frac{3}{4})^d \prod_{i=1}^{d} (1 - x_i^2) \mathbf{1}_{\{|x_i|<1\}}.$$

Pour le noyau sphérique, le noyau d'**Epanechnikov** sphérique est le noyau optimal :

$$K_0 = K_*^s(x) = \frac{1}{2} v_d^{-1} (d+2)(1 - x^T x) \mathbf{1}_{\{x^T x \leq 1\}},$$

où $v_d = (2\pi)^{\frac{d}{2}} / \{d\Gamma(\frac{d}{2})\}$ est le volume de la sphère unité d-dimensionnelle (pour plus de détails, voir **Müller**([66],1988), pp. 82-83).
L'efficacité relative d'un noyau multivarié K se mesure alors par :

$$E_R(K) = \{C_d(K)/C_d(K_0)\}^{\frac{d+4}{4}}.$$

En pratique, on choisit le noyau K en fonction de la facilité des calculs plutôt que de l'efficacité relative.

2.7 Paramétrisation optimale

En général, on choisit la matrice de lissage H dans l'ensemble F des matrices symétriques et définies positives. Cependant, des restrictions peuvent être imposés à la matrice H en la définissant comme un élément d'une classe particulière $M \subset F$. Ce qui entraîne une perte d'efficacité. On calcule cette perte d'efficacité par le critère asymptotique d'efficacité relative ARE (Asymptotic Relative Efficiency criterion).
Ainsi, pour une classe particulière M de matrices de lissage, le ARE de M comparé à la classe générale F est définie par :

$$ARE_f(F : M) = \{inf_{H \in F} AMISE(\hat{f}(., H)) / inf_{H \in M} AMISE(\hat{f}(., H))\}^{\frac{d+4}{4}}.$$

Pour chaque classe $M \in F$, on a $inf_{H \in M} AMISE(\hat{f}(., H)) = o(n^{\frac{-4}{d+4}})$.

Exemple 2.7.1. *Pour $d = 2$, si f est la densité bivariée $N(\mu, \Sigma)$ de coefficient de corrélation ρ et si K est le noyau Gaussien bivariée, alors :*

$$\inf_{H \in F} AMISE(H) = \{\frac{3}{8\pi}\} |\Sigma|^{-\frac{1}{2}} n^{-\frac{2}{3}}$$

et

$$\inf_{H\in D} AMISE(H) = \{\frac{3}{8\pi}\}\,|\Sigma|^{-\frac{1}{2}}\,[\frac{(2+\rho^2)}{2(1-\rho^2)}]^{\frac{1}{3}} \times n^{-\frac{2}{3}}.$$

$$ARE(F:D) = \{inf_{H\in F} AMISE(H)/inf_{H\in D} AMISE(H)\}^{\frac{3}{2}} = \{2(1-\rho^2)/(2+\rho^2)\}^{\frac{1}{2}}.$$

Remarque 2.7.1. *L'interprétation de la quantité $ARE(F:M)$ est que, pour une taille n de l'échantillon, l'erreur minimum en utilisant n observation avec $H \in M$ est la même en utilisant un nombre d'observations égal à $ARE(F:M) \times n$ avec $H \in F$.*

2.8 Conclusion

Ce deuxième chapitre est consacré à l'introduction de l'estimateur à noyau dans la cas multidimensionnel. Ainsi, nous avons présenté sa forme et étudié ses propriétés statistiques (Espérance, Biais, Variance et Erreur quadratique moyenne intégrée). Nous nous sommes intéressé tout particulièrement à la construction du noyau multidimensionnel à partir d'un noyau unidimensionnel et nous avons abordé les différents types de la paramétrisation. Des résultats élémentaires trouvés dans la littérature sont aussi exposés.

3

Choix de la matrice de lissage

3.1 Introduction

La qualité de l'estimateur à noyau de la densité de probabilité dépend crucialement du choix du paramètre de lissage. Dans le cas univarié, ceci revient à choisir un paramètre scalaire h strictement positif qui contrôle le degré de lissage. Par contre dans le cas multivarié, le paramètre de lissage est une matrice symétrique et définie positive H. Cette matrice contrôle, en même temps, le degré et la direction de lissage (l'orientation par rapport aux axes des coordonnées). Ce qui rend son choix plus difficile.
Jusqu'ici la majeure partie de l'effort de recherche a été déployée pour le choix automatique du paramètre de lissage optimal dans le cas univarié. Ainsi, un grand nombre de travaux existent dans la littérature sur ce sujet. Pour une synthèse générale, on peut consulter **M. C. Jones** et **al.**([58],1996).
Cependant, dans le cas multivarié, on trouve dans la littérature un nombre limité de travaux sur le choix optimal de la matrice de lissage H . **Sain** et **al.** ([76],1994) ont proposé une généralisation des méthodes cross validation et bootstrap pour le choix de la matrice H. Mais ces auteurs ont limité leur attention aux estimateurs à noyau produit. **Wand** et **Jones** ([98],1994) ont développé les méthodes plug-in (ré-injection). Ces deux auteurs ont montré qu'il est, en général, impossible d'avoir une expression explicite pour le choix de la matrice H en utilisant l'algorithme plug-in. Ce qui les a amenés à concentrer leurs travaux sur les matrices diagonales dans le cas bivarié. Dans ce contexte, ils ont réussi à développer des expressions plug-in explicites. **Duong** et **Hazelton** dans ([30],2003), ([31],2005a) et ([32],2005b) ont généralisé les résultats de **Sain** et **al.** ([76],1994) et ceux

obtenus par **Wand** et **Jones** ([98],1994). L'approche bayésienne pour le cas multivarié a été considérée par **Xibin Zhang** et **al.**([100],2006).
La décision d'un choix optimal pour le paramètre de lissage suppose la spécification d'un critère d'erreur qui puisse être optimisé. Bien sûr, l'optimalité n'est pas un concept absolu : elle est intimement liée aux choix du critère, qui peut faire intervenir à la fois la densité inconnue f et l'estimateur $\hat{f}(.,H)$ (donc H et le noyau K).
Dans ce chapitre, nous allons nous intéresser au problème de sélection de la matrice de lissage optimale H. Ainsi, deux classes de méthodes seront introduites. Il s'agit des méthodes plug-in (ré-injection) et des méthodes cross-validation.

3.2 Les méthodes plug-in (ré-injection)

Les méthodes plug-in sont très répandues et très utilisées dans l'estimation de la densité de probabilité univariée, car elles présentent de bonnes propriétés théoriques et pratiques. Ce sont en outre des méthodes qui convergent très rapidement. Dans le cas multivariée, peu d'attention a été donnée à ces méthodes. Nous allons présenter ici l'algorithme de **Wand** et **Jones** ([98],1994) et celui de **Duong** et **Hazelton** ([30],2003).

3.2.1 Matrice de lissage optimale

Afin de mesurer la qualité de l'estimateur $\hat{f}(x,H)$, nous utiliserons l'erreur quadratique moyenne intégré ($MISE$) définie par :

$$\begin{aligned} MISE\hat{f}(.,H) &= E\{ISE\hat{f}(.,H)\} \\ &= E\int[\hat{f}(x,H)-f(x)]^2dx. \end{aligned}$$

Notre but est de choisir la matrice de lissage qui minimise le $MISE$, c'est à dire :

$$H_{MISE} = argmin_{H\in F} MISE\hat{f}(.,H),$$

où F est l'ensemble des matrices carrées d'ordre d symétrique et définies positives. **Wand** et **Jones** ([99],1995) ont montré que sous certaines conditions, on a :

$$MISE\hat{f}(.,H) = AMISE\hat{f}(.,H) + o(n^{-1}\left|H\right|^{-\frac{1}{2}} + tr^2H), \tag{3.1}$$

où

$$AMISE\hat{f}(.,H) = n^{-1}\left|H\right|^{-\frac{1}{2}} R(K) + \frac{1}{4}\mu_2(K)^2(vech^TH)\Psi_4(vechH). \tag{3.2}$$

Ψ_4 est une matrice carrée d'ordre $\frac{1}{2}d(d+1)$ définie par :

$$\Psi_4 = \int vech\{2\chi_f(x) - dg\chi_f(x)\} \times vech^T\{2\chi_f(x) - dg\chi_f(x)\}dx.$$

Ces deux dernières formules donnent une approximation maniable de $MISE$ par le $AMISE$. Par conséquent, il est plus pratique de faire une analyse asymptotique. C'est-à-dire, on cherche à estimer :

$$H_{AMISE} = argmin_{H\in F} AMISE\hat{f}(., H),$$

au lieu de

$$H_{MISE} = argmin_{H\in F} MISE\hat{f}(., H).$$

Rappelons que les éléments de la matrice Ψ_4 s'écrivent sous la forme :

$$\psi_r = \int f^{(r)}(x)f(x)dx,$$

avec

$$f^{(r)}(x) = \frac{\partial^{|r|}}{\partial x_1^{r_1}, ..., \partial x_d^{r_d}} f(x).$$

$r = (r_1, ..., r_d)$ est un vecteur d'ordre d, dont les éléments sont des entiers non-négatifs et $|r| = \sum_{i=1}^{d} r_i$. Ce qui veut dire que le $AMISE$ est une fonctionnelle de la densité inconnue f, à travers les éléments de Ψ_4. Par conséquent, nous aurons besoin d'estimateurs pilotes pour les fonctionnelles ψ_r que nous pouvons ré-injecter afin d'avoir un estimateur $\widehat{AMISE}$ de $AMISE$ que nous pouvons numériquement minimiser pour obtenir la matrice de lissage optimale plug-in H_{opt}. Notons que la démarche est plus faciles si H est diagonale (voir **Wand** et **Jones** ([98],1994)).

3.2.2 Estimation fonctionnelle pilote

Posons :

$$\begin{aligned} \psi_r &= \int_{\Re^d} f^{(r)}(x)f(x)dx \\ &= Ef^{(r)}(X), \end{aligned}$$

où X est une variable aléatoire d-dimensionnelle de densité f. Alors l'estimateur naturel de ψ_r est la moyenne empirique de $\hat{f}^{(r)}(X)$, c'est à dire :

$$\begin{aligned} \widehat{\psi_r}(G) &= n^{-1}\sum_{i=1}^{n} \hat{f}^{(r)}(X^{(i)}, G) \\ &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n} K_G^{(r)}(X^{(i)} - X^{(j)}) \end{aligned} \tag{3.3}$$

où G est la matrice de lissage pilote différente de la matrice de lissage initiale H. Par définition, la matrice pilote G est symétrique et définie positive.

En suivant le même raisonnement développé dans le chapitre 2, On peut montrer que :

$$\begin{aligned} Biais\widehat{\psi_r}(G) &= n^{-1}K_G^{(r)}(0) + \frac{1}{2}\mu_2(K)\int_{\Re^d} tr(G\chi_f(x))f^{(r)}(x)dx + o(n^{-1}\left|G\right|^{-\frac{|r|}{2}} \\ &\quad + \|vechG\|). \end{aligned} \tag{3.4}$$

et

$$\begin{aligned} Var\widehat{\psi_r}(G) &= 2n^{-2}\psi_0\int_{\Re^d} K_G^{(r)}(x)^2dx + 4n^{-1}[\int_{\Re^d} f^{(r)}(x)^2f(x)dx - \psi_r^2] + o(n^{-2}\left|G\right|^{-\frac{1}{2}} \\ &\quad \times \left\|vechG^{-|r|}\right\| + n^{-1}). \end{aligned} \tag{3.5}$$

Ces deux expressions ont été données par **Wand** et **Jones** ([99],1995). De nouveau, on rencontre le problème du choix d'une paramétrisation pour $\widehat{\psi_r}(G)$, déja discuté pour $\hat{f}(.,H)$ dans le deuxième chapitre. Rappelons qu'une paramétrisation de la forme : h^2I, $h > 0$, a été considérée très restrictive pour $\hat{f}(.,H)$. Cependant **Wand** et **Jones** ([98],1994), **Duong** et **Hazelton** ([30],2003) ont adopté cette paramétrisation pour $\widehat{\psi_r}(G)$ et ceci pour avoir des formules maniables (faciles à manipuler). Ainsi, nous posons : $G = g^2I$, $g > 0$. Maintenant, ceci semble contredire les arguments avancés dans le deuxième chapitre sur le choix de la paramétrisation, alors que ce n'est pas le cas. Car, en premier lieu, la matrice de lissage pilote G n'a pas besoin d'être spécifier au même degré que la matrice de lissage initiale H. En second lieu, on peut éviter les effets de cette paramétrisation restrictive en faisant des pré-transformations appropriées pour les données et qu'on va developer plus tard. En troisième lieu, la paramétrisation de G n'affecte pas la vitesse de convergence de $\widehat{\psi_r}(G)$.

AMSE- Matrice pilote de lissage

Soit $G = g^2I$, $g > 0$. Et soit $|r| = j$, alors le biais devient :

$$Biais\widehat{\psi_r}(g) = n^{-1}g^{-d-j}K^{(r)}(0) + \frac{1}{2}g^2\mu_2(K)\sum_{i=1}^{d}\psi_{r+2e_i} + o(n^{-1}g^{-d-j} + g^2). \tag{3.6}$$

et la variance :

$$Var\widehat{\psi_r}(g) = 2n^{-2}g^{-d-2j}\psi_0 R(K^{(r)}) + o(n^{-2}g^{-d-2j}). \tag{3.7}$$

En supposons que $K^{(r)}$ est de carré-integrable, $g = g_n \longrightarrow 0$ quand $n \longrightarrow \infty$ et $n^{-1}g^{-d-2j} \longrightarrow 0$ quand $n \longrightarrow \infty$, on obtient :

$$AMSE\widehat{\psi_r}(g) = 2n^{-2}g^{-d-2j}\psi_0 R(K^{(r)}) + [n^{-1}g^{-d-j}K^{(r)}(0) + \frac{1}{2}g^2\mu_2(K)\sum_{i=1}^{d}\psi_{r+2e_i}]^2. \tag{3.8}$$

Nous allons chercher :

$$g_{r,AMSE} = argmin_{g>0} AMSE\widehat{\psi_r}(g).$$

Notons que la formule de $AMSE$ a été donnée par **Wand** et **Jones** ([98],1994). Pour la plupart des noyaux, y compris le noyau gaussien, si tous les éléments de r sont pairs alors $K^{(r)}(0)$ et ψ_{r+2e_i} sont de signes opposés, pour tous $i = 1, ..., d$. Dans ce cas, $g_{r,AMSE}$ prend la valeur de g qui annule le terme de Biais :

$$g_{r,AMSE} = [\frac{-2K^{(r)}(0)}{\mu_2(K)(\sum_{i=1}^{d}\psi_{r+2e_i})n}]^{\frac{1}{d+j+2}}. \tag{3.9}$$

Si l'un au moins des éléments de r est impair alors $K^{(r)}(0) = 0$ et le minimum de $AMSE$ est atteint pour g égale à :

$$g_{r,AMSE} = [\frac{2\psi_0(2\,|r|+d)R(K^{(r)})}{\mu_2(K)^2(\sum_{i=1}^{d}\psi_{r+2e_i})^2 n^2}]^{\frac{1}{d+2j+4}}. \tag{3.10}$$

Les expressions $g_{r,AMSE}$ dépend de fonctionnelles d'ordre supérieure ψ_{r+2e_i}. Ces fonctionnelles sont des éléments de Ψ_6. Nous pouvons estimer chaque élément de Ψ_6, c'est à dire chaque fonctionnelle ψ_{r+2e_i} avec $|r| = 4$, en utilisant un autre estimateur à noyau, mais on trouvera que sa matrice de lissage optimale dépend des éléments de Ψ_8. On retombe sur le même problème, ce qui nous amène à estimer Ψ_8 par la méthode de noyau, puis $\Psi_{10}, \Psi_{12}, ...$ On est amené ainsi à faire des estimations à noyau sans fin. Il faut donc trouver un autre moyen pour estimer Ψ_6. L'idée est de fixer le nombre maximum m d'estimation à noyau à faire et on estime ainsi successivement $\Psi_6, \Psi_8, ..., \Psi_{4+2m}$. Puis on estime les éléments de Ψ_{4+2m} par une approximation normale de référence, à savoir :

$$\widehat{\psi}_r^{NR} = (-1)^{|r|}\Phi_{2S}^{(r)}(0), \tag{3.11}$$

pour $|r| = 4 + 2m$.

Notons que $\Phi_\Sigma(x)$ est la densité normal multivariée de moyenne nulle et de matrice variance-covarianc Σ, évaluée en x. S est la matrice variance-covariance de l'échantillon. Cette méthode due à **Wand** et **Jones** ([98], 1994) suggère de choisir une matrice de lissage pilote G différente pour chaque élément de Ψ_4. Ceci n'est pas un problème pour la matrice de lissage H diagonale. Cependant, ceci peut causer un sérieux problème si H n'est pas diagonale. En effet, dans ce cas, on peut avoir une matrice $\widehat{\Psi}_4$ qui n'est pas définie positive. Injecter une telle matrice dans la formule de $AMISE\hat{f}(., H)$ produit un estimateur $\widehat{AMISE}\hat{f}(., H)$ qui a un minimum global non-finie (puisque il décroît strictement dans certaines directions). Comme on peut avoir une matrice $\widehat{\Psi}_4$ définie positive mais presque singulière, ce qui entraîne des instabilités numériques lorsqu'on cherche à minimiser le $AMISE$ (voir **Duong** et **Hazelton**, ([30],2003).

Ainsi, utiliser un estimateur différent pour chaque élément de la matrice Ψ_4 ne donne pas nécessairement un estimateur approprié $\widehat{\Psi_4}$ pour la matrice Ψ_4. Ce qui nous amène a chercher un nouveau estimateur pilote qui ne souffre pas de ces inconvénients, c'est à dire nous allons essayer d'estimer la matrice Ψ_4 dans son intégralité plutôt que d'estimer chaque élément tout seul. **Duong** et **Hazelton** ([30],2003) ont montré que l'utilisation d'une seule matrice pilote de lissage G pour toutes les fonctionnelles Ψ_4 nous amène nécessairement à une matrice $\widehat{\Psi_4}$ définie positive.

Lemme 3.2.1. *(**Duong** et **Hazelton**, ([30],2003))*
Si une seule matrice pilote de lissage G et le noyau Gaussien sont utilisés pour estimer tous les éléments de Ψ_4, alors $\widehat{\Psi}_4$ est définie positive.

Démonstration

Si on remplace $\hat{f}(.,\frac{1}{2}G)$ dans la formule de ψ_r, $|r|=4$, on obtient :

$$\begin{aligned}\int_{\Re^d}\hat{f}^{(r)}(x,\frac{1}{2}G)\hat{f}(x,\frac{1}{2}G)dx &= n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\int_{\Re^d}\Phi^{(r)}_{\frac{1}{2}G}(x-X_i)\Phi_{\frac{1}{2}G}(x-X_j)dx\\ &= (-1)^{|r|}n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}\Phi^{(r)}_{G}(X_i-X_j)\\ &= \widehat{\psi}_r(G).\end{aligned}$$

Ceci implique que $\widehat{\Psi}_4$ est obtenue en remplaçant f par $\hat{f}(.,\frac{1}{2}G)$ dans Ψ_4. Par définition la matrice Ψ_4 est définie positive pour toute densité f. Et comme $\hat{f}(.,\frac{1}{2}G)$ est une densité, alors $\widehat{\Psi}_4$ est définie positive.

SAMSE- Matrice pilote de lissage

Pour avoir une matrice $\widehat{\Psi}_4$ définie positive **Duong** et **Hazelton** ([30],2003) ont défini un nouveau critère d'erreur noté $SAMSE$ (somme asymptotique des erreurs quadratiques moyennes).
Pour un ordre de dérivation j, on définit le $SAMSE$ par :

$$SAMSE_j(G)=\sum_{r:|r|=j}AMSE\widehat{\psi}_r(G).$$

Pour les mêmes arguments développés précédemment, nous adopterons pour G la forme : g^2I, $g>0$. **Duong** et **Hazelton** ([30],2003) ont proposé ainsi de minimiser le $SAMSE$ afin de trouver :

$$g_{j,SAMSE}=argmin_{g>0}SAMSE_j(g).$$

On a :

$$\begin{aligned} SAMSE_j(G) &= \sum_{r:|r|=j} AMSE\widehat{\psi}_r(G) \\ &= \sum_{r:|r|=j} 2n^{-2}g^{-2j-d}R(K^{(r)}) + \sum_{r:|r|=j}[n^{-1}g^{-j-d}K^{(r)}(0) + \frac{1}{2}g^2\mu_2(K)\sum_{i=1}^{d}\psi_{r+2e_i}]^2 \\ &= 2n^{-2}g^{-2j-d}A_0 + n^{-2}g^{-2j-2d}A_1 + n^{-1}g^{-j-d+2}A_2 + \frac{1}{4}g^4A_3. \end{aligned}$$

où A_0, A_1, A_2 et A_3 sont indépendants de n et définies par :

$$\begin{aligned} A_0 &= \sum_{r:|r|=j} R(K^{(r)}) \\ A_1 &= \sum_{r:|r|=j} K^{(r)}(0)^2 \\ A_2 &= \mu_2(K)\sum_{r:|r|=j} K^{(r)}(0)(\sum_{i=1}^{d}\psi_{r+2e_i}) \\ A_3 &= \mu_2(K)^2\sum_{r:|r|=j}(\sum_{i=1}^{d}\psi_{r+2i})^2. \end{aligned}$$

On a A_0, A_1 et A_3 sont positifs par construction. A_2 est négatif car $K^{(r)}(0)$ et ψ_{r+2e_i} sont de signes opposés si r est pair et si r est impair, $K^{(r)}(0) = 0$.

L'expression de $SAMSE_j(G)$ peut être simplifier. En effet, le premier terme est un $o(n^{-2}g^{-2j-d})$ et le deuxième est un $o(n^{-2}+g^{-2j-2d})$. Ce qui signifie que le second terme domine toujours le premier. Si on enlève le premier terme (qui est la variance asymptotique), on obtient :

$$SAMSE_j(G) = n^{-2}g^{-2j-2d}A_1 + n^{-1}g^{-j-d+2}A_2 + \frac{1}{4}g^4A_3. \tag{3.12}$$

Ainsi, on a uniquement considéré la contribution du biais au carré. En dérivant par rapport à g, on trouve :

$$\begin{aligned} \frac{\partial}{\partial g}SAMSE_j(G) &= -(2j+2d)n^{-2}g^{-2j-2d-1}A_1 - (j+d-2)n^{-1}g^{-j-d+1}A_2 + g^3A_3 \\ &= g^3[-(2j+2d)n^{-2}g^{-2j-2d-4}A_1 - (j+d-2)n^{-1}g^{-1-d-2}A_2 + A_3]. \end{aligned}$$

Si on annule cette expression

$$-(2j+2d)n^{-2}g^{-2j-2d-4}A_1 - (j+d-2)n^{-1}g^{-1-d-2}A_2 + A_3 = 0,$$

c'est-à-dire :

$$-(2j+2d)[n^{-1}g^{-j-d-2}]^2A_1 - [n^{-1}g^{-1-d-2}](j+d-2)A_2 + A_3 = 0,$$

qui est une équation de deuxième degré en $[n^{-1}g^{-j-d-2}]$. Sa résolution donne :

$$g_{j,SAMSE} = [\frac{(4j+4d)A_2}{((-j-d+2)A_2 + \sqrt{(-j-d+2)^2A_2^2 + (8j+8d)A_1A_3})n}]^{\frac{1}{j+d+2}}. \quad (3.13)$$

On définit ainsi la $SAMSE$-matrice pilote de lissage d'ordre j ($|r| = j$). Cette méthode pour estimer Ψ_4 utilise la même matrice pilote de lissage $G = g_{j,SAMSE}I$ pour tout éléments de Ψ_4. Donc d'après le lemme (3.2.1), $\widehat{\Psi_4}$ est définie positive. L'autre principal avantage de la méthode $SAMSE$ est qu'elle est plus économique que la méthode $AMSE$ lorsqu'on compare le nombre de paramètres (pilotes et initiaux) de lissage à calculer pour chaque méthode. Ainsi, on a :

Avec la méthode AMSE :

- Si H est diagonale ($H \in D$), on calcule : $\sum_{i=1}^{m}\sum_{j=0}^{min(i,d-1)}\binom{i}{j}\binom{d}{j+1}$ paramètres pilotes plus d paramètres initiaux de lissage.
- Si H est complète ($H \in F$), on calcule : $v_m + \sum_{i=1}^{m}\sum_{j=0}^{min(2i,d-1)}\binom{2i+1}{j}\binom{d}{j+1}$ paramètres pilotes plus $\frac{1}{2}d(d+1)$ paramètres initiaux de lissage.
  Avec :
  $v_0 = 0$, $v_1 = 1$, $v_2 = 3$ et pour $m \geq 4$, on a :

$$v_m = \sum_{i=1}^{m-3}\sum_{j=0}^{min(i,d-1)}\binom{i}{j}\binom{d}{j+1}.$$

  Ces formules ont été données par **Wand** et **Jones** ([98], 1994).

Avec la méthode SAMSE :

- Si H est complète ($H \in F$), on calcule seulement m paramètres pilotes plus $\frac{1}{2}d(d+1)$ paramètres initiaux de lissage.
  Pour plus d'illustration, un calcul a été effectué pour $m = 2$ et $d = 1, 2, ..., 6$. Les résultats sont données par le tableau suivant :

		d=1	d=2	d=3	d=4	d=5	d=6
La méthode AMSE	$H \in D$	3	9	19	34	55	83
	$H \in F$	3	16	50	130	296	610
La méthode SAMSE	$H \in F$	3	5	8	12	17	23

TABLE 3.1 – Nombre de paramètres pilotes et initiaux de lissage à calculer.

3.2.3 Transformation initiale de données (Pre-scaling et pre-sphering)

Dans les paragraphes précédents nous avons adopté pour chaque estimateur à noyau des paramètres pilotes de lissage, une paramétrisation G tel que :

$$G = g^2 I.$$

Cette paramétrisation est beaucoup plus restrictive. Afin d'éviter les effets de cette restriction, nous devons transformer les données $X_1, X_2, ..., X_n$ avant chaque estimation pilote de lissage. La plus courante est la transformation pre-scaling, qui consiste à transformer les données pour obtenir une variance égale à 1 dans chaque direction des coordonnées. Soit X^* la version transformée par pre-scaling (scaled version) de X, i.e. $X^* = S_D^{-\frac{1}{2}} X$ avec : $S_D = dgS$ et S est la matrice variance-covariance de l'échantillon. Ceci signifie :

$$X^* = (S_1^{-1} X_1, S_2^{-1} X_2, ..., S_d^{-1} X_d),$$

où S_i^2 $(i = 1, ..., d)$ est la i^{eme} variance marginale de l'échantillon.
Soit S_D^* la variance de l'échantillon pour les données ainsi transformées (scaled data), alors :

$$S_D^* = \widehat{varX^*} = S_D^{-\frac{1}{2}} \widehat{varX} S_D^{-\frac{1}{2}} = S_D^{-\frac{1}{2}} S S_D^{-\frac{1}{2}} = \begin{pmatrix} 1 & \frac{S_{12}}{S_1 S_2} & .... & \frac{S_{1d}}{S_1 S_2} \\ . & 1 & .... & . \\ . & . & .... & . \\ . & . & .... & . \\ . & . & .... & . \\ \frac{S_{1d}}{S_1 S_d} & \frac{S_2 d}{S_2 S_d} & .... & 1 \end{pmatrix}.$$

Une autre transformation qu'on peut appliquer pour les données avant chaque estimation pilote est la transformation sphérique (pre-sphering). Dans ce cas, nous transformons les données pour avoir une matrice variance-covariance égale à la matrice identité. La version sphérique de X est $X^* = S^{-\frac{1}{2}} X$. Et la matrice variance-covariance des données sphériques (pre-sphered data) est :

$$S^* = \widehat{varX^*} = S^{-\frac{1}{2}} (\widehat{varX}) S^{-\frac{1}{2}} = S^{-\frac{1}{2}} S S^{-\frac{1}{2}} = I.$$

Une fois que nous avons pre-transformé les données, on peut trouver la matrice de lissage H^* pour les données transformées. Ainsi, pour obtenir la matrice de lissage H pour les données originaux, on utilise le deux lemmes suivants :

Lemme 3.2.2. *Si H est la matrice de lissage pour les données originaux et H^* est la matrice de lissage pour les données sphériques, alors :*

$$H = S^{\frac{1}{2}} H^* S^{\frac{1}{2}}.$$

Démonstration

$$\begin{aligned}\hat{f}^*(x^*, H^*) &= n^{-1}\sum_{i=1}^{n} K_{H^*}(x^* - X_i^*) \\ &= n^{-1}\left|H^*\right|^{-\frac{1}{2}}\sum_{i=1}^{n} K({H^*}^{-\frac{1}{2}}(x^* - X_i^*)) \\ &= n^{-1}\left|H^*\right|^{-\frac{1}{2}}\sum_{i=1}^{n} K((S^{\frac{1}{2}}{H^*}^{\frac{1}{2}})^{-1}(x - X_i))\end{aligned}$$

Rappel

Si A et B sont deux matrices symétriques et définies positives alors :

$$(B^{\frac{1}{2}}AB^{\frac{1}{2}})^{\frac{1}{2}} = B^{\frac{1}{2}}A^{\frac{1}{2}}. \tag{3.14}$$

En utilisant (3.14), on obtient :

$$\hat{f}^*(x^*, H^*) = n^{-1}\left|S\right|^{\frac{1}{2}}\left|S^{\frac{1}{2}}H^*S^{\frac{1}{2}}\right|^{-\frac{1}{2}}\sum_{i=1}^{n} K((S^{\frac{1}{2}}H^*S^{\frac{1}{2}})^{-\frac{1}{2}}(x - X_i)).$$

C'est-à-dire :

$$\hat{f}^*(x^*, H^*) = \left|S\right|^{\frac{1}{2}}\hat{f}(x, H).$$

Par conséquent :

$$H = S^{\frac{1}{2}}H^*S^{\frac{1}{2}}.$$

Lemme 3.2.3. *Si H est la matrice de lissage pour les données originaux et H^* est la matrice de lissage pour les données pré-scalées (pre-scaled data), alors :*

$$H = S_D^{\frac{1}{2}}H^*S_D^{\frac{1}{2}}.$$

Démonstration

Il suffit de remplacer dans la démonstration de lemme (3.2.2) la matrice S par S_D. On aura ainsi :

$$\hat{f}^*(x^*, H^*) = \left|S_D\right|^{\frac{1}{2}}\hat{f}(x, H).$$

D'où :

$$H = S_D^{\frac{1}{2}}H^*S_D^{\frac{1}{2}}.$$

3.2.4 Taux relatif de convergence des méthodes plug-in

Les méthodes plug-in nous donnent des estimateurs pour les fonctionnelles ψ_r et donc un estimateur $\widehat{\Psi}_4$ de Ψ_4. En remplaçant Ψ_4 par son estimateur $\widehat{\Psi}_4$ dans la formule de $AMISE\hat{f}(.,H)$, on obtient un estimateur de celui-ci, i.e : $\widehat{AMISE}\hat{f}(.,H)$ qui peut être minimisé pour obtenir un estimateur de H_{AMISE}, notée $\widehat{H}$.
La performance de $\widehat{H}$ peut être évaluée par son taux de convergence relatif.

Définition 3.2.1. *On dit qu'un estimateur $\widehat{H}$ converge vers H_{AMISE} avec un taux relatif égal à $n^{-\alpha}$ $(\alpha > 0)$ si :*

$$vech(\widehat{H} - H_{AMISE}) = O_p(J_{d'}n^{-\alpha})vechH_{AMISE},$$

où : $J_{d'}$ est la matrice carrée d'ordre $d' = \frac{1}{2}d(d+1)$ et dont les éléments sont égaux à 1.

Définition 3.2.2. *Soit $\{A_n\}_{n\in N}$ et $\{B_n\}_{n\in N}$ deux suites de matrices telles que, pour tout $n \in N$, A_n et B_n sont de mêmes dimensions. On dit que :*

1. *$A_n = O_p(B_n)$ si $[A_n]_{ij} = O_p([B_n]_{ij})$ pour tout élément $[A_n]_{ij}$ de A_n et $[B_n]_{ij}$ de B_n.*
2. *$A_n = o_p(B_n)$ si $[A_n]_{ij} = o_p([B_n]_{ij})$ pour tout élément $[A_n]_{ij}$ de A_n et $[B_n]_{ij}$ de B_n.*

Définition 3.2.3. *Soit $\{A_n\}_{n\in N}$ et $\{B_n\}_{n\in N}$ deux suites de variables aléatoires réelles.*

1. *On dit que $A_n = o_p(B_n)$ si pour tout $\xi > 0$,*

$$\lim_{n\to\infty} P(|A_n/B_n| > \xi) = 0$$

2. *On dit que $A_n = O_p(B_n)$ si pour tout $\xi > 0$ il existe λ et M tel que :*

$$P(|A_n/B_n| > \lambda) < \xi \text{ pour tout } n > M$$

Lemme 3.2.4. *(**Duong** et **Hazelton**([31],2005a))*
Supposons que :

- (A_1) *tout les éléments de $\chi_f(x)$ sont bornés, continus et carré-intégrables.*
- (A_2) *tout les éléments de H tendent vers zéro et $n^{-1}\left|H\right|^{-\frac{1}{2}} \to 0$ quand $n \to \infty$.*
- (A_3) *K est un noyau sphérique.*

Soit $\widehat{H} = argmin_{H\in F}\widehat{AMISE}(H)$ l'estimateur de H. On définit son erreur quadratique moyenne par :

$$MSE(vech\widehat{H}) = E[vech(\widehat{H} - H_{AMISE})vech^T(\widehat{H} - H_{AMISE})].$$

alors

$$MSE(vech\widehat{H}) = [I_{d'} + o(J_{d'})]AMSE(vech\widehat{H}),$$

où

$$AMSE(vech\widehat{H}) = AVar(vech\widehat{H}) + [ABiais(vech\widehat{H})][ABiais(vech\widehat{H})]^T$$

et

$$ABiais(vech\widehat{H}) = [D_H^2 AMISE(H_{AMISE})]^{-1}E[D_H(\widehat{AMISE} - AMISE)(H_{AMISE})]$$

$$\begin{aligned} AVar(vech\widehat{H}) &= [D_H^2 AMISE(H_{AMISE})]^{-1} Var[D_H(\widehat{AMISE} - AMISE)(H_{AMISE})] \\ &\quad \times [D_H^2 AMISE(H_{AMISE})]^{-1} \end{aligned}$$

D_H est l'opérateur différentiel par rapport à vechH et D_H^2 est l'opérateur Hessien correspondant.

- Ce lemme est connu dans la littérature sous le nom du lemme de AMSE.
- Notons que si

$$MSE(vech\widehat{H}) = O(J_{d'} n^{-2B})(vechH_{AMISE})(vech^T H_{AMISE}),$$

alors $\widehat{H}$ a un taux de convergence égal à n^{-B}.

- Le lemme (3.2.4) nous permet donc de calculer le taux de convergence de $\widehat{H}$ vers H_{AMISE} en connaissant l'espérance et la matrice variance-covariance de $D_H(\widehat{AMISE} - AMISE)(H_{AMISE})$. Rappelons que :

$$\widehat{AMISE}(H) = n^{-1} R(K) \, |H|^{-\frac{1}{2}} + \frac{1}{4} \mu_2(K)^2 (vech^T H) \widehat{\Psi}_4 (vechH)$$

d'où :

$$(\widehat{AMISE} - AMISE)(H) = \frac{1}{4} \mu_2(K)^2 (vech^T H)(\widehat{\Psi}_4 - \Psi_4)(vechH)[1 + o_p(1)].$$

On obtient ainsi

$$E[D_H(\widehat{AMISE} - AMISE)(H)] = \frac{1}{2} \mu_2(K)^2 [I_{d'} + o(J_{d'})](E\widehat{\Psi}_4 - \Psi_4)(vechH)$$

et

$$Var[D_H(\widehat{AMISE} - AMISE)(H)] = \frac{1}{4} \mu_2(K)^4 [I_{d'} + o(J_{d'})] Var[\widehat{\Psi}_4(vechH)]$$

Ceci nous conduit au résultat suivant :

Théorème 3.2.1. *(**Duong** et **Hazelton** ([31],2005a)*
Supposons que les conditions $(A_1), (A_2)$et (A_3) du lemme (3.2.4) sont satisfaites et que $K^{(r)}$ est carré-intégrable.
Supposons aussi que si $|r| = 4$, on a : $K^{(r)}(0) = 1$ si tous les éléments de r sont paires et $K^{(r)}(0) = 0$ sinon.
Si $\widehat{H}_{AMSE}$ et $\widehat{H}_{SAMSE}$ sont respectivement les matrices de lissage obtenues par la méthode AMSE et la méthode SAMSE, alors :

- *(i) Le taux relatif de convergence de $\widehat{H}_{AMSE}$ vers H_{AMISE} est $n^{-\frac{4}{d+12}}$.*
- *(ii) Le taux relatif de convergence de $\widehat{H}_{SAMSE}$ vers H_{AMISE} est $n^{-\frac{2}{d+6}}$.*

Remarque 3.2.1.

1. *Les conditions supplémentaires sur K sont satisfaites par la plupart des noyaux usuelles y compris le noyau gaussien.*
2. *Les propriétés asymptotiques de $\widehat{H}_{AMSE}$ sont meilleures que celles de $\widehat{H}_{SAMSE}$. Néanmoins, la différence dans le taux de convergence n'est pas important. Par conséquent, la comparaison entre les taux de convergence ne nous permet pas de préférer une méthode par rapport à une autre.*
3. ***Wand** et **Jones** ([98],1994) ont montré que le taux relatif de convergence pour la méthode AMSE-plug-in, si H est diagonale, est $n^{-\frac{min(8,d+4)}{2d+12}}$.*

3.2.5 Algorithmes de sélection des méthodes plug-in

Nous allons présenter deux algorithmes de sélection plug-in, il s'agit de l'algorithme m-étapes $AMSE$-matrice (diagonale ou complète) de lissage de **Wand** et **Jones** ([98],1994) et l'algorithme m-étapes $SAMSE$-matrice complète de lissage de **Duong** et **Hazelton** ([30],2003).

Algorithme de sélection m-étapes AMSE-matrice de lissage

1. Fixer m (le nombre d'étapes à faire).
2. Pour $J_{max} = 2m + 4$, donner l'approximation normale de référence $\widehat{\psi}_r^{NR}$ avec $|r| = J_{max}$. Puis injecter celle-ci dans la formule de $g_{r,AMSE}, |r| = J_{max} - 2$.
3. Pour $J = J_{max}-2,\ J_{max}-4, .., 6$:
   - a) Calculer les estimateurs à noyau pour les fonctionnelles ψ_r, $|r| = J$, en utilisant l'estimateur plug-in de $g_{r,AMSE}, |r| = J$
   - b) Remplacez les estimateurs $\widehat{\psi_r}$ dans les équation (3.9) et (3.10) pour avoir des estimateurs plug-in de $g_{r,AMSE}, |r| = J - 2$
4. Utiliser $g_{r,AMSE}, |r| = 4$ pour obtenir l'estimateur à noyau $\widehat{\Psi}_4$. Injecter cet estimateur dans l'équation (3.2) pour avoir $\widehat{AMISE}(H)$.
5. Pour avoir la matrice de lissage plug-in $\widehat{H}_{AMSE}$:
   - a) Si la matrice de lissage est diagonale et $d = 2$ alors on utilise :

$$h_{1,AMISE} = [\frac{\psi_{04}^{\frac{3}{4}} R(K)}{\mu_2(K)^2 \psi_{40}^{\frac{3}{4}} (\psi_{40}^{\frac{1}{2}} \psi_{04}^{\frac{1}{2}} + \psi_{22}) n}]^{\frac{1}{6}}.$$

$$h_{2,AMISE} = [\frac{\psi_{40}^{\frac{3}{4}} R(K)}{\mu_2(K)^2 \psi_{40}^{\frac{3}{4}} (\psi_{40}^{\frac{1}{2}} \psi_{40}^{\frac{1}{2}} + \psi_{22}) n}]^{\frac{1}{6}}.$$

   - b) sinon on minimise numériquement $\widehat{AMISE}(H)$.

Algorithme de sélection m-étapes SAMSE-matrice de lissage

1. Fixer m (le nombre d'étapes à faire).
2. Pour $J_{max} = 2m + 4$, donner l'approximation normale de référence $\widehat{\psi}_r^{NR}$ avec $|r| = J_{max}$. Puis injecter celle-ci dans la formule de $g_{J_{max}-2,SAMSE}$.
3. Pour $J = J_{max}-2,\ J_{max}-4, .., 6$:
   - a) Calculer les estimateurs à noyau pour les fonctionnelles ψ_r, d'ordre $J = |r|$, en utilisant l'estimateur plug-in de $g_{J,SAMSE}$
   - b) Remplacez les estimateurs $\widehat{\psi}_r$ dans l'équation (3.13) pour avoir un estimateur plug-in de $g_{J-2,SAMSE}$.
4. Utiliser $g_{4,SAMSE}$ pour obtenir l'estimateur à noyau $\widehat{\Psi}_4$. Injecter cet estimateur dans l'équation (3.2) pour avoir $\widehat{AMISE}(H)$.
5. Minimiser numériquement $\widehat{AMISE}(H)$ Pour avoir la matrice de lissage SAMSE-plug-in $\widehat{H}_{SAMSE}$.

Remarque 3.2.2. *Avant d'executer ces algorithmes, une pré-transformation des données doit être effectuée avant chaque estimation à noyau pilote. On obtient ainsi une matrice de lissage plug-in $\widehat{H}^*$ pour les données pré-transformées. La matrice des données originaux sera :*

$$H = \begin{cases} S^{\frac{1}{2}}\widehat{H}^* S^{\frac{1}{2}} & \textit{si les données sont pre-sphered} \\ S_D^{\frac{1}{2}}\widehat{H}^* S_D^{\frac{1}{2}} & \textit{si les données sont pre-scaled.} \end{cases}$$

3.3 Méthodes Cross validation (validation croisée)

Les méthodes de sélection Cross validation sont la principale alternative aux méthodes de sélection Plug-in. Ces méthodes ont été largement utilisées dans l'estimation de la densité de probabilité univariée. Dans ce contexte, ces méthodes comme les méthodes Plug-in d'ailleurs ont connu un développement considérable et leurs performances ont été largement étudiées. Cependant, leur utilisation est très limitée pour l'estimation de la densité de probabilité multivariée.

Il existe trois principales méthodes cross validation : validation croisée non-biaisée (UCV), validation croisée biaisée (BCV) et validation croisée lissé (SCV). Dans cette section, nous allons introduire les versions multivariées de ces trois méthodes. Pour le calcul de leurs taux relatifs de convergence, nous adopterons la même stratégie suivie dans la section précédente pour les méthodes Plug-in. Ainsi, les définitions (3.2.1), (3.2.2), (3.2.3) et le lemme (3.2.4) de AMSE données dans la section précédente restent valables et seront utilisées ici.

3.3.1 Validation croisée non-biaisée (UCV)

Présentation de la méthode

La version multivariée du critère de la méthode UCV est une simple généralisation de sa version univariée proposée par **Rudemo** ([73],1982) et **Bowman** ([12],1984) :

$$UCV(H) = \int_{\Re^d} \hat{f}(x;H)^2 \; dx \; - \; 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i} \; (X_i;H),$$

où :

$$\hat{f}_{-i} \; (x;H) = (n-1)^{-1} \sum_{j=1;j\neq i}^{n} K_H(x - X_j).$$

Notre but est de choisir la matrice de lissage qui minimise $UCV(H)$, c'est à dire :

$$\hat{H}_{UCV} = argmin_{H\in F} \; UCV(H).$$

Remarque 3.3.1. *La méthode UCV propose de minimiser une quantité proche de $ISE(H)$. En effet, on a :*

$$\begin{aligned} ISE(H) &= \int_{\Re^d} [f(x) - \hat{f}(x;H)]^2 \; dx \\ &= \int_{\Re^d} \hat{f}(x;H)^2 \; dx \; - \; 2 \int_{\Re^d} f(x) \; \hat{f}(x;H) \; dx \; + \; \int_{\Re^d} f(x)^2 \; dx \\ &= R(\hat{f}(x)) \; - \; 2 \int_{\Re^d} f(x) \; \hat{f}(x;H) \; dx \; + \; R(f(x)). \end{aligned}$$

Comme $R(f(x))$ est indépendant de H, ce terme peut être ignoré. Il reste alors à estimer $\int_{\Re^d} f(x) \; \hat{f}(x;H) \; dx$. Notons que :

$$\int_{\Re^d} f(x) \; \hat{f}(x;H) \; dx = E(\hat{f}(X;H)).$$

Son estimateur naturel est :

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}(X_i;H).$$

Pour garantir l'indépendance des variables, on utilise l'estimateur Jackknife suivant :

$$\hat{f}_{-i}(x;H) = (n-1)^{-1} \sum_{j=1;j\neq i}^{n} K_H(x - X_j).$$

On estime $\int_{\Re^d} f(x) \; \hat{f}(x;H) \; dx$ par :

$$\frac{1}{n} \sum_{i=1}^{n} \hat{f}_{-i}(X_i;H).$$

En remplaçant dans la formule de $ISE(H)$, on obtient le critère $UCV(H)$ à minimiser. Ce critère peut être développé pour avoir :

$$\begin{aligned} UCV(H) &= n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (K_H * K_H)(X_i - X_j) - 2n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j=1;j\neq i}^{n} K_H(X_i - X_j) \\ &= n^{-1}R(K)\,|H|^{-\frac{1}{2}} + n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j=1;j\neq i}^{n} (K_H * K_H - 2K_H)(X_i - X_j) \quad (3.15) \end{aligned}$$

Pour le noyau gaussien, cette dernière expression se simplifie beaucoup plus. En effet, on a : $\phi_H * \phi_H = \phi_{2H}$, d'où :

$$UCV(H) = n^{-}(4\pi)^{-\frac{d}{2}}\,|H|^{-\frac{1}{2}} + n^{-1}(n-1)^{-1} \sum_{i=1}^{n} \sum_{j=1;j\neq i}^{n} (\phi_{2H} - 2\phi_H)(X_i - X_j).$$

Des recherches ont été entreprises par ***Sain*** *et* ***al.*** *([76],1994) sur la version multivariée de la méthode de selection UCV, mais leur intérêt s'est porté uniquement sur le noyau produit qui est équivalent à utiliser une paramétrisation diagonale avec un noyau sphérique K^s. Ces auteurs ont calculé la vitesse de convergence relative pour la méthode UCV dans le cas où la paramétrisation est diagonale.* ***Duong*** *et* ***Hazelton*** *([32],2005b) ont généralisé ces résultats pour une matrice de lissage H symétrique et définie positive quelconque ($H \in F$).*

Taux relatif de convergence de la méthode UCV

En suivant le même procédé développé dans la section précédente, on peut calculer le taux de convergence relatif de $\hat{H}_{UCV}$ vers H_{AMISE}.

Lemme 3.3.1. *Supposons que les conditions $(A1)$, $(A2)$ et $(A1)$ du lemme de $AMSE$ (3.2.4) sont satisfaites et supposons aussi que le noyau K est normal, alors :*

$$ABiais(vech\hat{H}_{UCV}) = O(J_d n^{-\frac{2}{d+4}})vechH_{AMISE}.$$

$$AVar(vech\hat{H}_{UCV}) = O(J_d n^{-\frac{d}{d+4}})(vechH_{AMISE})(vech^T H_{AMISE}).$$

En combinant le lemme (3.2.4) de $AMSE$ et le lemme (3.3.1) précédent, on obtient le théorème suivant qui donne le taux relatif de convergence de la méthode de selection UCV.

Théorème 3.3.1. *Sous les conditions du Lemme(3.3.1), le taux relatif de convergence de $\hat{H}_{UCV}$ vers H_{AMISE} est $n^{-\frac{min(d,4)}{2d+8}}$.*

Ce taux de convergence est calculé pour une matrice de lissage quelconque $H \in F$. Le taux de convergence reste le même pour $H \in D$ ou $H \in A$.

3.3.2 Validation croisée biaisée (BCV)

Présentation de la méthode

L'approche basée sur la méthode de sélection validation croisée biaisée (BCV), consiste à minimiser un estimateur de AMISE :

$$AMISE\hat{f}(.;H) = n^{-1}R(K)\,|H|^{-\frac{1}{2}} + \frac{1}{4}\mu_2(K)^2(vech^T H)\Psi_4(vechH).$$

Comme pour la méthode de sélection Plug-in, on a besoin d'estimer Ψ_4. La méthode Plug-in utilise une matrice de lissage pilote G qui est indépendante de la matrice de lissage H. Pour la méthode BCV, on pose $G = H$ et on utilise des estimateurs un peu différents. Comme le $AMISE$ est un estimateur biaisé de $MISE$, l'estimateur BCV est aussi un estimateur biaisé de MISE (bien qu'il soit asymptotiquement non-biaisé). Ceci donne à la méthode BCV son nom : le biais est introduit pour réduire la variance.
Il existe deux versions pour la méthode BCV et ceci selon l'estimateur utilisé pour ψ_r, $|r| = 4$. On peut consulter **Sain** et **al.** ([76],1994) et **Jones** et **Kappenman** ([56],1992) sue ce sujet. Ainsi, on peut utiliser l'estimateur :

$$\check{\psi}_r(H) = n^{-2}\sum_{i=1}^{n}\sum_{j=1;j\neq i}^{n}(K_H^{(r)} * K_H)(X_i - X_j),$$

comme on peut considérer l'estimateur :

$$\tilde{\psi}_r(H) = n^{-1}\sum_{i=1}^{n}\hat{f}_{-i}^{(r)}(X_i;H) = n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{j=1;j\neq i}^{n}K_H^{(r)}(X_i - X_j).$$

Les estimateurs $\check{\Psi}_4$ et $\tilde{\Psi}_4$ sont obtenus en remplaçant dans Ψ_4, les fonctionnelles ψ_r par les estimateurs $\check{\psi}_r$ et $\tilde{\psi}_r$ respectivement.
La fonction $BCV1$ est la version de la BCV avec $\check{\Psi}_4$:

$$BCV1(H) = n^{-1}R(K)\,|H|^{-\frac{1}{2}} + \frac{1}{4}\mu_2(K)^2(vech^T H)\check{\Psi}_4(vechH). \tag{3.16}$$

La fonction $BCV2$ est la version de la BCV avec $\tilde{\Psi}_4$:

$$BCV2(H) = n^{-1}R(K)\,|H|^{-\frac{1}{2}} + \frac{1}{4}\mu_2(K)^2(vech^T H)\tilde{\Psi}_4(vechH). \tag{3.17}$$

La méthode de sélection BCV propose de trouver la matrice $\hat{H}_{BCV}$ qui minimise la fonction BCV appropriée. C'est à dire :

$$\hat{H}_{BCV1} = argmin_{H\in F}\ BCV1(H) \quad \texttt{et} \quad \hat{H}_{BCV2} = argmin_{H\in F}\ BCV2(H).$$

Sain et **al.**([76],1994) ont effectué des recherches sur la méthode de sélection BCV dans le cas d'une paramétrisation diagonale. Ils ont ainsi calculé le taux relatif de convergence de la BCV diagonale. La généralisation de cette méthode pour une matrice de lissage symétrique et définie positive quelconque ($H \in F$), a été donnée par **Duong** et **Hazelton** ([32],2005b).

Taux relatif de convergence de la méthode BCV

Les deux estimateurs $\check{\psi}_r$ et $\tilde{\psi}_r$ sont relativement similaires. Si on utilise le noyau Gaussien on obtient $\phi_H^{(r)} * \phi_H = (-1)^{|r|}\phi_{2H}^{(r)}$. Par conséquent, la seule différence est que $\check{\psi}_r$ utilise $2H$ et $\tilde{\psi}_r$ utilise H. Cette différence n'affecte pas le taux relatif de convergence, comme il n'affecte pas l'ordre de biais et de la variance asymptotique. Ainsi, on a seulement besoin de calculer le taux de $BCV2$.

Lemme 3.3.2. *Supposons que les conditions* $(A1)$, $(A2)$ *et* $(A3)$ *de lemme (3.2.4) de* $AMSE$ *sont satisfaites. Alors :*

$$ABiais(vech\hat{H}_{BCV}) = O(J_d n^{-\frac{2}{d+4}})vechH_{AMISE}.$$

et

$$AVar(vech\hat{H}_{BCV}) = O(J_d n^{-\frac{d}{d+4}})(vechH_{AMISE})(vech^T H_{AMISE}).$$

En combinant le lemme (3.2.4) de $AMSE$ et le lemme (3.3.2) précédent, on obtient la théorème suivant qui donne le taux relatif de convergence de la méthode de sélection BCV. C'est à dire le taux relatif de convergence de $\hat{H}_{BCV}$ vers H_{AMISE}.

Théorème 3.3.2. *Sous les conditions du lemme(3.3.2), le taux relatif de convergence de* $\hat{H}_{BCV}$ *vers* H_{AMISE} *est* $n^{-\frac{min(d,4)}{2d+8}}$.

Ce taux de convergence est identique à celui de la méthode de sélection UCV. **Sain** et **al.** ([76],1994) ont donné un taux de convergence relatif pour la méthode BCV diagonale égal à $n^{-\frac{d}{2d+8}}$, ce résultat est incorrect pour $d > 4$.

3.3.3 Validation croisée lissée (SCV)

Présentation de la méthode

La méthode de sélection SCV (Smoothed cross validation) peut être considérée comme une combinaison de la méthode UCV et de la méthode BCV. Cette méthode permet de lisser le critère UCV. La version multivariée du critère SCV est une simple généralisation de sa version univariée donnée par **Hall**, **Marron** et **Park**([45],1992). Elle s'écrit sous la forme :

$$SCV(H) = n^{-1}R(K)\,|H|^{-\frac{1}{2}} + n^{-2}\sum_{i=1}^{n}\sum_{j=1}^{n}(K_H * K_H * L_G * L_G - 2K_H * L_G * L_G + L_G * L_G)(X_i - X_j),$$

où L_G est le noyau pilote avec la matrice de lissage pilote G. le premier terme est l'estimateur de la variance asymptotique intégrée et le deuxième terme est l'estimateur de l'intégral du biais au carré. La méthode de sélection SCV propose de trouver la matrice $\hat{H}_{SCV}$ qui minimise le critère $SCV(H)$.

Notons que pour $G = O$ (matrice nulle), on obtient $SCV(H) = UCV(H)$. L_0 est la fonction delta de Dirac.

Si $K = L = \phi$ alors le critère SCV à une forme plus simple :

$$SCV(H) = n^{-1} |H|^{-\frac{1}{2}} (4\pi)^{-\frac{d}{2}} + n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} (\phi_{2H+2G} - 2\phi_{H+2G} + \phi_{2G})(X_i - X_j). \quad (3.18)$$

Matrice de lissage pilote optimale

Nous avons déjà soulevé ce problème pour les méthodes Plug-in : comment choisir la matrice pilote de lissage optimale ?

• **Sain** et **al.** ([76],1994) considèrent la matrice pilote égale à la matrice de lissage finale.

• **Jones** et **Al.**([57],1991), dans le cas univarié, proposent de prendre le paramètre pilote qui minimise l'erreur quadratique moyenne relative $RMSE$ (relative mean squared error). Pour un paramètre univarié $\hat{h}$, on a

$$RMSE(\hat{h}) = E[(\hat{h} - h_{AMISE})/h_{AMISE}]^2.$$

• **Duong** et **Hazelton** ([32],2005b) par contre, proposent de minimiser le $AMSE$. Notons que minimiser le $AMSE$ est équivalent à minimiser le $RMSE$ car le dénominateur de $RMSE$ ne dépend pas de paramètre de lissage.

Une généralisation de la version univariée de $MSE(\hat{h}) = E(\hat{h} - h_{AMISE})^2$ est :

$$trMSE(vech\hat{H}; G) = E[vech^T(\hat{H} - H_{AMISE})vech(\hat{H} - H_{AMISE})].$$

Cette formule exacte de MSE est difficile à calculer. **Duong** et **Hazelton**([32],2005b) proposent de travailler avec le $AMSE$. Si on pose $G = g^2 I$, alors on cherche :

$$g_0 = argmin_{g>0} tr\{AMSE(vech\hat{H}_{scv}; g)\}.$$

Pour calculer g_0 on aura besoin des résultats intermédiaires suivants dus à **Duong** et **Hazelton** ([32],2005b). Pour cela, on définit l'extension d'ordre supérieur de $AMISE$, le $AMISE'$ défini par :

$$AMISE' = AMISE(H) + \frac{1}{8} \int_{\Re^d} tr(H\chi_f(x)) tr(H^2 \chi_f^2(x)) dx.$$

On notera $\widehat{AMISE'}$ l'estimateur de $AMISE'$.

Lemme 3.3.3. *Supposons que les conditions* $(A1)$, $(A2)$ *et* $(A3)$ *de lemme de AMSE (3.2.4) sont satisfaites. Soit* $\hat{H} = argmin_{H \in F} \widehat{AMISE'}$, *alors :*

$$MSE(vech\hat{H}) = [I_{d'} + o(J_{d'})] AMSE'(vech\hat{H}).$$

Où l'extension d'ordre supérieur asymptotique de MSE s'écrit :

$$AMSE'(vech\hat{H}) = AVar'(vech\hat{H}) + [ABiais'(vech\hat{H})][ABiais'(vech\hat{H})]^T,$$

avec :

$$ABiais'(vech\hat{H}) = [D_H^2 AMISE(H_{AMISE})]^{-1} E[D_H(\widehat{AMISE}' - AMISE')(H_{AMISE})]$$

et

$$\begin{aligned} AVar'(vech\hat{H}) &= [D_H^2 AMISE(H_{AMISE})]^{-1} Var[D_H(\widehat{AMISE}' - AMISE')(H_{AMISE})] \\ &\quad \times [D_H^2 AMISE(H_{AMISE})]^{-1}. \end{aligned}$$

Lemme 3.3.4. *Supposons que les conditions de lemme (3.3.3) de $AMSE'$ sont satisfaites et supposons aussi :*

- *(S1) f admet des dérivées partielles jusqu'à l'ordre huit, bornées et continues,*
- *(S2) chaque élément de $\theta_6 = \int_{\Re^d} \chi_f^3(x) f(x) dx$ est finie,*
- *(S3) la suite des paramètres pilotes de lissage $g = g_n$ vérifie $g^{-2}H \to 0$ quand $n \to \infty$.*
- *(S3) K et L sont des noyaux gaussiens,*

alors

$$ABiais'(vech\hat{H}_{SCV}; g) = n^{-\frac{2}{d+4}} g^2 C_{\mu_1} + n^{-\frac{2}{d+4}} n^{-1} g^{-d-4} C_{\mu_2} + 0(J_{d'}(g^4 + n^{-1}g^{-d-6}))(vechH_{AMISE}).$$

où

$$C_{\mu_1} = \frac{1}{2} n^{\frac{2}{(d+4)}} D_d^T vec(\theta_6 H_{AMISE})$$

et

$$C_{\mu_2} = \frac{1}{8}(4\pi)^{-\frac{d}{2}} n^{\frac{2}{(d+4)}} [2D_d^T vecH_{AMISE} + (trH_{AMISE}) D_d^T vecI_d].$$

Lemme 3.3.5. *Supposons que les conditions de lemme (3.3.3) de $AMSE'$ sont satisfaites et sous les conditions de lemme (3.3.4), on a :*

$$AVar'(vech\hat{H}_{SCV}; g) = 0(J_{d'}(n^{-2}g^{-d-8} + n^{-1}))(vechH_{AMISE})(vech^T H_{AMISE}).$$

Les trois lemmes (3.3.3), (3.3.4), (3.3.5) nous conduisent au théorème suivant :

Théorème 3.3.3. *Sous les conditions du lemme (3.3.4) et (3.3.5), le paramètre pilote qui minimise la trace de $AMSE'(vech\hat{H}_{SCV}; g)$ pour $d > 1$ est :*

$$g_0 = \left\{ \frac{2(d+4) C_{\mu_2}^T C_{\mu_2}}{[-(d+2) C_{\mu_2}^T C_{\mu_1} + C_{\mu_0}^{\frac{1}{2}}] n} \right\}^{\frac{1}{(d+6)}}$$

où :

$$\begin{aligned} C_{\mu_0} &= (d+2)^2 (C_{\mu_2}^T C_{\mu_1})^2 + 8(d+4)(C_{\mu_1}^T C_{\mu_1})(C_{\mu_2}^T C_{\mu_2}), \\ C_{\mu_1} &= \frac{1}{2} n^{\frac{2}{(d+4)}} D_d^T vec(\theta_6 H_{AMISE}), \end{aligned}$$

et

$$C_{\mu_2} = \frac{1}{8}(4\pi)^{-\frac{d}{2}} n^{\frac{2}{(d+4)}} [2D_d^T vecH_{AMISE} + (trH_{AMISE}) D_d^T vecI_d].$$

Taux relatif de convergence de la méthode SCV

Le taux relatif de convergence de la méthode de sélection SCV est une conséquence immédiate de théorème (3.3.3) et de lemme (3.3.3) de $AMSE'$. Notons que si

$$trMSE(vech\hat{H}) = 0(n^{-2\alpha}\,||vechH_{AMISE}||^2),$$

alors $\hat{H}$ a un taux relatif de convergence vers H_{AMISE} égal à $n^{-\alpha}$.

Théorème 3.3.4. *Sous les conditions de lemme (3.3.4) et (3.3.5), pour $d > 1$ le taux relatif de convergence de $\hat{H}_{SCV}$ vers H_{AMISE} est $n^{-\frac{2}{(d+6)}}$.*

3.3.4 Algorithmes de sélection des méthodes cross validation

Nous allons présenter les algorithmes des trois méthodes UCV, BCV et SCV donnés par **Duong** et **Hazelton**([32],2005b). La méthode de sélection SCV est beaucoup plus complexe que les méthodes UCV et BCV. En effet, la SCV a besoin d'estimer une matrice pilote de lissage en utilisant la méthode Plug-in. Elle exige aussi des pré-transformations de données.

Algorithme pour la méthode de sélection UCV

1.Minimiser numériquement l'équation (3.15) de $UCV(H)$.

Algorithme pour la méthode de sélection BCV

1.Minimiser numériquement

(a) L'équation (3.16) de $BCV1(H)$, ou
(b) L'équation (3.17) de $BCV2(H)$.

Algorithme de sélection m-étape de la méthode SCV

1. Fixer m (le nombre d'étapes à faire).
2. Pour $J_{max} = 2m + 4$, donner l'approximation normale de référence $\hat{\psi}_r^{NR}$ avec $|r| = J_{max}$. Puis injecter celle-ci dans la formule de $g_{J_{max}-2,SAMSE}$.
3. Pour $J = J_{max}-2,\ J_{max}-4, .., 6$:
   - (a) Calculer les estimateurs à noyau pour les fonctionnelles ψ_r, d'ordre $J = |r|$, en utilisant l'estimateur plug-in de $g_{J,SAMSE}$
   - b) Remplacer les estimateurs $\hat{\psi}_r$ dans l'équation (3.13) pour obtenir des estimateurs plug-in de $g_{J-2,SAMSE}$.
4. Utiliser $g_{6,SAMSE}$ pour avoir l'estimateur à noyau $\hat{\theta}_6$ de θ_6.

5. Utiliser $g_{4,SAMSE}$ pour obtenir l'estimateur à noyau $\widehat{\Psi}_4$. Injecter cet estimateur dans l'équation (3.2) pour avoir $\widehat{AMISE}(H)$.
6. Minimiser numériquement $\widehat{AMISE}(H)$ pour avoir la matrice de lissage $SAMSE$-plug-in $\hat{H}_{SAMSE}$.
7. Utiliser $\hat{H}_{SAMSE}$ et $\hat{\theta}_6$ pour obtenir l'estimateur $\hat{g}_0$ de théorème (3.3.3).
8. Remplacer $\hat{g}_0$ dans l'équation (3.18) pour obtenir la formule de $SCV(H)$.
9. Minimiser numériquement $SCV(H)$.

3.4 Conclusion

Dans ce chapitre, on a présenté les différentes méthodes pour le choix de la matrice de lissage optimale H qui intervient dans l'estimation de la densité de probabilité dans le cas multidimensionnel. Dans le chapitre suivant, on va comparer sur des densités cibles à plusieurs modes, ces différentes méthodes de sélection de la matrice de lissage optimale H.

4
Simulation et résultats numériques

4.1 Introduction

Nous présentons dans ce chapitre le travail de simulation effectué pour étayer les différents aspects théoriques abordés dans les chapitres précédents. Cet étude portera essentiellement sur :

- La comparaison des différents algorithmes de sélection de la matrice de lissage ;
- L'étude de la performance de ces algorithmes ;
- L'étude de l'influence de la taille de l'échantillon sur ces différents algorithmes.

4.2 Plan de simulation

Pour notre plan de simulation, nous allons restreindre notre étude au cas bivarié, et les résultats obtenus se généralisent facilement aux cas multivariés. Nous nous contenterons de faire des simulations et d'observer le comportement asymptotique de l'estimateur à noyau. Nous utiliserons pour les simulations des échantillons de lois connues de tailles de plus en plus grandes (100,...,1000) et nous réaliserons 40 répétitions d'expérience pour chaque type de densité et pour chaque taille d'échantillon.
Afin d'illustrer les performances des méthodes de sélection présentées dans le troisième chapitre, nous utiliserons des densité cibles connues. C'est -à-dire des densités dont le $MISE$ a une forme explicite facile à calculer (closed forme), ceci va nous permettre de faire des comparaisons entre les valeurs réelles de H et de $MISE$ et les valeurs obtenues

par simulation. Les densités suivant une loi mélange gaussienne vérifient cette condition. Nous avons choisi des densités présentant différents aspects :

- La densité (A) : Loi gaussienne bivariée :

$$N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{4} & 0 \\ 0 & 1 \end{bmatrix}).$$

- La densité (B) : Loi mélange gaussienne :

$$\frac{1}{2}N(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{4}{9} & 0 \\ 0 & \frac{4}{9} \end{bmatrix}) + \frac{1}{2}N(\begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{4}{9} & 0 \\ 0 & \frac{4}{9} \end{bmatrix}).$$

- La densité (c) : Loi mélange gaussienne :

$$\frac{1}{2}N(\begin{bmatrix} \frac{3}{2} \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{16} & 0 \\ 0 & 1 \end{bmatrix}) + \frac{1}{2}N(\begin{bmatrix} -\frac{3}{2} \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{1}{16} & 0 \\ 0 & 1 \end{bmatrix}).$$

- La densité (D) : Loi mélange gaussienne :

$$\frac{1}{2}N(\begin{bmatrix} 1 \\ -1 \end{bmatrix}, \begin{bmatrix} \frac{4}{9} & \frac{14}{45} \\ \frac{14}{45} & \frac{4}{9} \end{bmatrix}) + \frac{1}{2}N(\begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} \frac{4}{9} & 0 \\ 0 & \frac{4}{9} \end{bmatrix}).$$

- La loi (E) : Loi mélange gaussienne :

$$\frac{3}{7}N(\begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} \frac{9}{25} & \frac{63}{250} \\ \frac{63}{250} & \frac{49}{100} \end{bmatrix}) + \frac{3}{7}N(\begin{bmatrix} 1 \\ \frac{2}{\sqrt{3}} \end{bmatrix}, \begin{bmatrix} \frac{9}{25} & 0 \\ 0 & \frac{49}{100} \end{bmatrix} + \frac{1}{7}N(\begin{bmatrix} 1 \\ -\frac{2}{\sqrt{3}} \end{bmatrix}, \begin{bmatrix} \frac{9}{25} & 0 \\ 0 & \frac{49}{100} \end{bmatrix}).$$

- La densité (F) : Loi gaussienne bivariée :

$$N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \frac{9}{10} \\ \frac{9}{10} & 1 \end{bmatrix}).$$

Les contours-plots des densités A, B, C, D, E et F sont donnés par la figure (4.1) suivante :

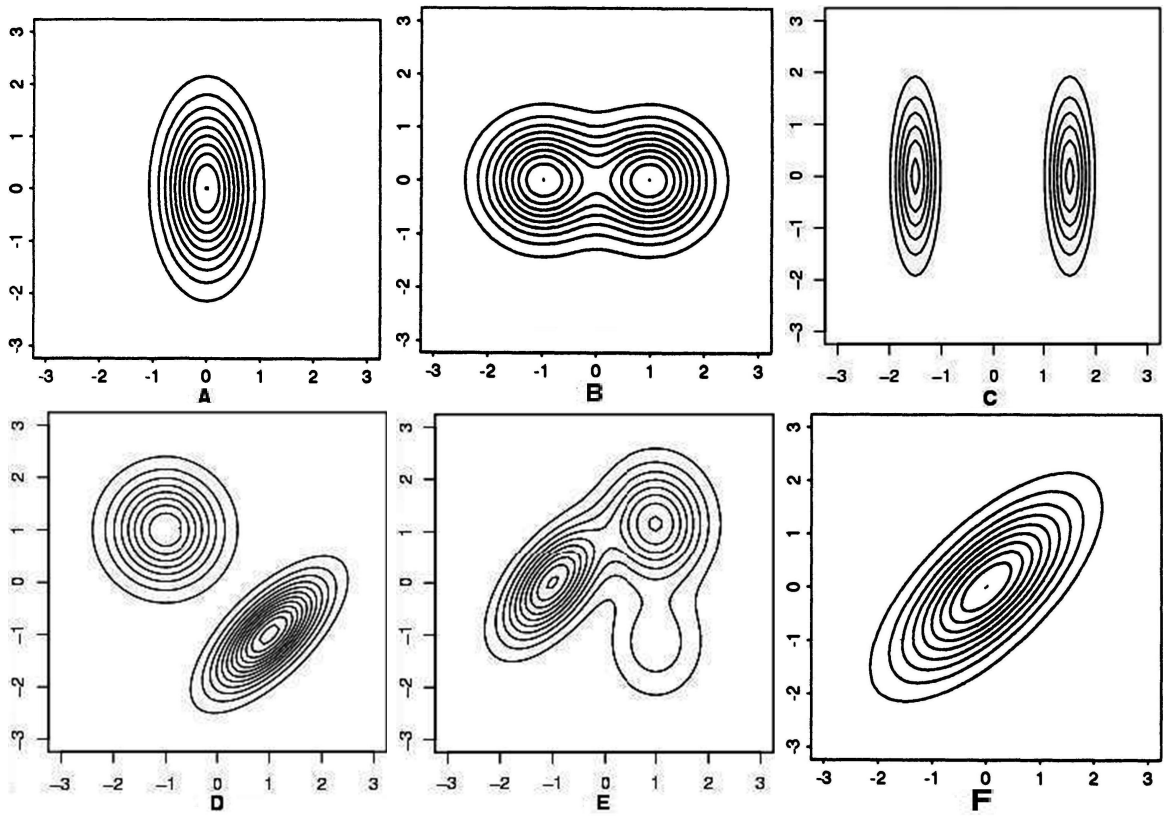

FIGURE 4.1 – Contours-plots des densités A, B, C, D, E et F.

Notons que :

- Les contours-plots de la densité A sont orientés suivant les axes des coordonnées. Par contre les contours-plots de F ont une orientation différente de celle définie par les axes des coordonnées. Ces deux densités sont unimodales.
- Les densités B, C et D sont bimodales. Les densités C et D ont des modes séparés.
- La densité E est trimodale.

Dans la programmation des méthodes de sélection de la matrice de lissage H, nous avons choisi comme noyau, le noyau gaussien, afin d'avoir une forme du $MISE$ explicite et facile à calculer.

Nous allons suivre les étapes suivantes pour la simulation :

- Simuler un échantillon de taille n pour la densité cible choisie.
- Calculer le H_{MISE} qui minimise le $MISE$.
- Calculer la matrice de lissage optimale H_{opt} en utilisant les différentes méthodes de sélection :

- L'algorithme de sélection plug-in (AMSE) à 2 étapes, noté F1 pour les données pre-scaling et F2 pour les données pre-sphéring .
- L'algorithme de sélection plug-in (SAMSE) à 2 étapes, noté S1 pour les données pre-scaling et S2 pour les données pre-sphéring.
- L'algorithme de sélection UCV, noté U.
- Les deux algoritmes de sélection BCV1 et BCV2, noté B1 et B2 respectivement.
- L'algorithme de sélection SCV, noté SC1 pour les données pré-scaling- et SC2 pour les données pré-sphéring.

• Construire l'estimateur par la méthode du noyau à partir des observations.

• Calculer le $MISE(H_{MISE})$ et le $MISE(H_{opt})$.

• Calculer le critère de comparaison (Crit) pour les différentes méthodes utilisées. Le critère utilisé (Crit) est l'efficacité moyenne de l'estimateur de H définie par : $\frac{MISE(H_{MISE})}{MISE(H_{opt})}$.

Les simulations et les graphiques ont été réalisés à l'aide du logiciel R (package ks). Nous avons utilisé la version 2.8.0 pour la programmation. R est un système d'analyse statistique et graphique crée par **Ross Ihaka** et **Robert Gentlelman**. Il est à la fois un langage et un logiciel qui comporte de nombreuses fonctions pour les analyses statistiques et les graphiques.

4.3 Résultats de la simulation

Les résultats de la simulation sont données sous forme de tableaux. Les tableaux : (4.1), (4.2), (4.3), (4.4), (4.5), (4.6), (4.7), (4.8), donnent les valeurs de H_{MISE} et H_{opt} pour les différentes densités cibles, avec les différentes méthodes de sélection utilisées, pour une taille n de l'échantillon. Et les tableaux : (4.9), (4.10), (4.11), (4.12), (4.13), (4.14), (4.15), (4.16), donnent les valeurs de $MISE(H_{MISE})$, de $MISE(H_{opt})$ ainsi que la valeur du critère de comparaison pour chaque densité cible, avec les différentes méthodes de sélection, pour un échantillon n donnée.
Pour les résultats, on a considéré la médiane des 40 simulations réalisées.

4.4 Performance des estimateurs

La lecture des résultats obtenus dans les tableaux (4.1),...,(4.16), indique que :

4.4.1 Pour les deux lois (A) et (F) (unimodales)

La lecture des tableaux, nous permet d'affirmer que :

- Les méthodes de sélection plug-in et validation croisée fournissent une matrice des paramètres de lissage très proche de H_{MISE}. Les résultats indiquent que toutes les méthodes de sélection estiment correctement la matrice des paramètres de lissage H.
- La taille de l'échantillon n influe sur les valeurs de $MISE$, H_{AMISE}, H_{opt}. On constate que l'augmentation de n entraîne la décroissance de $MISE(H_{MISE})$ et de $MISE(H_{opt})$.
- La taille de l'échantillon n influe aussi sur la valeur de $Crit$. L'efficacité moyenne de l'estimateur augmente lorsque la taille de l'échantillon n augmente.
- Les plus grandes valeurs de l'efficacité moyenne sont obtenues pour les méthodes plug-in en générale, et particulièrement avec les données pré-scaled, où par exemple pour $n = 1000$ la valeur de $Crit$ est 0.94 pour $AMSE$ plug-in et 0.96 pour $SAMSE$ plug-in.
- Les autres méthodes (validation croisée) se sont également révélées performantes dans ce cas. Les résultats obtenus sont proches des résultats obtenues par les méthodes plug-in.

4.4.2 Pour les loi (B), (C), (D) et (E) (multimodales)

L'analyse des résultats, indique que :

- La matrice de lissage optimale H_{opt} estimée par les différentes méthodes de sélection se rapproche de H_{MISE} lorsque la taille de l'échantillon n augmente.
- La taille de l'échantillon influe aussi sur les valeurs de $MISE$ et de H_{MISE}, l'augmentation de n entraîne la décroissance de $MISE(H_{MISE})$ et de $MISE(H_{opt})$.
- Le critère d'efficacité moyenne $Crit$ s'améliore lorsque n augmente.
- Les résultats du critère d'efficacité moyenne obtenus pour les densités multimodales (B), (C), (D) et (E) indiquent que les performances des méthodes s'amoindrissent en présence de cibles plus complexes. Les meilleurs résultats sont obtenus par les méthodes plug-in. Pour la densité (B), les résultats obtenus par les méthodes cross-validation sont proches de ceux obtenus par les méthodes plug-in. Pour les autres densités, les méthodes cross-validation sont moins performantes que les méthodes plug-in.

4.5 Conclusion

Notre travail de simulation aborde un problème crucial, bien connu dans l'estimation à noyau de la densité multivariée. Il s'agit du problème de choix de la matrice de lissage optimale pour l'estimateur. Les résultats obtenus dans le contexte bivarié confirment ceux

déja obtenus dans la cas unidimensionnel et se généralisent facilement aux cas multidimensionnels. En effet :

- La performance de la méthode de sélection de la matrice de lissage varie en fonction de la densité à estimer.
- Les résultats montrent qu'il n'existe pas de méthode de sélection de la matrice de lissage qui soit meilleure que toutes les autres.
- La matrice de lissage optimale H_{opt} obtenue par les méthodes de sélection plug-in ou validation croisée n'est pas toujours adaptée lorsque on est en présence de densités cibles complexes (multimodales par exemple).
- Les méthodes plug-in imposent des restrictions sur la densité à estimer f (f par exemple doit être suffisamment lisse et régulière) souvent difficiles à vérifier.
- L'utilisation d'un nombre d'observations de plus en plus important, ne garantit pas une bonne estimation.

f	H_{MISE}	F1	F2
A	$\begin{bmatrix} 0.063 & 0 \\ 0 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.049 & -0.015 \\ -0.015 & 0.150 \end{bmatrix}$	$\begin{bmatrix} 0.050 & -0.003 \\ -0.003 & 0.127 \end{bmatrix}$
B	$\begin{bmatrix} 0.201 & 0 \\ 0 & 0.135 \end{bmatrix}$	$\begin{bmatrix} 0.196 & -0.001 \\ -0.001 & 0.072 \end{bmatrix}$	$\begin{bmatrix} 0.241 & -0.010 \\ -0.010 & 0.117 \end{bmatrix}$
C	$\begin{bmatrix} 0.021 & 0 \\ 0 & 0.335 \end{bmatrix}$	$\begin{bmatrix} 0.091 & -0.001 \\ -0.001 & 0.287 \end{bmatrix}$	$\begin{bmatrix} 0.091 & 0.033 \\ 0.033 & 0.025 \end{bmatrix}$
D	$\begin{bmatrix} 0.136 & 0.071 \\ 0.071 & 0.136 \end{bmatrix}$	$\begin{bmatrix} 0.101 & 0.021 \\ 0.021 & 0.112 \end{bmatrix}$	$\begin{bmatrix} 0.164 & -0.109 \\ -0.109 & 0.07 \end{bmatrix}$
E	$\begin{bmatrix} 0.140 & 0.073 \\ 0.073 & 0.185 \end{bmatrix}$	$\begin{bmatrix} 0.198 & 0.047 \\ 0.047 & 0.185 \end{bmatrix}$	$\begin{bmatrix} 0.205 & 0.051 \\ 0.051 & 0.140 \end{bmatrix}$
F	$\begin{bmatrix} 0.252 & 0.227 \\ 0.227 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.241 & 0.106 \\ 0.106 & 0.240 \end{bmatrix}$	$\begin{bmatrix} 0.200 & 0.150 \\ 0.150 & 0.160 \end{bmatrix}$

TABLE 4.1 – Résultats des simulations effectuées avec les méthodes AMSE-plug-in pour un échantillon de taille n=100.

f	H_{MISE}	S1	S2
A	$\begin{bmatrix} 0.063 & 0 \\ 0 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.029 & 0.011 \\ 0.011 & 0.183 \end{bmatrix}$	$\begin{bmatrix} 0.041 & -0,018 \\ -0.018 & 0.142 \end{bmatrix}$
B	$\begin{bmatrix} 0.201 & 0 \\ 0 & 0.135 \end{bmatrix}$	$\begin{bmatrix} 0.206 & 0.017 \\ 0.017 & 0.065 \end{bmatrix}$	$\begin{bmatrix} 0.290 & -0,010 \\ -0.010 & 0.114 \end{bmatrix}$
C	$\begin{bmatrix} 0.021 & 0 \\ 0 & 0.335 \end{bmatrix}$	$\begin{bmatrix} 0.106 & -0.003 \\ -0.003 & 0.172 \end{bmatrix}$	$\begin{bmatrix} 0.110 & 0,006 \\ 0.006 & 0.180 \end{bmatrix}$
D	$\begin{bmatrix} 0.136 & 0.071 \\ 0.071 & 0.136 \end{bmatrix}$	$\begin{bmatrix} 0.124 & 0.017 \\ 0.017 & 0.115 \end{bmatrix}$	$\begin{bmatrix} 0.138 & -0,051 \\ -0.051 & 0.141 \end{bmatrix}$
E	$\begin{bmatrix} 0.140 & 0.073 \\ 0.073 & 0.185 \end{bmatrix}$	$\begin{bmatrix} 0.186 & 0.042 \\ 0.042 & 0.152 \end{bmatrix}$	$\begin{bmatrix} 0.198 & 0,071 \\ 0.071 & 0.133 \end{bmatrix}$
F	$\begin{bmatrix} 0.252 & 0.227 \\ 0.227 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.150 & 0.103 \\ 0.103 & 0.141 \end{bmatrix}$	$\begin{bmatrix} 0.127 & 0,126 \\ 0.126 & 0.160 \end{bmatrix}$

TABLE 4.2 – Résultats des simulations effectuées avec les méthodes SAMSE-plug-in pour un échantillon de taille n=100.

f	H_{MISE}	F1	F2
A	$\begin{bmatrix} 0.026 & 0 \\ 0 & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.027 & -0.001 \\ -0.001 & 0.088 \end{bmatrix}$	$\begin{bmatrix} 0.026 & -0.002 \\ -0.002 & 0.084 \end{bmatrix}$
B	$\begin{bmatrix} 0.073 & 0 \\ 0 & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.070 & 0.000 \\ 0.000 & 0.042 \end{bmatrix}$	$\begin{bmatrix} 0.070 & 0.002 \\ 0.002 & 0.042 \end{bmatrix}$
C	$\begin{bmatrix} 0.010 & 0 \\ 0 & 0.140 \end{bmatrix}$	$\begin{bmatrix} 0.020 & -0.000 \\ -0.000 & 0.144 \end{bmatrix}$	$\begin{bmatrix} 0.021 & 0.002 \\ 0.002 & 0.140 \end{bmatrix}$
D	$\begin{bmatrix} 0.056 & 0.030 \\ 0.030 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 0.060 & 0.027 \\ 0.027 & 0.063 \end{bmatrix}$	$\begin{bmatrix} 0.050 & -0.030 \\ -0.030 & 0.053 \end{bmatrix}$
E	$\begin{bmatrix} 0.053 & 0.027 \\ 0.027 & 0.072 \end{bmatrix}$	$\begin{bmatrix} 0.054 & 0.025 \\ 0.025 & 0.077 \end{bmatrix}$	$\begin{bmatrix} 0.060 & 0.024 \\ 0.025 & 0.078 \end{bmatrix}$
F	$\begin{bmatrix} 0.011 & 0.099 \\ 0.099 & 0.011 \end{bmatrix}$	$\begin{bmatrix} 0.020 & 0.092 \\ 0.092 & 0.019 \end{bmatrix}$	$\begin{bmatrix} 0.009 & 0.081 \\ 0.081 & 0.010 \end{bmatrix}$

TABLE 4.3 – Résultats des simulations effectuées avec les méthodes AMSE-plug-in pour un échantillon de taille n=1000.

f	H_{MISE}	S1	S2
A	$\begin{bmatrix} 0.026 & 0 \\ 0 & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.027 & -0.001 \\ -0.001 & 0.090 \end{bmatrix}$	$\begin{bmatrix} 0.024 & -0,001 \\ -0.001 & 0.010 \end{bmatrix}$
B	$\begin{bmatrix} 0.073 & 0 \\ 0 & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.089 & 0.002 \\ 0.002 & 0.042 \end{bmatrix}$	$\begin{bmatrix} 0.095 & 0,004 \\ 0.004 & 0.042 \end{bmatrix}$
C	$\begin{bmatrix} 0.010 & 0 \\ 0 & 0.140 \end{bmatrix}$	$\begin{bmatrix} 0.026 & 0.000 \\ 0.000 & 0.077 \end{bmatrix}$	$\begin{bmatrix} 0.026 & 0,003 \\ 0.003 & 0.080 \end{bmatrix}$
D	$\begin{bmatrix} 0.056 & 0.030 \\ 0.030 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 0.050 & 0.017 \\ 0.017 & 0.053 \end{bmatrix}$	$\begin{bmatrix} 0.050 & -0,004 \\ -0.004 & 0.044 \end{bmatrix}$
E	$\begin{bmatrix} 0.053 & 0.027 \\ 0.027 & 0.072 \end{bmatrix}$	$\begin{bmatrix} 0.056 & 0.013 \\ 0.013 & 0.055 \end{bmatrix}$	$\begin{bmatrix} 0.061 & 0,024 \\ 0.024 & 0.072 \end{bmatrix}$
F	$\begin{bmatrix} 0.011 & 0.099 \\ 0.099 & 0.011 \end{bmatrix}$	$\begin{bmatrix} 0.022 & 0.096 \\ 0.096 & 0.023 \end{bmatrix}$	$\begin{bmatrix} 0.001 & 0,100 \\ 0.100 & 0.001 \end{bmatrix}$

TABLE 4.4 – Résultats des simulations effectuées avec les méthodes SAMSE-plug-in pour un échantillon de taille n=1000.

f	H_{MISE}	U	B1	B2
A	$\begin{bmatrix} 0.063 & 0 \\ 0 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.104 & -0.051 \\ -0.051 & 0.667 \end{bmatrix}$	$\begin{bmatrix} 0.031 & -0,010 \\ -0.010 & 0.220 \end{bmatrix}$	$\begin{bmatrix} 0.060 & 0,007 \\ 0.007 & 0.0216 \end{bmatrix}$
B	$\begin{bmatrix} 0.201 & 0 \\ 0. & 0.135 \end{bmatrix}$	$\begin{bmatrix} 0.352 & 0.164 \\ 0.0164 & 0.171 \end{bmatrix}$	$\begin{bmatrix} 0.287 & 0.023 \\ 0.023 & 0.101 \end{bmatrix}$	$\begin{bmatrix} 0.310 & 0.011 \\ 0.011 & 0.102 \end{bmatrix}$
C	$\begin{bmatrix} 0.021 & 0 \\ 0 & 0.335 \end{bmatrix}$	$\begin{bmatrix} 0.010 & -0.004 \\ -0.004 & 0.302 \end{bmatrix}$	$\begin{bmatrix} 0.020 & -0,003 \\ -0.003 & 0.054 \end{bmatrix}$	$\begin{bmatrix} 0.0.480 & 0,017 \\ 0.017 & 0.225 \end{bmatrix}$
D	$\begin{bmatrix} 0.136 & 0.071 \\ 0.071 & 0.136 \end{bmatrix}$	$\begin{bmatrix} 0.170 & 0.106 \\ 0.106 & 0.197 \end{bmatrix}$	$\begin{bmatrix} 0.075 & 0,023 \\ 0.023 & 0.0501 \end{bmatrix}$	$\begin{bmatrix} 0.276 & 0,128 \\ 0.128 & 0.250 \end{bmatrix}$
E	$\begin{bmatrix} 0.140 & 0.073 \\ 0.073 & 0.185 \end{bmatrix}$	$\begin{bmatrix} 0.372 & 0.007 \\ 0.007 & 0.201 \end{bmatrix}$	$\begin{bmatrix} 0.006 & 0,009 \\ 0.009 & 0.273 \end{bmatrix}$	$\begin{bmatrix} 0.287 & 0,133 \\ 0.133 & 0.261 \end{bmatrix}$
F	$\begin{bmatrix} 0.252 & 0.227 \\ 0.227 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.187 & 0.191 \\ 0.191 & 0.271 \end{bmatrix}$	$\begin{bmatrix} 0.192 & 0,179 \\ 0.179 & 0.170 \end{bmatrix}$	$\begin{bmatrix} 0.223 & 0,195 \\ 0.195 & 0.211 \end{bmatrix}$

TABLE 4.5 – Résultats des simulations effectuées avec les méthodes UCV, BCV1 et BCV2 pour un échantillon de taille n=100.

f	H_{MISE}	SC1	SC2
A	$\begin{bmatrix} 0.063 & 0 \\ 0 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.0.066 & -0.002 \\ -0.002 & 0.350 \end{bmatrix}$	$\begin{bmatrix} 0.064 & -0,003 \\ -0.003 & 0.347 \end{bmatrix}$
B	$\begin{bmatrix} 0.201 & 0 \\ 0. & 0.135 \end{bmatrix}$	$\begin{bmatrix} 0.199 & -0.001 \\ -0.001 & 0.140 \end{bmatrix}$	$\begin{bmatrix} 0.330 & -0.028 \\ -0.028 & 0.121 \end{bmatrix}$
C	$\begin{bmatrix} 0.021 & 0 \\ 0 & 0.335 \end{bmatrix}$	$\begin{bmatrix} 0.142 & 0.073 \\ 0.073 & 0.142 \end{bmatrix}$	$\begin{bmatrix} 0.151 & -0,003 \\ -0.003 & 0.401 \end{bmatrix}$
D	$\begin{bmatrix} 0.136 & 0.071 \\ 0.071 & 0.136 \end{bmatrix}$	$\begin{bmatrix} 0.200 & 0.070 \\ 0.070 & 0.200 \end{bmatrix}$	$\begin{bmatrix} 0.229 & -0,007 \\ -0.007 & 0.177 \end{bmatrix}$
E	$\begin{bmatrix} 0.140 & 0.073 \\ 0.073 & 0.185 \end{bmatrix}$	$\begin{bmatrix} 0.273 & 0.090 \\ 0.090 & 0.235 \end{bmatrix}$	$\begin{bmatrix} 0.370 & 0,210 \\ 0.210 & 0.300 \end{bmatrix}$
F	$\begin{bmatrix} 0.252 & 0.227 \\ 0.227 & 0.252 \end{bmatrix}$	$\begin{bmatrix} 0.120 & 0.067 \\ 0.067 & 0.113 \end{bmatrix}$	$\begin{bmatrix} 0.356 & 0,320 \\ 0.320 & 0.350 \end{bmatrix}$

TABLE 4.6 – Résultats des simulations effectuées avec les méthodes SC1 et SC2 pour un échantillon de taille n=100.

f	H_{MISE}	U	B1	B2
A	$\begin{bmatrix} 0.026 & 0 \\ 0 & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.035 & -0.003 \\ -0.003 & 0.102 \end{bmatrix}$	$\begin{bmatrix} 0.027 & 0,001 \\ -0.001 & 0.096 \end{bmatrix}$	$\begin{bmatrix} 0.026 & 0,002 \\ 0.002 & 0.093 \end{bmatrix}$
B	$\begin{bmatrix} 0.073 & 0 \\ 0. & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.075 & 0.002 \\ 0.002 & 0.051 \end{bmatrix}$	$\begin{bmatrix} 0.130 & 0.001 \\ 0.001 & 0.035 \end{bmatrix}$	$\begin{bmatrix} 0.144 & 0.001 \\ 0.001 & 0.0043 \end{bmatrix}$
C	$\begin{bmatrix} 0.010 & 0 \\ 0 & 0.140 \end{bmatrix}$	$\begin{bmatrix} 0.011 & -0.010 \\ -0.010 & 0.144 \end{bmatrix}$	$\begin{bmatrix} 0.011 & -0,001 \\ -0.001 & 0.042 \end{bmatrix}$	$\begin{bmatrix} 0.024 & 0,000 \\ 0.000 & 0.023 \end{bmatrix}$
D	$\begin{bmatrix} 0.056 & 0.030 \\ 0.030 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 0.051 & 0.030 \\ 0.030 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 0.031 & 0,016 \\ 0.016 & 0.033 \end{bmatrix}$	$\begin{bmatrix} 0.161 & 0,016 \\ 0.016 & 0.162 \end{bmatrix}$
E	$\begin{bmatrix} 0.053 & 0.027 \\ 0.027 & 0.072 \end{bmatrix}$	$\begin{bmatrix} 0.032 & 0.030 \\ 0.030 & 0.104 \end{bmatrix}$	$\begin{bmatrix} 0.041 & 0,031 \\ 0.031 & 0.134 \end{bmatrix}$	$\begin{bmatrix} 0.023 & 0,060 \\ 0.060 & 0.014 \end{bmatrix}$
F	$\begin{bmatrix} 0.011 & 0.099 \\ 0.099 & 0.011 \end{bmatrix}$	$\begin{bmatrix} 0.136 & 0.141 \\ 0.141 & 0.152 \end{bmatrix}$	$\begin{bmatrix} 0.100 & 0,094 \\ 0.094 & 0.102 \end{bmatrix}$	$\begin{bmatrix} 0.104 & 0,095 \\ 0.095 & 0.111 \end{bmatrix}$

TABLE 4.7 – Résultats des simulations effectuées avec les méthodes UCV, BCV1 et BCV2 pour un échantillon de taille n=1000.

f	H_{MISE}	SC1	SC2
A	$\begin{bmatrix} 0.026 & 0 \\ 0 & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.029 & 0.001 \\ 0.001 & 0.127 \end{bmatrix}$	$\begin{bmatrix} 0.029 & 0,000 \\ 0.000 & 0.127 \end{bmatrix}$
B	$\begin{bmatrix} 0.073 & 0 \\ 0. & 0.108 \end{bmatrix}$	$\begin{bmatrix} 0.121 & -0.011 \\ -0.011 & 0.062 \end{bmatrix}$	$\begin{bmatrix} 0.113 & 0.002 \\ 0.028 & 0.067 \end{bmatrix}$
C	$\begin{bmatrix} 0.010 & 0 \\ 0 & 0.140 \end{bmatrix}$	$\begin{bmatrix} 0.035 & 0.000 \\ 0.000 & 0.122 \end{bmatrix}$	$\begin{bmatrix} 0.033 & -0,000 \\ -0.000 & 0.141 \end{bmatrix}$
D	$\begin{bmatrix} 0.056 & 0.030 \\ 0.030 & 0.056 \end{bmatrix}$	$\begin{bmatrix} 0.100 & 0.031 \\ 0.031 & 0.101 \end{bmatrix}$	$\begin{bmatrix} 0.102 & 0,010 \\ 0.010 & 0.101 \end{bmatrix}$
E	$\begin{bmatrix} 0.053 & 0.027 \\ 0.027 & 0.72 \end{bmatrix}$	$\begin{bmatrix} 0.104 & 0.036 \\ 0.036 & 0.101 \end{bmatrix}$	$\begin{bmatrix} 0.102 & 0,033 \\ 0.033 & 0.083 \end{bmatrix}$
F	$\begin{bmatrix} 0.011 & 0.099 \\ 0.099 & 0.011 \end{bmatrix}$	$\begin{bmatrix} 0.106 & 0.100 \\ 0.100 & 0.105 \end{bmatrix}$	$\begin{bmatrix} 0.124 & 0,114 \\ 0.114 & 0.115 \end{bmatrix}$

TABLE 4.8 – Résultats des simulations effectuées avec les méthodes SC1 et SC2 pour un échantillon de taille n=1000.

f	$MISE(H_{MISE})$	F1	Crit	F2	Crit
A	0.096	0.104	0.92	0.106	0.91
B	0.008	0.014	0.58	0.015	0.52
C	0.06	0.17	0.35	0.19	0.32
D	0.013	0.028	0.46	0.031	0.42
E	0.014	0.035	0.42	0.036	0.39
F	0.087	0.097	0.90	0.098	0.89

TABLE 4.9 – Le MISE(Hopt) pour les méthodes AMSE plug-in pour un échantillon n=100.

f	$MISE(H_{MISE})$	F1	Crit	F2	Crit
A	0.071	0.0755	0.94	0.0736	0.93
B	0.0034	0.0055	0.62	0.0058	0.59
C	0.042	0.072	0.58	0.076	0.55
D	0.011	0.0166	0.66	0.0204	0.54
E	0.009	0.013	0.68	0.015	0.61
F	0.066	0.0717	0.92	0.0725	0.91

TABLE 4.10 – Le $MISE(H_{opt})$ pour les méthodes AMSE plug-in pour un échantillon n=1000.

f	$MISE(H_{MISE})$	S1	Crit	S2	Crit
A	0.096	0.107	0.90	0.107	0.90
B	0.008	0.0133	0.60	0.0126	0.63
C	0.06	0.18	0.33	0.19	0.32
D	0.013	0.024	0.54	0.027	0.48
E	0.014	0.032	0.44	0.034	0.41
F	0.087	0.099	0.88	0.101	0.86

TABLE 4.11 – Le $MISE(H_{opt})$ pour les méthodes SAMSE plug-in pour un échantillon n=100.

f	$MISE(H_{MISE})$	S1	Crit	S2	Crit
A	0.071	0.0739	0.96	0.0771	0.92
B	0.0034	0.0046	0.74	0.0048	0.71
C	0.0042	0.0763	0.55	0.0750	0.56
D	0.011	0.0189	0.58	0.0200	0.55
E	0.009	0.0145	0.62	0.0155	0.58
F	0.066	0.0709	0.93	0.0725	0.91

TABLE 4.12 – Le $MISE(H_{opt})$ pour les méthodes SAMSE plug-in pour un échantillon n=1000.

f	$MISE(H_{MISE})$	U	Crit	B1	Crit	B2	Crit
A	0.096	0.110	0.87	0.108	0.89	0.105	0.91
B	0.008	0.014	0.58	0.013	0.60	0.014	0.59
C	0.060	0.182	0.33	0.194	0.31	0.200	0.30
D	0.013	0.0216	0.60	0.0220	0.59	0.0220	0.59
E	0.014	0.042	0.33	0.035	0.40	0.037	0.38
F	0.087	0.101	0.86	0.099	0.88	0.099	0.88

TABLE 4.13 – Le $MISE(H_{opt})$ pour les méthodes UCV, BCV1 et BCV2 pour un échantillon n=100.

f	$MISE(H_{MISE})$	U	Crit	B1	Crit	B2	Crit
A	0.071	0.0788	0.90	0.0788	0.90	0.0763	0.93
B	0.0034	0.0056	0.61	0.0056	0.61	0.0057	0.60
C	0.042	0.113	0.37	0.116	0.36	0.127	0.33
D	0.011	0.0172	0.64	0.0177	0.62	0.0180	0.61
E	0.009	0.0264	0.34	0.0209	0.43	0.0214	0.42
F	0.066	0.0742	0.89	0.0733	0.90	0.0725	0.91

TABLE 4.14 – Le $MISE(H_{opt})$ pour les méthodes UCV, BCV1 et BCV2 pour un échantillon n=1000.

f	$MISE(H_{MISE})$	SC1	Crit	SC2	Crit
A	0.096	0.109	0.88	0.109	0.88
B	0.008	0.0123	0.65	0.0127	0.63
C	0.06	0.15	0.41	0.12	0.49
D	0.013	0.031	0.42	0.030	0.43
E	0.014	0.033	0.43	0.036	0.39
F	0.087	0.099	0.88	0.102	0.85

TABLE 4.15 – Le $MISE(H_{opt})$ pour les méthodes SC1 et SC2 pour un échantillon n=100.

f	$MISE(H_{MISE})$	SC1	Crit	SC2	Crit
A	0.071	0.0797	0.89	0.0788	0.90
B	0.0034	0.005	0.68	0.0053	0.64
C	0.042	0.089	0.47	0.082	0.51
D	0.011	0.025	0.44	0.025	0.44
E	0.009	0.0188	0.48	0.0214	0.42
F	0.066	0.0733	0.90	0.0759	0.87

TABLE 4.16 – Le $MISE(H_{opt})$ pour les méthodes SC1 et SC2 pour un échantillon n=1000.

Conclusion générale et perspectives

Ce mémoire est une introduction à l'estimation fonctionnelle dans le cas multidimensionnel, nous nous sommes intéressé tout particulièrement à l'estimation non-paramétrique de la densité de probabilité multivariée. Dans une première étape, nous avons fixé le cadre général de l'estimation fonctionnelle. Ceci est fait à travers la présentation d'un certain nombre de résultats classiques bien connus des spécialistes. Dans la deuxième partie, nous avons introduit l'estimateur à noyau de la densité de probabilité multidimensionnelle, cet estimateur est fonction de deux paramètres : une fonction multidimensionnelle K appelée noyau multidimensionnel et une matrice H symétrique définie positive appelée matrice des paramètres de lissage. La matrice H contrôle le degré de lissage et l'orientation par rapport aux axes des coordonnées des contours de la densité cible à estimer. Ainsi, la construction du noyau multidimensionnel K à partir d'un noyau unidimensionnel symétrique k et les différentes formes de la paramétrisation ont été étudiées. Quand au troisième chapitre, il a été consacré au problème du choix de la matrice de lissage optimale. Dans ce chapitre, nous avons exposé les différentes méthodes de sélection de la matrice de lissage H : les méthode plug-in (ré-injection) qui repose sur l'estimation de la matrice inconnue Ψ_4, et l'autre classe de méthodes dite validation croisée (cross-validation). Enfin, afin de tester l'efficacité des différentes méthodes de sélection de la matrice de lissage, nous avons simulé des densités de probabilité tests bivariées présentant différents aspects. Nous avons estimé la matrice de lissage H_{opt} par les différentes méthodes de sélection, calculer le H_{MISE}, construit l'estimateur de la densité de probabilité par la méthode du noyau et comparé l'efficacité des différentes méthodes de sélection en calculant le critère d'efficacité moyenne. Les résultats numériques obtenus indiquent que : les différentes méthodes plug-in et les différentes méthodes de validation croisée donnent de très bons résultats pour des densités simples (unimodales). Cependant, les performances des méthodes s'amoindrissent en présence de cibles plus complexes (multimodales). La simulation montre aussi qu'il n'existe pas de méthode de sélection de la matrice de lissage meilleure que toutes les autres.

Le travail effectué dans le cadre de ce mémoire nous a permis de dégager plusieurs perspectives de recherches très intéressantes. En effet :

1. Ce travail pourrait donner lieu à des prolongements sur le choix de la matrice de lissage optimale qui reste d'actualité :

   • Il serait intéressant dans un travail futur d'effectuer des simulations sur d'autres formes de densités cibles et dans des dimensions plus élevées de l'espace.

   • Il est intéressant aussi de développer d'autres résultats sur l'estimation pilote. Nous avons vue que **Wand** et **Jones** ([98],1994), ainsi que **Duong** et **Hazelton** dans ([30],2003) et ([32],2005b)ont adopté pour la matrice pilote de lissage G une paramétrisation de la forme : $G = gI_d$; $g > 0$. Cette paramétrisation est beaucoup plus restrictive et afin d'éviter les effets de cette restriction, ces auteurs ont proposé de faire des pré-transformations de données avant chaque estimation pilote. **Chacón** et **Duong** ([19], 2010) ont montré que ces pré-transformations de données ne sont pas adaptés pour toutes les formes de densités. Ainsi, l'idéal est de développer une méthode avec une paramétrisation pilote complète sans restriction, i.e : prendre G comme une matrice symétrique définie positive quelconque.

2. Appliquer d'autres méthodes comme : l'estimation par projections orthogonales et l'estimateur spline, et explorer d'autres approches récentes comme l'approche bayésienne et l'estimateur à base d'ondellettes.

3. Essayer de développer d'autres résultats théoriques et pratiques en utilisant d'autres critères.

Bibliographie

[1] B. Abdous and R. Berlinet, Pointwise improvement of multivariate kernel density estimates, Journal of Multivariate Analysis 65, 109-128, 1998.

[2] I. S. Abramson, On bandwidth variation in kernel estimates-a square root law, The Annals of Statistics 10, 1217-1223, 1982

[3] H. Akaike, An approximation to the density function, Ann. Inst. Statist. Math. 6, 127-132, 1954.

[4] M. J. Baxter, C. C. Beardah et S. Westwood, Sample size and related issues in the analysis of lead isotope data, J. Archaeological Science, 27, 973-980, 2000.

[5] J. Bleuez et D. Bosq, Conditions nécessaires et suffisantes de convergence pour une classe d'estimateurs de la densité. C. R. Acad. Sci. Paris A, 282, 63-66, 1976.

[6] J. Bleuez et D. Bosq, Etude d'une classe d'estimateurs non paramétriques de la densité, Ann. Inst. H.Poincaré, 14, 479-498, 1978.

[7] D. Bosq, Sur l'estimation d'un paramètre à valeurs dans un espace de Hilbert. Publ. int. univ. Lille I, n° 70, 1976.

[8] D. Bosq, Sur l'estimation sans biais d'un paramètre fonctionnel, Publ.Int. U.E.R. Math. Lille I n° 104, 19 p, 1977a.

[9] D. Bosq, Sur l'estimation sans biais d'un paramètre à valeurs dans $L^2(\mu)$ et sur le choix d'un estimateur de la densité, C.R. Acad. Sci. Paris A, 284, 85-88, 1977b.

[10] D. Bosq et J. P. Lecoutre, Théorie de l'estimation fonctionnelle, Economica, Paris, 1987.

[11] T. Bouezmarni and J. Rombouts, Nonparametric Density Estimation For Multivariate Bounded Data, ECORE discussion paper, 85, 2007.

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

[13] A. W. Bowman and A. Azzalini, Applied Smoothing Techniques for Data Analysis, Oxford University Press, Oxford, 1997.

[14] L. Breiman, W. Meisel and E. Purcell, Variable Kernel estimates of probability density estimates, Technometrics 19, 135-144, 1977.

[15] A. Bugeau and P. Pérez, Bandwidth selection for kernel estimation in mixed multi-dimensional spaces, Technical report, INRIA, Rennes, RR-6282, 2007.

[16] T. Cacoullos, Estimation of a multivariate density, Annals of the institute of statistical Mathematics 18, 179-189, 1966.

[17] R. Cao, A. Cuevas and W. G. Manteiga, A comparative study of several smoothing methods in density estimation, Computational Statistics and Data Analysis 17, 153-176, 1994.

[18] N. N. Cencov, Evaluation of an unknown distribution density from observations, Soviet. Math. 3, 1559-1562, 1962.

[19] J. E. Chacón and T. Duong, Multivariate plug-in bandwidth selection with unconstrained pilot bandwidth matrices, Test, 19, 375-398, 2010.

[20] M. Cheng, J. Fan and J. Marron, On Automatic Boundary Corrections, Annals of Statistics, 25, 1991-1708, 1997.

[21] S. T. Chiu, Bandwidth selection for kernel density estimation, The Annals of Statistics 19, 1983-1905, 1991.

[22] S. T. Chiu, A comparative review of bandwidth selection for kernel density estimation, Statistica Sinica 6, 126-145, 1996.

[23] J. Cwik and J. Kronacki, A combined adaptive-mixtures/plug-in estimator of multivariate probability densities, Computational Statistics and Data Analysis 26, 199-218, 1997a.

[24] J. Cwik and J. Kronacki, Multivariate density estimation : A comparative study, Neural Computing and Applications 6, 173-185, 1997b.

[25] P. Deheuvels, estimation non paramétrique de la densité par histogrammes généralisés, Rev. Statist. Appl. 25, 5-42, 1977a.

[26] P. Deheuvels, estimation non paramétrique de la densité par histogrammes généralisés (II), Publications de l'institut de Statistique de l'Université de Paris 22, 1-23, 1977b.

[27] l. Devroy and L. Györfi, Nonparametric Density Estimation : the L_1 View, John Wiley and Sons, New York, 1985 .

[28] J. Dinardo, N. Fortin and T. Lemieux, Labor market institutions and the distribution of wages, 1973-1992 : A semiparametric approach. Econometrica, 64, 1001-1044, 1996.

[29] R. C. Duda and P. E. Hart, Pattern Classification and Scene Analysis, John Wiley and Sons, New York, 1973.

[30] T. Duong et M. L. Hazelton, Plug-in bandwidth matrices for bivariate kernel density estimation, Journal of Nonparametric Statistics, 15, 17-30, 2003.

[31] T. Duong et M. L. Hazelton, Convergence rates for unconstrained bandwidth matrix selectors in multivariate kernel density estimation, Journal of Multivariate Analysis, 93, 417-433, 2005a.

[32] T. Duong et M. L. Hazelton, Cross-validation bandwidth matrices for multivariate kernel density estimation, Scand. J. Stat., 32, 485-506, 2005b.

[33] V. A. Epanechnikov, Non-parametric estimation of a multivariate probability density, Theory of Probability and its Applications 14, 153-158, 1969.

[34] B. S. Everitt, Clusters analysis, 3rd edn, Edward Arnold, london, 1993.

[35] J. J. Faraway and M. Jhun, Bootstrap choice of bandwidth for density estimation, journal of the American Statistical Association 85, 1119-1122, 1900.

[36] R. A. Ferreyra, G. P. Podesta, C. D. Messina, D. Letson, J. Dardanelli, E. Guevara and S. Meira A linked-modeling framework to estimate maize production risk associated with ENSO-related climate variability in Argentina, Agricultural and Forest Meteorology, 107, 177-192, 2001.

[37] E. Fix and J. L. Hodges, Discriminatory analysis - non-parametric discrimination : consistency properties, Report N°. 4, Project no. 21-29-004, USAF School of Aviation Medicine, Randolph Field, Texas, 1951.

[38] A. Földes and P. Révész, A general method for density estimation, Studia Sci. Math. Hungar. 9, 82-92, 1974.

[39] P. Foster, A comparative study of some bias correction techniques for kernel-based density estimators, Journal of Statistical Computation and Simulation 51, 137-152, 1995.

[40] K. Fukunaga, Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972.

[41] A. D. Gordon, Classification, 2nd edn, Chapman and Hall/CRC, London, 1999.

[42] B. Grund, P. Hall and J. S. Marron, Loss and risk in smoothing parameter selection, Journal of nonparametic Statistics 4, 107-132, 1994.

[43] P. Hall and J. S. Marron Extent to which least-squares croos-validation minimises integrated square error in nonparametric density estimation, Probability Theory and Related Fields 74, 567-581, 1987.

[44] P. Hall and J. S. Marron Lower bounds for bandwidth selection in density estimation, Probability Theory and Related Fields 90, 149-173, 1991.

[45] P. Hall, J. S. Marron and B. U. Park Smoothed cross-validation, Probability Theory and Related Fields 92, 1-20, 1992.

[46] P. Hall, S. J. Sheather, M. C. Jones and J. S. Marron, on optimal data-based bandwidth selection in kernel density estimation, Biometrika 78, 263-269, 1991.

[47] P. Hall and M. P. Wand, On nonparametric discrimination using density differences, Biometrika 75, 541-547,1988.

[48] D. J. Hand, Kernel discriminant analysis, vol.2 of Electronic and Electrical Engineering Research Studies : Pattern Recognition and Image Processing Series, Research Studies Press 5John Wiley and Sons], Chichester, 1982.

[49] M. L. Hazelton, Bandwidth selection for local density estimators, Scandinavian Journal of Statistics. Theory anq Applications 23, 221-232, 1996.

[50] M. L. Hazelton, An optimal local Bandwidth selection for kernel density estimation, Journal of Statistical Planning and Inference 77, 37-50, 1999.

[51] D. V. Hinkley, On the ratio of two correlated normal random variables, Biometrika 56, 635-639, 1969.

[52] M. L. Hazelton and M. C. Jones, Variable kernel density estimates and variable kernal density estimates, The Australian Journal of Statistics 32, 361-371, 1990.

[53] H. V. Henderson and S. R. Searle, Vec and vech operators for matrices, with some uses in Jacobians and multivariate statistics, Canad. J. Statist., 7, 65-81, 1979.

[54] M. C. Jones, The roles of ISE and MISE in density estimation, Statstics and Probability Letters 12, 51-56, 1991.

[55] M. C. Jones, Potential for automatic bandwidth choice in variations on kernel density estimation, Statistics and Probability letters 13, 351-356, 1992.

[56] M. C. Jones and R. F. Kappenman, On a class of Kernel density estimate bandwidth selectors, Scandinavian Journal of Statistics. Theory and Applications 19, 337-349, 1992.

[57] M. C. Jones, J. S. Marron and B. U. Park, A simple root n bandwidth selector, The Annals of Statistics 19, 1919-1932, 1991.

[58] M. C. Jones, J. S. marron and S. J. Sheather, A brief survey of bandwidth selection for debsity estimation, Journal of the American Statistical Association, 91, 401-407, 1996.

[59] K. D. Kim and J. H. Heo, Comparative study of flood quantiles estimation by nonparametric models, J. Hydrology, 260, 176-193, 2002.

[60] C. R. Loader, Bandwidth selection : classical or plug-in ?, The Annals of Statistics 27, 415-438,1999.

[61] D. O. Loftsgaarden and C. P. Quesenberry, A nonparametric estimate of a multivariate density function, Annals of Mathematical Statistics 36, 1049-1051, 1965.

[62] J. R. Magnus and H. Neudecker, Matrix Differential Calculus with Applications in Statistics and Econometrics, John Wiley and Sons Ltd., Chichester, 1988.

[63] G. M. Maniya, Remarks on nonparametric estimates of a bivariate probability density, Soobshch. Akad. Nauk Gruz. SSR, 27, 4, pp. 385-400, 1961.

[64] D. J. Marchette, C. E. Priebe, G. W. Rogers and J. L. Solka Filtered kernel density estimation, Computational Statistics 11, 95-112, 1996.

[65] J. S. Marron and A. B. Tsybakov, Visual error criteria for qualitative smoothing, Journal of American Statistical Association 90, 499-507, 1995.

[66] H. G. Müller, Nonparametric Regression Analysis of longitudinal Data, Springer-Verlag, Berlin, 1988.

[67] H. G. Müller, Smooth Optimum Kernel Estimators near Endpoints, Biometrika, 78, 521-530, 1991.

[68] E. A. Nadaraya, Estimation of a bivariate probability density, Soobshch. Akad. Nauk Gruz. SSR, 36, 2, pp. 267-268, 1964.

[69] B. U. Park and J. S. Marron, Comparaison of data-driven bandwidth selectors, Journal of the American Statistical Society 85, 66-72, 1990.

[70] E. Parzen, On the estimation of a probability density function and the mode. Ann. Math. Statist. 33,1065-1076, 1962.

[71] O. Paulsen and P. Heggelund, Quantal properties of spontaneous EPSCs in neurones of the Guinea-pig dorsal lateral geniculate nucleus, J. Physiology, 496, 759-772, 1996.

[72] M. Rosenblatt, Remarks on some nonparametric estimates of a density function, The Annals of Mathematical Statistics 27, 832-837, 1956.

[73] M. Rudemo, Empirical choise of histograms and kernel density estimators, Scandinavian Journal of statistics. Theory and applications 9, 65-78, 1982.

[74] N. Saadi and S. Adjabi, On the estimation of the probability density by trigonometric series. Communications in statistics- Theory and methodes, 38 : 19, 3583-3595, 2009.

[75] S. R. Sain, Multivariate loccally adaptive density estimation, computational Statistics and Data Analysis 39, 165-186, 2002.

[76] 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.

[77] M. G. Schimek, Smoothing and Regression, John Wiley and Sons Inc., New York, 2000.

[78] E. Schuster, Incorporating Support Constraints into Nonparametric Estimators of Densities, Communications in Statistics - Theory and Methods, 14, 1123-1136, 1985.

[79] D. W. Scott, Multivariate Density Estimation : Theory, Practice and Visualization, John Wiley and Sons Inc., New York, 1992.

[80] D. W. Scott and G. R. Terrell, Biased and unbiased cross-validation in density estimation, Journal of the American Statistical Association 82, 1131-1146, 1987.

[81] D. W. Scott and M. P. Wand, Feasibility of multivariate density estimates, Biometrika 78, 197-206, 1991.

[82] G. A. F. Seber, Linear Regression Analysis, Wiley, New York, 1977.

[83] M. R. Segal and J. L. Wiemels, Clustering of translocation breakpoints, J. Amer. Statist. Assos., 97, 66-76, 2002.

[84] S. J. Sheather and M. C. Jones, A reliable data-based bandwidth selection method for kernel density estimation, Journal of the Royal statistical Society. Series B. Methodological 53, 683-690, 1991.

[85] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, 1986.

[86] J. S. Simonoff, Smoothing Methods in Statistics, Springer-Verlag, New York, 1996.

[87] R. S. Singh, Nonparametric estimation of mixed partial derivatives of a multivariate density. J. Multivariate Anal. 6, 111-122, 1976.

[88] C. J. Stone, An asymptotically optimal window selection rule for kernel density estimates, The Annals of Statistics 12, 1285-1297, 1984.

[89] M. E. Tarter and M. D. lock, Model-Free Curve Estimation, Chapman and Hall, London, 1993.

[90] G. R. Terrell, The maximal smoothing principle in density estimation, Journal of the American Statistical association 85, 470-477, 1990.

[91] G. R. Terrell and D. W. Scott, Variable kernel density estimation, the Annals of statistics 20, 1236-1265, 1992.

[92] J. Tiago De Oliveira, Estatistica de densidades : resultados assintoticos. Rev. Fac. Ci. Univ. lisboa A9, 65-171, 1963.

[93] E. Tortosa-Ausina, Financial costs, operating costs, and specialization of Spanish banking firms as distribution dynamics, Applied Economics, 34, 2165-2176, 2002.

[94] A. B. Tsybakov, Introduction à l'estimation non-paramétrique, Springer-Verlag, New York, 2004.

[95] B. Turlach, Bandwidth selection in kernel density estimation : a review, Discussion paper 9317. Institut de statistique, Voie du Roman Pays, B-1348, Louvain-la-Neuve,1993.

[96] M. P. Wand, Error analysis for general multivariate kernel estimators, Journal of Nonparametric Statistics 2, 1-15, 1992a.

[97] M. P. Wand and M. C. Jones, Comparison of smoothing parametrizations in bivariate kernel density estimation. J. Amer. Statist. Assoc. 88, 520-528, 1993.

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

[99] M. P. Wand and M. C. Jones, Kernel smoothing, Chapman and Hall Ltd., London, 1995.

[100] Xibin Zhang, M. L. King et R. J. Hyndman, A Bayesian approach to bandwidth selection for multivariate kernel density estimation, Computational Statistics and Data Analysis, Elsevier, Vol. 50(11), p. 3009-3031, Jul, 2006.