RÉPUBLIQUE ALGÉRIENNE DÉMOCRATIQUE ET POPULAIRE
Ministère de L'enseignement Supérieur et de la Recherche Scientifique
UNIVERSITÉ IBN KHALDOUN TIARET
FACULTÉ DE MATHÉMATIQUES ET DE L'INFORMATIQUES
Département de Mathématiques

MÉMOIRE DE MASTER

Spécialité :
« Mathématiques »
Option :
«Analyse fonctionnelle et »
équations différentielles

Présenté Par :

BOUMAZA Meriem & CHADLI Ikram

Intitulé :

Estimation par noyaux de la fonction de densité dans le cas indépendantes et identiquement distribuées

Soutenu publiquement le 15 / 06 / 2023
à Tiaret devant le jury composé de :

Mr. MAAZOUZ Kada	MCA	U. Ibn Khaldoun Tiaret	Président
Mr. BENALLOU Mohamed	MCB	U. Ibn Khaldoun Tiaret	Encadreur
Mr. REZZOUG Nadir	MCB	U. Ibn Khaldoun Tiaret	Examinateur

Année universitaire :2022/2023

Dédicace

Je dédie cet humble travail

A mes chers parents qui toujours été présent pour moi , et qui m'ont donné la vie, Pour de millions des raisons, ils m'ont donné jour après jour autant d'amour et de confiance, ils ont veillé à m'encourager tout au long de ma vie, à me donner de l'aide et à protéger . Que le Dieu me les protège.

A mes chers frères et sœurs, ma source de joie et de bonheur

A tous mes amis.

A tous mes proches.

A tous mes professeurs et tous ceux qui ont contribué à mon éducation.

A mes camarades de promotion 2022/2023.

A tous ceux qui m'aiment.

A tous ceux que j'aime.

Meriem

Dédicace

Je dédie ce mémoire

A mes chers parents ma mère et mon père, pour leur patience, leur amour, leur soutien et leur encouragement .

Mon chère père qui nous a quitté voilà 6 ans dieu accorde la paix à son âme et grâce à lui que je suis là

Ma mère qui m'a encourager à avancer et qui m'a donner tout son amour pour mes études .

A mes sœurs et mes frères .

A mes amis et mes camarades.

Sans oublier tout mes professeurs que ce soit du primaire ,moyen , secondaire ou de l'enseignement supérieur

Une spécial dédicace à une personne qui a été paternaliste avec moi : **jamila**

Ikram

Remerciements

Au nom de ALLAH Clément et Miséricordieux.

En premier lieu, nous remercions ALLAH, le tout puissant et miséricordieux, qui nous a donné la force, le courage et la patience d'accomplir ce modeste travail

Nos remerciements et nos profondes gratitudes vont à notre Encadreur Monsieur BENALLOU MOHAMED pour ses précieux conseils et pour tout le soutien et l'orientation. D'avoir bien voulu diriger notre travail, d'avoir donner le meilleur de son savoir, de son aide, surtout d'avoir fait preuve de beaucoup de patience, son aide durant toute la période du travail.

Nous tenons aussi à remercier les membres du jury pour leur précieux temps accordé à l'étude de notre mémoire.

Nous remercions nos enseignants pour leurs efforts , nos parents et nos proches pour l'amour et le soutien contant qu'ils nous ont témoigné tout au long de notre parcours. Merci à toutes et touts nos amis pour leurs encouragements.

Résumé

Dans ce mémoire, nous proposons d'étudier l'estimateurs de la densité de probabilité, dans ce cadre, nous commençons par rappeler d'abord les notions essentielles, à savoir, les variables aléatoires, rappel des notations de base de la statistique mathématique comme : l'échantillon, l'estimateur et leurs propriétés, ensuite, nous étudions les deux types d'estimation paramétrique et non-paramétrique et quelques théorèmes et types de convergences. Ensuite on va présenter l'approches non paramétriques : approche basée sur l'histogramme qui était introduit par **John Graunt en 1962**, et approche basée sur le noyau introduit par **Rosenblatt en 1956**, puis amélioré par **Parzen en 1962**. Avec la notion de mesure de risque on étudie le comportement des estimations appliquées aux densités usuelles et faire une comparison entre les deux approches.

Le but de ce travail est de montrer l'efficacité de la méthode du noyau pour estimer la densité.

Le long de ce mémoire les variables sont indépendantes et identiquement distribuées (iid).

Table des matières

Table des figures

Introduction

0.1 Historique

L'estimation d'une densité de probabilité sous-jacente à un ensemble fini d'observations est un problème fondamental en statistique qui a fait l'objet d'une vaste littérature. On retrouve cette problèmatique dans de nombreux domaines des sciences et techniques tels que le traitement du signal et des images, la mécanique, la robotique, etc.

La modélisation de phénomènes aléatoires est traditionnellement réalisée par une distribution de probabilité qui est généralement inconnue. On doit donc l'identifier à partir d'un nombre fini d'observations, identification qui consiste généralement à estimer la densité de probabilité en tout point de l'espace considéré. La tache de l'estimation de densité revient à produire, à partir d'un ensemble de données, un modèle probabiliste. On trouve deux types d'approches d'estimation de la densité de probabilité : estimation paramétrique et non paramétrique. L'approche paramétrique consiste à supposer que la densité de probabilité f appartient à une famille de densités qui peuvent être décrites par un petit nombre (connu) de paramètres réels. Le statisticien qui opte pour une telle approche possède une bonne connaissance à priori du phénomène aléatoire. Il sait, par intuition ou par expérience, que la variable aléatoire X suit une loi f, tout en ignorant la valeur de son espérance ou de sa variance. Dans un tel contexte, l'estimation de la densité se réduit alors à un problème d'estimation de paramètres. Alors l'approche paramétrique a comme inconvénient principal la connaissance au préalable de la loi du phénomène étudié. Différents méthodes existent pour l'estimation notamment : la méthode du maximum de vraisemblance et la méthode des moments. Pour pallier les insuffisances et les défauts des familles paramétriques, une seconde approche dite non paramétrique propose de laisser parler les données, sans spécifier au préalable de forme sur f, ainsi cette approche estime la densité à partir de l'information disponible et ne pas nécessiter d'hypothèses à priori sur l'appartenance de cette densité à une famille de lois connues. Si la qualité de l'estimation paramétrique est fortement liée à la validité de l'hypothèse faite sur la loi de probabilité, celle de l'approche non paramétrique dépend du nombre d'observations et de certains paramètres (noyau, bande lissage, etc.).

L'outil d'estimation non paramétrique nous est fourni par l'histogramme : une fois les données regroupées en classes de valeurs, les fréquences empiriques sont représentées par des aires rectangulaires dont les bases correspondent aux classes elles mêmes. L'histogramme convient bien pour des analyses relativement grossières, un problème vient du fait que l'histogramme donne une fonction qui n'est pas continue.

Il existe d'autre méthodes non paramétriques plus éfficace que la méthode du l'histogramme

est celle la méthode du noyau consiste à retrouver la continuité et qui est la plus utilisée vu sa simplicité et la qualité de l'estimation qu'elle assure. **Rosenblatt (1956**), et **Parzen (1962)** sont les premiers à proposer un estimateur à noyau d'une densité univariée. Cet estimateur est basé sur un échantillon d'une population statistique et permet d'estimer la densité en tout point du support.

Dans ce mémoire nous étudions l'estimation de la densité de probabilité à partir d'un échantillon indépendant et identiquement distribué **(i.i.d).** Nous nous intéressons aux approches non paramétriques : approche basée sur l'histogramme et approche basée sur le noyau, en premier lieu la méthode d'estimation par histogramme, puis la méthode d'estimation par noyau, ainsi que leurs propriétés asymptotiques. Le principe de base de la méthode des noyaux s'apparente relativement bien à la notion d'histogramme couramment utilisé pour l'analyse exploratoire d'un échantillon. L'histogramme donne une idée de la forme de la distribution empirique d'un échantillon en calculant la proportion d'observations se trouvant dans chacun des intervalles de largeur h. Le choix de la largeur h des fenêtres de l'histogramme est déterminant. Le concept de fenêtre de l'histogramme est aussi présent dans la méthode des noyaux. On utilise alors le terme du "paramètre de lissage" pour désigner la fenêtre h. Comme dans le cas de l'histogramme, l'estimation de la fonction de densité par la méthode des noyaux est principalement conditionnée par le paramètre de lissage.

Pour estimer f il faut choisir le noyau K et le paramètre h ; si le choix du noyau n'est pas un problème, il n'est pas de même pour le choix de la largeur de la fenêtre h qui dépend essentiellement de la taille n de l'échantillon. En effet, dans cette méthode d'estimation se pose acuité le problème du choix du paramètre de lissage, il existe beaucoup de méthodes parmi lesquels : méthode de validation croisées. Cet estimateur est utilisé dans l'estimation des fonctions de régression, de quantiles et de densité conditionnelle.

0.2 Organisation du mémoire

Ce mémoire s'organise en trois chapitres. Les deux premiers chapitres sont principales et lues l'une aprés l'autre.

Le premier chapitre, dont l'objectif est de présenter des notions de base et des définitions nécéssaires, à savoir, les variables aléatoires, on rappels leurs définitions, ses moments, qui sont des grandeurs nécessaires pour l'analyse statistique des données, les lois de probabilité ect...,

Au seconde chapitre on fait le point sur l'estimation en générale, nous commençons par rappeler d'abord les notions essentielles d'estimation. Nous examinons par la suite les propriétés des estimateurs plus précisément le biais, la variance et les erreurs quadratiques moyennes, puis on va voir la convergence avec les différents types et le taux de convergence, on termine par la présentation des deux types d'approches, à savoir, l'approche paramétrique et l'approche non paramétrique avec quelques méthodes d'estimation .

Le dernier chapitre est consacré à l'estimation non paramétrique de la densité de probabilité, aprés la présentation de l'erreur quadratique moyenne intégrée ou le mesure de risque on va commençer par la méthode de l'histogramme, on rappelle ces propriétés fondamentales et

aborder le problème de choix optimal de paramètre de lissage (la valeur qui minimise le risque intégré), puis la méthode du noyau qui est la méthode la plus connue, aprés la restrinction de la méthode précédente on va montrer que les estimateurs à noyau sont optimaux en termes de taux de convergence.

Chapitre 1

Variables aléatoires et Lois statistiques

Ce chapitre permet de reprendre certaines notions de base des variables aléatoires, et leurs propriétés.

1.1 Notations et généralités

1.1.1 Échantillon

En statistique, un échantillon est un ensemble d'individus extraits d'une population étudiée de manière à ce qu'il soit représentatif de cette population, au moins pour l'objet de l'étude. Pour ce faire, on peut le tirer de façon aléatoire, par un ensemble de méthodes mathématiquement très contraignantes, ou quand ces méthodes se révèlent impossibles à appliquer, par des méthodes pratiques comme la méthode des quotas.

- En traitement des signaux, on parle alors d'échantillonnage de signal.
- Chez les boulangers, l'échantillon accompagné de sa souche, était la baguette de bois qui permettait de comptabiliser le nombre de pains livrés au client mais non encore payés.

Nous allons voir que si une variable aléatoire suit une certaine loi, alors ses réalisations (sous forme d'échantillons) sont encadrées avec des probabilités de réalisation. Par exemple, lorsque l'on a une énorme urne avec une proportion p de boules blanches alors le nombre de boules blanches tirées sur un échan-tillon de taille n est parfaitement défini. En pratique, la fréquence observée varie autour de p avec des probabilités fortes autour de p et plus faibles lorsqu'on s'éloigne de p.

Nous allons chercher à faire l'inverse : **l'inférence statistique** consiste à induire les caractéristiques inconnues d'une population à partir d'un échantillon issu de cette population. Les caractéristiques de l'échantillon, une fois connues, reflètent avec une certaine marge d'erreur possible celles de la population.

1.1.2 Espace probabilisé

Une expérience est appellé "aléatoire" s'il est impossible de prévoir à l'avance son résultat et si elle est répétée dans des conditions identiques. On appelle ensemble associé à une expérience aléatoire l'énsemble fondamentale Ω ={tous résultat possibles de cette expérience}. Un espace de probabilité ou espace probabilisé est la donnée d'une probabilité à tout événement, il permet la modélisation quantitative de l'expérience aléatoire étudiée. Formellement, c'est un triplet $(\Omega, F, \mathbb{P})$, F est un ensemble des événements ou tribu sur Ω, et $\mathbb{P}$ est une probabilité sur F, comme définie ci-dessous,(pour une étude plus détaillée voir,par exemple, Dusart [4] et Veysseyre[17].

1.1.3 Variable aléatoire

Une variable aléatoire est une fonction définie sur l'ensemble des éventualités, c'est-à-dire l'ensemble des résultats possibles d'une expérience aléatoire.

Une variable aléatoire est souvent à valeurs réelles (gain d'un joueur dans un jeu de hasard, durée de vie) et on parle alors de variable aléatoire réelle : $X : \Omega \to X(\Omega) \in \mathbb{R}$. La variable aléatoire peut aussi associer à chaque éventualité un vecteur de $\mathbb{R}^n$ ou $\mathbb{C}^n$, et on parle alors de vecteur aléatoire : $X : \Omega \to X(\Omega) \in \mathbb{R}^n$ ou $X : \Omega \to X(\Omega) \in \mathbb{C}^n$. La variable aléatoire peut encore associer à chaque éventualité une valeur qualitative (couleurs, Pile ou Face), ou même une fonction (p.e. une fonction de $C(\mathbb{R}_+, \mathbb{R}^d)$, et on parlera alors de processus stochastique.

Ce furent les jeux de hasard qui amenèrent à concevoir les variables aléatoires, en associant à une éventualité (résultat du lancer d'un dé, d'un tirage à pile ou face, d'une roulette, ...) un gain. Cette association éventualité-gain a donné lieu par la suite à la conception d'une fonction de portée plus générale. Le développement des variables aléatoires est associé à la théorie de la mesure.

Définition 1.1 *Soient $(\Omega, F, \mathbb{P})$ un espace probabilisé et $(E, \mathcal{E})$ un espace mesurable. On appelle variable aléatoire de Ω vers E, toute fonction mesurable X de Ω vers E.*

Cette condition de mesurabilité de X assure que l'image réciproque par X de tout élément B de la tribu $\mathcal{E}$ possède une probabilité et permet ainsi de définir, sur $(E, \mathcal{E})$, une mesure de probabilité, notée $\mathbb{P}_X$, par $\mathbb{P}_X(B) = \mathbb{P}(\mathbb{X}^{-1}(B)) = \mathbb{P}(X \in B)$.

La mesure $\mathbb{P}_X$ est l'image, par l'application X, de la probabilité $\mathbb{P}$ définie sur (Ω, F).

Définition 1.2 *La probabilité $\mathbb{P}_X$ est appelée loi de probabilité de la variable aléatoire X .*

Remarque 1.1 – *La loi d'une variable aléatoire réelle décrit en détail la répartition des valeurs de cette variable. La loi de la variable X contient toutes les informations nécessaires pour calculer sa fonction de répartition, son espérance et plus généralement ses moments, sa fonction caractéristique, sa médiane et ses quantiles.*
– *En d'autres termes, si deux variables aléatoires réelles X et Y ont même loi de probabilité, alors elles ont même fonction de répartition, même espérance et plus généralement mêmes moments, même fonction caractéristique, même médiane et mêmes quantiles.*

Remarque 1.2 *Les variables aléatoires réelles sont les variables aléatoires les plus couramment étudiées, ce qui conduit certains auteurs à omettre l'adjectif réel, et à parler de variable aléatoire tout court. Une variable aléatoire peut être discrète ou continue, et nous verrons en détail ses deux types.*

Quelques variables aléatoires réelles

En guise d'introduction aux définitions concernant les variables aléatoires réelles, il semble intéressant de présenter brièvement une famille de variables très utilisées.

Outre la variable certaine qui prend une valeur donnée avec une probabilité égale à 1, la variable aléatoire réelle la plus simple est appelée **variable de Bernoulli**. Celle-ci peut prendre deux états, qu'il est toujours possible de coder 1 et 0, avec les probabilités p et $1-p$. Une interprétation simple concerne un jeu de dé dans lequel on gagnerait 100 dinars en tirant le six ($p = 1/6$). Sur une séquence de parties, la moyenne des gains tend vers p lorsque le nombre de parties tend vers l'infini.

Si on considère qu'une partie est constituée par n tirages au lieu d'un seul, le total des gains est une réalisation d'une **variable binomiale** qui peut prendre toutes les valeurs entières de 0 à n. Cette variable a pour moyenne le produit np.

Variable aléatoire réelle discrète

Définition 1.3 *On dit qu'une v.a.r* ***discrète*** *si elle ne prend qu'un nombre fini ou infini dénombrable de valeurs, formellement*

$$X \in \{x_i, i \in K \subset \mathbb{N}\}$$

Ainsi le résultat d'un lancer de dé cubique est une variable aléatoire réelle discrète car elle ne peut prendre que 6 valeurs : $1, 2, 3, 4, 5, 6$. Le résultat de deux lancers de dés cubiques est une variable aléatoire discrète car elle ne peut prendre que 36 valeurs possibles : les couples $(1,1), (1,2), \ldots, (2,1), (2,2), \ldots, (6,5), (6,6)$.De même, la variable aléatoire donnant le nombre minimal de lancers nécessaires pour obtenir un premier 6 avec un dé cubique est une variable aléatoire discrète car on peut obtenir le premier 6 au premier lancer ($X = 1$), au second ($X = 2$), au 20^e ($X = 20$), $\ldots$, au n^e ($X = n$), $\ldots$ L'ensemble des valeurs possibles pour X est donc infini et dénombrable.

Dans ce cas,la loi de la variable aléatoire X est la loi de probabilité sur l'ensemble des valeurs possibles de X qui affecte la probabilité $\mathbb{P}(X = x_i)$ au singleton $\{x_i\}$. En pratique, l'ensemble des valeurs que peut prendre X est $\mathbb{N}$ ou une partie de $\mathbb{N}$.

1. L'espérance mathématique (moment d'ordre 1) de la v.a.r discrète X, notée $\mathbb{E}(X)$ est définie par (si la série $\sum_{i \in K} x_i \mathbb{P}(x_i)$.est absolument converge ou lorsque K est fini) :

$$\mathbb{E}(X) = \sum_{i \in K} x_i \mathbb{P}(X = x_i).$$

2. Le nombre :

$$Var(X) = \mathbb{E}[(X - E(X))^2],$$

lorsqu'il existe, est appelé variance de X, et l'écart type de X est :

$$\sigma(X) = \sqrt{Var(X)}.$$

Variables aléatoires réelles continues

Définition 1.4 *Une v.a.r prend des valeurs sur un ensemble infini non dénombrable des points, est dit continues si elle existe une fonction f non négative,définie pour toute valeur x appartenant à $\mathbb{R}$ et vérifiant, pour toute partie A de $\mathbb{R}$,la propriété :*

$$\mathbb{P}(X \in A) = \int_A f(x)dx,$$

et telle que :

$$\int_{\mathbb{R}} f(x)dx = 1$$

La fonction f est appelée la densité de probabilité de la variable aléatoire X.

1. L'espérance mathématique de la v.a.r continue X, définie sur l'espace probabilisé $(\Omega, F, \mathbb{P})$,est donnée par l'integrale, si elle converge :

$$\mathbb{E}(X) = \int_\Omega X d\mathbb{P} = \int_\Omega x d\mathbb{P} dx,$$

que l'on peut écrire, si f est la densité de probabilité de X

$$\mathbb{E}(X) = \int_{\mathbb{R}} x f(x)dx.$$

2. La variance, ou carré de l'écart -type σ, est donnée par l'intégrale si cette intégrale et la densité f existent

$$\mathbb{E}[(X - \mathbb{E}(X))^2] = Var(X) = \sigma^2 = \int_\Omega [x - \mathbb{E}(X)]^2 f(x)dx.$$

1.1.4 Notions de base

La fonction de répartition

En théorie des probabilités ou en statistiques, la fonction de répartition d'une variable aléatoire réelle caractérise la loi de probabilité de cette variable qu'elle soit discrète ou continue. La fonction de répartition de la variable aléatoire réelle X est la fonction F_X qui à tout réel x associe

$$F_X(x) = \mathbb{P}(X \leqslant x),\ x \in \mathbb{R},$$

où le membre de droite représente la probabilité que la variable aléatoire réelle X prenne une valeur inférieure ou égale à x. La probabilité que X se trouve dans l'intervalle $]a;b]$ est donc, si $a < b$,

$$\mathbb{P}(a < X \leqslant b) = F_X(b) - F_X(a)$$

La fonction de répartition d'une mesure de probabilité $\mathbb{P}$ définie sur la tribu borélienne $\mathcal{B}(\mathbb{R})$ est la fonction F qui à tout réel x associe

$$F(x) = \mathbb{P}(\,]-\infty; x]).$$

Remarque 1.3 *Une fonction de répartition doit vérifier un certain nombre de propriétés suivantes :*

- $0 \leqslant F_X(x) \leqslant 1$.
- $F_X(x)$ tend vers 0 en $-\infty$ et vers 1 en $+\infty$.
- $F_X(x)$ est croissante et continue à droite.
- $\forall x \in \mathbb{R}$, $\mathbb{P}(X > x) = 1 - \mathbb{P}(X \leqslant x) = 1 - F_X(x)$.
- Si X est une variable aléatoire discrète alors :

$$\forall x \in \mathbb{R},\ F_X(x) = \mathbb{P}(X \leqslant x) = \sum_{x_i \leq x} \mathbb{P}(x_i)$$

Dans ce cas F_X est une fonction en escalier présentant des sauts.

- Si X est une variable aléatoire continue alors :

$$\forall x \in \mathbb{R},\ F_X(x) = \mathbb{P}(X \leqslant x) = \int_{-\infty}^{x} f(t)dt \text{ et } f(x) = \frac{\partial}{\partial x} F_X(x).$$

Dans ce cas F_X est une fonction continue et f la fonction de densité.

Densité de probabilité d'une variable continue

Une variable continue possède souvent une fonction de répartition continue en tout point et dérivable par morceaux. Il est alors commode de la dériver pour obtenir la densité de probabilité $f : \mathbb{R} \to \mathbb{R}$ positive (en tout $x \in \mathbb{R}$), intégrable sur $\mathbb{R}$ vérifiant :

$$\int_{-\infty}^{+\infty} f(x)dx = 1.$$

Pour tout intervalle $]a;b] \subset \mathbb{R}$, on a alors :

$$\mathbb{P}(X \in]a;b]) = \mathbb{P}(a < X \leq b) = \int_a^b f(x)dx.$$

On reconstruit la fonction de répartition par la relation :

$$\forall x \in \mathbb{R},\ F_X(x) = \int_{-\infty}^{x} f(t)dt$$

Densité de probabilité d'une variable discrète

La loi d'une variable discrète est déterminée par l'ensemble des probabilités de ses valeurs nommé fonction de probabilité (mass function en anglais). Si l'on suppose qu'elle prend des valeurs entières (de signe quelconque), cela s'écrit :

$$\mathbb{P}_X(i) = \mathbb{P}(X = i), \quad (i \in \mathbb{Z}).$$

On reconstruit la fonction de répartition (dont les valeurs sont alors appelées probabilités cumulées) par la relation :

si $n \leq x < n+1$, alors $F_X(x) = \sum_{k=-\infty}^{n} \mathbb{P}_X(k)$.

En considérant la fonction de répartition comme une somme d'échelons ou fonctions de Heaviside, sa dérivée peut s'interpréter comme une somme d'impulsions ou fonctions de Dirac. En posant $\mathbb{P}_X(i) = \mathbb{P}_i$ elle s'écrit :

$$\mathbb{P}_X(x) = \sum_{k=-\infty}^{\infty} \mathbb{P}_i \delta(x-i).$$

Cette « densité de probabilité » présente un intérêt dans un problème particulier. Lorsqu'une intégrale porte sur une densité de probabilité, la propriété fondamentale de la fonction de Dirac permet de transformer l'intégrale en une simple somme impliquant la fonction de probabilité.

1.1.5 Grandeurs observées sur les échantillons

Espérance mathématique

L'espérance mathématique d'une variable aléatoire réelle X se définit comme la valeur de cette variable pondérée par sa probabilité. Pour une variable continue, la formule est :

$$\mathbb{E}[X] = \int_{-\infty}^{+\infty} x\ \mathbb{P}_X(x)\ dx.$$

Cette quantité est plus connue sous le nom de moyenne.

X étant une variable aléatoire réelle, une fonction f supposée régulière définit une nouvelle variable aléatoire $f \circ X$ notée $f(X)$ dont l'espérance, lorsqu'elle existe, s'écrit en remplaçant x par $f(X)$ dans la formule précédente (théorème de transfert).

$$\mathbb{E}\left[f(X)\right] = \int_{-\infty}^{+\infty} f(x)\ \mathbb{P}_X(x)\ dx.$$

Pour une variable discrète, la « densité de probabilité » conduit, sous réserve de sommabilité, à

$$\mathbb{E}\left[f(X)\right] = \sum_{k=-\infty}^{\infty} f(k)\ \mathbb{P}_X(k)\ .$$

Fonction caractéristique

Si la densité de probabilité d'une variable aléatoire réelle X possède une transformée de Fourier, celle-ci (ou, plus précisément, la transformée inverse), fonction à valeurs complexes définie sur $\mathbb{R}$

$$\phi_X(t) = \mathbb{E}\left[e^{itX}\right]$$

s'appelle fonction caractéristique de la variable.

Fonction génératrice des moments

La fonction génératrice des moments d'une variable aléatoire X est définie par :

$$M_X(t) = \mathbb{E}\left(e^{tX}\right),\ t \in \mathbb{R}.$$

lorsque son espérance existe. Cette fonction, comme son nom l'indique, est utilisée afin de générer les moments associés à la distribution de probabilités de la variable aléatoire X. Elle permet en outre de déterminer l'additivité d'une loi.

Moments

Si la fonction caractéristique (ou la fonction génératrice) d'une variable aléatoire est développable en série, celle- fait apparaître les moments de celle-ci, le moment d'ordre k étant défini comme

$$m_k = \mathbb{E}(X^k)$$

Dans le cas, important pratiquement, d'une variable assez régulière, celle-ci peut donc être caractérisée par la suite de ses moments, sa fonction caractéristique ou sa fonction génératrice, sa densité de probabilité ou, éventuellement, sa fonction de probabilité ou par sa fonction de répartition.

Dans le cas général, seuls les premiers moments peuvent exister.

1.1.6 Outils pratiques

Moments et moments centrés

Le moment d'ordre un, espérance ou moyenne de la variable,

$$\mu = m_1 = \mathbb{E}(X)$$

est un indicateur de tendance centrale.

Les moments d'ordre supérieur éliminent ce paramètre de position en considérant la variable centrée par soustraction de sa moyenne.

Le moment centré d'ordre deux (**variance**)

$$\sigma^2 = m_2' = \mathbb{E}\left[(X-\mu)^2\right],$$

est un indicateur de dispersion appelé variance. Sa racine carrée σ, grandeur homogène à la grandeur de base, s'appelle écart type. Lorsque la variable aléatoire est une valeur à un instant donné d'un processus aléatoire, l'expression moyenne quadratique est généralement préférée.

Ces deux moments fournissent une partie importante de l'information sur la variable, la totalité si celle-ci peut être considérée comme normale.

Les moments d'ordre supérieur, qui apportent pour les autres variables des précisions supplémentaires sur la forme de la distribution, portent sur la variable centrée réduite, rendue adimensionnelle par division par son écart type.

Le moment d'ordre trois de la variable centrée réduite,

$$m_3' = \mathbb{E}\left[\left(\tfrac{X-\mu}{\sigma}\right)^3\right],$$

est un indicateur d'asymétrie.

Le moment d'ordre quatre de la variable centrée réduite,

$$m_4' = \mathbb{E}\left[\left(\tfrac{X-\mu}{\sigma}\right)^4\right],$$

est un indicateur d'aplatissement des extrêmes des distributions appelé kurtosis.

Médiane et quantiles

On appelle médiane d'une variable aléatoire X, un réel m tel que

$$\mathbb{P}(X \leq m) \geq 1/2 \leq \mathbb{P}(X \geq m)$$

Dans le cas d'une variable aléatoire discrète, cette définition est peu intéressante car elle permet l'existence de plusieurs médianes.

Si X est le numéro apparaissant sur la face supérieure d'un dé à 6 faces parfaitement équilibré, pour tout réel m strictement compris entre 3 et 4, on a :

$$\mathbb{P}(X \leq m) = \mathbb{P}(X \geq m) = 1/2$$

ou bien l'existence d'une médiane qui ne donne pas une probabilité de 0,5

Si X est la somme obtenue en lançant deux dés à 6 faces parfaitement équilibrés. X ne possède qu'une seule médiane 7 mais $\mathbb{P}(X \leq 7) = 21/36$.

Dans le cas d'une variable continue, si la fonction de répartition est strictement croissante, la définition est équivalente à la suivante :

la médiane de X est le réel unique m tel que $F_X(m) = 0,5$

Le fait que la fonction de répartition soit continue, et supposée strictement croissante, à valeurs dans $]0;1[$, assure l'existence et l'unicité de la médiane.

Si la médiane a comme valeur $m = 0.5$, il est possible cependant de s'intéresser à d'autres valeurs de m (que l'on nomme les quantiles) :

- Quartile : $m = 0,25;\ 0,75$.
- Décile : $m = 0,1;\ 0,2;\ 0,3...$
- Centile : $m = 0,01;\ 0,02...$

1.2 Lois usuelles

1.2.1 Loi normale ou loi de Gauss

Une variable aléatoire réelle X suit une loi normale (ou loi gaussienne, loi de Laplace-Gauss) d'espérance μ et d'écart type σ (nombre strictement positif, car il s'agit de la racine carrée de la variance σ^2) si cette variable aléatoire réelle X admet pour densité de probabilité la fonction f définie, pour tout nombre réel x, par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{X-\mu}{\sigma}\right)^2}$$

Une telle variable aléatoire est alors dite variable gaussienne.

Une loi normale sera notée de la manière suivante $\mathcal{N}(\mu, \sigma)$ car elle dépend de deux paramètres μ (la moyenne) et σ (l'écart-type). Ainsi si une variable aléatoire X suit $\mathcal{N}(\mu, \sigma)$ alors

$$\mathbb{E}(X) = \mu \quad \text{et } V(X) = \sigma^2.$$

Lorsque la moyenne μ vaut 0, et l'écart-type σ vaut 1, la loi sera notée $\mathcal{N}(0, 1)$ et sera appelée loi normale standard ou centrée réduite. Sa fonction caractéristique vaut $e^{-t^2/2}$. Seule la loi $\mathcal{N}(0, 1)$ est tabulée car les autres lois (c'est-à-dire avec d'autres paramètres) se déduise de celle-ci à l'aide du théorème suivant : Si Y suit $\mathcal{N}(\mu, \sigma)$ alors $Z = \frac{Y-\mu}{\sigma}$ suit $\mathcal{N}(0, 1)$.

On note Φ la fonction de répartition de la loi normale centrée réduite :

$$\Phi(x) = \mathbb{P}(Z < x)$$

avec Z une variable aléatoire suivant $\mathcal{N}(0, 1)$.

Propriétés et Exemples :

$\Phi(-x) = 1 - \Phi(x)$

$\Phi(0) = 0,5$; $\Phi(1,645) \approx 0,95$; $\Phi(1,960) \approx 0,9750$.

Remarque 1.4 *La somme de deux variables gaussiennes indépendantes est elle-même une variable gaussienne (stabilité) : Soient X et Y deux variables aléatoires indépendantes suivant respectivement les lois $\mathcal{N}(\mu_1, \sigma_1)$ et $\mathcal{N}(\mu_2, \sigma_2)$. Alors, la variable aléatoire $X + Y$ suit la loi normale $\mathcal{N}\left(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2}\right)$.*

1.2.2 Loi du $\mathcal{X}^2$(khi-deux)

Définition 1.5 *Soit $Z_1, Z_2, \cdots, Z_v$ une suite de variables aléatoires indépendantes de même loi $\mathcal{N}(0, 1)$. Alors la variable aléatoire $\sum_{i=1}^{v} Z_i^2$ suit une loi appelée loi du Khi-deux à v degrés de liberté, notée $\mathcal{X}^2(v)$.*

Proposition 1.1

1. *Sa fonction caractéristique est $(1 - 2it)^{-v/2}$.*

2. *La densité de la loi du $\mathcal{X}^2(v)$ est* $f_v(x) = \begin{cases} \frac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2} & \text{pour } x > 0 \\ 0 \quad \text{sinon} \end{cases}$

où Γ est la fonction Gamma d'Euler définie par $\Gamma(r) = \int_0^\infty x^{r-1} e^{-x} dx$.

3. *L'espérance de la loi du $\mathcal{X}^2(v)$ est égale au nombre v de degrés de liberté et sa variance est $2v$.*

4. *La somme de deux variables aléatoires indépendantes suivant respectivement $\mathcal{X}^2(v_1)$ et $\mathcal{X}^2(v_2)$ suit aussi une loi du $\mathcal{X}^2$ avec $v_1 + v_2$ degrés de liberté.*

1.2.3 Loi de Student

Définition 1.6 *Soient Z et Q deux variables aléatoires indépendantes telles que Z suit $\mathcal{N}(0,1)$ et Q suit $\mathcal{X}^2(v)$. Alors la variable aléatoire*
$T = \frac{Z}{\sqrt{Q/v}}$
suit une loi appelée ***loi de Student*** *à degrés de liberté, notée $St(v)$.*

Proposition 1.2
1. *La densité de la loi de Student à v degrés de liberté est*

$$f(x) = \frac{1}{\sqrt{\pi v}} \frac{\Gamma\left(\frac{v+1}{2}\right)}{\Gamma\left(\frac{v}{2}\right)} \frac{1}{(1+x^2/v)^{\frac{1+v}{2}}}$$

2. *L'espérance n'est pas définie pour $v = 1$ est vaut 0 si $v \geq 2$. Sa variance n'existe pas pour $v \leq 2$ et vaut $v/(v-2)$ pour $v \geq 3$.*

3. *La loi de Student converge en loi vers la loi normale centrée réduite.*

Remarque 1.5 *Pour $v = 1$, la loi de Student s'appelle loi de* ***Cauchy****, ou loi de* ***Lorentz.***

1.2.4 Loi de Fisher-Snedecor

Définition 1.7 *Soient Q_1 et Q_2 deux variables aléatoires indépendantes telles que Q_1 suit $\mathcal{X}^2(v_1)$ et Q_2 suit $\mathcal{X}^2(v_2)$ alors la variable aléatoire*
$F = \frac{Q_1/v_1}{Q_2/v_2}$
suit une loi de ***Fisher-Snedecor*** *à (v_1, v_2) degrés de liberté, notée $F(v_1, v_2)$.*

Proposition 1.3 *La densité de la loi $F(v_1, v_2)$ est*

$$f(x) = \frac{\Gamma\left(\frac{v_1+v_2}{2}\right)}{\Gamma(v_1/2)\Gamma(v_2/2)} \left(\frac{v_1}{v_2}\right)^{v_1/2} \frac{x^{v_1/2-1}}{\left(1+\frac{v_1}{v_1}x\right)^{\frac{v_1+v_2}{2}}} \text{ si } x > 0 \text{ (0 sinon)}$$

Son espérance n'existe que si $v_2 \geq 3$ et vaut $\frac{v_2}{v_2-2}$. Sa variance n'existe que si $v_2 \geq 5$ et vaut $\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-2)^2(v_2-4)}$.

Proposition 1.4 ***1.*** *Si F suit une loi de Fisher $F(v_1, v_2)$ alors $\frac{1}{F}$ suit une loi de Fisher $F(v_2, v_1)$.*

2. *Si T suit une loi de Student à degrés de liberté v alors T^2 suit une loi de Fisher $F(1, v)$.*

Chapitre 2

L'estimation et la convergence

L'estimation consiste à donner des valeurs approximatives aux paramètres d'une population à l'aide d'un échantillon de n observations issues de cette population. On peut se tromper sur la valeur exacte, mais on donne la "meilleure valeur" possible que l'on peut supposer. Alors on cherche à ce qu'un estimateur soit sans biais, convergent, efficace et robuste.

Exemple d'estimateurs

Si l'on cherche à évaluer la taille moyenne des enfants de 10 ans, on peut effectuer un sondage sur un échantillon de la population des enfants de 10 ans (par exemple en s'adressant à des écoles réparties dans plusieurs milieux différents). La taille moyenne calculée sur cet échantillon, appelée moyenne empirique, sera un estimateur de la taille moyenne des enfants de 10 ans.

Si l'on cherche à déterminer le pourcentage d'électeurs décidés à voter pour le candidat A, on peut effectuer un sondage sur un échantillon représentatif. Le pourcentage de votes favorables à A dans l'échantillon est un estimateur du pourcentage d'électeurs décidés à voter pour A dans la population totale.

Si l'on cherche à évaluer la population totale de poissons dans un lac, on peut commencer par ramasser n poissons, les baguer pour pouvoir les identifier ultérieurement, les relâcher, les laisser se mélanger aux autres poissons. On tire alors un échantillon de poissons du lac, on calcule la proportion p de poissons bagués. La valeur n/p est un estimateur de la population totale de poissons dans le lac. S'il n'y a aucun poisson bagué dans l'échantillon, on procède à un autre tirage.

Un estimateur est très souvent une moyenne, une population totale, une proportion ou une variance.

Définition 2.1 ***(Définition formelle)***
Un estimateur du paramètre θ d'un modèle ou loi de probabilité est une fonction qui fait correspondre à une suite d'observations issues du modèle ou loi de probabilité la valeur $\widehat{\theta}$, que l'on nomme estimé ou estimation

$$\widehat{\theta} = f(x_1, x_2, \cdots, x_n)$$

Remarque 2.1 *Un estimateur ne doit évidemment jamais dépendre de θ , il ne dépend que des observations empiriques.*

2.1 Qualité d'un estimateur

Un estimateur est une valeur $\widehat{\theta}$ calculée sur un échantillon tiré au hasard, la valeur $\widehat{\theta}$ est donc une variable aléatoire possédant une espérance $\mathbb{E}\left(\widehat{\theta}\right)$ et une variance $Var\left(\widehat{\theta}\right)$. On comprend alors que la valeur $\widehat{\theta}$ puisse fluctuer selon l'échantillon. Elle a de très faibles chances de coïncider exactement avec la valeur exacte θ qu'elle est censée représenter. L'objectif est donc de maîtriser l'erreur commise en prenant la valeur x pour la valeur X.

2.1.1 Biais

Une variable aléatoire fluctue autour de son espérance. On souhaite donc que l'espérance de $\widehat{\theta}$ soit égale à θ, soit qu'en "moyenne" l'estimateur ne se trompe pas.

Définition 2.2 $Biais\left(\widehat{\theta}\right) = \mathbb{E}\left(\widehat{\theta}\right) - \theta.$

Lorsque l'espérance de l'estimateur coincide avec la vraie valeur, l'estimateur est dit sans biais.

Exemple 2.1 *L'estimateur choisi précédemment sur la taille moyenne des enfants de 10 ans est un estimateur sans biais mais celui des poissons comporte un biais : le nombre de poissons estimé est en moyenne supérieur au nombre de poissons réels.*

2.1.2 Erreur quadratique moyenne

Définition 2.3 *L'erreur quadratique moyenne est l'espérance du carré de l'erreur entre la vraie valeur et sa valeur estimée.*

$$MSE\left(\widehat{\theta}\right) = \mathbb{E}\left(\left(\widehat{\theta} - \theta\right)^2\right).$$

Remarque 2.2 *L'erreur quadratique moyenne peut être écrite comme une somme de la variance et du carré du biais de l'estimateur*

$$MSE\left(\widehat{\theta}\right) = Var(\widehat{\theta}) + \left(\mathbb{E}\left(\widehat{\theta}\right) - \theta\right)^2$$

2.1.3 Comparaison d'estimateurs

Définition 2.4 *On dit que l'estimateur $\widehat{\theta}_1$ domine l'estimateur $\widehat{\theta}_2$ si pour tout $\theta \in \Theta$, $MSE\left(\widehat{\theta}_1\right) \leq MSE\left(\widehat{\theta}_2\right)$.*

Définition 2.5 *On dit qu'un estimateur est admissible s'il n'existe aucune estimateur le dominant.*

Définition 2.6 *Soit $\widehat{\theta}_1, \widehat{\theta}_2$ deux estimateurs sans biais de θ, $\widehat{\theta}_1$ est dit plus efficace que $\widehat{\theta}_2$ si :*

$Var\ (\widehat{\theta}_1) < Var\ (\widehat{\theta}_2)$, $\forall \theta \in \Theta$.

2.2 Convergence

On souhaite aussi pouvoir, en augmentant la taille de l'échantillon, diminuer l'erreur commise en prenant $\widehat{\theta}_n$ à la place de $\widehat{\theta}$. Si c'est le cas, on dit que l'estimateur est convergent (on voit aussi consistant), c'est-à-dire qu'il converge vers sa vraie valeur. La définition précise en mathématique est la suivante :

Définition 2.7 *L'estimateur $\widehat{\theta}_n$ est convergent s'il* ***converge en probabilité*** *vers θ, soit :* $\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| > \varepsilon\right) = 0$, $\forall \varepsilon > 0$.

On l'interprète comme le fait que la probabilité de s'éloigner de la valeur à estimer de plus de ε tend vers 0 quand la taille de l'échantillon augmente. Cette définition est parfois écrite de manière inverse :

Définition 2.8 *L'estimateur $\widehat{\theta}_n$ est convergent s'il* ***converge en probabilité*** *vers θ, soit :* $\lim_{n\to\infty} \mathbb{P}\left(\left|\widehat{\theta}_n - \theta\right| \leq \varepsilon\right) = 1$, $\forall \varepsilon > 0$.

Il existe enfin un type de convergence plus forte, la convergence presque sûre, définie ainsi pour un estimateur :

Définition 2.9 *L'estimateur $\widehat{\theta}_n$ est fortement convergent s'il* ***converge presque sûrement*** *vers θ, soit :* $\mathbb{P}\left(\lim_{n\to\infty} \widehat{\theta}_n = \theta\right) = 1$.

Exemple 2.2 *La moyenne empirique est un estimateur convergent de l'espérance d'une variable aléatoire. La loi faible des grands nombres assure que la moyenne converge en probabilité vers l'espérance et la loi forte des grands nombres qu'elle converge presque sûrement.*

Lois des Grands Nombres

Ces lois décrivent le comportement asympotique de la moyenne de l'échantillon. Elles sont de deux types : **lois faibles** mettant en jeu la convergence en probabilité et **lois fortes** relatives à la convergence presque sûre.

Théorème 2.1 *Si $(X_1, \cdots, X_n)$ est un échantillon d'une v.a.r X tel que $\mathbb{E} \mid X \mid < \infty$, alors*

$$\begin{array}{ll} \text{loi faible} & \overline{X}_n \xrightarrow{P} \mu \quad \text{quand } n \longrightarrow \infty \\ \text{loi forte} & \overline{X_n} \xrightarrow{p.s} \mu \quad \text{quand } n \longrightarrow \infty \end{array}$$

ou $\mu := \mathbb{E}(X), \overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

2.2.1 Taux de convergence

Efficience

La variable aléatoire fluctue autour de son espérance. Plus la variance $Var\left(\widehat{\theta}_n\right)$ est faible, moins les variations sont importantes. On cherche donc à ce que la variance soit la plus faible possible. C'est ce qu'on appelle l'efficience d'un estimateur.

Robustesse

Il arrive que lors d'un sondage, une valeur extrême et rare apparaisse (par exemple un enfant de 10 ans mesurant $1,80m$). On cherche à ce que ce genre de valeur ne change que de manière très faible la valeur de l'estimateur. On dit alors que l'estimateur est robuste.

Exemple 2.3 *En reprenant l'exemple de l'enfant, la moyenne n'est pas un estimateur robuste car ajouter l'enfant très grand modifiera beaucoup la valeur de l'estimateur. La médiane par contre n'est pas modifiée dans un tel cas.*

2.3 Quelque estimateur classique

Soit X une variable aléatoire de moyenne μ et d'écart-type σ

1. On prend en général comme estimateur de la moyenne μ la moyenne empirique $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ qui est un estimateur sans biais. Son estimation $\overline{x}$ est la moyenne observée dans une réalisation de l'échantillon.

2. $\widetilde{S}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$ est un estimateur consistant de σ^2(mais biaisé).

3. $S^2 = \frac{n}{n-1}\widetilde{S}^2$est un estimateur sans biais et consistant de σ^2. Son estimateur est $s^2 = \frac{n}{n-1}\sigma_e^2$ où $\sigma_e$est l'ecart-type observé dans une réalisation de l'échantillon.

Remarque 2.3 *: Si la moyenne μ de X est connue, $T = \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)^2$ est un meilleur estimateur de σ^2 que S^2.*

L'estimation est habituellement divisée en deux composantes principales : l'estimations paramétrique et l'estimation non-paramétrique.

2.4 Estimation paramètrique

Un estimateur de T est une fonction $T_n : x \to T_n(x, X_1, \cdots, X_n)$ mesurable par rapport à l'observation $(X_1, \cdots, X_n)$. Si l'on sait à priori que T appartient à une famille paramètrique $\{T(x, \theta), \theta \in \Theta\}$ où $\Theta \in \mathbb{R}^d$ et $T(\cdot, \cdot)$ est une fonction continue, on parle alors d'estimation paramètrique, car estimer T revient à estimer le paramètre fini dimensionnel θ.

Au phénomène étudié, nous associons maintenant un modèle statistique P_θ qui dépend d'un paramètre θ. Pour se faire une idée de la valeur inconnue du paramètre θ, a partir des observation $(X_1, \cdots, X_n)$ qui sont i.i.d, on calcule ensuite une certaine valeur numérique, que l'on considérera comme valeur approchée de θ qu'on appellera une estimateur de θ.

Dans ce cas ou'il n'y a pas d'estimateur évident, on cherche un estimateur par la méthode de vraisemblance, ou par la méthode des moments,...etc

1. Soit un échantillon issu d'une v.a.r X normale de fonction de densité f qui dépend de deux paramètres qui sont inconnus (μ, σ^2) :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-1}{2}(\frac{x-\mu}{\sigma})^2}$$

Pour estimer la fonction de densité f il se fait d'estimer le paramètre inconnue $\theta = (\mu, \sigma^2)$, où μ est la moyenne de X et σ^2 sa variance.

On a dans ce cas :

$$\widehat{\mu} = \overline{X} \text{ et } \widehat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2,$$

et l'estimateur de $\theta = (\mu, \sigma^2)$ est $\widehat{\theta} = (\widehat{\mu}, \widehat{\sigma^2}) = (\overline{X}, \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2)$.

Donc l'estimateur $\widehat{f}$ de f est donné :

$$\widehat{f}(x) = \frac{1}{\sqrt{2\pi}\sqrt{\widehat{\sigma^2}}} \exp\left[\frac{-1}{2}\left(\frac{x-\widehat{\mu}}{\sqrt{\widehat{\sigma^2}}}\right)^2\right]$$

$$= \frac{1}{\sqrt{2\pi}\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}} \exp\left[\frac{-1}{2}\left(\frac{x-\overline{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2}}\right)^2\right].$$

2.Soit X un v.a.r suit la loi exponentielle de fonction de densité $f(x) = \lambda e^{-\lambda x}$ avec $\lambda > 0$ paramètre inconnu.

Pour estimer la fonction de densité f il se fait d'estimer le paramètre inconnue λ.

L'estimateur de λ est $\widehat{\lambda} = \frac{1}{\overline{X}}$.

Donc l'estimateur $\widehat{f}$ de f est donnée :

$$\widehat{f}(x) = \widehat{\lambda} e^{-\widehat{\lambda} x} = \frac{1}{\overline{X}} e^{-\frac{1}{\overline{X}} x}.$$

Problème : comment obtenir, en général, des estimateurs possèdant les qualités enoncées précédemment ?

Nous allons maintenant présenter deux méthodes générales permettant d'obtenir des estimateurs

2.4.1 Méthode des moments :

Une première approche, appelée méthode des moments consiste a tirer profit de la loi des grands nombres qui nous dit (sous certaines hypothèses) que les moments empiriques d'ordre k

$$\overline{X^k} = \frac{1}{n} \sum_{i=1}^{n} X_i^k$$

convergent vers les moments théorique

$$m_k = \mathbb{E}\left[X^k\right]$$

La méthode des moments consiste a exprimer les p premiers moments de la loi de X en fonction des p paramètres $\theta = (\theta_1, \theta_2, \cdots, \theta_p)$ inconnus. Cela amène a un système de p equation

$$\begin{aligned} m_1 &= h_1(\theta_1, \theta_2, \cdots, \theta_p) \\ &\cdots \\ m_p &= h_p(\theta_1, \theta_2, \cdots, \theta_p) \end{aligned}$$

que l'on cherche à résoudre afin d'obtenir l'expression de chaque paramètre θ_j en fonction des moments $m_1, \cdots, m_p$. On obtient ensuite des estimateurs en remplaçant les moments théoriques m_l par les moments empiriques $\overline{X^l}$ dans les expressions obtenues pour les θ_j.

Remarque 2.4 *il peut arriver que l'on ne considère pas les p premiers moments afin d'obtenir un système d' équations permettant d'aboutir a une unique solution. car en général les premiers moments suffisent.*

Exemple 2.4 *La méthode que nous avons utilisée au début pour l'estimation paramétrique de la densité pour la loi exponentielle et la loi normale est un exemple de l'utilisation de la méthode des moments.*

2.4.2 Méthode du Maximum de vraisemblance :

La méthode des moments décrite dans la section précédente est assez intuitive et permet de proposer un certain nombre d'estimateurs convergents. Cependant, ces estimateurs n'ont pas toujours d'aussi bonnes propriétés que ceux que l'on obtient en considérant les estimateurs obtenus par la méthode du maximum de vraisemblance que nous allons voir maintenant.

La méthode permet d'aboutir dans de nombreux cas à des estimateurs eficaces. Son principe consiste a choisir comme estimation du paramètre θ, la valeur la plus vraisemblable, c'est-a-dire celle qui a la plus forte probabilité de provoquer les valeurs observées dans l'échantillon.

La loi d'une variable X est caractérisée par la probabilité des valeurs possibles (cas discret) ou par sa densité (cas continu). De la même façon, si X suit une loi de paramètre θ, la loi du

n- échantillon (composé de variables indépendantes) est caractérisée par $\Pi_{i=1\cdots n}\mathbb{P}_\theta(s_i)$ (variables discrêtes) ou $\Pi_{i=1\cdots n}f_\theta(s_i)$ (variables continues), pour toutes les valeurs $(s_1\cdots,s_n)$ prises par les variables $X_1,\cdots,X_n$. Dans les deux cas, ce produit est une fonction des valeurs s_i et du paramètre θ, que l'on notera $L\left(s_1\cdots,s_n;\theta\right)$.

Définition 2.10 *On appelle estimateur du maximum de vraisemblance de θ une valeur (de t) qui maximise la fonction de vraisemblance*

$$t\mapsto L\left(X_1\cdots,X_n;t\right)$$

L'estimation de θ par le maximum de vraisemblance revient donc a chercher le paramètre avec lequel $L\left(x_1\cdots,x_n;t\right)$ est maximale, c'est a dire "la valeur du param etre avec laquelle on avait le plus de chance d'obtenir ce qu'on a obtenu".

Remarque 2.5 *pour des raisons de commodité de calcul, on utilise souvent la fonction de logvraisemblance*

$$LL:t\mapsto\ln\left(L\left(X_1\cdots,X_n;t\right)\right)$$

qui est le logarithme népérien de la fonction de vraisemblance, elles sont maximales en même temps (car le logarithme népérien est strictement croissant sur $\mathbb{R}_+^*$

Exemple 2.5 *Si X suit une loi exponentielle de paramètre θ, sa densité est :*

$$f(x) = \theta e^{-\theta x} \mathbf{1}_{[0;+\infty[}(x)\,.$$

la fonction de vraisemblance est :

$$L(\overrightarrow{x}, \theta) = \theta e^{-\theta x_1} \times \cdots \times \theta e^{-\theta x_n} = \theta e^{-\theta \Sigma x_i}.$$

La log-vraisemblance est donc

$$n \ln(\theta) - \theta \sum x_i.$$

Pour maximiser la log-vraisemblance qui est concave, on dérive par rapport a θ et on annule la dérivée ce qui donne $\frac{n}{\theta}$ -$\sum x_i = 0$ et donc

$$\widehat{\theta} = \frac{n}{\sum x_i} = \frac{1}{\overline{x}}.$$

Remarque 2.6 *Sous des hypothèses très générales on peut montrer que l'estimateur du maximum de vraisemblance est efficace ou asymptotiquement efficace et que sa distribution d"échantillonnage est asymptotiquement normale.*

2.5 Estimation non-paramètrique

Par opposition, en statistique non paramétrique, le modèle n'est pas décrit par un nombre fini de paramètres. Divers cas de figures peuvent se présenter, comme par exemple :

- On s'autorise toutes les distributions possibles, i.e. on ne fait aucune hypothèse sur la forme/nature/type de la distribution des variables aléatoires
- On travaille sur des espaces fonctionnels, de dimension infinie. Exemple : les densités continues sur $[0, 1]$, ou les densités monotones sur $\mathbb{R}$.
- Le nombre de paramètres du moèle n'est pas fixé et varie (augmente) avec le nombre d'observations.
- Le support de la distribution est discret et varie (augmente) avec le nombre d'observations.

L'avantage principale de l'estimation non-paramètrique à un ensemble fini d'observation est de ne pas nécessiter d'hypothèses à priori sur l'appartenence à une famille de lois connues. L'estimation ne concerne pas les paramètres permettant de sélectionner une loi, mais directement la fonction elle même (d'où le terme non-paramètrique)

Définition 2.11 *Un modèle non-paramètrique est un modèle qui ne peut pas être d'écrit par un nombre fini de paramètres. On a quelques exemples de modèles non-paramètriques les plus connus : la fonction de répartition, la fonction caractéristique et la fonction de quantile*

Estimation non paramètrique de la fonction de répartition

Pour tout $x \in \mathbb{R}$, on appelle valeur de la fonction de répartition empirique en x, la statistique, notée $F_n(x)$, définie par :

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1]-\infty, x](x_i),$$

où $\mathbf{1}(-\infty, x]$ est la fonction indicatrice de l'intervalle $(-\infty, x]$, à savoir $1(-\infty, x](u) = 1$ si $u \in (-\infty, x]$ et 0 sinon.

En d'autre terme $F_n(x)$ est la v.a ‹‹ proportion›› des n observations $X_1, ..., X_n$ indépendantes et identiquement distribuées $(i.i.d)$ prenant unr valeur inférieur ou égale à x. Chaque X_i ayant une probabilité $F(x)$ d'être inférieure ou égale à x.

$$F_n(x) = \frac{\text{nombre d'observation} \leq x}{n} = \frac{1}{n} \sum_{i=1}^{n} 1\{X_i \leq x\}$$

$$= \begin{cases} 0 & \text{si} \quad x \leq X_1 \\ \frac{k}{n} & \text{si} \quad X_k \leq x \leq X_{k+1} \quad k = 1, ..., n-1 \\ 1 & \text{si} \quad x \geq X_n \end{cases}$$

$nF(x)$ suit une loi binomiale $B(n, F(x))$. En conséquence $F(x)$ est une v.a discrète prenant les valeurs $\frac{k}{n}$, où $k = 0, ..., n$ avec probabilités :

$$\mathbb{P}(F_n(x) = \tfrac{k}{n}) = \mathbb{P}(nF_n(x) = k) = \binom{n}{k} [F(x)]^k [1 - F(x)]^{n-k}$$

$F_n(x)$ est un estimateur simple de $F(x)$. Il s'avère que cette fonction est un très bon estimateur de F.

Les propriétés de la fonction de répartition empirique :

- Pour tout $x \geq \max\{x_1, \cdots x_n\}$, $F_n(x) = 1$. de même, Pour tout $x < \min\{x_1, \cdots x_n\}$, $F_n(x) = 0$. On a donc clairement
  $\lim_{x \to -\infty} F_n(x) = 0$ et $\lim_{x \to +\infty} F_n(x) = 1$.
- L'espérance de $F_n(x)$ et $\mathbb{E}[F_n(x)] = F(x)$.
- La variance de $F_n(x)$ et $Var(F_n(x)) = \frac{F(x)(1-F(x))}{n}$.
- Le biais de $F_n(x)$ est : $Biais(F_n(x)) = \mathbb{E}[F_n(x)] - F(x) = 0$, donc $F_n(x)$ est un estimateur sans biais de $F(x)$.
- Le MSE de $F_n(x)$ est donnée par : $MSE(F_n(x)) = Var(F_n(x)) + (Biais(F_n(x)))^2$.
- (**Théoreme de Glivernko-Cantelli**) La convergence uniforme presque sûre de $F_n(x)$ vers $F(x)$ définie par :

$$\sup_{x \in \mathbb{R}} |F_n(x) - F(x)| \to 0 \quad ps.$$

2.5.1 Quelques méthodes d'estimation non paramétriques

Dans cette section, nous décrivons quelques méthodes classiques d'estimation non paramétrique, sans entrer dans autant de détails, et on va etudier au chapitre 3 la méthodes de noyau pour estimer la densité.

- **Estimateur de Nadaraya{Watson}**
- **Estimation par polynômes locaux**
- **Estimation par plus proches voisins**
- **Estimation par splines**
- **Estimation par projection**

Chapitre 3

Estimation de la densité

3.1 Le problème et les mesures de risque L_2

Considérons une variable aléatoire X qui admet une fonction de densité $f : \mathbb{R} \to \mathbb{R}^+$.

Rappelons que ceci signifie que la probabilité que X prenne une valeur dans un borélien $\mathcal{B}$ arbitraire peut être evaluée par la relation :

$$\mathbb{P}(X \in \mathcal{B}) = \int_{\mathcal{B}} f(x)dx.$$

Dans ce chapitre, nous considérons le problème de l'estimation de f sur la base de n copies indépendantes de X, c'est- a-dire sur la base d'observations $X_1, \cdots, X_n$ qui sont mutuellement indépendantes et admettent chacune la même densité f que X.

Le plus souvent, on considèrera des mesures de risque L_2 pour mesurer la qualité d'un estimateur $\widehat{f_n}(x)$ de $f(x)$.Ce risque L_2, qui correspond au MSE définie au chapitre2, qui admet une décomposition en termes de biais et de variance.

Il est désirable que le risque L_2 tende vers zéro lorsque la taille d'échantillon n diverge vers l'infini. Par définition, ceci signifie que $\widehat{f_n}(x)$ converge vers $f(x)$ en moyenne quadratique, ce qui implique que $\widehat{f_n}(x)$ converge vers $f(x)$ en probabilité.

La mesure de risque L_2 ci-dessus, qui mesure la qualité de l'estimation de $f(x)$ par l'estimateur $\widehat{f_n}(x)$, est adéquate si on est spécifiquement intéressé par l'estimation de f au point x. Le plus souvent, cependant, on désire estimer la densité f dans son ensemble, auquel cas le risque suivant est plus naturel.

Définition 3.1 *Le risque L_2 intégré de $\widehat{f_n}$ est :*

$$R\left(\widehat{f_n}, f\right) = \mathbb{E}\left[\int_{-\infty}^{+\infty} \left\{\widehat{f_n}(x) - f(x)\right\}^2 dx\right],$$

où l'espérance est calculée en utilisant le fait que $\widehat{f_n}$ est basée sur un échantillon d'observations aléatoires $X_1, \cdots, X_n$ indépendantes et admettant la densité f .

On parlera parfois d'erreur quadratique moyenne intégrée (ou $MISE$ en anglais, pour Mean Integrated Square Error). En permutant espérance et intégrale, il vient

$$R\left(\widehat{f_n}, f\right) = \left[\int_{-\infty}^{+\infty} \mathbb{E}\left[\left\{\widehat{f_n}(x) - f(x)\right\}^2\right] dx\right],$$

ce qui montre que le risque intégré $R\left(\widehat{f_n}, f\right)$ est l'intégrale du risque ponctuel MSE.

A titre d'illustration dans ce chapitre, nous considérerons souvent la densité

$$f(x) = \tfrac{1}{2}\phi_{0,1}(x) + \tfrac{1}{10}\sum_{j=0}^{4}\phi_{\frac{j}{2}-1,\frac{1}{100}}(x),$$

qui est une "densité de mélange" a six composantes gaussiennes initiales (ici, ϕ_{μ,σ^2} est la densité de la loi gaussienne de moyenne μ et de variance σ^2. Le graphe de

cette densité, appelée dans la densité de **Bart Simpson**, est donnée a la Figure (3.1)**.**

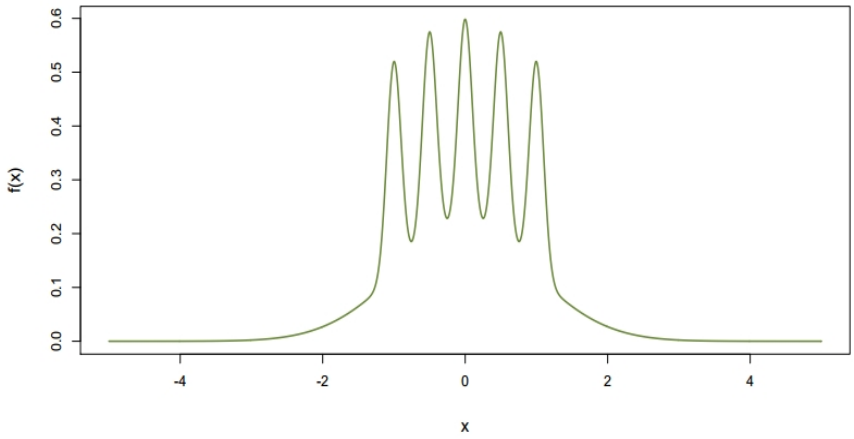

Figure (3.1)-*Graphe de la densité de Bart Simpson*

L'approche la plus classique pour estimer f est de nature paramétrique et consiste a supposer que f appartient a une classe de fonctions $\mathcal{F} = \left\{f_\theta : \theta \in \Theta \subset \mathbb{R}^k\right\}$ indicée par un paramètre θ de dimension finie. Par exemple, on pourrait supposer que X est de loi gaussienne, auquel cas $\mathcal{F} = \left\{\phi_{\mu,\sigma^2} \ : \begin{pmatrix} \mu \\ \sigma^2 \end{pmatrix} \in \mathbb{R} \times \mathbb{R}_+^* \subset \mathbb{R}^2\right\}$ est la collection de toutes les densités gaussiennes. Bien entendu, ceci ramène l'estimation de f a l'estimation du paramètre θ, ce qui est un problème classique en statistique, pour lequel on dispose d'un large eventail de méthodes standards (méthodes du maximum du vraisemblance, méthode des moments, méthodes des moindres carrés, etc.) Dans le cas du modèle gaussien, la densité estimée par la méthode du maximum de vraisemblance sera

$$\widehat{f}_n(x) = \phi_{\widehat{\mu}_n, \widehat{\sigma}_n^2}(x) = \frac{1}{\sqrt{2\pi\widehat{\sigma}_n^2}} \exp\left(-\frac{(x-\widehat{\mu}_n)^2}{2\widehat{\sigma}_n^2}\right), \tag{3.1}$$

où $\widehat{\mu}_n = \overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ et $\widehat{\sigma}_n^2 = s_n^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^2$ sont respectivement les estimateurs du maximum de vraisemblance de μ et σ^2.

Si f est efectivement gaussienne ou est proche d'une loi gaussienne, alors cet estimateur est adéquat. Si, par contre, f s'éloigne fortement d'une densité gaussienne, comme c'est le cas par exemple de la densité de Bart Simpson, alors $\widehat{f}_n$ n'est pas un estimateur satisfaisant ; ceci est illustré à la Figure (3.2). Si f n'est pas gaussienne, le risque intégré $R\left(\widehat{f}_n, f\right)$ ne tendra pas vers zéro quand n diverge vers l'infini.

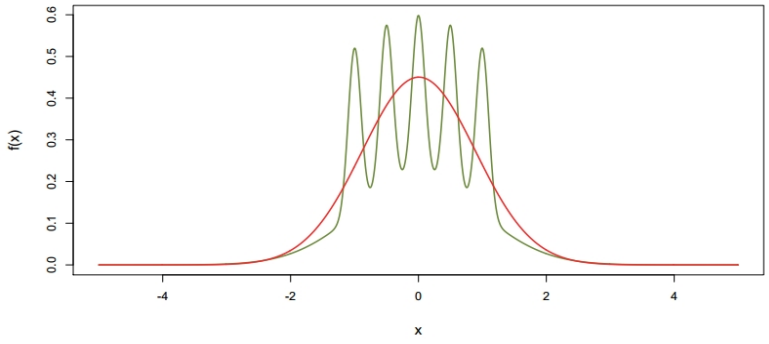

Figure (3.2)- *Graphe de l'estimateur parametrique gaussien (en rouge) calcule sur une échantillon aléatoire de taille* $n = 1000$ *engendre depuis la densité de Bart Simpson(en vert)*

Ceci constitue une motivation importante pour introduire des estimateurs de densité non paramétriques, pour lesquels le risque intégré tendra vers zéro sous des hypothèses minimes sur la densité f qu'on estime.

3.2 Estimation par histogramme

Supposons que la densité f à estimée soit nulle en dehors de l'intervalle $\mathcal{I} = [a; b]$ (si un tel intervalle $[a; b]$ n'existe pas, comme c'est le cas pour la densité de Bart Simpson, on prendra l'intervalle sufisamment grand pour que f soit presque nulle en dehors de l'intervalle ; nous renvoyons a la fin de cette section pour une autre approche). Pour un entier strictement positif m fixé, partitionnons $\mathcal{I}$ en les m sous-intervalles ou cellules

$$\mathcal{I}_1 = [a; a+h[\,, \mathcal{I}_1 = [a+h; a+2h[\,, \cdots, \mathcal{I}_1 = [a+(m-1)\,h; a+mh = b[\,,$$

où $h = (b-a)/m$ est la taille de la cellule.

L'estimateur par histogramme de f associé au nombre de cellules m est donné par :

$$\widehat{f}_n(x) = \frac{1}{h}\sum_{j=1}^{m} \widehat{P}_j \mathbf{1}\left[x \in \mathcal{I}_j\right], \tag{3.2}$$

où $\widehat{P}_j = Y_j/n$ est basé sur le nombre $Y_j = \sum_{i=1}^{n} \mathbf{1}\left[X_i \in \mathcal{I}_j\right]$ d'observations appartenant à $\mathcal{I}_j$.

L'estimateur par histogramme en (3.2) se réécrit

$$\widehat{f}_n(x) = \begin{cases} \widehat{P}_1/h & \text{si } x \in \mathcal{I}_1 \\ \vdots & \\ \widehat{P}_m/h & \text{si } x \in \mathcal{I}_m \end{cases}$$

ce qui implique que $\widehat{f}_n$ est une fonction constante par morçeaux, et donc discontinue. La Figure (3.3) montre l'estimateur $\widehat{f}_n$ associé à $a = -5$, $b = 5$ et $m = 100$ calculé sur un echantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson. Clairement, pour ces valeurs des paramètres a, b et m, l'histogramme fournit une reconstruction assez satisfaisante de f .

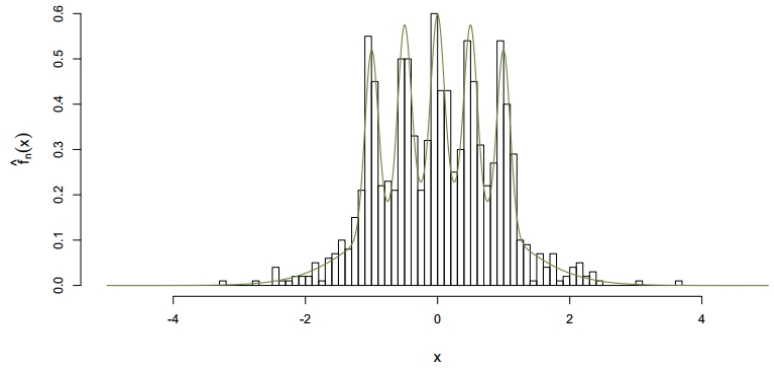

Figure (3.3)- *Graphe de l'estimateur par histogramme $\widehat{f}_n$ associé à $a = -5$, $b = 5$ et $m = 100$ calculé sur une chantillon aléatoire de taille $n = 1000$ engendre depuis la densité de Bart Simpson (vert)*

Nous donnons maintenant un argument heuristique suggérant que l'estimateur par histogramme est une procédure pour laquelle les risques L_2 introduits ci-dessus pourraient efectivement tendre vers zéro quand n diverge vers l'infini. Notons d'abord que la loi des grands nombres assure que $\widehat{P}_j$ converge presque sûrement vers

$$P_j = \mathbb{E}\left[\mathbf{1}\left[X_1 \in \mathcal{I}_j\right]\right] = \mathbb{P}\left[X_1 \in \mathcal{I}_j\right] = \int_{\mathcal{I}_j} f(x)dx.$$

Par ailleurs, si f est continue, alors le théorème de la moyenne montre que, pour un certain $\theta_j \in \mathcal{I}_j$, on a $P_j/h = f(\theta_j)$, de sorte que si m diverge vers l'infini, on a :

$$P_j/h\mathbf{1}\left[x \in \mathcal{I}_j\right] \to f(x)\mathbf{1}\left[x \in \mathcal{I}_j\right].$$

Ceci suggère que, pour n grand,

$$\widehat{f}_n(x) = \sum_{j=1}^{m} \widehat{P}_j/h\mathbf{1}\left[x \in \mathcal{I}_j\right] \approx \sum_{j=1}^{m} P_j/h\mathbf{1}\left[x \in \mathcal{I}_j\right] \approx \sum_{j=1}^{m} f(x)\mathbf{1}\left[x \in \mathcal{I}_j\right] = f(x).$$

Ce qui tendrait a montrer que $\widehat{f}_n(x)$ est un estimateur convergent de f , pour que $m = m_n \to \infty$. Nous insistons sur le caractère heuristique de cet argument (en particulier, l'usage de la loi des grands nombres ci-dessus requiert que m soit fixé, alors que la suite de l'argument demande que m diverge vers l'infini). Le résultat principal pour cette section est le théorème ci-dessous, qui etablit un résultat précis sur le risque intégré de l'estimateur par histogramme. Son importance réside aussi dans le fait qu'il nous permettra de préciser l'impact du paramètre-clé à choisir quand on utilise cet estimateur, à savoir le nombre m de cellules.

Théorème 3.1 *Supposons que f soit nulle en dehors de $[a;b]$ et soit de classe C^2 sur $[a;b]$. Soit $m = m_n$ une suite d'entiers strictement positifs qui diverge vers l'infini et soit $h_n = (b-a)/m_n$. Alors, le risque intégré de l'estimateur par histogramme satisfait*

$$R\left(\widehat{f}_n, f\right) = \frac{h_n^2}{12}\left\|f'\right\|_2^2 + \frac{1}{nh_n} + \circ\left(h_n^2\right) + \circ\left(\frac{1}{nh_n}\right) \tag{3.3}$$

quand n diverge vers l'infini, où $\|g\|_2 = \left(\int_a^b \left(g(x)\right)^2 dx\right)^{1/2}$ est la norme L2 de g sur $[a;b]$.

ce théorème a de nombreuses conséquences intéressantes. Tout d'abord, il montre que, pour une densité f de classe C^2, le risque intégré $R\left(\widehat{f}_n, f\right)$ de l'estimateur par histogramme tend vers zéro si et seulement si $m = m_n$ mène à

$$(i)\ h_n \to 0 \ \text{ et } \ (ii)\ \tfrac{1}{nh_n} \to 0\ .$$

(ou, formulé en termes de la suite (m_n), si et seulement si (i) $m_n \to \infty$ et (ii) $m_n/n \to 0$). Ceci indique que (i) le nombre de cellules doit bien entendu diverger vers l'infini pour obtenir un estimateur convergent, mais que (ii) ce nombre de cellules ne peut pas diverger trop vite vers l'infini. Le point (ii) est egalement raisonnable : pour que l'estimation de la masse de probabilité dans chaque cellule se fasse de façon convergente (par le mécanisme de la loi des grands nombres), il faut que le nombre d'observations par cellule diverge vers l'infini plus vite que le nombre de cellules (si l'inverse se produit, les cellules seront finalement vides quand

n diverge vers l'infini, ce qui exclut qu'on puisse estimer la densité dans chaque cellule). Pour montrer que le risque intégré peut être minimisé en choisissant $m = m_n$ de façon adéquate, nous avons conduit l'expérience empirique suivante. Pour chaque taille d'échantillon $n \in \{300; 600; 1000; 2000\}$ et chaque nombre de cellules $m \in \{10; 20; 30; \cdots, 590; 600\}$, nous avons estimé le risque intégré $R\left(\widehat{f}_n, f\right)$, (nous avons pris $a = -5$ et $b = 5$).

On peut tenter d'identifier la valeur de m qui minimise le risque intégré apparaissant dans le Théorème 3.1. Ceci fournit le résultat suivant

Théorème 3.2 *Sous les conditions du Théorème* 3.1*, la valeur de $h = h_n$ qui minimise le risque intégré est*

$$h_n^{opt} = C_f n^{-1/3},$$

où $C_f = \frac{1}{6}\left(\|f'\|_2^2\right)^{-1/3}$, (la valeur optimale correspondante de $m = m_n$ et donc $m_n^{opt} = (b-a) \,/\, h_n^{opt}$. Le risque intégré minimal associé est de la forme

$$R^{opt}\left(\widehat{f}_n, f\right) = D_f n^{-2/3} + \circ\left(n^{-2/3}\right),$$

quand n diverge vers l'infini, où $D_f = \left(\frac{9}{16}\|f'\|_2^2\right)^{1/3}$.

Ce résultat montre en particulier que, lorsqu'il est fondé sur la valeur optimale de m_n (de façon equivalente, de h_n), le risque intégré de l'estimateur par histogramme converge vers zéro a la vitesse $n^{-2/3}$.

3.3 Estimation par noyau

L'estimateur par histogramme souffre d'un certain nombre de défauts importants. En particulier, il prend comme valeur une fonction discontinue, et ce même si la densité f qu'il estime est trés lisse. Par ailleurs, comme annoncé ci-dessus, il ne fournit pas le plus petit risque intégré possible, même lorsque le nombre de cellules utilisé a eté choisi de façon optimale. Ceci fournit une motivation importante pour présenter et etudier les *estimateurs a noyau,* qui ne souffriront pas des défauts.

3.3.1 Définition

Les estimateurs à noyau sont construits en exploitant le fait que, en tout point x, la fonction de densité f est la dérivée de la fonction de répartition correspondante $x \to F(x) = \mathbb{P}(X \leq x)$ au point x, et satisfait donc

$$f(x) = F'(x) = \lim_{\substack{> \\ h \to 0}} \frac{F(x+h) - F(x-h)}{2h} \tag{3.3}$$

L'idée naturelle consiste a définir un estimateur $\widehat{f}_n$ de $f(x)$ en remplaçant F dans (3.3) par la fonction de répartition empirique $\widehat{F}_n$, mais ceci est exclu car $\widehat{F}_n$ n'est pas dérivable (ceci ménerait à un estimateur $\widehat{f}_n$ qui serait nul partout sauf en un nombre fini de valeurs de x pour lesquelles $\widehat{f}_n(x)$ ne serait pas défini).

Par contre, (3.3) implique que, pour $h > 0$ petit, on a

$$f(x) \approx \frac{F(x+h) - F(x-h)}{2h} \tag{3.4}$$

ce qui suggère d'estimer $f(x)$ par

$$\begin{aligned} f(x) &= \tfrac{\widehat{F}_n(x+h)-\widehat{F}_n(x-h)}{2h} \\ &= \tfrac{1}{2h}\left(\tfrac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[X_i \le x+h\right] - \tfrac{1}{n}\sum_{i=1}^{n}\mathbf{1}\left[X_i \le x-h\right]\right) \\ &= \tfrac{1}{2nh}\sum_{i=1}^{n}\mathbf{1}\left[x-h < X_i \le x+h\right] \\ &= \tfrac{1}{nh}\sum_{i=1}^{n}K\left(\tfrac{x-X_i}{h}\right), \end{aligned}$$

où h est petit et où $z \to K(z) = \frac{1}{2}\mathbf{1}\left[-1 < z \le 1\right]$ est la fonction de densité de la loi uniforme sur $[1;1)$ (ou $[1;1]$). Si cet estimateur $\widehat{f}_n$ va hériter du caracère discontinu de la fonction de densité K ci-dessus, on peut penser a remplacer K par une autre densité continue, ce qui rendra continu l'estimateur $\widehat{f}_n$ associé.

Définition 3.2 *Un estimateur a noyau de f est un estimateur de la forme :*

$$\widehat{f}_n = \frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-X_i}{h}\right) \tag{3.5}$$

où K est un noyau, c'est- a-dire une fonction de $\mathbb{R}$ dans $\mathbb{R}$ telle que

(i) $\int_{-\infty}^{\infty} K(z)dz = 1$

(ii) $\int_{-\infty}^{\infty} zK(z)dz = 0$

(iii) $\int_{-\infty}^{\infty} K^2(z)dz < \infty$.

Outre le noyau uniforme ci-dessus, les noyaux les plus communément utilisés sont présentés à la Figure (3.4). On vérifie directement que, quelles que soient les valeurs prises par les variables aléatoires $X_1, \cdots, X_n$, on a

$$\int_{-\infty}^{\infty} \widehat{f}_n(x)dx = 1.$$

Par ailleurs, si $K(z) \geq 0$ pour tout z , alors on a trivialement $\widehat{f}_n(x) \geq 0$ pour tout x, auquel cas $\widehat{f}_n$ est donc une fonction de densité (ce qui est naturel, car on estime alors une fonction de densité par une fonction de densité). Comme on le verra par la suite, utiliser un noyau K négatif par endroit peut cependant présenter certains avantages en termes d'eficacité. Finalement, comme indiqué ci-dessus, si le noyau K est continu, alors $\widehat{f}_n$ l'est egalement ; plus généralement, $\widehat{f}_n$ hérite de la régularité (C^k, C^∞,ect.) du noyau K utilisé. Ceci est illustré a la Figure (3.5), qui représente, pour le noyau gaussien et le noyau uniforme, les estimateurs à noyau obtenus pour un echantillon de taille $n = 1000$ engendré depuis la densité de Bart Simpson. Le noyau gaussien fournit un estimateur à noyau lisse, tandis que le noyau uniforme fournit un estimateur a noyau discontinu.

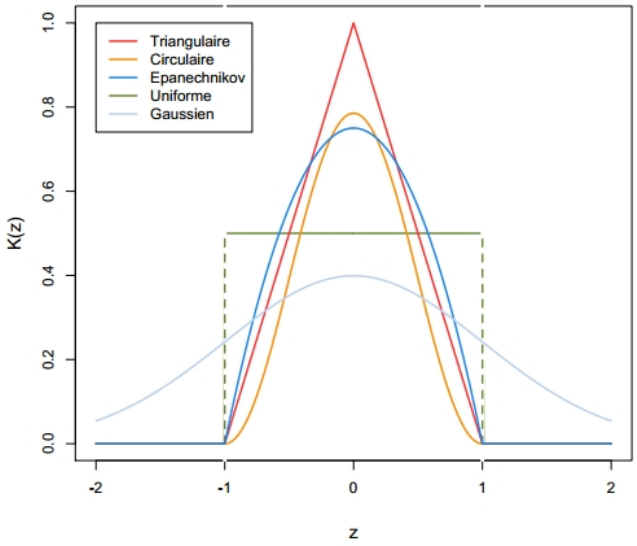

Figure (3.4) - *Graphes des noyau triangulaire* $(1 - |z|)1\left[-1 < z \leq 1\right]$, *noyau circulaire* $\frac{\pi}{4}\cos(\frac{\pi}{2}z)1\left[-1 < z \leq 1\right]$, *noyau d'Epanechnikov* $\frac{3}{4}\left(1 - z^2\right)1\left[-1 < z \leq 1\right]$, *noyau uniforme* $\frac{1}{2}1\left[-1 < z \leq 1\right]$, *et noyau gaussien* $K(z) = (2\pi)^{-1/2}\exp\left(-z^2/2\right)$.

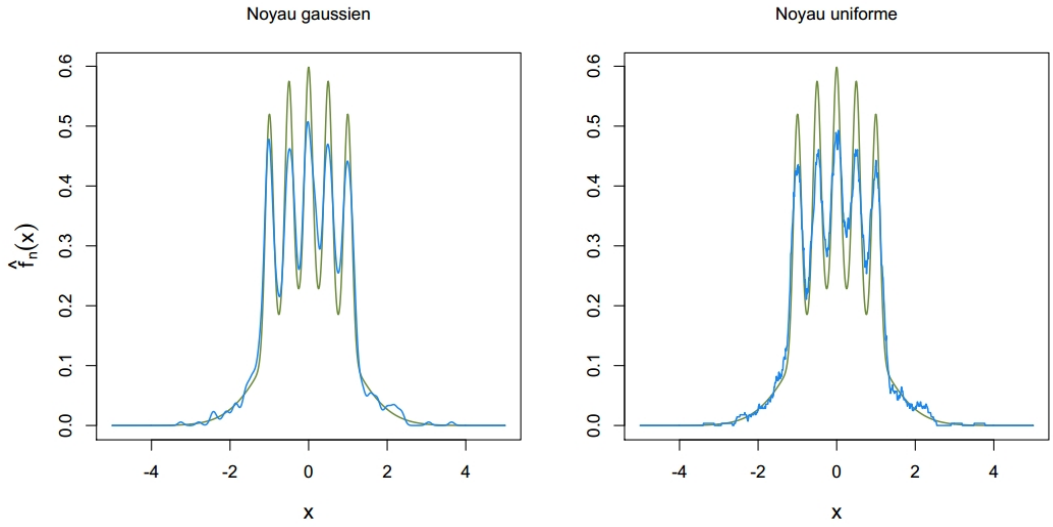

Figure (3.5)- *Graphes de l'estimateur à noyau $\widehat{f_n}$ calculé sur un échantillon aléatoire de taille $n = 1000$ engendré depuis la densité de Bart Simpson, en utilisant le noyau gaussien et $h = .07$ (gauche) ou le noyau uniforme et $h = .14$ (droite).*

Il devrait être clair que pour que $\widehat{f_n}$ soit un estimateur convergent, il faut non seulement que le nombre d'observations n diverge vers l'infini, mais aussi qu'on utilise une taille de fenêtre $h = h_n$ qui tende vers zéro (par valeurs plus grandes, puisque h doit être strictement positif). Cette convergence ne peut pas avoir lieu de manière arbitraire, comme le suggère l'argument heuristique suivant. Supposons, pour simplifier, que le noyau utilisé soit à support compact, et plus spécifiquement, que $\{z \in \mathbb{R} : K(z) > 0\} =]-c; c[$, disons. Dans l'estimation de $f(x)$ par $\widehat{f_n}$, l'observation X_i va alors apporter une contribution (au sens où le $i$ème terme du membre de droite de (3.5) va prendre une valeur strictement positive) si et seulement si

$$\left|\frac{x-X_i}{h_n}\right| < c,$$

c'est- a-dire si et seulement si $X_i \in]x - ch_n; x + ch_n[$ .Une observation X_i est donc "active" si et seulement si elle prend sa valeur dans une "fenêtre" centrée en x et de largeur $2ch_n$, ce qui explique la terminologie taille de la fenêtre. Si f est continue en x, alors le nombre moyen d'observations actives est :

$$\mathbb{E}\left[\sum_{i=1}^{n} \mathbf{1}\left[X_i \in]x - ch_n; x + ch_n[\right]\right] = \sum_{i=1}^{n} \mathbb{P}\left[X_i \in]x - ch_n; x + ch_n[\right].$$

$$= n\mathbb{P}\left[X_1 \in]x - ch_n; x + ch_n[\right] = n \int_{x-ch_n}^{x+ch_n} f(z)dz = 2cnh_n\left(f(x) + \circ(1)\right),$$

quand n diverge vers l'infini (la dernière egalité est obtenue en utilisant le théorème de la moyenne puis la continuité de f en x). Clairement, une estimation convergente requiert que ce nombre moyen d'observations actives diverge vers l'infini, ce qui impose que nh_n diverge vers l'infini. La taille de la fenêtre doit ainsi converger vers zéro, mais pas trop vite. Que ceci est une condition nécessaire pour la convergence de $\widehat{f}_n(x)$ sera confirmé théoriquement ci-dessous.

3.3.2 Risques ponctuel et intégré

Contrairement a l'estimateur par histogramme, l'estimateur a noyau permet d'obtenir de façon simple une expression du risque L_2 ponctuel.

Théorème 3.3 *Supposons que f soit de classe C^3 sur $\mathbb{R}$ et que $\|f'\|_\infty$ et $\|f'''\|_\infty$ soient finies (où $\|g\|_\infty = \sup_{x\in\mathbb{R}} |g(x)|$). soit K un noyau tel que $\int_{-\infty}^{\infty} |z|^3 |K(z)|\, dz < \infty$ et $\int_{-\infty}^{\infty} |z|\, K^2(z)dz < \infty$. soit $h = h_n$ une suite qui est $\circ(1)$ et telle que nh_n diverge vers l'infini. Soit $\widehat{f}_n$ l'estimateur a noyau de f associé a K et à h_n. posons $d_K := \int_{-\infty}^{\infty} z^2 K(z)dz$ et $c_k := \|K\|_2^2$. Alors, pour tout $x \in \mathbb{R}$, on a*

quand n diverge vers l'infini,

$$Biais\left(\widehat{f}_n\right) = \tfrac{h_n^2}{2} f''(x) d_K + \circ\left(h_n^2\right) \text{ et } Var\left(\widehat{f}_n\right) = \tfrac{1}{nh_n} f(x) c_K + \circ\left(\tfrac{1}{nh_n}\right).$$

De sorte que

$$MSE\left(\widehat{f}_n\right) = \frac{h_n^4}{4}\left(f''(x)\right)^2 d_K^2 + \frac{1}{nh_n} f(x) c_K + \circ\left(h_n^4\right) + \circ\left(\frac{1}{nh_n}\right) \tag{3.6}$$

Ce théorème permet d'appréhender l'impact de la taille de la fenêtre h_n sur les propriétés de biais et de variance de l'estimateur a noyau. Clairement, il faut choisir h_n petit pour obtenir un petit biais, mais cela conduira a une grande variance ; au contraire, il faut prendre h_n grand pour obtenir une petite variance, mais cela ménera a un biais important. Ce comportement en h, qui est illustré a la Figure (3.6), suggère qu'il ne faut choisir h_n ni trop petit ni trop grand pour obtenir un risque optimal, réalisant un équilibre entre biais et variance. Ceci est confirmé a la Figure (3.7), qui graphe un risque (intégré) estimé en fonction de h_n pour les estimateurs a noyau fondés sur un noyau gaussien et un noyau uniforme. Il est à noter que le risque n'est pas nécessairement convexe en h_n, comme le montre l'exemple du noyau uniforme.

Un calcul direct montre que le risque ponctuel (MSE) en (3.6) est minimisé lorsque

$$h_n = \left(\frac{f(x) c_K}{\left(f''(x)\right)^2 d_K^2}\right)^{1/5} n^{-1/5} \tag{3.7}$$

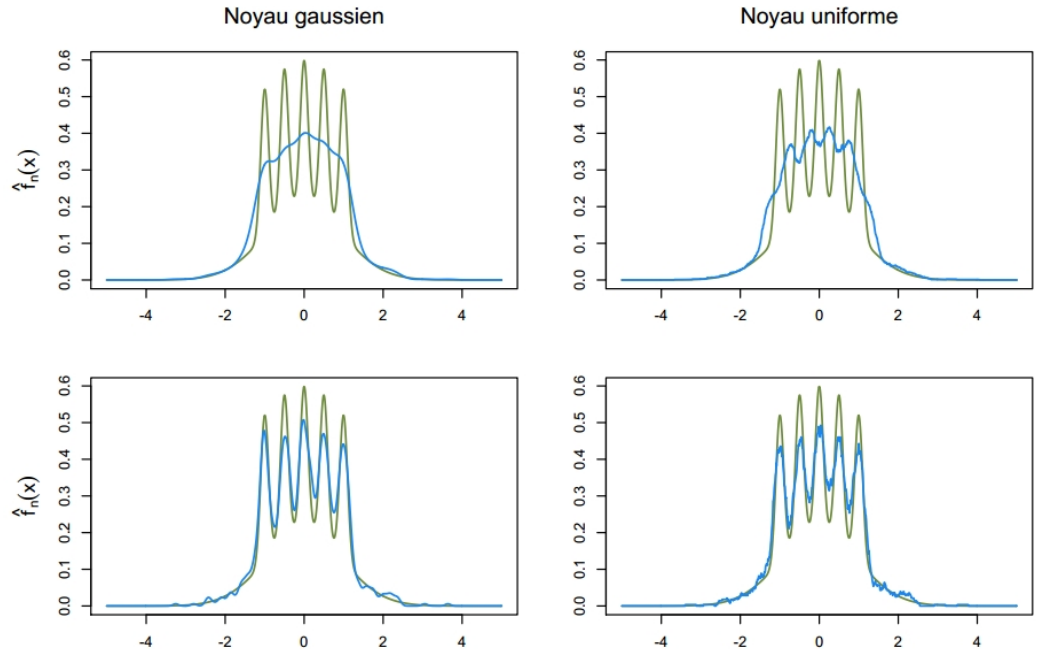

Figure (3.6) - *(Gauche :) Graphes des estimateurs a noyau* $\widehat{f_n}$ *calculés sur un échantillon aléatoire de taille* $n = 1000$ *engendré depuis la densité de Bart Simpson, en utilisant le noyau gaussien . (Droite :) Les graphes correspondants obtenus en utilisant le noyau uniforme et et, pour* $h_1 = 0.21$ *(haut),* $h_2 = 0.07$ *(bas).*

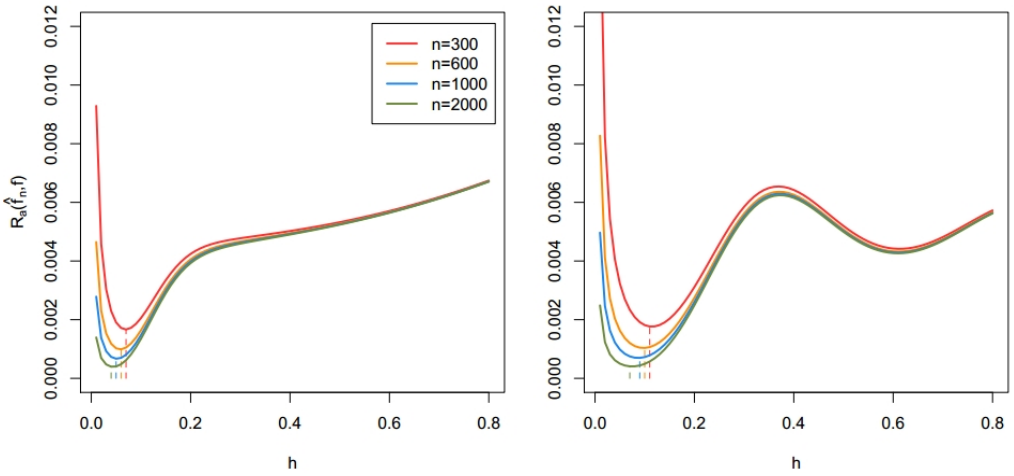

Figure (3.7) - *Pour différentes tailles d'échantillon* n*, graphes, en fonction de* h*, de l'estimation* $MSE\left(\widehat{f_n}\right)$*, le noyau gaussien (gauche) ou le noyau uniforme (droite), la valeur de* h *fournissant le risque minimal est mise en evidence par un trait en pointillés.*

Nous considérons maintenant le risque intégré de l'estimateur a noyau, qui nous permettra de comparer les performances de cet estimateur avec celles de l'estimateur par histogramme.

Théorème 3.4 *Supposons que f soit de classe C^2sur $\mathbb{R}$ et que $\left\|f''\right\|_2 < \infty$. Soit K un noyau non négative et tel que $\int_{-\infty}^{\infty} z^2 K(z)dz < \infty$. Soit $\widehat{f}_n$ l'estimateur à noyau de f associé à K et à h. Posons $d_K := \int_{-\infty}^{\infty} z^2 K(z)dz$ et $c_k := \|K\|_2^2$. Alors, on a :*

$$(i) \quad R\left(\widehat{f}_n, f\right) \leq \tfrac{h^4}{3}\left\|f''\right\|_2^2 d_K^2 + \tfrac{1}{nh} c_K,$$

(ii) Par conséquent, si $h = h_n = \zeta n^{-1/5}$, on a avec $F_M := \left\{f : \left\|f''\right\|_2 \leq M\right\}$,

$$\sup_{f \in \mathcal{F}_M} R\left(\widehat{f}_n, f\right) \leq C n^{-4/5} \tag{3.8}$$

pour une constante C qui ne dépend que de ζ, de K et de M.

Ce théorème indique que le risque intégré de l'estimateur a noyau tend vers zéro a la vitesse $n^{-4/5}$, qui est plus rapide que la vitesse $n^{-2/3}$ réalisée par l'estimateur par histogramme (voir théorème (3.2). L'estimateur a noyau jouit donc d'un taux de convergence meilleur. Une question naturelle est celle de l'optimalité de ce taux de convergence. En d'autres termes, existe-t-il un estimateur de densité pour lequel le risque intégré convergerait plus vite vers zéro que $n^{-4/5}$?. Comme l'indique le résultat suivant, la réponse est négative.

Théorème 3.5 *Il existe une constante $D = D_M$ telle que, avec $\mathcal{F}_M := \left\{f : \left\|f''\right\|_2 \leq M\right\}$, on ait :*

$$\sup_{f \in \mathcal{F}_M} R\left(\widehat{f}_n, f\right) \geq D n^{-4/5},$$

quel que soit l'estimateur de densité $\widehat{f}_n$.

Remarque 3.1 *Nous ne présentons pas la preuve de ce résultat, qui est très technique (le lecteur intèressè pourra trouver une preuve dans* [1]*). Ce résultat, qui garantit que les estimateurs à noyau sont optimaux en termes de taux de convergence, explique, bien entendu, le succés important de ces estimateurs.*

Remarque 3.2 *Le Théorème* 3.5 *peut essentiellement être lu en disant que si on se restreint a la collection des densités de classe C^2, alors le taux de convergence optimal est $n^{-4/5}$. On peut montrer que si on se restreint à des collections de densités plus régulières, le taux de convergence optimal s'améliore. On a le résultat suivant.*

Théorème 3.6 *Supposons que f soit de classe C^ksur $\mathbb{R}$ et que $\left\|f^{(k)}\right\|_2 < \infty$. Soit K un noyau non négative et tel que $\int_{-\infty}^{\infty} z^r K(z)dz = 0$, pour tout $r = 1, \cdots, k-1$, et*

$\int_{-\infty}^{\infty} |z|^k K(z)dz < \infty$. *Soit* $\widehat{f}_n$ *l'estimateur à noyau de* f *associé à* K *et à* h. *Alors, on a :*
$(i) \quad R\left(\widehat{f}_n, f\right) \leq C\left(h^{2k} + \frac{1}{nh}\right),$
pour une certain constante C *, de sorte que si* $h = h_n = \zeta n^{-1/(2k+1)}$, *on a avec*

$$F_{k,M} := \left\{f : \left\|f^{(k)}\right\|_2 \leq M\right\}, \sup_{f \in F_{k,M}} R\left(\widehat{f}_n, f\right) \leq C n^{-2k/(2k+1)}.$$

(ii) *Il existe une constante* $D = D_M$ *telle qu'on ait*

$$\sup_{f \in \mathcal{F}_{k,M}} R\left(\widetilde{f}_n., f\right) \geq D n^{-2k/(2k+1)},$$

quel que soit l'estimateur de densité $\widetilde{f}_n$.

Pour les densités de classe C^k, le taux de convergence optimal est donc $n^{-2k/(2k+1)}$, qui devient arbitrairement proche du taux de convergence param etrique n^{-1} quand k diverge vers l'infini. Pour $k > 2$, réaliser le taux de convergence $n^{-2k/(2k+1)}$ requiert un noyau d'ordre supérieur, au sens où K doit annuler des intégrales du type $\int_{-\infty}^{\infty} z^r K(z)dz$ pour $r \geq 2$. Ceci nécessite bien entendu que K soit négatif par endroit, ce qui pourrait, dans certains cas, donner lieu à des estimateurs a noyau $\widehat{f}_n$ qui sont aussi négatifs par endroit.

3.3.3 Choix de K et de h_n

Estimer f par un estimateur a noyau demande de choisir deux quantités : le noyau K et la taille de la fenêtre $h = h_n$. Dans cette section, nous décrivons des procédures qui permettent d'effectuer ces choix, en particulier pour le paramètre crucial h_n.

Choix de K

Considérons d'abord le choix du noyau K. Comme on l'a vu en (3.6), le risque ponctuel (MSE) associé à l'estimateur a noyau utilisant le noyau K et la taille de fenêtre h_n optimale est donné par :

$$MSE\left(\widehat{f}_n\right) = \frac{5}{4}\left(f^2(x)f''(x)c_K^2 d_K\right)^{2/5} n^{-4/5} + r_n, \tag{3.9}$$

où r_n est un terme de reste. Il est important de noter que quand on construit un estimateur à noyau, les noyaux $K(z)$ et $K_\sigma(z) = \frac{1}{\sigma}K(\frac{z}{\sigma})$ sont equivalents pour tout $\sigma > 0$: si on prend $h_\sigma = \frac{h}{\sigma}$, on a en effet :

$$\widehat{f}_n^{K_\sigma, h_\sigma}(x) = \frac{1}{nh_\sigma}\sum_{i=1}^{n} K_\sigma\left(\frac{x - X_i}{h_\sigma}\right)$$

$$= \frac{1}{nh_\sigma\sigma}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h_\sigma\sigma}\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) = \widehat{f}_n^{K,h}(x),$$

ce qui montre que l'estimateur a noyau $\widehat{f}_n$ fondé sur $(K;h)$ coïncide avec celui fondé sur $(K_\sigma;h_\sigma)$ Plutôt qu'un noyau K, c'est donc une classe de noyaux K, définis à une échelle σ près, qu'il convient de choisir. En ligne avec ceci, on vérifiera facilement que le facteur dépendant du noyau dans le terme principal du risque en (3.9) satisfait $c^2_{K_\sigma} d_{K_\sigma} = c^2_K d_K$. Par conséquent, il n'y a pas de perte de généralité à supposer que l' échelle de K a été fixée de sorte que $d_K = 1$. En se restreignant à de tels noyaux, le risque ci-dessus s'écrit :

$$MSE\left(\widehat{f}_n\right) = \tfrac{5}{4}\left(f^2(x)f''(x)c^2_K\right)^{2/5} n^{-4/5} + r_n.$$

Choisir le noyau de façon optimale consiste donc a minimiser c_K dans la classe des noyaux satisfaisant $d_K = 1$, ce qui mène au problème variationnel visant a minimiser $c_K = \int_{-\infty}^{\infty} z^2 K(z)dz$ dans la classe des fonctions $K : \mathbb{R} \to \mathbb{R}$ satisfaisant $\int_{-\infty}^{\infty} K(z)dz = 1$, $\int_{-\infty}^{\infty} zK(z)dz = 0$ et $d_K = \int_{-\infty}^{\infty} z^2 K(z)dz = 1$. Les outils du calcul variationnel permettent de montrer que le noyau optimal est donné par :

$$K(z) = \tfrac{3}{4\sqrt{5}}\left(1 - \tfrac{z^2}{5}\right)\mathbf{1}\left[-\sqrt{5} \le z \le \sqrt{5}\right],$$

qui, a un facteur d'échelle près, est le noyau d'Epanechnikov. On peut donc conclure que le noyau d'Epanechnikov est optimal.

Au vu de ce résultat d'optimalité, il peut sembler étonnant, a première vue, qu'il soit de pratique courante d'utiliser d'autres noyaux (comme le noyau gaussien par exemple). La raison est que, si le noyau d'Epanechnikov fournira efectivement le plus petit risque, le gain en termes de risque reste minimal. Pour le montrer, revenons a (3.9), qui rend clair que l'impact du noyau sur le risque ponctuel est mesuré par le facteur $\left(c^2_K d_K\right)^{2/5}$. Par conséquent, le bénéfice à utiliser le noyau d'Epanechnikov $K = K_{Epan}$ plutôt qu'un noyau K peut être mesuré par le ratio :

$$\frac{\left(c^2_{K_{Epan}} d_{K_{Epan}}\right)^{2/5}}{\left(c^2_K d_K\right)^{2/5}}$$

Ce ratio prend les valeurs 0.989 pour le noyau triangulaire, 0.999 pour le noyau circulaire, 0.943 pour le noyau uniforme, et 0.961 pour le noyau gaussien. Le gain en termes de risque à utiliser le noyau d'Epanechnikov plutôt que le noyau gaussien est donc de moins de 4%, ce qui explique qu'en pratique, on préfèrera souvent les estimateurs très lisses qui résultent de l'utilisation du noyau gaussien.

Choix de h_n

Si le noyau d'Epanechnikov est optimal, le gain par rapport aux autres noyaux est marginal, et le choix du noyau n'est donc pas prépondérant. Au contraire, le choix de h_n est crucial car il a un impact énorme sur l'estimateur a noyau $\widehat{f}_n$. Il est donc capital de disposer d'une méthode permettant de faire ce choix en pratique.

La plupart des méthodes disponibles dans la littérature visent a choisir h_n de façon à minimiser le risque intégré $R\left(\widetilde{f}_n., f\right)$. Si le Théorème (3.4) ne donne qu'une borne supérieure sur ce risque, il est facile de montrer que, sous les hypothèses du Théorème 3.3,

$$R\left(\widetilde{f}_n., f\right) = \frac{h_n^4}{4} \left\|f''\right\|_2^2 d_K^2 + \frac{1}{nh_n} c_K + \circ\left(h_n^4\right) + \circ\left(\frac{1}{nh_n}\right) \tag{3.10}$$

quand n diverge vers l'infini. On vérifie directement que la taille de fenêtre qui minimise ce risque intégré est donnée par :

$$h_n = \left(\frac{c_k}{\left\|f''\right\|_2^2 d_K^2}\right)^{1/5} n^{-1/5} \tag{3.11}$$

Le fait que f soit inconnue empêche d'utiliser cette formule en pratique. Nous présentons deux méthodes permettant de faire le choix de h_n.

(A) La première méthode consiste a se comporter comme si la densité f inconnue était gaussienne, de moyenne μ et de variance σ^2, disons. Dans ce cas, on vérifie que $\left\|f''\right\|_2^2 = 3/\left(8\sqrt{\pi}\sigma^5\right)$, ce qui mène à estimer (3.11) par :

$$h_n = \widehat{\sigma}\left(\frac{8\sqrt{\pi}c_k}{3d_K^2}\right)^{1/5} n^{-1/5},$$

où $\widehat{\sigma}$ est un estimateur de σ Pour le noyau gaussien, ceci donne :

$$h_n \approx 1,06\widehat{\sigma}n^{-1/5}.$$

Il est naturel d'estimer σ par l' ecart-type empirique $\widehat{\sigma}$ où $\widehat{\sigma}^2 \frac{1}{n}\sum_{i=1}^n \left(X_i - \overline{X}\right)^2$, avec $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$.

. Néanmoins, cet estimateur est peu robuste. Il est alors courant d'utiliser

$$\widetilde{\sigma} = \min\left(\widehat{\sigma}, \tfrac{IR}{1.34}\right),$$

où IR désigne la longueur de l'intervalle inter-quartile, c'est- à-dire la différence entre les quantiles d'ordre 75% et d'ordre 25% de $X_1, \cdots, X_n$ (le facteur 1.34 assure que $IR/1.34$ estime bien l'écart-type population si la loi sous-jacente est effectivement normale). La taille de fenêtre finale qui en résulte est donc

$$h_n \approx 1,06 \min\left(\widehat{\sigma}, \frac{IR}{1.34}\right) n^{-1/5}, \tag{3.12}$$

Cette procédure repose explicitement sur le fait que f est normale, ce qui est incompatible avec l'approche non paramétrique. Elle donnera de bons résultats si la densité inconnue n'est pas trop différente d'une loi normale.

(B) La seconde méthode est connue sous le nom de validation croisée (en anglais, cross-validation). En notant par $\widehat{f}_{n,h}$ l'estimateur a noyau utilisant la taille de fenetre h, elle vise a choisir la valeur de h qui minimise l'erreur quadratique intégrée (en anglais, Integrated Square Error)

$$ISE_n(h) = \int_{-\infty}^{\infty} \left\{ \widehat{f}_{n,h}(x) - f(x) \right\}^2 dx$$
$$= \int_{-\infty}^{\infty} \widehat{f}_{n,h}^2(x)dx - 2\int_{-\infty}^{\infty} \widehat{f}_{n,h}(x)f(x)dx + \int_{-\infty}^{\infty} f^2(x)dx,$$

ou, de manière équivalente (puisque le dernier terme ne dépend pas de h),

$$ISE_{n0}(h) = \int_{-\infty}^{\infty} \widehat{f}_{n,h}^2(x)dx - 2\int_{-\infty}^{\infty} \widehat{f}_{n,h}(x)f(x)dx = T_{n1} - 2T_{n2} \tag{3.13}$$

En principe, T_{n1} peut être evalué sur la base des observations $X_1, \cdots, X_n$, mais il sera plus commode de réécrire ce terme sous une forme qui ne demande pas d'évaluer une intégrale. Pour ce faire, on ecrit

$$T_{n1} = \int_{-\infty}^{\infty} \widehat{f}_{n,h}^2(x)dx = \int_{-\infty}^{\infty} \left(\frac{1}{(nh)^2} \sum_{i,j=1}^{n} K\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_j}{h}\right) \right) dx$$
$$= \frac{1}{(nh)^2} \sum_{i,j=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{x-X_i}{h}\right) K\left(\frac{x-X_j}{h}\right) dx,$$

ce qui, en faisant le changement de variable $x = X_i - hz$ donne

$$T_{n1} = \frac{1}{n^2 h} \sum_{i,j=1}^{n} \int_{-\infty}^{\infty} K\left(\frac{X_i - X_j}{h} - z\right) K(-z)dz = \frac{1}{n^2 h} \sum_{i,j=1}^{n} \left(K * \overline{K}\right)\left(\frac{X_i - X_j}{h}\right),$$

où $(K*L)(x) = \int_{-\infty}^{\infty} K(x-z)L(z)dz$ est la convolution des fonctions K, L, et où $\overline{K} = K(-z)$ (on peut vérifier que pour le noyau gaussien, on a $\left(K * \overline{K}\right)(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-z^2/4\right)$, tandis que pour le noyau uniforme, on a $\left(K * \overline{K}\right)(z) = \frac{1}{4}\left\{(2+Z)\,\mathbf{1}\left[-2 < z < 0\right] + (2-Z)\,\mathbf{1}\left[0 < z < 2\right]\right\}$). La quantité T_{n2} fait par contre intervenir la densité f inconnue et doit donc être estimée. Pour ce faire, notons que

$$T_{n2} = \int_{-\infty}^{\infty} \widehat{f}_{n,h}(x)f(x)dx = \mathbb{E}\left[\widehat{f}_{n,h}(X)|X_1, \cdots, X_n\right],$$

où X est indépendant de $X_1, \cdots, X_n$ et admet aussi la densité f . Par conséquent, en notant $\widehat{f}_{n,h}^{(-i)}$ l'estimateur a noyau fondé sur $X_1, \cdots, X_{i-1}, X_{i+1}, \cdots, X_n$ et la taille de fenêtre h, on peut estimer cette quantité par

$$\widehat{T}_{n2} = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_{n,h}^{(-i)}(X_i) = \frac{1}{n}\sum_{i=1}^{n} \left\{ \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j\neq i}}^{n} K\left(\frac{X_i - X_j}{h}\right) \right\}$$
$$= \frac{1}{n(n-1)h} \sum_{\substack{i,j=1 \\ j\neq i}}^{n} K\left(\frac{X_i - X_j}{h}\right).$$

La principe de validation croisée consiste donc a choisir la taille de fenêtre

$$\widehat{h}_n^{CV} = \arg\min_{h>0} \widehat{ISE}_{n0}(h) \tag{3.14}$$

où

$$\widehat{ISE}_{n0}(h) = \frac{1}{n^2 h} \sum_{i,j=1}^{n} \left(K * \overline{K}\right)\left(\frac{X_i - X_j}{h}\right) - \frac{2}{n(n-1)h} \sum_{\substack{i,j=1 \\ j\neq i}}^{n} K\left(\frac{X_i - X_j}{h}\right).$$

De façon remarquable, il a et e prouvé (voir [2]) que, si la densité f est bornée, alors

$$\frac{\widehat{ISE}_n(\widehat{h}_n^{CV})}{\min_{h>0} ISE_n(h)} \to 1.$$

presque sûrement quand n diverge vers l'infini. Ceci indique que la procédure de validation croisée livre une taille de fenêtre qui fournit une erreur quadratique intégrée qui coïncide asymptotiquement avec l'erreur quadratique intégrée minimale.

Conclusion

L'estimateur d'une densité de probabilité par la méthode du noyau a connu un très grand succés parmi les estimateurs non paramétriques, ceci est dû sa simplicité et sa convergence vers la densité f pour tous les modes (convergence dans $L1$, presque sûre, en probabilité et en moyenne quadratique). L'estimation à noyau est une méthode non paramétrique basée sur l'utilisation d'une fonction appelée noyau et d'un paramètre de lissage ou fenêtre. On remarque que le choix sur le noyau qui n'a pas une grande influence pour cette estimation, par contre le choix du paramètre de lissage a un impact important, et qui est en effet, beaucoup plus déterminant pour l'obtention des bons estimateurs.

Bibliographie

[1] van der Vaart, A.W. (1998). Asymptotic Statistics. Cambridge University Press, New York.

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

[3] Rosenblatt, M. (1956). Estimation of a probability density-function and mode. Ann Math Statist, 27, 832-837.

[4] Parzen, E. (1962). On estimation of a probability density function and mode. The annals of mathematical statistics, 33(3), 1065-1076.

[5] Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Chap-man Hall, London.

[6] Veysseyre, R., 2006 : Aide-mémoire Statistique et probabilités pour l'ingénieur. Dunod.

[7] Dusart, P., 2015 : Cours de Statistiques inférentielles.Licence 2-S4 SI-Mass.

[8] Belahcene, I., 2017 :Mémoire Master Estimation non naramétrique de La fonction densité de probabilité avec un noyau. Université de Ouargla.

[9] Nadaraya, E. A., 1964. On Estimating Regression.Theory of Probability and its Applications, 1964 ; 9 : 1 ; 141 142.

[10] STAT., 2413 ; 2002 2003 : Chapitre 3 Estimation non paramétrique d'une fonction de répartition et d'une densité.