République Algérienne Démocratique et Populaire

Ministère de l'Enseignement Supérieur et de la Recherche Scientifique

Université A. Mira de Béjaïa

Faculté des Sciences Exactes

Département de Mathématiques

Mémoire de Fin d'Etudes

En vue de l'obtention du diplôme de Master en Mathématiques.
Spécialité Probabilités Statistique et Applications

Thème

Estimation non paramétrique de la fonction densité par la méthode des k plus proches voisins

---

Réalisé par M^{elle} :

FEDDAL Ihssane

Soutenu le 29 Juin devant le jury composé de :

Président :	*M. BOURAINE* Mohand
Examinatrice :	*Mme TABTI* Hadjila
Encadrante :	*Mme TIMERIDJINE* Karima

Promotion 2024 - 2025.

Remerciements

Je vous remercie M^{me} TIMRIDJINE Karima de m'avoir consacré de votre temps. Vos recommandations m'ont permises la réalisation de ce modeste travail.
Je tiens à remercier également chacun des membres du jury pour l'intérêt porté à ce travail et d'avoir accepté de l'évaluer.

Dédicaces

Je dédie ce modeste travail à ma famille et à mes amis qui m'ont porté et soutenu.
Je suis redevable à mes chers parents pour leur soutien moral et matériel et aussi pour leur amour.

Table des matières

Table des figures

Liste des tableaux

Liste des abréviations

— **IQR** : Inter Quartile Range

— **kNN** : k-plus proches voisins (k-Nearest Neighbors)

— **ACP** : Analyse en Composantes Principales (Principal Component Analysis, **PCA**)

— **LDA** : Analyse Discriminante Linéaire (Linear Discriminant Analysis)

— **t-SNE** : t-distributed Stochastic Neighbor Embedding

— **UMAP** : Uniform Manifold Approximation and Projection

— **LMNN** : Large Margin Nearest Neighbor

— **MSE** : Mean Squared Error

— **EQM** : Erreur quadratique moyenn

— **MISE** : Mean Integrated Squared Error

Résumé

L'estimation non paramétrique constitue une approche flexible permettant d'estimer une fonction de densité de probabilité sans supposer de forme paramétrique préalable. Deux méthodes majeures sont étudiées : la méthode des k-plus proches voisins et la méthode à noyau. Après une présentation des fondements théoriques de ces techniques, l'accent est mis sur leur application dans le cadre de l'estimation de densité. Une analyse comparative est ensuite menée afin d'examiner leurs caractéristiques, leurs conditions d'utilisation, ainsi que leurs limites. L'étude repose sur un cadre méthodologique rigoureux et une exploration progressive des concepts, dans le but de mieux comprendre les apports et spécificités de chaque méthode dans le contexte de l'estimation non paramétrique.

Mots clés : Estimation non paramétrique, densité de probabilité, méthode des k-plus proches voisins, estimation par noyau, hyperparamètres.

Abstract

Nonparametric estimation provides a flexible approach to estimating a probability density function without assuming a predefined parametric form. Two major methods are studied : the k-nearest neighbors method and the kernel-based method. After introducing the theoretical foundations of these techniques, the focus is placed on their application in the context of density estimation. A comparative analysis is then conducted to examine their characteristics, usage conditions, and limitations. The study is based on a rigorous methodological framework and a progressive exploration of the concepts, with the aim of better understanding the contributions and specificities of each method in nonparametric estimation.

Keywords : Nonparametric estimation, probability density, k-nearest neighbors method, kernel estimation, hyperparameters.

Introduction générale

Dans de nombreux domaines scientifiques et techniques, la modélisation des données repose sur la connaissance de la loi de probabilité sous-jacente à une variable aléatoire. Cependant, dans la pratique, cette loi est souvent inconnue, ce qui rend nécessaire le recours à des méthodes d'estimation. Parmi celles-ci, les approches non paramétriques occupent une place essentielle, car elles ne supposent pas de forme prédéfinie pour la densité recherchée.

L'estimation non paramétrique de densité permet ainsi de reconstruire la structure probabiliste des données à partir d'un échantillon sans hypothèse forte sur la distribution. Cette flexibilité la rend particulièrement utile dans les situations où les modèles paramétriques classiques échouent ou s'avèrent trop restrictifs [19, 18].

Dans ce mémoire, nous nous intéressons à deux méthodes non paramétriques majeures : la méthode des k-plus proches voisins (k-NN) et la méthode à noyau. La première repose sur l'idée que la densité en un point peut être estimée à partir de la concentration des observations dans son voisinage. La seconde utilise une fonction de lissage (noyau) pour pondérer les observations en fonction de leur distance au point visée. Chacune de ces méthodes présente des avantages et des limites qui méritent une étude comparative approfondie [5, 14].

L'objectif de ce travail est double : d'une part, comprendre les fondements théoriques de ces méthodes et leurs propriétés statistiques ; d'autre part, analyser leur comportement en pratique à travers des simulations, en étudiant l'impact de paramètres tels que la taille de l'échantillon, la forme de la loi sous-jacente et la sensibilité au choix des hyperparamètres.

Ce mémoire est structuré comme suit : le premier chapitre présente la méthode à noyau et ses principales caractéristiques ; le deuxième chapitre est consacré à la méthode des k-plus proches voisins en classification ; le troisième chapitre introduit cette même

méthode dans le cadre de l'estimation de densité ; enfin, le quatrième chapitre propose une application comparative sur des données simulées, permettant d'évaluer les performances des deux approches.

Chapitre 1

Méthode à noyau

Sommaire

1.1 Introduction

L'estimation non paramétrique d'une densité de probabilité constitue une alternative fondamentale aux méthodes paramétriques classiques, en particulier lorsque la forme de la loi sous-jacente est inconnue ou difficile à spécifier. Parmi les approches non paramétriques, l'estimateur à noyau occupe une place centrale. Introduit indépendamment par Rosenblatt et Parzen, il permet d'estimer une fonction de densité à partir d'un échantillon sans imposer de structure fonctionnelle rigide à la distribution cible [19, 18, 5].

La méthode repose sur l'utilisation d'une fonction noyau et d'un paramètre de lissage (ou bande passante), qui influe sur la forme de l'estimation. Elle est particulièrement appréciée pour sa simplicité d'application et sa capacité à produire des estimations lisses, même dans des contextes complexes [19, 18].

Par rapport à des techniques plus rudimentaires comme l'histogramme, l'estimation à noyau présente plusieurs avantages : une meilleure continuité, une convergence plus rapide. Elle est également capable de révéler des structures fines dans les données, telles que la multimodalité, souvent invisibles avec des méthodes paramétriques [14, 13]. La méthode à noyau est également utilisée dans le cas fonctionnel (telque la fonction de régression).

Dans la suite de ce chapitre, nous présenterons les principes fondamentaux de cette méthode, sa formulation mathématique, ses propriétés statistiques, les méthodes de choix de la bande passante, ainsi qu'une application illustrant son comportement sur des données simulées.

1.2 Principes généraux de l'estimation à noyau

1.2.1 Motivation et cadre général de l'estimation

L'estimation par noyau a été proposée indépendamment par Rosenblatt (1956), dans le cadre de l'analyse de densité, et par Parzen (1962), dans un contexte plus général d'estimation non paramétrique. Elle est une méthode non paramétrique destinée à estimer

une fonction de densité de probabilité à partir d'un échantillon d'observations $X_1, \ldots, X_n$, d'une variable aléatoire X de distribution selon une densité de probabilité inconnue f [19, 18, 5].

Contrairement aux approches paramétriques qui imposent une forme prédéfinie à la densité (comme la loi normale, exponentielle, etc.), l'estimation à noyau ne repose sur aucune hypothèse explicite sur la structure de f. Elle construit l'estimateur en superposant une *fonction de lissage* (appelée noyau K) centrée sur chaque observation de l'échantillon, puis en moyennant ces contributions avec un facteur d'échelle h, appelé *bande passante* ou *paramètre de lissage* [19, 18].

Le principe intuitif peut être comparé à la création d'un « nuage de courbes » centrées sur les points observés. Le noyau joue le rôle d'une « petite cloche », tandis que le paramètre h règle la largeur de cette cloche [23, 3].

L'estimateur à noyau s'écrit :

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1.1}$$

où :

- K est une fonction réelle intégrable, souvent symétrique, satisfaisant les conditions :
  - $K \geq 0, \quad \int_{-\infty}^{\infty} K(u)\, du = 1$;
  - K est souvent symétrique : $K(u) = K(-u)$;
  - K décroît lorsque $|u|$ augmente.
- $h > 0$ est un paramètre de lissage, aussi désigné sous le nom de fenêtre ou bande passante, qui règle l'étendue de la zone autour de chaque observation sur laquelle le noyau agit, influençant ainsi la régularité de l'estimation [19, 18, 5].

Le choix de h a une influence directe sur le comportement de l'estimateur :

- Si h est *trop petit*, l'estimation devient très sensible aux détails des données, avec un faible biais mais une forte variance ;
- Si h est *trop grand*, l'estimation devient trop lissée, avec un biais élevé mais une variance réduite.

Ce compromis entre biais et variance est au cœur de l'efficacité de la méthode, et fera l'objet d'une analyse plus détaillée dans les sections suivantes [5].

1.2.2 Choix typiques de noyaux (Gaussien, uniforme, Epanechnikov, etc.)

Parmi les noyaux les plus courants, on trouve :

— **Gaussien** :

$$K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2), u \in \mathbb{R} \tag{1.2}$$

Support infini, lissage doux.

Noyau Gaussien

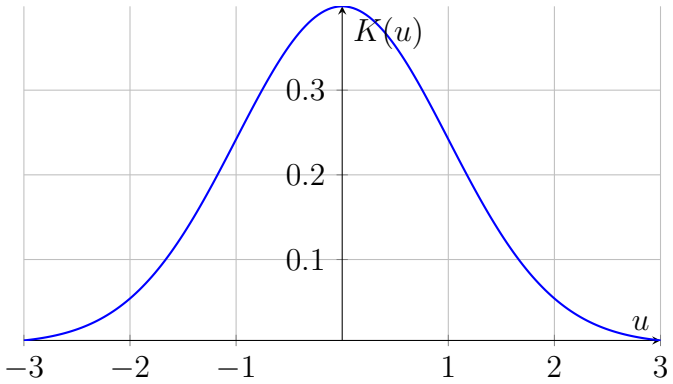

FIGURE 1.1 – Fonction noyau Gaussien

— **Uniforme** :

$$K(u) = \tfrac{1}{2} \mathbf{1}_{|u|\leq 1}. \tag{1.3}$$

Support compact, fenêtre plate.

Noyau uniforme

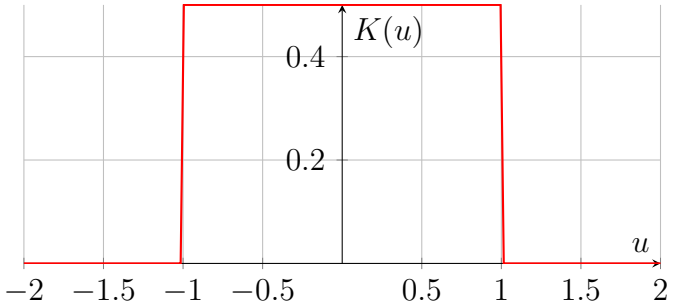

FIGURE 1.2 – Fonction noyau uniforme

— **Epanechnikov** :

$$K(u) = \tfrac{3}{4} (1 - u^2) \mathbf{1}_{|u|\leq 1}. \tag{1.4}$$

Optimal pour la MISE asymptotique.

Noyau d'Epanechnikov

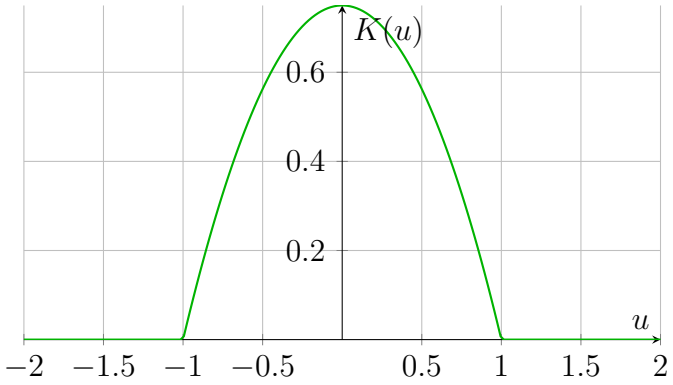

FIGURE 1.3 – Fonction noyau d'Epanechnikov

— **Biweight, Triweight, etc.** :

Utilisent plusieurs moments pour affiner le compromis biais-variance [19, 18, 5].

Remarque : Quel que soit le noyau, c'est le choix de la bande passante h qui a l'impact majeur sur la qualité de l'estimation [23].

1.3 Définition formelle de l'estimateur à noyau

1.3.1 Méthode de construction (approche de Rosenblatt)

On considère un échantillon $X_1, \ldots, X_n$ de variables aléatoires réelles, de densité f, et de fonction de répartition empirique donnée par :

$$\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}. \tag{1.5}$$

L'idée est d'approximer la densité f en utilisant la définition :

$$f(x) = \lim_{h \to 0} \frac{F(x+h) - F(x-h)}{2h}, \tag{1.6}$$

qui représente une dérivée centrée de la fonction de répartition. Pour une petite valeur de $h > 0$, on a l'approximation suivante :

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

En remplaçant F par $\hat{F}_n$, on propose l'estimateur suivant :

$$\hat{f}_n(x) = \frac{1}{2h}\left(\hat{F}_n(x+h) - \hat{F}_n(x-h)\right).$$

En utilisant l'expression de $\hat{F}_n$, cela donne :

$$\hat{f}_n(x) = \frac{1}{2h} \cdot \frac{1}{n}\sum_{i=1}^{n}\left[\mathbf{1}_{\{X_i \leq x+h\}} - \mathbf{1}_{\{X_i \leq x-h\}}\right] = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2h}\mathbf{1}_{\{x-h<X_i\leq x+h\}}.$$

On reconnaît ici un noyau uniforme donné par :

$$K(u) = \frac{1}{2}\mathbf{1}_{\{|u|\leq 1\}}. \tag{1.7}$$

On peut alors réécrire l'estimateur à noyau sous la forme classique (formule (1.1)) [19, 18].

Cette construction montre que l'estimateur à noyau peut être obtenu comme une approximation de la dérivée centrée de la fonction de répartition, à l'aide d'un noyau symétrique de largeur h. Elle illustre également le rôle du noyau comme fonction de pondération centrée sur x, qui attribue un poids à chaque observation en fonction de sa distance à ce point.

1.3.2 Formule générale en dimension 1

Dans le cas univarié, c'est-à-dire lorsque les données sont réelles, l'estimateur à noyau de la densité de probabilité f, à partir d'un échantillon $X_1, X_2, \ldots, X_n$, est défini par la formule (1.1).

L'interprétation intuitive de la formule est la suivante : on place une fonction K centrée sur chaque observation X_i, redimensionnée selon la largeur h, puis on moyenne les contributions de toutes les observations [23, 3].

Le rôle du noyau K est de pondérer les observations en fonction de leur distance à x.

Les noyaux usuels incluent le *noyau gaussien*, *uniforme*, *Epanechnikov*, etc. Leur choix a une influence modérée, contrairement au paramètre h qui a un impact déterminant sur la qualité de l'estimation [5].

1.3.3 Extension au cas multivarié

L'estimation à noyau peut être généralisée au cas multivarié, c'est-à-dire à des variables aléatoires vectorielles dans $\mathbb{R}^d$, pour $d \geq 2$. Cette extension est essentielle dans de nombreuses applications où les données sont multidimensionnelles, notamment en apprentissage automatique, en économie, en finance ou en traitement d'image [19, 18, 3].

Soit une suite $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ de vecteurs aléatoires indépendants et identiquement distribués, issus d'une densité de probabilité inconnue $f : \mathbb{R}^d \to \mathbb{R}$. L'estimateur à noyau multivarié de la densité f en un point $x \in \mathbb{R}^d$ est défini par [19, 18, 23, 3] :

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{|H|^{1/2}} K\left(H^{-1/2}(x - X_i)\right), \tag{1.8}$$

où :

— $K : \mathbb{R}^d \to \mathbb{R}$ est une fonction noyau multivariée, généralement symétrique ;

— H est une matrice de bande passante symétrique définie positive de taille $d \times d$, qui contrôle le degré de lissage dans chaque direction ainsi que les corrélations éventuelles entre les variables ;

— $|H|$ désigne le déterminant de la matrice H, et $H^{-1/2}$ sa racine carrée inverse.

Cas particulier : matrice H diagonale ou isotrope

Un choix fréquent consiste à prendre une matrice de bande passante isotrope : $H = h^2 I_d$, où $h > 0$ est un paramètre de lissage scalaire commun à toutes les directions et I_d est la matrice identité. Dans ce cas, la formule se simplifie en :

$$\hat{f}_n(x) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right). \tag{1.9}$$

Cette forme est souvent suffisante lorsque les variables sont supposées indépendantes ou lorsqu'aucune direction privilégiée n'est connue [3, 23].

Choix du noyau multivarié

Le noyau K est généralement de la forme radiale :

$$K(u) = k(\|u\|), \tag{1.10}$$

où $k : \mathbb{R}^+ \to \mathbb{R}$ est une fonction décroissante, et $\|u\|$ désigne une norme sur $\mathbb{R}^d$, souvent la norme euclidienne. Le noyau gaussien multivarié est un choix standard [5, 3, 20] :

$$K(u) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\|u\|^2\right). \tag{1.11}$$

Remarques importantes

— L'extension multivariée est sujette à la *malédiction de la dimension* : lorsque d augmente, la quantité de données nécessaires pour estimer correctement la densité croît exponentiellement (voir section 1.7.2).

— Le choix de la matrice H est crucial : une mauvaise sélection peut entraîner un lissage excessif ou insuffisant dans certaines directions, ou ignorer des corrélations importantes entre les variables.

— Des techniques adaptatives, telles que l'utilisation de bande passante variable ou l'intégration de méthodes de réduction de dimension (PCA, t-SNE, etc.), peuvent être envisagées pour améliorer la performance de l'estimation dans des contextes à haute dimension [19, 18, 5, 3, 20].

1.3.4 Propriétés ponctuelles de l'estimateur

L'estimateur à noyau est dit *ponctuel* (ou local) dans le sens où il évalue la densité en un point donné x, en s'appuyant essentiellement sur les observations proches de ce point. Cette propriété découle naturellement de la forme de l'estimateur, où la contribution des observations est pondérée par un noyau centré en x, de largeur déterminée par la matrice de bande passante H [3, 25].

— La **contribution principale** à l'estimation en x provient des observations X_i proches de x, au sens de la distance induite par H. Plus précisément, si $K(u)$ est négligeable en dehors d'un voisinage de 0, alors $K(H^{-1/2}(x - X_i))$ est négligeable dès que X_i s'éloigne trop de x.

— Sous des hypothèses de **régularité** sur la densité f, les **quantités d'erreur** comme le biais et la variance s'expriment en fonction des **dérivées de f en x** (gradient, hessienne) et des **moments du noyau K**. Ces résultats sont détaillés dans la section suivante (cf. section 1.4).

— Enfin, l'estimateur est également localisé au sens où sa fenêtre de lissage est restreinte à la **zone de support effectif** du noyau, c'est-à-dire là où la fonction $u \mapsto K(u)$ est non nulle.

Cette localité est essentielle pour permettre à l'estimateur de s'adapter aux variations locales de la densité, notamment dans les régions où celle-ci change rapidement [3, 20, 25, 23].

1.4 Propriétés statistiques

1.4.1 Biais et variance

L'objectif de cette section est d'analyser le comportement de l'estimateur à noyau $\hat{f}_n(x)$ en un point fixé $x \in \mathbb{R}^d$, notamment à travers son biais et sa variance. Pour cela, on suppose que la densité f est suffisamment régulière (au moins deux fois continûment dérivable) et on utilise un développement en série de Taylor (voir [3, 25, 8]).

Définition de l'estimateur à noyau

Considérons l'estimateur donnée en (1.9), pour un noyau symétrique :

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

Espérance de $\hat{f}_n(x)$

Par linéarité de l'espérance et indépendance des X_i, on a :

$$\mathbb{E}[\hat{f}_n(x)] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \frac{1}{h^d} K\left(\frac{x - X_i}{h}\right)\right] = \mathbb{E}\left[\frac{1}{h^d} K\left(\frac{x - X_1}{h}\right)\right]. \tag{1.12}$$

En écrivant cette espérance sous forme intégrale :

$$\mathbb{E}[\hat{f}_n(x)] = \int_{\mathbb{R}^d} \frac{1}{h^d} K\left(\frac{x - y}{h}\right) f(y) dy. \tag{1.13}$$

On effectue le changement de variable $u = \frac{x-y}{h}$, ce qui donne :

$$\mathbb{E}[\hat{f}_n(x)] = \int_{\mathbb{R}^d} K(u) f(x - hu) du. \tag{1.14}$$

Développement de Taylor

Si f est deux fois continûment dérivable, on développe $f(x - hu)$ autour de x :

$$f(x - hu) = f(x) - hu^\top \nabla f(x) + \frac{h^2}{2} u^\top \nabla^2 f(x) u + o(h^2). \tag{1.15}$$

Par symétrie du noyau K, les termes linéaires s'annulent :

$$\int K(u) u du = 0. \tag{1.16}$$

Ainsi,

$$\mathbb{E}[\hat{f}_n(x)] = f(x) + \frac{h^2}{2} \sum_{j,k=1}^{d} \frac{\partial^2 f}{\partial x_j \partial x_k}(x) \int u_j u_k K(u) du + o(h^2). \tag{1.17}$$

En notant $\mu_2(K) = \int \|u\|^2 K(u) du$ et $\nabla f(x) = \sum_{j=1}^{d} \frac{\partial^2 f}{\partial x_j^2}(x)$, on obtient :

$$\mathbb{E}[\hat{f}_n(x)] = f(x) + \frac{h^2}{2} \mu_2(K) \nabla f(x) + o(h^2). \tag{1.18}$$

Cette analyse du biais de l'estimateur à noyau repose sur les développements classiques de Parzen et Rosenblatt [19, 18], et a été approfondie dans les travaux ultérieurs comme ceux de Collomb [5, 3, 25, 8].

Biais

Le biais est défini par

$$\text{Biais}[\hat{f}_n(x)] = \mathbb{E}[\hat{f}_n(x)] - f(x) = \frac{h^2}{2} \mu_2(K) \nabla f(x) + o(h^2). \tag{1.19}$$

Variance

L'estimateur étant une moyenne de variables indépendantes, on a

$$\text{Var}[\hat{f}_n(x)] = \frac{1}{n} \text{Var}\left(\frac{1}{h^d} K\left(\frac{x - X_1}{h} \right) \right). \tag{1.20}$$

On calcule :

$$\text{Var}\left(\frac{1}{h^d} K\left(\frac{x - X_1}{h} \right) \right) = \mathbb{E}\left[\left(\frac{1}{h^d} K\left(\frac{x - X_1}{h} \right) \right)^2 \right] - \left(\mathbb{E}\left[\frac{1}{h^d} K\left(\frac{x - X_1}{h} \right) \right] \right)^2 \tag{1.21}$$

Par changement de variable $u = \frac{x-y}{h}$ et sous hypothèses régulières, on approxime :

$$\mathrm{Var}[\hat{f}_n(x)] \approx \frac{1}{nh^d} f(x) R(K), \tag{1.22}$$

avec

$$R(K) = \int K^2(u) du. \tag{1.23}$$

Cette approximation est également discutée dans [19, 18, 25].

Erreur quadratique moyenne (EQM)

L'erreur quadratique moyenne combine le biais et la variance :

$$\mathrm{EQM}[\hat{f}_n(x)] = \Big(\mathrm{Biais}[\hat{f}_n(x)]\Big)^2 + \mathrm{Var}[\hat{f}_n(x)]. \tag{1.24}$$

En utilisant les développements précédents :

$$\mathrm{EQM}[\hat{f}_n(x)] \approx \left(\frac{h^2}{2}\mu_2(K)\nabla f(x)\right)^2 + \frac{1}{nh^d} f(x) R(K) + o(h^4) + o\left(\frac{1}{nh^d}\right). \tag{1.25}$$

Cette expression met en évidence le compromis classique : réduire h diminue le biais mais augmente la variance, et inversement [3, 25, 8].

Résumé et compromis biais/variance

En un point x, on a les approximations asymptotiques suivantes :

$$\mathrm{Biais}[\hat{f}_n(x)] \sim C_1 h^2 \tag{1.26}$$

et

$$\mathrm{Var}[\hat{f}_n(x)] \sim \frac{C_2}{nh^d}, \tag{1.27}$$

où C_1 et C_2 sont des constantes dépendant de f et K [3, 25].

Le choix du paramètre de lissage h est donc crucial pour équilibrer les deux sources d'erreur et minimiser l'erreur globale [19, 18, 5, 25].

1.4.2 Consistance et convergence

Nous étudions ici la convergence de l'estimateur à noyau $\hat{f}_n(x)$ vers la vraie densité $f(x)$, lorsque la taille de l'échantillon n tend vers l'infini. Cette propriété est essentielle pour garantir la consistance de l'estimation à grande échelle [27, 12].

Consistance ponctuelle

On dit que l'estimateur $\hat{f}_n(x)$ est *consistant en probabilité* en un point x si :

$$\hat{f}_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} f(x),$$

c'est-à-dire que pour tout $\varepsilon > 0$,

$$\mathbb{P}\Big(|\hat{f}_n(x) - f(x)| > \varepsilon\Big) \to 0$$

quand $n \to \infty$.

Ici, h_n désigne une *bande passante* (ou paramètre de lissage) qui dépend de la taille de l'échantillon n et contrôle le voisinage autour du point x utilisé pour l'estimation.

Cette propriété découle du fait que le biais de $\hat{f}_n(x)$ tend vers 0 lorsque $h_n \to 0$, et que la variance tend également vers 0 lorsque $nh_n^d \to \infty$.

Ainsi, sous les conditions suivantes :

$$h_n \to 0, \quad \text{et} \quad nh_n^d \to \infty \quad \text{quand } n \to \infty,$$

l'estimateur $\hat{f}_n(x)$ est consistant en probabilité [3, 20].

Consistance uniforme sur un compact

Sous des hypothèses supplémentaires sur la régularité de f et K, on peut démontrer une convergence uniforme sur un compact $A \subset \mathbb{R}^d$:

$$\sup_{x\in A} |\hat{f}_n(x) - f(x)| \xrightarrow[n\to\infty]{\mathbb{P}} 0.$$

Cela signifie que la convergence n'est pas seulement valable point par point, mais de manière uniforme sur l'ensemble A [3, 20].

Convergence en moyenne quadratique (MSE)

En utilisant l'expression asymptotique de l'erreur quadratique moyenne (1.28), on a :

$$\mathbb{E}[(\hat{f}_n(x) - f(x))^2] \to 0,$$

si $h_n \to 0$ et $nh_n^d \to \infty$. On dit alors que $\hat{f}_n(x)$ est *convergent en moyenne quadratique* vers $f(x)$ (voir [23, 8, 27]).

Remarque sur le choix du pas de lissage h_n (largeur de bande)

Les conditions $h_n \to 0$ et $nh_n^d \to \infty$ sont nécessaires pour assurer la convergence de l'estimateur à noyau. Elles reflètent le compromis biais-variance suivant :

— Si h_n est trop petit, l'estimateur présente une variance élevée.

— Si h_n est trop grand, l'estimateur devient biaisé [23, 25].

Dans le cas de fonctions densité deux fois dérivables et sous des hypothèses régulières sur le noyau, on peut démontrer que l'erreur quadratique moyenne (EQM) de $\hat{f}_n(x)$ est asymptotiquement de l'ordre :

$$\text{EQM}[\hat{f}_n(x)] = \mathcal{O}(h_n^4) + \mathcal{O}\left(\frac{1}{nh_n^d}\right).$$

Le choix du pas de lissage h_n qui équilibre ces deux termes est alors donné par :

$$h_n \propto n^{-1/(d+4)},$$

ce qui permet de minimiser l'ordre asymptotique de l'erreur quadratique moyenne [23, 25, 20].

1.5 Choix du paramètre de lissage (bande passante h)

Le paramètre de lissage h, ou bande passante, est un élément clé dans l'estimation de la densité par la méthode à noyau. Il contrôle la largeur de la fenêtre dans laquelle les observations X_i sont pondérées par le noyau K pour estimer $f(x)$ [23].

1.5.1 Critère de sélection optimal

Un objectif central en estimation de densité est de choisir une largeur de bande h qui équilibre correctement biais et variance. Un critère fréquemment utilisé pour évaluer cette performance est l'erreur quadratique moyenne (EQM) :

$$\mathrm{EQM}(x) = \mathbb{E}[(\hat{f}_n(x) - f(x))^2] = \mathrm{Biais}^2[\hat{f}_n(x)] + \mathrm{Var}[\hat{f}_n(x)]. \tag{1.28}$$

Sous certaines hypothèses de régularité sur f et le noyau K, cette erreur peut être approximée par :

$$\mathrm{EQM}(x) \approx C_1 h^4 + \frac{C_2}{nh^d}, \tag{1.29}$$

où C_1 et C_2 sont des constantes positives. La minimisation de cette expression conduit à une largeur de bande asymptotiquement optimale donnée par :

$$h_n \propto n^{-1/(d+4)}, \tag{1.30}$$

(voir [23, 25, 20]). Ce résultat fournit un point de départ théorique, mais ne constitue pas le seul critère pertinent en pratique.

1.5.2 Méthodes pratiques de choix de h

En pratique, plusieurs approches permettent de sélectionner automatiquement la bande passante h à partir des données :

— **Validation croisée (leave-one-out) :**

Elle consiste à retirer chaque observation tour à tour, estimer la densité au point supprimé avec les données restantes, puis choisir h en minimisant l'erreur globale. Une version populaire est la validation croisée aux moindres carrés (LSCV), dont le critère s'écrit :

$$\mathrm{LSCV}(h) = \int \hat{f}_h^2(x)\, dx - \frac{2}{n}\sum_{i=1}^{n} \hat{f}_{h,-i}(X_i), \tag{1.31}$$

où $\hat{f}_{h,-i}$ est l'estimateur obtenu sans la i-ème donnée.

— **Méthodes plug-in (Sheather–Jones) :**

Ces approches utilisent une approximation de l'EQM théorique en intégrant une es-

timation préalable de certaines quantités comme la dérivée seconde de la densité. La méthode de Sheather–Jones est réputée pour sa précision mais reste plus complexe à mettre en œuvre.

— **Règles heuristiques :**

Elles proposent des formules simples à base de statistiques comme l'écart-type ou l'intervalle interquartile. Pour $d = 1$, la règle de Silverman donne :

$$h_{\text{Silverman}} = 0.9 \times \min\left(\hat{\sigma}, \frac{\text{IQR}}{1.34}\right) \times n^{-1/5}, \tag{1.32}$$

où $\hat{\sigma}$ est l'écart-type et $\text{IQR} = Q_3 - Q_1$ l'intervalle interquartile. Cette règle est robuste face aux valeurs extrêmes.

Ainsi, si la règle asymptotique fournit un repère théorique utile, le choix effectif de h doit tenir compte du contexte, du volume de données, et du critère d'optimisation retenu (voir [23, 25, 22]).

1.5.3 Effet de h sur le biais et la variance

Comme vu précédemment :

— **Le biais** de l'estimateur à noyau est en $\mathcal{O}(h^2)$,

— **La variance** est en $\mathcal{O}\left(\frac{1}{nh}\right)$.

Ainsi, le choix optimal de la bande passante h, dans un cadre théorique idéal (fonction régulière, noyau bien choisi), consiste à équilibrer le biais et la variance.

On en déduit que l'ordre optimal de h est :

$$h_{\text{opt}} \propto n^{-1/5}, \tag{1.33}$$

ce qui reflète la vitesse optimale de convergence pour l'estimateur à noyau univarié [25, 23].

1.6 Application illustrative

Cette section illustre l'impact du choix de la bande passante h dans l'estimation de densité par noyau, à travers un exemple classique inspiré de l'ouvrage de B. W. Silver-

man [23].

1.6.1 Estimation sur données simulées

Un échantillon de taille $n = 200$ est généré à partir d'une densité bimodale définie par le mélange suivant :

$$f(x) = 0{,}5 \cdot \mathcal{N}(-1,\, 0{,}3^2) + 0{,}5 \cdot \mathcal{N}(1,\, 0{,}3^2) \tag{1.34}$$

Cette distribution sert à illustrer comment l'estimateur à noyau peut ou non révéler la structure réelle des données, selon la valeur choisie pour h.

1.6.2 Comparaison visuelle selon différentes valeurs de h

L'estimation de la densité est réalisée à l'aide d'un noyau gaussien, pour différentes valeurs de la bande passante h, afin de visualiser son influence :

— h **petit** : L'estimateur suit de très près les données, produisant une courbe irrégulière avec un faible biais, mais une variance élevée, typique d'un surajustement.

— h **optimal** : La forme bimodale est bien restituée, sans sur-ajustement. L'estimation présente un bon équilibre entre biais et variance.

— h **grand** : La densité est fortement lissée, les deux pics sont fusionnés. Cela indique un biais important et une perte de structure.

Pour illustrer visuellement ces effets, nous représentons trois courbes d'estimation obtenues avec un noyau gaussien pour trois valeurs de h : trop petite, optimale, et trop grande, appliquées à la densité bimodale simulée ci-dessus.

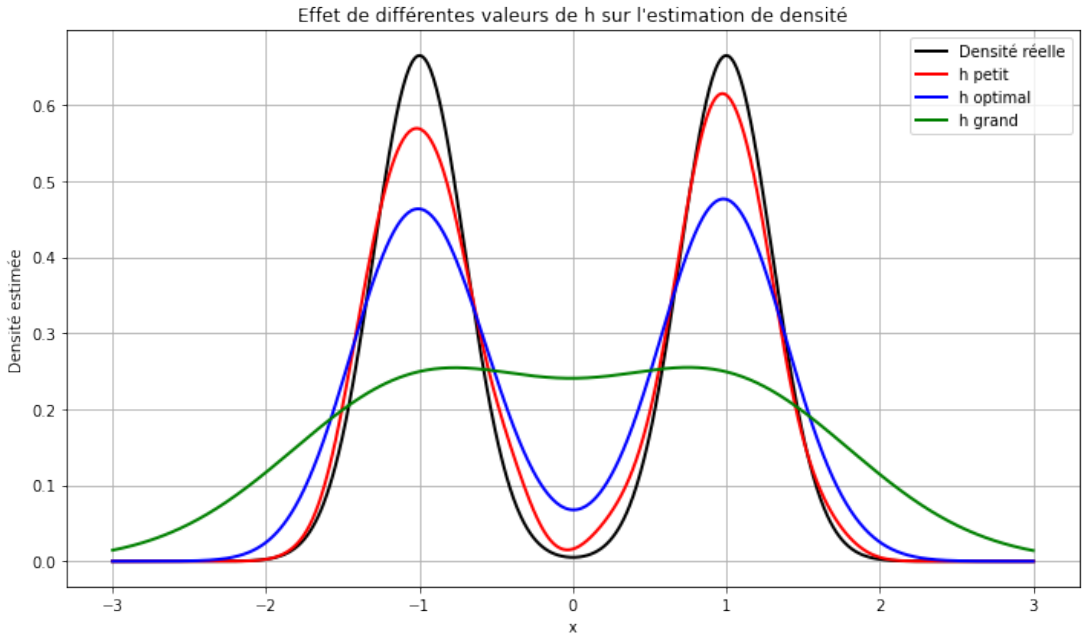

FIGURE 1.4 – Comparaison visuelle de l'effet de différentes valeurs de h sur l'estimation de densité

Interprétation : La figure 1.4 met en évidence l'effet de la bande passante sur la qualité de l'estimation :

— Un h trop petit (courbe rouge) produit une estimation très sensible aux données individuelles, avec une variance élevée.

— Un h bien choisi (courbe bleue) capture correctement la structure bimodale, avec une estimation stable et représentative.

— Un h trop grand (courbe verte) entraîne un excès de lissage, masquant la véritable forme de la densité.

Cette démonstration visuelle souligne l'importance cruciale du choix de la bande passante pour obtenir une estimation fiable, qui reflète correctement la structure sous-jacente tout en maintenant un bon compromis entre biais et variance.

1.7 Limites et remarques pratiques

Bien que la méthode à noyau soit largement utilisée pour l'estimation non paramétrique de densité, elle présente certaines limites qu'il est important de connaître. Nous présentons ici deux aspects pratiques majeurs : la sensibilité au choix de la bande passante, et la difficulté d'utilisation en haute dimension.

1.7.1 Sensibilité au choix de h

La qualité de l'estimation repose fortement sur le paramètre de lissage h. Un choix inadéquat peut entraîner :

— un *sur-ajustement* (bande passante trop petite) où l'estimation est trop fluctuante, capturant le bruit aléatoire des données,

— ou un *sous-ajustement* (bande passante trop grande) où les structures fines de la densité sont effacées.

Bien que des méthodes automatiques existent pour choisir h, telles que la règle de Silverman ou la validation croisée, leur efficacité peut être limitée dans certaines situations : petites tailles d'échantillon, distributions asymétriques ou multimodales, ou présence d'outliers. Dans la pratique, une inspection visuelle des courbes estimées, accompagnée d'une expertise métier, demeure souvent indispensable pour un ajustement pertinent [23, 21].

1.7.2 Difficulté en haute dimension (malédiction de la dimension)

La méthode à noyau souffre d'un problème bien connu en statistique : la "*malédiction de la dimension*" (*curse of dimensionality*). Lorsque la dimension d de l'espace augmente :

— Le volume de l'espace croît de manière exponentielle, ce qui rend les données extrêmement dispersées. L'information devient alors plus diffuse et difficile à exploiter.

— En haute dimension, la concentration des données autour d'un point donné devient très faible, ce qui rend l'estimation peu fiable.

— Il devient nécessaire de disposer d'un nombre exponentiellement croissant d'observations pour obtenir une estimation précise.

Ce phénomène entraîne une augmentation rapide de la variance de l'estimateur à noyau lorsque d augmente, même si le biais peut rester limité. La qualité globale de l'estimation se détériore fortement en haute dimension [21, 17].

Pour limiter cet effet, plusieurs stratégies peuvent être envisagées en s'appuyant sur des hypothèses structurelles supplémentaires :

— **Hypothèse d'indépendance conditionnelle** : certaines variables sont supposées ne pas interagir fortement entre elles, ce qui permet de simplifier le modèle.

— **Sparsité des dimensions pertinentes** : on suppose que seules quelques variables ont un impact significatif sur la densité à estimer, les autres pouvant être négligées.

— **Réduction de dimension** : on projette les données dans un espace de plus faible dimension, à l'aide de techniques comme l'Analyse en Composantes Principales (ACP), tout en conservant l'essentiel de l'information.

En l'absence de telles hypothèses ou techniques, la performance des estimateurs non paramétriques, comme l'estimateur à noyau, se dégrade fortement avec la dimension [3, 25, 22, 17].

1.8 Conclusion

Dans ce chapitre, nous avons exploré les fondements des méthodes à noyau pour l'estimation de densité, en mettant particulièrement l'accent sur le rôle central de la bande passante h. Nous avons présenté différentes approches pour le choix de cette bande passante : méthodes heuristiques, plug-in et validation croisée, en discutant leurs avantages, inconvénients et conditions d'application.

Nous avons également analysé l'effet du paramètre h sur le biais et la variance de l'estimateur, soulignant le compromis fondamental entre sous-lissage et sur-lissage. Ce compromis est au cœur des méthodes d'optimisation du lissage, et influence directement la précision de l'estimation finale.

Ainsi, le choix de h joue un rôle déterminant dans la performance de l'estimation, et sa sélection ne doit pas être négligée. Une bonne estimation de la bande passante permet d'obtenir une densité lissée proche à la distribution réelle, tout en évitant le surajustement ou le sous-ajustement [3].

Ce cadre théorique constitue une base solide pour aborder l'analyse de données réelles à l'aide d'estimateurs à noyau, et ouvre la voie à des développements plus complexes tels que l'estimation adaptative ou multidimensionnelle.

Chapitre 2

Méthode des k-plus proches voisins en classification

Sommaire

2.1 Introduction

La méthode des k-plus proches voisins (k-NN) constitue un algorithme d'apprentissage supervisé non paramétrique largement utilisé en classification, en raison de sa simplicité conceptuelle et de son interprétation géométrique intuitive. Contrairement aux approches paramétriques, cette technique ne repose sur aucune hypothèse à priori concernant la distribution des données, mais s'appuie sur le principe que des observations proches dans l'espace des caractéristiques sont susceptibles d'appartenir à la même classe [6]. L'efficacité de cette méthode dépend néanmoins de manière critique du choix de ses hyperparamètres, en particulier du nombre de voisins k et de la métrique de distance utilisée [14].

Les applications de la méthode k-NN couvrent des domaines variés, tels que la reconnaissance visuelle [29], le diagnostic biomédical, ou encore la détection de fraudes, où sa capacité à modéliser des frontières de décision complexes est particulièrement appréciée. Toutefois, cette approche présente certaines limitations, notamment sa sensibilité aux valeurs aberrantes, à la dimension élevée des données, ainsi qu'aux déséquilibres entre classes [2]. Ces difficultés ont conduit au développement de nombreuses améliorations, telles que l'introduction de mécanismes de pondération adaptative, l'utilisation de noyaux ou l'apprentissage de métriques optimisées [15].

Ce chapitre propose une analyse détaillée des fondements théoriques de la méthode k-NN en classification, en s'appuyant tant sur les travaux fondateurs de Fix et Hodges [10] que sur des recherches plus récentes portant sur ses propriétés asymptotiques [31]. Une attention particulière sera portée au compromis biais-variance induit par le choix de k, illustré à travers des cas d'usage classiques comme le jeu de données Iris [9].

2.2 Principe fondamental de la méthode k-NN

L'algorithme des k-NN est une méthode non paramétrique largement utilisée en classification supervisée. Il repose sur une hypothèse intuitive : les observations proches dans l'espace des caractéristiques sont susceptibles d'appartenir à la même classe. Ce principe en fait un outil simple, mais puissant, ne nécessitant ni modélisation probabiliste ni apprentissage explicite, ce qui est particulièrement utile dans des contextes où les relations entre variables sont complexes ou inconnues [14, 6].

Contrairement aux méthodes linéaires classiques comme l'analyse discriminante de Fisher [9], k-NN s'appuie exclusivement sur la distance entre les observations pour prendre des décisions de classification.

2.2.1 Définition et règle de majorité

L'idée centrale de l'algorithme k-NN est d'associer à une observation non étiquetée $x \in \mathbb{R}^d$ la classe majoritaire parmi ses k plus proches voisins dans l'ensemble d'apprentissage. La notion de proximité est généralement mesurée par la distance euclidienne, bien que d'autres métriques puissent être utilisées selon le contexte (Manhattan, Minkowski, Mahalanobis, etc.) [13, 15].

Soit $D = \{(X_i, Y_i)\}_{i=1}^n$ un échantillon d'apprentissage avec $X_i \in \mathbb{R}^d$ et $Y_i \in \mathcal{Y}$, où $\mathcal{Y}$ est un ensemble fini de classes (classe de chaque X_i). On note $N_k(x)$ l'ensemble des k observations les plus proches de x. L'étiquette prédite pour x est :

$$\hat{Y}(x) = \arg\max_{y \in \mathcal{Y}} \sum_{i \in N_k(x)} \mathbb{I}_{Y_i = y} \tag{2.1}$$

où $\mathbb{I}_{Y_i=y}$ est l'indicatrice de l'appartenance de Y_i à la classe y.

Le classifieur 1-NN (cas particulier avec $k = 1$) est historiquement le premier à avoir été introduit pour la reconnaissance de motifs [10]. Cependant, il est très sensible aux anomalies et aux erreurs de mesure, ce qui justifie l'utilisation de $k > 1$ pour stabiliser la prédiction [29].

Des variantes modernes de k-NN proposent des améliorations telles que l'apprentissage de métriques de distance optimisées, ou encore l'intégration du classifieur dans des schémas semi-supervisés ou robustes [15, 31].

2.2.2 Impact du paramètre k

Le choix du paramètre k influence de manière significative les performances du classifieur k-NN. Un faible k rend l'algorithme très sensible au bruit, tandis qu'un k élevé augmente le risque d'inclure des observations éloignées appartenant à d'autres classes [14, 13].

Ce comportement illustre le compromis classique entre biais et variance :

— Un petit k induit un faible biais mais une variance élevée (modèle surajusté) ;

— Un grand k accroît le biais tout en réduisant la variance (modèle plus stable) [14].

La sélection optimale de k s'effectue généralement de manière empirique par validation croisée, en testant plusieurs valeurs et en retenant celle qui minimise l'erreur de classification sur un ensemble de validation [29].

D'un point de vue théorique, les travaux de Biau *et al.* [2] ont montré que, sous certaines conditions (notamment $k \to \infty$ et $k/n \to 0$), le classifieur k-NN converge vers la règle de Bayes, laquelle constitue le classifieur optimal en termes d'erreur.

Enfin, des extensions modernes du classifieur k-NN, telles que les variantes bayésiennes ou les approches hybrides incorporant des principes de pondération adaptative ou d'apprentissage métrique, permettent de mieux gérer la variabilité des données tout en renforçant la robustesse face au bruit et au déséquilibre entre classes [15, 31].

2.3 Mesures de distance

Dans l'algorithme des k-plus proches voisins (k-NN), il est important de pouvoir mesurer à quel point deux données sont proches l'une de l'autre. Pour cela, on utilise une mesure d'approximité appelée "*distance*". Le choix de cette distance est très important, car il détermine quels voisins seront considérés comme les plus proches. Si les données sont très différentes entre elles (par exemple, si certaines variables ont des unités ou des valeurs très différentes), ou si on a beaucoup de variables, ce choix peut avoir un grand impact sur les résultats [14, 13].

Les distances les plus couramment utilisées sont la distance euclidienne et la distance de Manhattan, toutes deux appartenant à la famille des distances de Minkowski. D'autres mesures, telles que la distance de Mahalanobis, peuvent être plus adaptées selon la nature des données [15, 24].

2.3.1 Distance euclidienne

La distance euclidienne, induite par la norme L_2, est définie pour deux vecteurs $x = (x_1, \ldots, x_d)$ et $x' = (x'_1, \ldots, x'_d) \in \mathbb{R}^d$ par :

$$d_{\text{Eucl}}(x, x') = \sqrt{\sum_{j=1}^{d} (x_j - x'_j)^2} \tag{2.2}$$

Elle correspond à la longueur du segment reliant deux points dans l'espace euclidien. Bien qu'intuitive et simple à implémenter, cette mesure est très sensible aux valeurs extrêmes, car elle élève les différences au carré. Elle suppose également que les dimensions ont la même échelle et une importance comparable. Par conséquent, une normalisation ou standardisation préalable des données est souvent nécessaire [14, 13].

2.3.2 Distance de Manhattan

La distance de Manhattan (induite par la norme L_1) sur $\mathbb{R}^d$ est définie par :

$$d_{\text{Man}}(x, x') = \sum_{j=1}^{d} |x_j - x'_j| \tag{2.3}$$

Elle mesure la somme des écarts absolus selon chaque dimension, à la manière d'un trajet dans une grille urbaine. Moins sensible aux valeurs extrêmes que la distance euclidienne, elle est bien adaptée aux contextes où les effets des dimensions sont additifs et indépendants. Elle est également pertinente lorsque les distributions sont asymétriques ou bruitées [13, 24].

En pratique, le choix de la distance doit être guidé par la structure des données. Des étapes de prétraitement (centrage, réduction, normalisation) sont souvent indispensables. Par ailleurs, des approches modernes d'apprentissage de métriques permettent d'adapter automatiquement la mesure de distance aux données, en optimisant la performance du classifieur [15, 24].

2.4 Limites et améliorations simples

L'algorithme des k-plus proches voisins (k-NN), malgré sa simplicité et sa bonne performance dans de nombreux contextes, présente certaines limites, notamment en présence de bruit, de redondance dans les données ou dans des espaces de grande dimension. Plusieurs améliorations simples peuvent être envisagées pour atténuer ces problèmes [14, 13, 24].

2.4.1 Sensibilité au bruit et à la dimension

L'un des principaux inconvénients de k-NN est sa sensibilité au bruit. Lorsqu'un ou plusieurs voisins proches sont mal étiquetés (erreurs ou bruit), ils peuvent facilement

fausser la prédiction, surtout si k est petit. Un k trop grand, en revanche, peut diluer l'information spécifique au voisinage en intégrant des points de classes différentes [14].

Pour remédier à cela, des méthodes de réduction de dimension telles que l'analyse en composantes principales (ACP), l'analyse discriminante linéaire (LDA), ou des techniques non linéaires comme t-SNE ou UMAP peuvent être utilisées. Certaines variantes avancées incluent l'apprentissage supervisé de métriques adaptées, comme le *Large Margin Nearest Neighbor* (LMNN) [11, 26].

2.4.2 Pondération des voisins

Une amélioration naturelle du classifieur k-NN consiste à **pondérer l'influence des voisins** selon leur distance à l'observation à classer. Autrement dit, plus un voisin est proche de x, plus son poids dans la décision est important [26, 29].

Soit une nouvelle observation $x \in \mathbb{R}^d$ que l'on souhaite classer, et un ensemble d'apprentissage constitué de vecteurs $X_i \in \mathbb{R}^d$. On commence par identifier les k plus proches voisins de x, notés $N_k(x)$ [29, 26].

L'intuition est que les voisins très proches de x sont plus susceptibles d'appartenir à la même classe, tandis que les plus éloignés apportent une information moins fiable. Il est donc pertinent de leur attribuer des poids dégressifs avec la distance [13, 11, 16].

Une pondération usuelle consiste à définir, pour chaque voisin $X_i \in N_k(x)$, un poids :

$$w_i = \frac{1}{d(x, X_i) + \varepsilon},$$

où $\varepsilon > 0$ est ajoutée pour éviter la division par zéro [13, 11]. La classe prédite pour x est alors celle qui maximise la décision pondérée :

$$\hat{Y}(x) = \arg\max_{y \in \mathcal{Y}} \sum_{i \in N_k(x)} w_i \cdot \mathbb{I}_{Y_i = y},$$

où $\mathbb{I}_{Y_i=y}$ est la fonction indicatrice égale à 1 si $Y_i = y$, et 0 sinon.

Ce schéma réduit l'influence des observations éloignées ou bruitées, tout en conservant la simplicité du modèle de base. Il améliore souvent les performances, notamment dans les régions où la densité des points varie dans l'espace des caractéristiques [11, 14].

Enfin, cette stratégie constitue une première étape vers des variantes plus avancées, comme l'apprentissage de métriques ou les modèles adaptatifs, dans lesquels les poids sont

optimisés automatiquement en fonction des données [14].

2.5 Application illustrative (1)

Dans cette section, nous illustrons le fonctionnement de la méthode des k-plus proches voisins (k-NN) sur un exemple concret de classification supervisée. Le jeu de données choisi est l'un des plus classiques en apprentissage automatique : le *jeu de données Iris* [9]. Ce jeu contient 150 observations réparties équitablement entre trois espèces de fleurs (*Setosa*, *Versicolor*, *Virginica*), chacune décrite par quatre variables numériques : la longueur et la largeur des sépales et des pétales.

L'objectif est de prédire l'espèce d'une fleur à partir de ses mesures morphologiques. Le jeu est particulièrement adapté à des approches comme k-NN, car il comporte à la fois des classes bien séparées (comme *Setosa*) et d'autres plus proches (comme *Versicolor* et *Virginica*), mettant en lumière les effets du choix du paramètre k.

2.5.1 Jeu de données Iris

Le jeu Iris est disponible dans de nombreuses bibliothèques de science des données, notamment `scikit-learn` en Python. Il se compose de :

— $n = 150$ observations ;

— 3 classes cibles ;

— 4 variables explicatives continues : longueur/largeur des pétales et des sépales.

Une représentation graphique, comme un nuage de points basé sur deux variables explicatives, permet de visualiser la séparation entre les classes et d'anticiper la performance de l'algorithme.

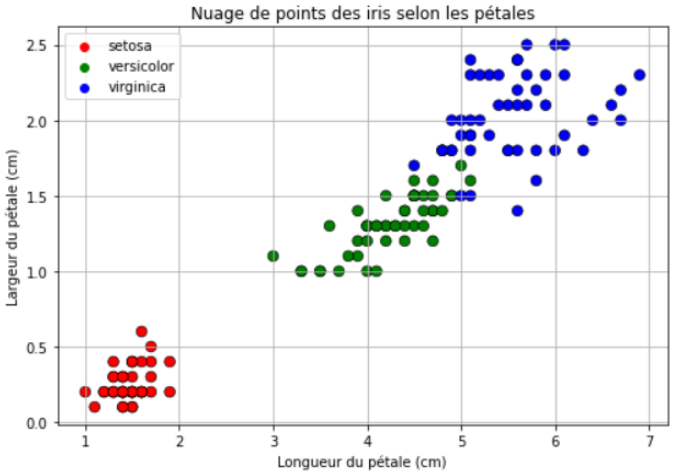

FIGURE 2.1 – Nuage de points des observations selon deux variables du jeu Iris. Les couleurs indiquent les classes.

2.5.2 Implémentation et résultats

L'algorithme k-NN a été appliqué au jeu de données Iris en utilisant uniquement deux variables : la longueur et la largeur des pétales, connues pour leur fort pouvoir discriminant. Avant l'apprentissage, les données ont été normalisées (centrées et réduites) afin de rendre les échelles des variables comparables — étape indispensable pour garantir une mesure de distance cohérente entre les observations [13].

Nous avons testé plusieurs valeurs du paramètre k ($k = 1, 3, 7, 50$), en évaluant la performance à l'aide d'une validation croisée à 5 plis (5-fold cross-validation). Cette procédure permet de réduire la variance des estimations de performance, en moyennant les scores obtenus sur plusieurs sous-échantillons du jeu de données.

Le tableau **??** présente les taux moyens de classification correcte obtenus pour chaque valeur de k :

k	Taux de précision moyen (%)
1	96.7
3	96.0
7	94.7
50	90.0

TABLE 2.1 – Taux de classification correcte selon k (validation croisée 5-fold)

Ces résultats montrent que les meilleures performances sont obtenues pour de petites

valeurs de k, en particulier $k = 1$. Ce constat est cohérent avec la structure du jeu de données Iris, où les classes sont bien séparées dans l'espace des caractéristiques sélectionnées.

On observe que la précision diminue progressivement lorsque k augmente. À partir de $k = 50$, une part importante des voisins pris en compte pour la décision appartient à des classes différentes, en particulier dans les zones proches des frontières. Cela entraîne un effet de sur-lissage, où les différences fines entre classes sont gommées au profit d'une décision plus « globale », mais potentiellement moins précise. Ce phénomène illustre le compromis classique entre biais et variance évoqué précédemment : un petit k capture bien les détails mais est sensible au bruit ; un grand k fournit des décisions plus stables, mais peut altérer la séparation entre classes [14, 9, 7].

Enfin, une visualisation graphique du nuage de points (cf. figure 2.1) confirme la bonne séparabilité des classes en fonction des caractéristiques choisies. Cela justifie l'efficacité de k-NN dans ce contexte, tout en soulignant l'importance du choix de k selon la structure des données.

2.5.3 Visualisation de la frontière de décision

En projetant les données sur deux dimensions — par exemple, la longueur et la largeur des pétales — il est possible de visualiser les frontières de décision induites par le classifieur k-NN. Les figures suivantes illustrent ces frontières pour différentes valeurs du paramètre k ($k = 1, 3, 7, 50$). Chaque graphique met en évidence la manière dont la valeur de k influence la forme et la complexité des régions de décision.

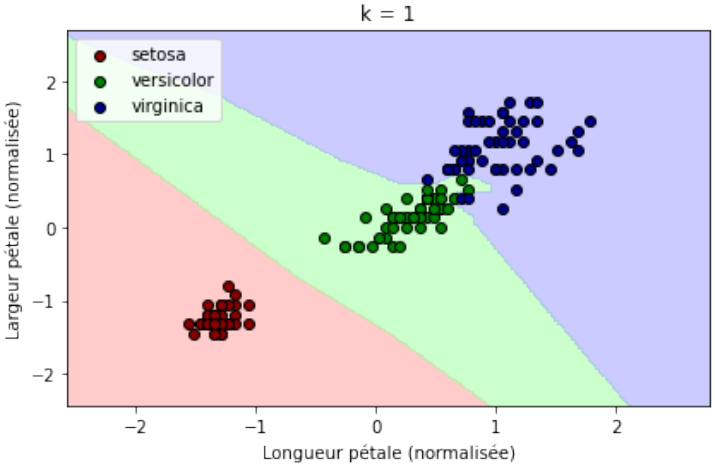

FIGURE 2.2 – Frontière de décision du classifieur k-NN pour $k = 1$

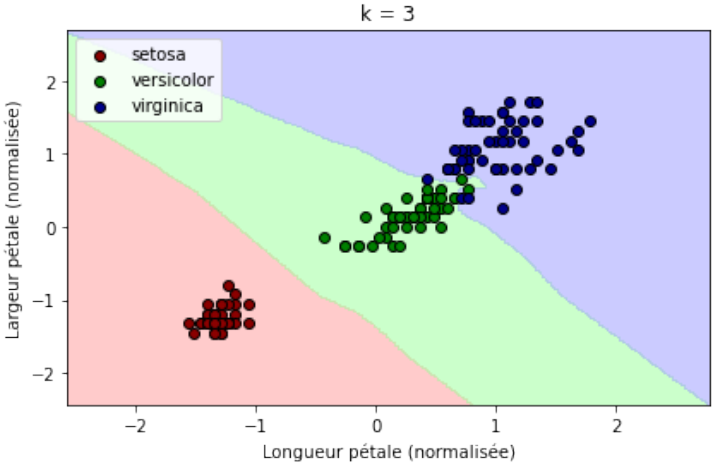

FIGURE 2.3 – Frontière de décision du classifieur k-NN pour $k = 3$

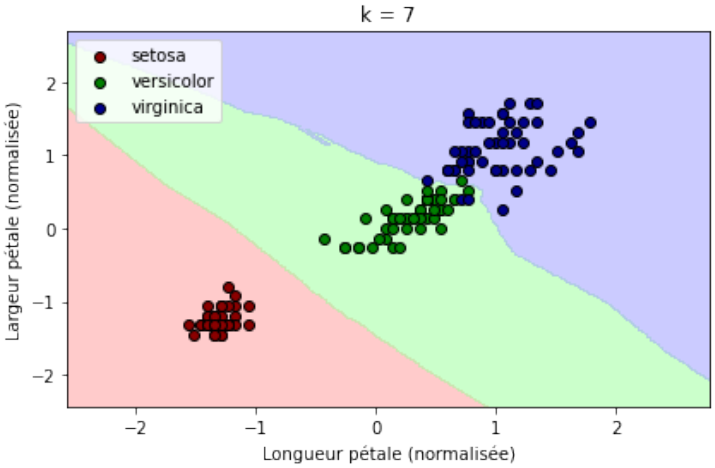

FIGURE 2.4 – Frontière de décision du classifieur k-NN pour $k = 7$

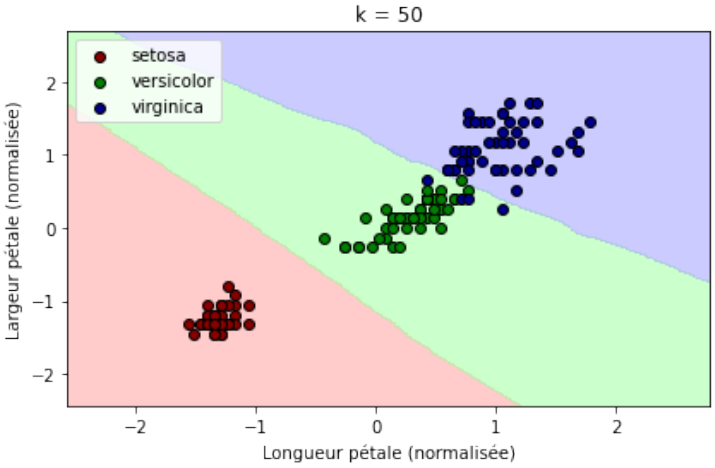

FIGURE 2.5 – Frontière de décision du classifieur k-NN pour $k = 50$

On observe l'évolution des frontières de décision en fonction de k :

— Pour $k = 1$, les frontières sont très fragmentées et s'adaptent très bien aux données d'entraînement. Cela peut entraîner un surajustement en cas de bruit ou d'observations atypiques.

— Pour $k = 3$, les frontières sont plus régulières tout en respectant la structure des groupes. Ce paramètre offre souvent un bon compromis entre stabilité et précision.

— Pour $k = 7$, les frontières deviennent encore plus lissées, au risque d'intégrer des points de classes différentes à proximité des frontières, ce qui peut dégrader la performance dans les zones de chevauchement.

— Pour $k = 50$, le lissage est très important. Les frontières tendent à être trop générales, ce qui peut entraîner une perte d'information fine et des erreurs accrues dans les régions proches des classes voisines.

2.5.4 Remarques

Cette expérimentation visuelle illustre clairement l'impact du paramètre k sur la complexité des frontières de décision et sur la performance globale du classifieur. Dans un contexte bien structuré comme le jeu Iris, une petite valeur de k peut suffire à obtenir de très bons résultats. Cependant, en présence de bruit, de classes déséquilibrées ou de données plus complexes, un ajustement plus soigneux du paramètre k est nécessaire. Des techniques complémentaires, comme la réduction de dimension, la pondération des voisins ou l'apprentissage de métriques, peuvent alors s'avérer utiles pour améliorer la robustesse du modèle [13, 7].

2.6 Application illustrative (2)

Afin d'illustrer le fonctionnement de la méthode des k-plus proches voisins (k-NN) en classification, considérons un exemple simple en deux dimensions. Supposons un jeu de données contenant deux classes distinctes, notées **Classe A** et **Classe B**, représentées par des points dans le plan $\mathbb{R}^2$. L'objectif est de prédire la classe d'un nouvel individu à partir des données existantes.

Données simulées

On suppose disposer de l'ensemble d'apprentissage suivant :

Observation	x_1	x_2	**Classe**
1	1.0	2.0	A
2	1.5	1.8	A
3	2.0	2.2	A
4	3.0	3.0	B
5	3.5	2.5	B
6	4.0	3.2	B

TABLE 2.2 – Ensemble d'apprentissage simulé en dimension 2.

On souhaite classer un nouveau point d'entrée $x = (2.5, 2.0)$.

Étapes de la classification avec k-NN

1. **Calcul des distances** : on commence par calculer la distance (euclidienne, par exemple) entre le point x à prédire et chacun des points de l'ensemble d'apprentissage. La distance euclidienne entre deux points $x = (x_1, x_2)$ et $x' = (x'_1, x'_2)$ est donnée par :
$$d(x, x') = \sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2}$$
2. **Classement des voisins** : on trie les distances obtenues en ordre croissant pour identifier les k plus proches voisins.
3. **Décision majoritaire** : la classe prédite est celle qui apparaît le plus fréquemment parmi les k voisins sélectionnés.

Résultat pour $k = 3$

Calculons les distances :

Observation	**Coordonnées**	**Distance à** $(2.5, 2.0)$	**Classe**
1	(1.0, 2.0)	1.50	A
2	(1.5, 1.8)	1.02	A
3	(2.0, 2.2)	0.54	A
4	(3.0, 3.0)	1.12	B
5	(3.5, 2.5)	1.12	B
6	(4.0, 3.2)	2.06	B

TABLE 2.3 – Distances euclidiennes entre le point à classer $(2.5, 2.0)$ et les observations de données simulées, pour $k = 3$ en classification k-NN.

Les 3 plus proches voisins sont les observations 3, 2 et 4. Leurs classes respectives sont : A, A, B.

Selon le choix de la majorité : **Classe A** (2 occurrences contre 1) $\Rightarrow$ Le point $x = (2.5, 2.0)$ est donc classé comme appartenant à la classe **A**.

Représentation graphique (schématique)

Une illustration de ce cas pourrait montrer :

— Les points de la classe A en bleu, ceux de la classe B en rouge.

— Le point à prédire en noir.

— Des cercles ou des lignes reliant les 3 plus proches voisins au point cible.

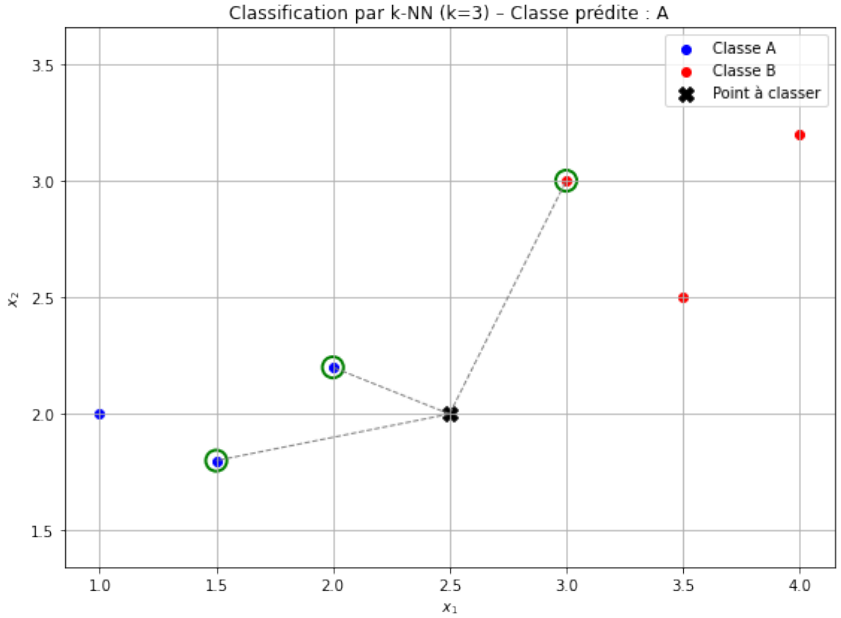

FIGURE 2.6 – Illustration de la classification par k-NN avec $k = 3$

Influence du paramètre k

On peut noter que si $k = 5$, les voisins les plus proches seraient alors :

— Observations : 3 (A), 2 (A), 4 (B), 5 (B), 1 (A)

— Classes : A, A, B, B, A $\Rightarrow$ Classe A (3 contre 2)

Le résultat reste le même. Cependant, si l'on avait pris $k = 4$, il y aurait eu égalité (2 A et 2 B), ce qui pose un problème de décision. Cela montre l'importance du choix du paramètre k.

2.7 Conclusion

La méthode des k-plus proches voisins (k-NN) se distingue par sa simplicité d'application et son efficacité dans des tâches de classification lorsque les données présentent une structure géométrique exploitable. En s'appuyant uniquement sur des mesures de distance entre les observations, elle permet d'établir des décisions sans modélisation probabiliste, ce qui en fait une méthode particulièrement adaptée dans des contextes où peu d'informations a priori sont disponibles sur les distributions [6].

Nous avons examiné dans ce chapitre ses fondements théoriques, son fonctionnement pratique, ainsi que les principaux paramètres qui influencent ses performances, en parti-

culier le choix du nombre de voisins k et la métrique utilisée. L'étude empirique réalisée sur le jeu de données *Iris* a mis en lumière les effets du paramètre k sur la stabilité et la précision du classifieur, confirmant le compromis classique entre biais et variance [14].

Malgré ses avantages, k-NN présente certaines limites, notamment une sensibilité aux dimensions élevées, au bruit et aux déséquilibres de classes. Plusieurs pistes d'amélioration, comme la pondération des voisins ou l'apprentissage de métriques adaptées, permettent d'en renforcer la robustesse [13].

Dans le chapitre suivant, nous étudierons l'utilisation de cette même méthode dans un cadre différent : l'estimation non paramétrique de densité, afin d'élargir la perspective sur ses applications en statistiques.

Chapitre 3

L'approche des k-plus proches voisins dans l'estimation de la fonction densité

Sommaire

3.1 Introduction

La méthode des k-plus proches voisins (k-NN), initialement développée pour la classification, a été étendue à l'estimation de la densité. Le principe repose sur l'idée intuitive que la densité d'un point peut être approximée à partir du volume minimal qui contient ses k plus proches voisins. Cette approche repose donc uniquement sur les données observées et la mesure de proximité choisie, ce qui la rend particulièrement attractive dans les contextes où peu d'informations sont disponibles sur la forme réelle de la distribution [1].

Plusieurs études, dont celles compilées dans [14], ont montré que cette méthode, bien que simple, possède de bonnes propriétés asymptotiques. Son efficacité dans des contextes réels et robustes a également été mise en évidence dans des travaux appliqués récents [1].

3.2 Principe fondamental de l'estimation par k-NN

3.2.1 Définition formelle de l'estimateur

L'estimation de densité par la méthode des k-plus proches voisins (k-NN) est une approche non paramétrique visant à approximer la densité de probabilité d'une variable aléatoire à partir d'un échantillon de données observées [1].

Soit $X_1, X_2, \ldots, X_n \in \mathbb{R}^d$ un échantillon issu d'une variable aléatoire X de densité f. Pour un point $x \in \mathbb{R}^d$, on note $R_k(x)$ la distance entre x et son k-ème plus proche voisin parmi les X_i. On définit la boule centrée en x et de rayon $R_k(x)$ par :

$$B(x, R_k(x)) = \left\{ y \in \mathbb{R}^d \ : \ \|y - x\| \leq R_k(x) \right\} \tag{3.1}$$

La méthode consiste à compter le nombre de points contenus dans cette boule, puis à estimer la densité $f(x)$ comme étant inversement proportionnelle au volume de cette boule. La formule de l'estimateur est donnée par :

$$\hat{f}_n(x) = \frac{k}{n \cdot V_d \cdot R_k(x)^d} \tag{3.2}$$

où

$V_d = \frac{\pi^{d/2}}{\Gamma\left(\frac{d}{2}+1\right)}$ *représente le volume de la boule unité en dimension* d ;

n *la taille de l'échantillon* ;

et d *la dimension de l'espace* [1, 30].

3.2.2 Interprétation géométrique et probabiliste

L'estimateur de densité par k-NN peut être interprété géométriquement comme une mesure de la concentration des observations autour d'un point donné x. Lorsque le rayon $R_k(x)$, défini comme la distance au k-ième plus proche voisin, est réduit, cela signifie que les données sont fortement regroupées près de x [1].

D'un point de vue probabiliste, cette méthode repose sur l'idée que la probabilité d'observer une valeur dans une région centrée en x peut être estimée empiriquement par la proportion $\frac{k}{n}$. L'estimateur de densité s'obtient alors en divisant cette proportion par le volume de la région considérée :

$$f(x) \approx \frac{p(X \in B(x, R_k(x)))}{\text{Vol}(B(x, R_k(x)))}$$

En pratique, cette probabilité est approchée par $\frac{k}{n}$, ce qui donne :

$$\hat{f}_n(x) = \frac{k}{n} \cdot \frac{1}{\text{Vol}(B(x, R_k(x)))}$$

En remplaçant le volume de la boule par son expression explicite dans $\mathbb{R}^d$, on retrouve la formule usuelle (3.2) [30, 2].

3.3 Propriétés statistiques

3.3.1 Biais

Pour analyser le biais, supposons que f est deux fois continûment différentiable. On approxime la densité dans une petite boule centrée en x par un développement de Taylor :

$$f(y) \approx f(x) + (y-x)^\top \nabla f(x) + \frac{1}{2}(y-x)^\top H_f(x)(y-x)$$

où $\nabla f(x)$ est le gradient de f en x, et $H_f(x)$ sa matrice hessienne.

En intégrant cette expression sur la boule $B(x, R_k(x))$, les termes linéaires s'annulent à cause de la symétrie de la boule autour de x. Il reste alors :

$$\mathbb{E}[\hat{f}_n(x)] = f(x) + C_1 \cdot R_k(x)^2 + o(R_k(x)^2) \tag{3.3}$$

où C_1 est une constante proportionnelle à la **trace de la matrice hessienne**, c'est-à-dire à la somme de ses éléments diagonaux (également appelée *laplacien* de f en x).
Sachant que $R_k(x)^d \approx \dfrac{k}{nf(x)V_d}$, on a :

$$R_k(x) \approx \left(\frac{k}{nf(x)V_d}\right)^{1/d}$$

et donc :

$$\boxed{\text{Biais} = \mathbb{E}[\hat{f}_n(x)] - f(x) = o\left(\left(\frac{k}{n}\right)^{2/d}\right)} \tag{3.4}$$

3.3.2 Variance

La variance de l'estimateur provient de la fluctuation du rayon $R_k(x)$, donc de la variation du volume $R_k(x)^d$.

En supposant que les observations sont indépendantes et identiquement distribuées, on peut montrer que :

$$\text{Var}[\hat{f}_n(x)] \approx \frac{f(x)}{n \cdot V_d^2 \cdot \mathbb{E}[R_k(x)^{2d}]} \tag{3.5}$$

Puisqu'on a approximativement :

$$\mathbb{E}[R_k(x)^{2d}] \propto \left(\frac{k}{n}\right)^2$$

on en déduit :

$$\boxed{\text{Var}[\hat{f}_n(x)] = o\left(\frac{1}{k}\right)} \tag{3.6}$$

Pour une démonstration détaillée de ces résultats, voir par exemple [23, 30, 2].

3.3.3 Erreur quadratique moyenne

En combinant le biais et la variance, l'erreur quadratique moyenne (MSE) s'écrit :

$$\boxed{\mathrm{MSE}[\hat{f}_n(x)] = o\left(\left(\frac{k}{n}\right)^{4/d} + \frac{1}{k}\right)} \tag{3.7}$$

Ce compromis entre biais et variance est central dans le choix du paramètre k, conditionnant la performance globale de l'estimateur [2, 1].

3.4 Consistance et convergence

3.4.1 Conditions de convergence de l'estimateur

L'estimateur $\hat{f}_n(x)$ est dit *consistant* s'il converge en probabilité vers la densité réelle $f(x)$ lorsque la taille de l'échantillon n tend vers l'infini. Pour que cela soit assuré dans le cadre du k-NN, il est nécessaire que le nombre de voisins $k = k_n$ vérifie les conditions suivantes :

$$k_n \to \infty \quad \text{et} \quad \frac{k_n}{n} \to 0 \quad \text{lorsque} \quad n \to \infty, \tag{3.8}$$

Autrement dit, le voisinage utilisé doit s'élargir avec n pour réduire la variance, tout en restant suffisamment petit pour que l'approximation de $f(x)$ reste proche à sa véritable valeur en ce point, limitant ainsi le biais [2, 6].

3.4.2 Types de convergence et théorèmes associés

Sous les conditions ci-dessus, on obtient la *convergence en probabilité* :

$$\hat{f}_n(x) \xrightarrow[n\to\infty]{\mathbb{P}} f(x), \tag{3.9}$$

ce qui constitue un résultat fondamental de consistance ponctuelle [2, 1].
Avec des hypothèses supplémentaires (continuité uniforme de f, support borné), on peut aussi démontrer une *convergence uniforme sur les compacts* :

$$\sup_{x\in\mathcal{C}} |\hat{f}_n(x) - f(x)| \xrightarrow[n\to\infty]{\mathbb{P}} 0, \tag{3.10}$$

pour tout compact $\mathcal{C} \subset \mathbb{R}^d$ [1]. La convergence presque sûre est également atteignable dans des contextes asymptotiques renforcés (voir [1, 4, 28]) :

$$\hat{f}_n(x) \xrightarrow[n\to\infty]{\text{a.s.}} f(x). \tag{3.11}$$

3.4.3 Choix quasi-optimal de k et malédiction de la dimension

Pour minimiser l'erreur quadratique moyenne, le choix quasi-optimal de k est donné asymptotiquement par :

$$k_n \propto n^{\frac{4}{4+d}}, \tag{3.12}$$

ce qui met en évidence l'impact de la dimension d. En effet, plus d est élevé, plus la taille du voisinage k doit être importante, ce qui ralentit la convergence et reflète la malédiction de la dimension [14, 2].

3.5 Choix du paramètre k

Le paramètre k joue un rôle central dans l'estimation de la densité par la méthode des k-plus proches voisins. Il contrôle le nombre d'observations utilisées autour du point x et agit ainsi comme un levier de régularisation, influençant directement le biais et la variance de l'estimateur.

3.5.1 Influence de k sur l'estimation

Le paramètre k module la taille du voisinage utilisé pour l'estimation :

— Lorsque k est petit, le rayon $R_k(x)$ est faible. L'estimation repose alors sur peu de points, ce qui permet de mieux suivre les variations locales de la densité, mais rend l'estimateur plus sensible au bruit, ce qui se traduit par une variance élevée.

— En revanche, lorsque k est grand, le rayon $R_k(x)$ augmente. L'estimation devient plus stable en moyenne, car elle intègre davantage d'observations, mais au prix d'un lissage excessif qui peut masquer les détails fins de la structure des données [14, 13, 24].

Ainsi, k contrôle un compromis classique :

$$\text{Biais} \uparrow \quad \text{et} \quad \text{Variance} \downarrow \quad \text{lorsque} \quad k \uparrow$$

3.5.2 Stratégies de sélection du paramètre k

Le choix optimal de k dépend de la structure des données, et plusieurs méthodes sont disponibles :

— **Validation croisée** : consiste à tester différentes valeurs de k pour minimiser l'erreur sur un ensemble de validation [13, 23].

— **Heuristiques pratiques** : formules simples comme $k = \sqrt{n}$ ou $k = \log n$, utilisées pour obtenir une valeur raisonnable sans optimisation coûteuse [23].

— **Méthodes adaptatives** : ajuster k en fonction de la densité estimée autour de x ; dans les régions à faible densité, on choisit généralement une valeur de k plus grande afin de garantir la stabilité de l'estimation [1, 4].

— **Critères théoriques** : les conditions de convergence imposent $k \to \infty$ et $k/n \to 0$ quand $n \to \infty$ [2, 1, 30]. Une relation quasi-optimale est donnée par $k \propto n^{\frac{4}{4+d}}$.

— **Méthodes avancées** : optimisation basée sur des critères comme la MISE, la vraisemblance, ou des approches bayésiennes [23, 2].

3.5.3 Effet du choix de k sur le biais et la variance

Le paramètre k contrôle directement le compromis biais-variance, illustré dans la décomposition du risque quadratique moyen (MSE) :

$$\text{MSE} = \text{Biais}^2 + \text{Variance} \tag{3.13}$$

Pour un petit k :

— Faible biais ;

— Forte variance : estimation instable et bruitée.

Pour un grand k :

— Biais élevé : sur-lissage ;

— Variance réduite : estimation plus stable.

Ainsi, le choix optimal de k est celui qui minimise la MSE [14, 13, 1].

Un petit k permet de capturer la structure fine des données, mais avec un risque élevé de sur-apprentissage.

Un grand k donne une estimation plus stable, mais avec un biais accru dû au lissage excessif [14, 13, 23].

Ce compromis est classique en estimation non paramétrique et constitue un point clé dans le paramétrage des algorithmes statistiques [14, 23, 24].

Remarque : Il faut noter que la dimension d joue un rôle critique ici : en haute dimension, le voisinage requis pour couvrir une région de densité suffisante devient rapidement très large, ce qui impose d'augmenter k pour éviter des estimateurs trop instables. Ce phénomène est une manifestation classique de la malédiction de la dimension [14, 13].

3.6 Application illustrative

Afin d'illustrer concrètement l'utilisation de l'estimateur de densité par la méthode des k-plus proches voisins, nous présentons ici un exemple simple inspiré de [4].

Considérons un échantillon unidimensionnel donné par :

$$\mathcal{X} = \{1,\, 2,\, 6,\, 11,\, 13,\, 14,\, 20,\, 33\}$$

Nous souhaitons estimer la densité en $x = 5$ en utilisant la méthode k-NN pour deux valeurs de k : $k = 2$ et $k = 5$. L'échantillon contient $n = 8$ observations, et la dimension est $d = 1$.

a) Cas $k = 2$

On commence par calculer la distance entre $x = 5$ et chacun des points de $\mathcal{X}$:

$$\{|5-1|, |5-2|, |5-6|, |5-11|, |5-13|, |5-14|, |5-20|, |5-33|\} = \{4, 3, 1, 6, 8, 9, 15, 28\}$$

Les deux plus proches voisins de 5 sont donc les points à distance 1 et 3, ce qui donne :

$$R_2(5) = 3$$

L'estimateur de densité s'écrit alors :

$$\widehat{f}_{\text{knn}}(5) = \frac{k}{n \cdot 2R_k(5)} = \frac{2}{8 \cdot 2 \cdot 3} = \frac{1}{24}$$

b) Cas $k = 5$

Cette fois, le 5$^{\text{e}}$ plus proche voisin de $x = 5$ se trouve à une distance de 8, donc :

$$R_5(5) = 8$$

d'où :

$$\widehat{f}_{\text{knn}}(5) = \frac{5}{8 \cdot 2 \cdot 8} = \frac{5}{128}$$

Interprétation

On observe que le choix de k influence fortement la valeur estimée de la densité. Un petit k conduit à un rayon plus réduit, ce qui donne une estimation plus sensible aux variations fines des données. En revanche, un k plus grand agrandit le voisinage utilisé, ce qui conduit à un effet de lissage plus important. Ce comportement est comparable à celui du paramètre de bande dans les méthodes à noyau, et justifie l'importance du choix adapté de k selon la taille de l'échantillon et la structure des données.

3.7 Limites et remarques pratiques

La méthode des k-plus proches voisins (k-NN) est simple, intuitive et efficace dans de nombreux cas. Toutefois, elle présente certaines limites importantes, notamment en haute dimension ou lorsque le choix du paramètre k est mal adapté.

Problèmes liés à la dimension

Lorsque la dimension des données augmente, la méthode devient moins performante. Ce phénomène est appelé la *malédiction de la dimension* [14, 2].

Explication : Pour inclure k voisins autour d'un point x, il faut agrandir la zone de recherche. En dimension d, le volume d'une boule de rayon r est proportionnel à r^d.

Ainsi, pour une base de données de taille n, il faut environ [13, 14] :

$$V_d \cdot R_k(x)^d \approx \frac{k}{n} \quad \Rightarrow \quad R_k(x) \approx \left(\frac{k}{nV_d}\right)^{1/d}$$

Plus d est grand, plus $R_k(x)$ s'élargit, ce qui rend l'estimation moins ciblée [14, 2]. Les données deviennent clairsemées, les distances augmentent, et la notion de "proche voisin" perd de sa pertinence. De plus, des points isolés peuvent recevoir une densité élevée, même dans des régions quasi vides [23].

Choix délicat de k

Le paramètre k contrôle le compromis biais-variance [14, 23] :

— Un k trop petit : estimation instable, sensible au bruit (variance élevée),

— Un k trop grand : l'estimation est trop lissée, perd en précision (biais élevé).

Exemple : Avec $n = 1000$, $d = 5$ et $k = 20$, on obtient :

$$R_k(x) \approx \left(\frac{20}{1000}\right)^{1/5} \approx 0{,}55$$

Cela signifie qu'il faut explorer une grande portion de l'espace, ce qui nuit à la précision de l'estimation autour de x [2].

Autres limites pratiques

— **Pas de lissage explicite** : contrairement à la méthode à noyau, k-NN n'utilise pas de fonction de pondération continue ; tous les voisins ont le même poids [14, 23].

— **Coût computationnel** : calculer les distances pour chaque point devient coûteux lorsque n est grand. Des structures de données comme les *kd-trees* ou *ball-trees* (arbres qui partitionnent l'espace pour accélérer la recherche des voisins) peuvent réduire ce coût, mais leur efficacité décroît en grande dimension [23].

— **Sensibilité au bruit** : les observations atypiques ont un impact fort lorsque k est petit [29, 23].

— **Estimation centrée sur les données voisines** : l'estimateur k-NN repose uniquement sur les observations proches, sans modèle global explicite [14].

Solutions proposées

Pour surmonter certaines de ces limites, plusieurs solutions ont été proposées :

— k **adaptatif** : faire varier dynamiquement k selon la densité des données autour du point considéré [1].

— **Distances apprises** : utiliser des métriques plus adaptées, comme la distance de Mahalanobis ou des distances apprises par apprentissage automatique [1, 26].

— **Méthodes hybrides** : combiner k-NN avec d'autres techniques comme les noyaux ou les méthodes à partitionnement [14, 23].

3.8 Conclusion

Dans ce chapitre, nous avons étudié l'approche des k-plus proches voisins (k-NN) appliquée à l'estimation de la fonction de densité. Nous avons tout d'abord présenté les fondements théoriques de la méthode, en mettant en évidence son fonctionnement, ses propriétés, ainsi que les conditions sous lesquelles elle fournit des estimations cohérentes.

Nous avons ensuite introduit plusieurs variantes et améliorations destinées à pallier certaines limitations du modèle de base, telles que l'ajustement adaptatif de k, l'usage de métriques apprises ou encore les combinaisons avec d'autres techniques non paramétriques.

Une application concrète sur des données univariées a permis d'illustrer la méthode et de visualiser le comportement de l'estimateur en fonction du choix de k. Enfin, nous avons discuté les principales limites pratiques de cette méthode, notamment sa sensibilité à la dimension, aux points atypiques, au choix du paramètre k, et l'absence de lissage explicite.

Ces constats soulignent l'intérêt de considérer des approches alternatives ou complémentaires pour l'estimation de densité, en particulier la méthode à noyau, déjà introduite dans le chapitre 1, et qui fera l'objet d'une comparaison pratique dans le chapitre 4.

Chapitre 4

Application et comparaison avec la méthode à noyau

Sommaire

4.1 Introduction

Ce chapitre est consacré à l'application pratique des méthodes d'estimation non paramétrique abordées précédemment, en particulier la méthode des k-NN et la méthode à noyau. À travers des exemples numériques et des comparaisons graphiques, nous analysons leurs performances respectives, notamment en termes de précision et de sensibilité aux paramètres. Cette étude vise à mettre en évidence les avantages et limites de chaque méthode dans un cadre appliqué.

4.2 Présentation de l'étude

Dans cette section, nous présentons le cas d'étude utilisé pour illustrer les méthodes d'estimation non paramétriques développées dans ce qui suit. Les données sont simulées selon deux lois de probabilité, chacune avec deux configurations de paramètres :

- La loi normale $\mathcal{N}(\mu, \sigma^2)$:
  - Cas 1 : $\mu = 0,\ \sigma = 1$;
  - Cas 2 : $\mu = 1,\ \sigma = 1$.
- La loi Gamma, notée $\Gamma(\alpha, \beta)$:
  - Cas 1 : $\alpha = 2,\ \beta = 2$;
  - Cas 2 : $\alpha = 5,\ \beta = 1$.

Pour chaque configuration, des échantillons de tailles $n = 50$, 100 et 1000 sont générés.

L'estimation de la densité de la variable aléatoire X est réalisée à l'aide de deux approches non paramétriques :

- La méthode des k-plus proches voisins (k-NN), avec différentes valeurs de k choisies en fonction de la taille de l'échantillon :
  - $n = 50 : k \in \{10,\ 20,\ 50\}$;
  - $n = 100 : k \in \{20,\ 50,\ 100\}$;
  - $n = 1000 : k \in \{20,\ 50,\ 100,\ 500,\ 1000\}$.
- La méthode à noyau, en utilisant deux types de noyaux (noyau gaussien et noyau gamma). Les estimations sont effectuées pour différentes valeurs de la bande passante :

$$h \in \{0{,}05,\ 0{,}1,\ 0{,}2,\ 0{,}5,\ 1{,}0\}.$$

La qualité des estimations est évaluée à l'aide de l'erreur quadratique intégrée moyenne (MISE), estimée numériquement sur une grille de points en comparant la densité estimée à la densité réelle.

Remarque : L'objectif de cette étude n'est pas de rechercher les valeurs optimales de k ou h. Ces paramètres sont fixés à des valeurs choisies uniquement dans un but comparatif. Leur optimisation ne fait pas partie du cadre de ce travail.

Méthodologie pratique d'implémentation

Les simulations et estimations ont été réalisées en **Python**, à l'aide des bibliothèques `NumPy`, `SciPy` et `Matplotlib`. Les échantillons ont été simulés selon les configurations décrites précédemment. Les estimations de densité ont été obtenues en appliquant directement la méthode des k-NN pour les valeurs de k fixées, ainsi que la méthode à noyau pour les valeurs de bande passante h spécifiées, sans recherche d'optimisation. Les performances ont été comparées par le calcul numérique du MISE et par l'analyse graphique des densités estimées.

4.3 Estimation de la densité par la méthode des k-NN

Nous appliquons ici la méthode k-NN pour estimer la densité de la variable aléatoire X, en utilisant les échantillons simulés selon les lois et configurations déjà décrites.

Les valeurs de k choisies varient selon la taille de l'échantillon, afin d'évaluer l'effet du paramètre sur la qualité de l'estimation. Pour chaque configuration, nous comparons graphiquement la densité estimée à la densité réelle.

Les figures ci-dessous regroupent les résultats pour chaque loi considérée.

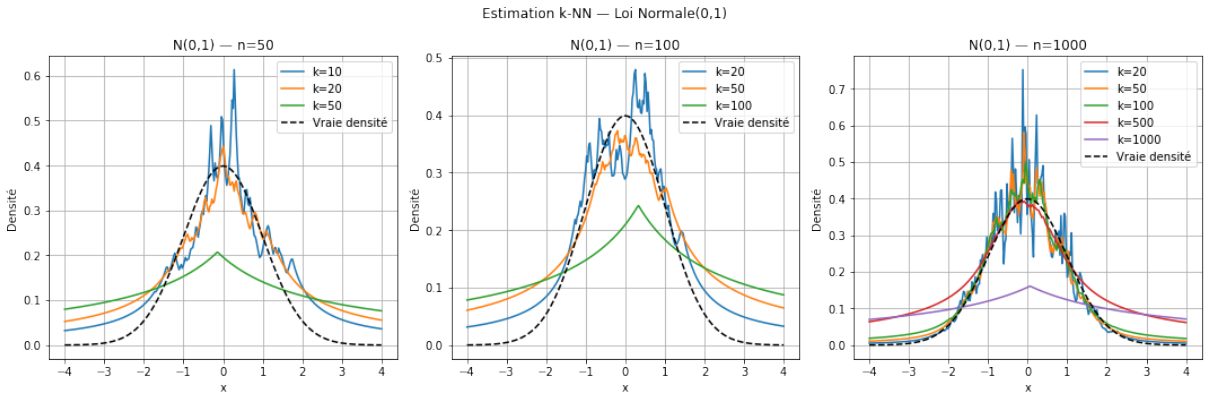

FIGURE 4.1 – Estimation k-NN de la densité — Loi $\mathcal{N}(0,1)$ pour $n = 50$, $n = 100$, $n = 1000$

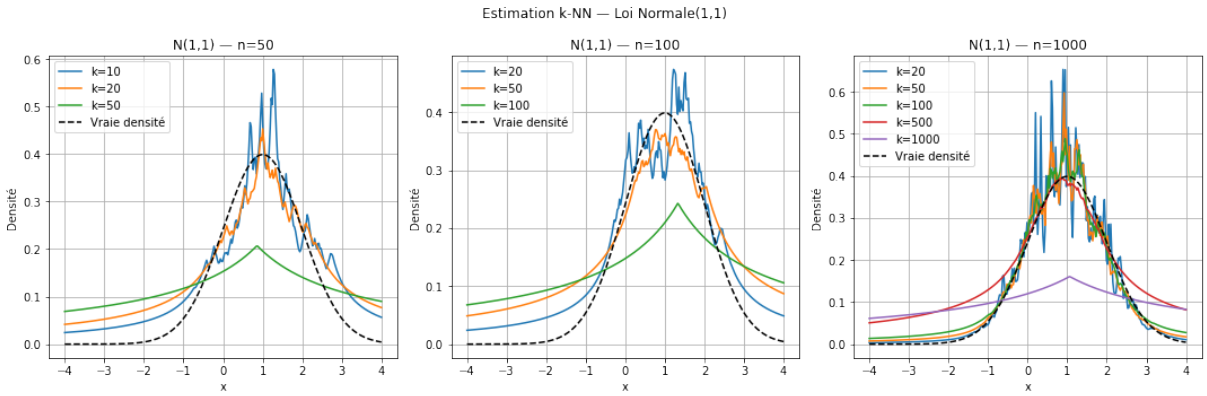

FIGURE 4.2 – Estimation k-NN de la densité — Loi $\mathcal{N}(1,1)$ pour $n = 50$, $n = 100$, $n = 1000$

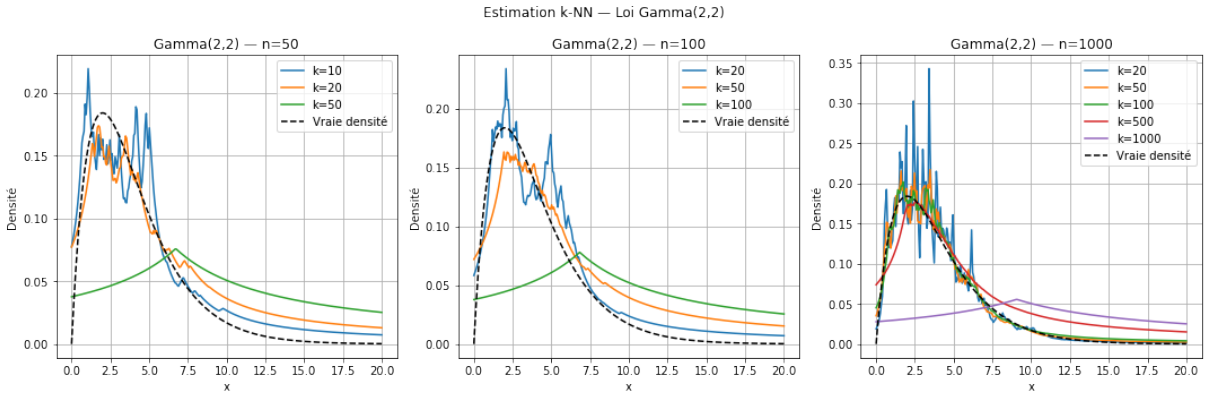

FIGURE 4.3 – Estimation k-NN de la densité — Loi $\Gamma(2,2)$ pour $n = 50$, $n = 100$, $n = 1000$

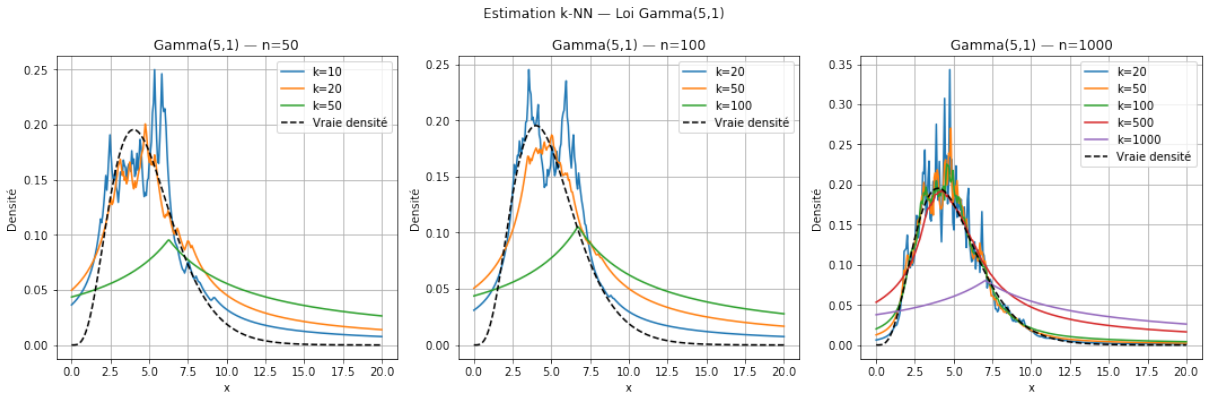

FIGURE 4.4 – Estimation k-NN de la densité — Loi $\Gamma(5,1)$ pour $n = 50$, $n = 100$, $n = 1000$

Interprétation

— **Figure 4.1 et Figure 4.2** : Estimation de la densité pour des lois normale $\mathcal{N}(0,1)$ et $\mathcal{N}(1,1)$ avec différentes tailles d'échantillon ($n = 50$, $n = 100$, $n = 1000$).

  — Pour $n = 50$, l'estimation peut être très irrégulière en raison du manque de données.

  — Pour $n = 1000$, l'estimation devrait se rapprocher de la vraie densité (courbe plus lisse et plus précise).

  — Le choix de k influence la régularité : un k trop petit donne une estimation bruitée, un k trop grand lisse trop la densité (l'estimateur s'aplatie).

— **Figures 4.3 et 4.4** : Estimation pour des lois Gamma $\Gamma(2,2)$ et $\Gamma(5,1)$.

  — La méthode k-NN peut avoir des difficultés à capturer les queues de distribution des lois Gamma.

  — Pour $n = 1000$, l'estimation devrait mieux suivre la forme théorique, surtout si k est bien choisi.

4.4 Estimation de la densité par la méthode à noyau

Cette section présente les résultats obtenus en appliquant la méthode à noyau à différents échantillons simulés. Pour chaque configuration, nous traçons l'estimation obtenue et la comparons à la densité.

Les figures illustrent l'influence de la taille de l'échantillon, du paramètre de lissage h et le choix de la fonction noyau sur la précision de l'estimation.

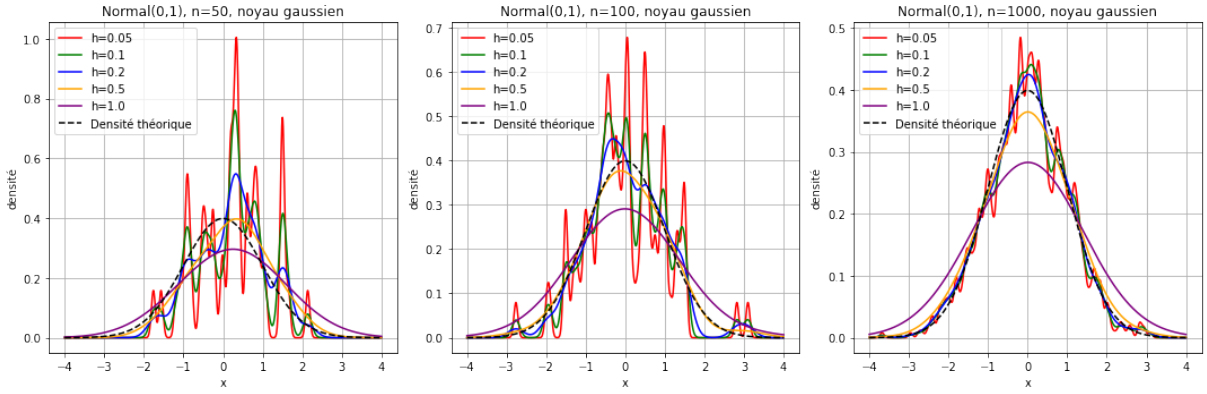

FIGURE 4.5 – Estimation de densité pour la loi $\mathcal{N}(0,1)$ avec noyau gaussien et différentes tailles d'échantillon.

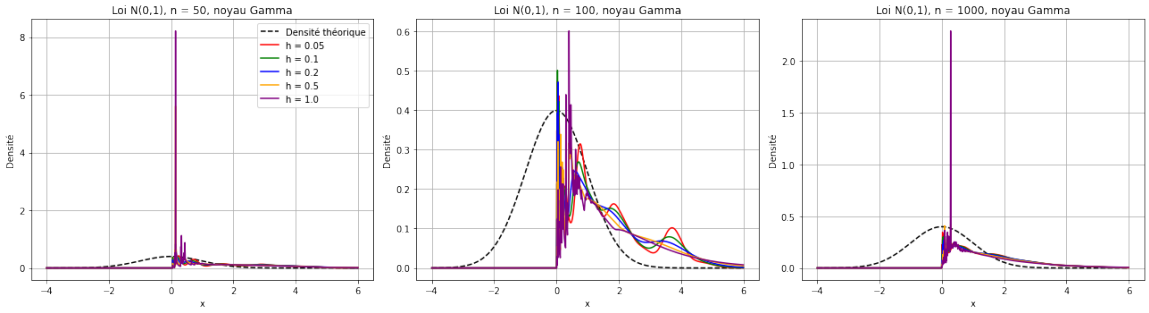

FIGURE 4.6 – Estimation de densité de la loi $\mathcal{N}(0,1)$ par la méthode à noyau avec noyau Gamma, pour $n = 50$, 100, et 1000.

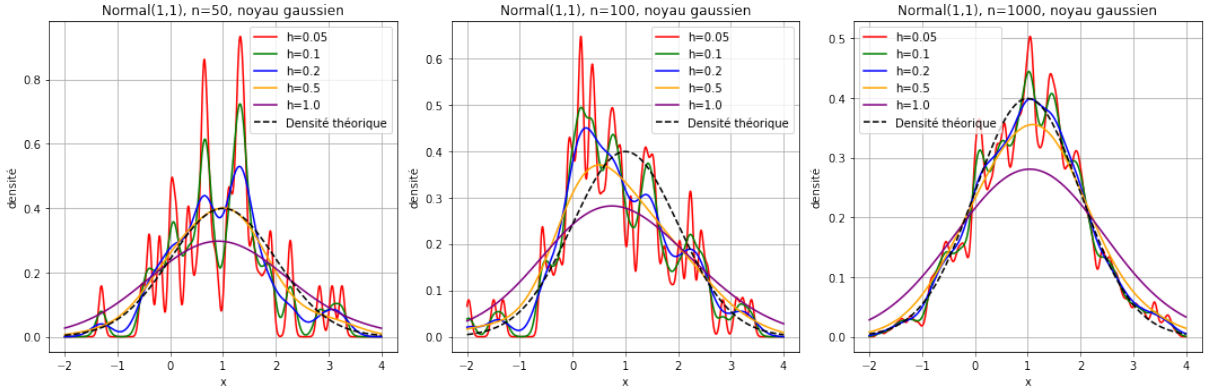

FIGURE 4.7 – Estimation de densité pour la loi $\mathcal{N}(1,1)$ avec noyau gaussien et différentes tailles d'échantillon.

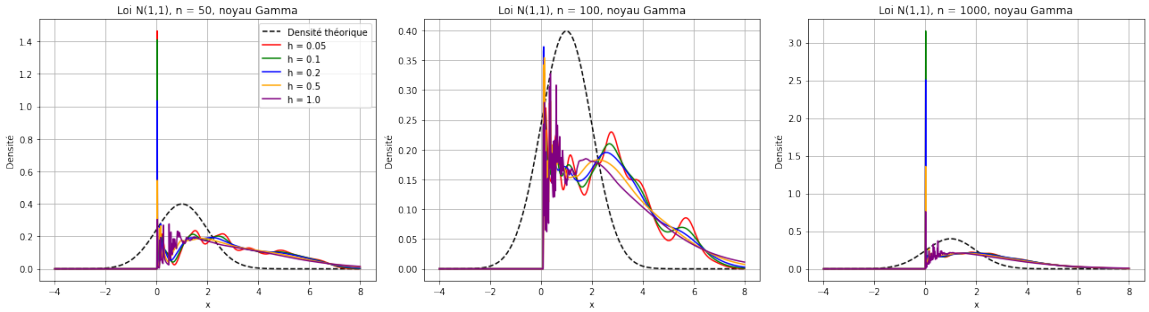

FIGURE 4.8 – Estimation de densité de la loi $\mathcal{N}(1,1)$ par la méthode à noyau avec noyau Gamma, pour $n = 50$, 100, et 1000.

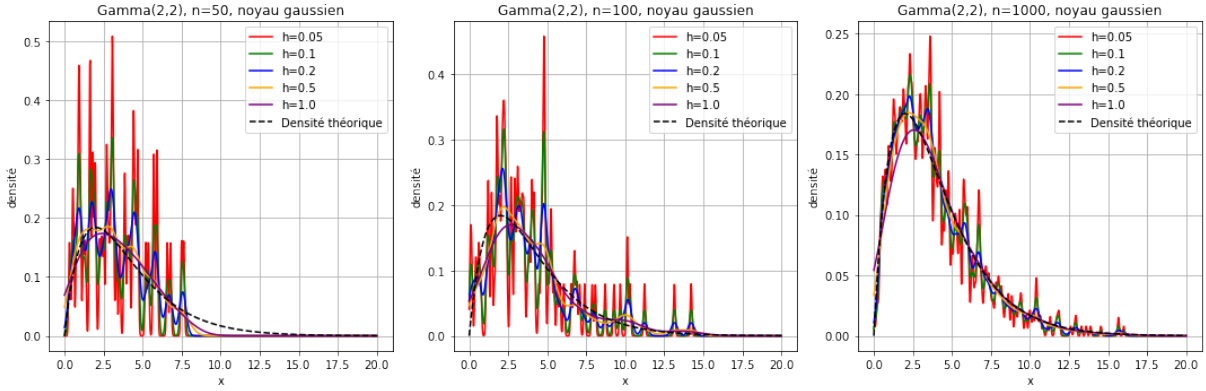

FIGURE 4.9 – Estimation de densité pour la loi Gamma$(2, 2)$ avec noyau gaussien.

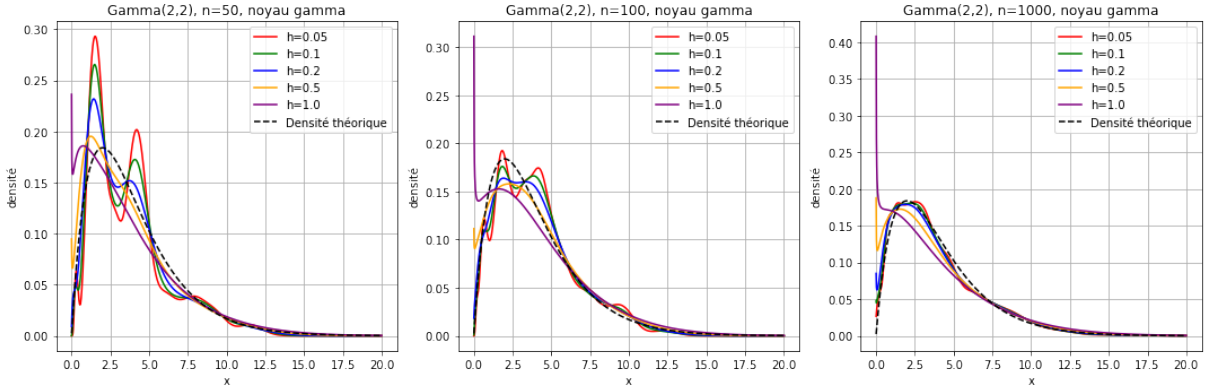

FIGURE 4.10 – Estimation de densité pour la loi Gamma$(2, 2)$ avec noyau gamma.

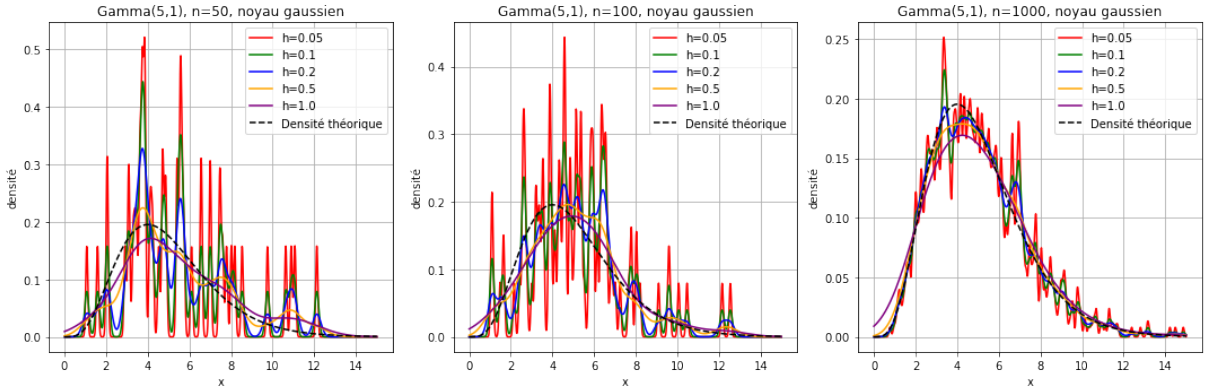

FIGURE 4.11 – Estimation de densité pour la loi Gamma$(5, 1)$ avec noyau gaussien.

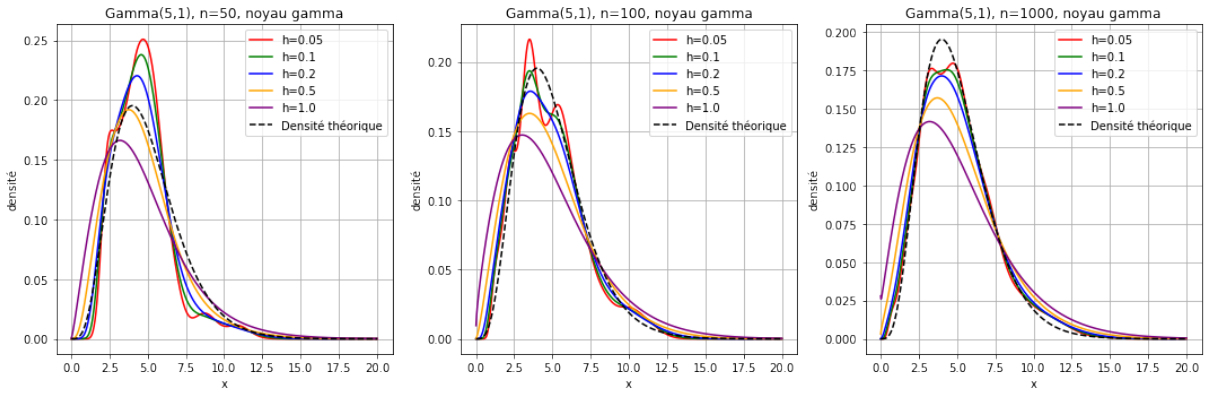

FIGURE 4.12 – Estimation de densité pour la loi Gamma$(5, 1)$ avec noyau gamma.

Interprétation

1. Loi Normale $\mathcal{N}(0, 1)$ et $\mathcal{N}(1, 1)$

Noyau Gaussien (Figures 4.5, 4.7) :

— Pour $n = 50$: Estimation très irrégulière (nombreux pics étroits), **précisement pour un h trop petit**.

— Pour $n = 1000$: Courbe lisse et proche de la vraie densité (**avec un bon choix de** h).

Discussion : Le noyau Gaussien convient aux lois normales, mais son efficacité est étroitement liée au réglage de h et à la taille d'échantillon n.

Noyau Gamma (Figure 4.6, 4.8) :

— L'estimation s'annule pour $x < 0$, ce qui est incompatible avec une loi Normale.

2. Loi Gamma $\Gamma(2, 2)$ et $\Gamma(5, 1)$

Noyau Gaussien (Figures 4.9, 4.11) :

— Pour $n = 50$: Estimation instable avec des pics mal calibrés et une forme globale déformée (notamment pour h petit).

— Pour $n = 1000$: Estimation améliorée, notamment pour bon choix de h.

Discussion : Bien que le noyau Gaussien donne des résultats acceptables pour $\Gamma(5, 1)$, il reste inadapté à $\Gamma(2, 2)$ en raison de son support.

Noyau Gamma (Figures 4.10, 4.12) :

- **Gamma(2,2)** : Estimation précise avec $h = 0.2$ (bleu), mais très bruitée pour $n = 50$ et $h = 0.05$ (rouge).
- **Gamma(5,1)** : Meilleure stabilité. Les valeurs de h entre 0.2 et 0.5 (bleu / orange) donnent de bons résultats même pour $n = 50$.
- **Impact de n** : Avec $n = 50$, l'estimation est trop fluctuante ; avec $n = 1000$ et $h = 0.2$, le noyau Gamma donne des résultats fiables et cohérents.
- **Choix de h** :
  - h trop petit (0.05), ce qui entraîne un sur-ajustement (l'estimateur s'adapte excessivement aux données, en suivant les moindres variations aléatoires).
  - h trop grand (1.0), ce qui provoque un sous-ajustement (l'estimateur est trop lisse et masque la structure réelle de la densité).
- **Noyau Gamma** : Adapté aux deux distributions, mais plus critique pour la loi Gamma(2,2) (asymétrique).

4.5 Comparaison globale des deux méthodes

Dans cette section, nous proposons une comparaison globale entre la méthode des k-NN et l'estimation par noyau. Cette comparaison repose à la fois sur les MISE, et sur l'analyse visuelle des graphiques d'estimation.

4.5.1 Comparaison qualitative des méthodes

Cette section s'appuie sur les graphiques générés précédemment, qui illustrent les courbes de densité estimées par les méthodes k-NN et à noyau, selon la taille de l'échantillon et la loi de probabilité sous-jacente.

- **La méthode k-NN** est plus flexible pour des lois asymétriques (comme Gamma), et plus stable même pour de petits échantillons, à condition de bien choisir k.
- **La méthode à noyau** donne de bons résultats pour des lois symétriques (comme la normale), mais est plus sensible au choix de h, surtout avec peu de données.

— **En résumé**, la performance dépend fortement du bon réglage de k ou h, ce qui souligne l'importance de méthodes de sélection automatique (comme la validation croisée).

4.5.2 Comparaison quantitative (MISE)

Nous présentons ci-dessous les tableaux synthétiques des MISE pour différentes tailles d'échantillons ($n = 50, 100, 1000$), en fonction des paramètres k (pour k-NN) et h (pour la méthode à noyau), pour chaque loi étudiée.

Méthode	**n**	**k**	10	20	50	100	1000
k-NN	50	MISE	0.005669	0.004376	0.009822	–	–
	100	MISE	0.005322	0.002501	0.003598	0.010539	–
	1000	MISE	0.004790	0.001777	0.000757	0.000494	0.012469
Noyau Gaussien		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.014079	0.005623	0.002431	0.000985	0.002629
	100	MISE	0.006008	0.003173	0.001336	0.000556	0.002561
	1000	MISE	0.000649	0.000289	0.000161	0.000341	0.002583
Noyau Gamma		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.013813	0.006113	0.004574	0.012698	0.021097
	100	MISE	0.005899	0.003653	0.003498	0.010604	0.021468
	1000	MISE	0.000764	0.000860	0.002709	0.010708	0.022144

TABLE 4.1 – MISE pour l'estimation de la densité de Normale$(0, 1)$

Interprétation

Le noyau gaussien donne les meilleures performances pour la loi $\mathcal{N}(0, 1)$, avec une MISE minimale de 0,000161 pour $n = 1000$ et $h = 0{,}2$. À l'inverse, le noyau Gamma est moins adapté, tandis que la méthode k-NN reste compétitive pour des valeurs modérées de k, mais se dégrade lorsque k devient trop grand.

Méthode	**n**	**k**	10	20	50	100	1000
k-NN	50	MISE	0.004291	0.003133	0.008867	–	–
	100	MISE	0.004397	0.001855	0.003512	0.008890	–
	1000	MISE	0.003995	0.001571	0.000545	0.000448	0.010623
Noyau Gaussien		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.108013	0.050438	0.020153	0.008789	0.022225
	100	MISE	0.053248	0.026940	0.013387	0.005474	0.020973
	1000	MISE	0.005609	0.002809	0.001180	0.002795	0.020995
Noyau Gamma		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.130295	0.072607	0.050319	0.099665	0.340073
	100	MISE	0.075711	0.044317	0.034340	0.092029	0.315219
	1000	MISE	0.016286	0.015113	0.020415	0.079085	0.306808

TABLE 4.2 – MISE pour l'estimation de la densité de Normale$(1, 1)$

Interprétation

Pour $\mathcal{N}(1, 1)$, la méthode k-NN donne la meilleure MISE (0,000448 pour $k = 100$, $n = 1000$), surpassant nettement les méthodes à noyau. Les performances limitées des noyaux, surtout gamma.

Méthode	**n**	**k**	10	20	50	100	1000
k-NN	50	MISE	0.000781	0.000493	0.003213	–	–
	100	MISE	0.000725	0.000346	0.000369	0.003458	–
	1000	MISE	0.001061	0.000396	0.000152	0.000083	0.004314
Noyau Gaussien		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.111521	0.052022	0.025506	0.008849	0.005167
	100	MISE	0.054927	0.027093	0.012671	0.005146	0.003580
	1000	MISE	0.005626	0.002815	0.001318	0.000786	0.002118
Noyau Gamma		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.108358	0.055769	0.027502	0.020041	0.055780
	100	MISE	0.055983	0.026525	0.013360	0.012242	0.050204
	1000	MISE	0.005405	0.002765	0.001620	0.005624	0.043834

TABLE 4.3 – MISE pour l'estimation de la densité de Gamma$(2, 2)$

Interprétation

Pour la loi Gamma$(2, 2)$, la méthode k-NN atteint la meilleure précision (MISE minimale de 0,000083 pour $k = 100$, $n = 1000$), surpassant nettement les méthodes à noyau. Les noyaux gaussien et gamma sont moins adaptés à cette distribution asymétrique, avec des erreurs bien plus élevées, surtout pour les petits échantillons.

Méthode	**n**	**k**	10	20	50	100	1000
k-NN	50	MISE	0.000927	0.000745	0.002496	–	–
	100	MISE	0.000922	0.000412	0.000683	0.003193	–
	1000	MISE	0.001086	0.000443	0.000159	0.000123	0.004033
Noyau Gaussien		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.115078	0.053730	0.030912	0.009243	0.005115
	100	MISE	0.055166	0.027835	0.012050	0.004547	0.003239
	1000	MISE	0.005513	0.002751	0.001169	0.000660	0.001790
Noyau Gamma		**h**	0.05	0.1	0.2	0.5	1.0
	50	MISE	0.106484	0.056265	0.027096	0.008776	0.006129
	100	MISE	0.054787	0.026134	0.014005	0.004419	0.004829
	1000	MISE	0.005556	0.002645	0.001291	0.000580	0.002510

TABLE 4.4 – MISE pour l'estimation de la densité de Gamma$(5, 1)$

Interprétation

Pour la loi Gamma$(5, 1)$, la méthode k-NN obtient la meilleure MISE (0,000123 pour $k = 100$, $n = 1000$), surpassant les noyaux. Les méthodes à noyau restent compétitives mais moins précises, surtout pour les petits échantillons.

4.5.3 Analyse globale des performances des deux méthodes

La comparaison des MISE selon plusieurs configurations statistiques (lois, effectifs, paramètres) permet de mieux cerner les domaines d'efficacité respectifs des méthodes k-NN et à noyau.

— **Comparaison des méthodes selon la loi de probabilité**

— Pour la loi $\mathcal{N}(0, 1)$, la **méthode à noyau** est systématiquement supérieure, avec des MISE nettement plus faibles, quel que soit n.

— Pour la loi $\mathcal{N}(1, 1)$, les résultats sont inversés : c'est la **méthode k-NN** qui donne de meilleures estimations, particulièrement pour les petits et moyens échantillons.

— Pour les lois **Gamma** asymétriques $\Gamma(2, 2)$ et $\Gamma(5, 1)$, la **méthode k-NN surpasse** la méthode à noyau.

- **Impact de la taille de l'échantillon (n)**
  - Lorsque n augmente, la performance des deux méthodes s'améliore globalement (MISE décroissante).
- **Sensibilité aux paramètres (k et h)**
  - Pour les deux méthodes, le choix des paramètres est **déterminant** pour obtenir une estimation de qualité.
  - Les meilleurs résultats sont souvent obtenus pour $k = 100$ ou $k = 20$ en k-NN, et pour $h = 0{,}2$ ou $h = 0{,}5$ dans la méthode à noyau.
  - Un mauvais choix de paramètre conduit rapidement à une MISE élevée (ex. : $h = 0{,}05$ ou $k = 10$).
- **Comportement par défaut des méthodes :**
  - La méthode à noyau semble mieux adaptée aux lois symétriques, en particulier la loi normale centrée réduite.
  - La méthode k-NN s'adapte mieux aux lois asymétriques et symétriques.

En résumé, aucune méthode n'est meilleure de manière absolue. Le choix entre k-NN et noyau dépend :

- du type de loi (symétrique vs. asymétrique),
- de la taille de l'échantillon,
- et du bon choix des paramètres (k ou h optimal).

4.6 Conclusion

L'étude comparative a montré que les méthodes k-NN et à noyau ont des performances variables selon la loi de probabilité, la taille de l'échantillon et le choix des paramètres. La méthode à noyau est plus adaptée aux lois symétriques, tandis que le k-NN excelle sur des lois asymétriques comme les lois Gamma. Aucun estimateur n'est supérieur en toutes circonstances : le choix dépend du contexte et des caractéristiques des données.

Conclusion générale

Ce travail a été consacré à l'étude de méthodes non paramétriques d'estimation de densité, en particulier la méthode des k-plus proches voisins et la méthode à noyau. Ces approches permettent de construire une estimation de la densité d'une variable aléatoire à partir d'un échantillon, sans supposer de forme particulière pour la loi sous-jacente. Cela les rend particulièrement utiles dans des situations où la distribution réelle des données est inconnue ou difficile à modéliser de manière paramétrique.

Dans une première partie, nous avons présenté les fondements théoriques de la méthode à noyau, en insistant sur le rôle crucial du choix du noyau et de la bande passante h, qui influence directement le biais et la variance de l'estimateur. Nous avons également exploré ses propriétés statistiques, telles que la consistance et la convergence.

La seconde partie du mémoire a porté sur la méthode des k-plus proches voisins, abordée d'abord dans un contexte de classification, puis appliquée à l'estimation de densité. Cette méthode, facile à implémenter, repose sur l'idée intuitive que les observations proches d'un point contiennent une information pertinente sur la densité au voisinage de ce point. Là encore, le choix du paramètre k s'est révélé essentiel pour la qualité des résultats.

Enfin, une étude comparative sur des données simulées nous a permis d'analyser les performances des deux méthodes selon plusieurs critères : loi sous-jacente, taille de l'échantillon, et sensibilité aux paramètres.

Ce travail ouvre la voie à plusieurs perspectives intéressantes. D'une part, il serait pertinent de développer des stratégies d'optimisation automatique des paramètres k et h (par validation croisée, critères adaptatifs, etc.), afin de garantir la performance sans avoir à tester manuellement différentes valeurs. D'autre part, une extension vers des contextes multidimensionnels permettrait d'analyser l'impact de la dimension sur les performances. Enfin, l'étude pourrait être enrichie par la comparaison avec d'autres approches non pa-

ramétriques comme les réseaux de neurones ou les arbres de décision.

Ce mémoire nous a permis d'approfondir notre compréhension de l'estimation de densité en statistique, en mettant en évidence l'intérêt et la flexibilité des méthodes non paramétriques, ainsi que l'importance du compromis biais-variance et du choix des paramètres pour garantir de bonnes performances.

Bibliographie

[1] N. Belatreche. *Estimation non paramétrique de quelques fonctions robustes par la méthode des k plus proches voisins*. PhD thesis, Université Kasdi Merbah Ouargla, (2024). Thèse de doctorat, Faculté des Mathématiques et des Sciences de la Matière, Département de Mathématiques, Spécialité : Probabilités et Statistique.

[2] G. Biau, L. Devroye, and G. Lugosi. Consistency of random forests and other averaging classifiers. *ESAIM : Probability and Statistics*, (2005).

[3] José E. Chacón and Tarn Duong. *Multivariate Kernel Smoothing and Its Applications*, volume 160 of *Monographs on Statistics and Applied Probability*. CRC Press, Taylor & Francis Group, Boca Raton, FL, (2018).

[4] Y.-C. Chen. Lecture 7 : Density estimation – k-nearest neighbor and basis approach. STAT 425 : Introduction to Nonparametric Statistics, Winter 2018, Department of Statistics, University of Washington, (2018).

[5] Gérard Collomb. Estimation non paramétrique de la régression par la méthode du noyau : propriété de convergence asymptotiquement normale indépendante. *Annales scientifiques de l'Université de Clermont-Ferrand 2, Série Mathématiques*, 65(15), (1977).

[6] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, (1967).

[7] Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers : 2nd edition (with python examples). *arXiv preprint arXiv :2004.04523*, (2020).

[8] Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Chapman & Hall, (1996).

[9] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, (1936).

[10] E. Fix and J. L. Hodges. Discriminatory analysis. nonparametric discrimination : Consistency properties. Technical report, USAF School of Aviation Medicine, (1951).

[11] C. Garcia and M. Delakis. A neural architecture for face detection in images. In T. Pajdla and J. Matas, editors, *Proceedings of the 6th European Conference on Computer Vision (ECCV)*. Springer, (2004).

[12] Evarist Giné, Vladimir Koltchinskii, and Joel Zinn. Weighted uniform consistency of kernel density estimators. *The Annals of Probability*, 32(3B), (2004).

[13] J. Gou et al. A survey on the k-nearest neighbor algorithm. *Pattern Recognition*, (2019).

[14] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer, (2009).

[15] B. Kulis. Metric learning : A survey. *Foundations and Trends in Machine Learning*, (2013).

[16] W. Maleika. Inverse distance weighting method optimization in the process of digital terrain model creation based on data collected from a multibeam echosounder. *Applied Geomatics*, 12(6), (2020).

[17] Thomas Nagler and Claudia Czado. Evading the curse of dimensionality in nonparametric density estimation with simplified vine copulas. *arXiv preprint arXiv :1505.05179*, (2015).

[18] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33, (1962).

[19] M. Rosenblatt. Remarks on some nonparametric estimates of a density function. *The Annals of Mathematical Statistics*, 27(3), (1956).

[20] D. W. Scott. *Multivariate Density Estimation : Theory, Practice and Visualization*. John Wiley & Sons, 2 edition, (2015).

[21] David W. Scott. *Multivariate Density Estimation : Theory, Practice, and Visualization*. Wiley, (1992).

[22] Simon 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*, 53(3) :683–690, (1991).

[23] B. W. Silverman. *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, (1986).

[24] S. Theodoridis and K. Koutroumbas. *Pattern Recognition*. Academic Press, 4 edition, (2009).

[25] M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman & Hall/CRC, (1995).

[26] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, (2009).

[27] Dominik Wied and Rafael Weißbach. Consistency of the kernel density estimator : A survey. *Statistical Papers*, 53(1), (2012).

[28] Andrzej Woznica and Nicolas Paparoditis. Consistent estimation of the multivariate probability density function using k-nearest neighbor methods. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 42(3), (2006).

[29] D. Zhang and Z.-H. Zhou. Ml-knn : A lazy learning approach to multi-label learning. *Neural Networks*, (2007).

[30] P. Zhao and L. Lai. Analysis of knn density estimation. Technical report, Department of Electrical and Computer Engineering, University of California, Davis, (2020).

[31] P. Zhao and L. Lai. Bayesian k-nearest neighbors for classification. *Journal of Machine Learning Research*, (2021).