

Cours de Statistique : Estimation de Densité Non-Paramétrique

Note de cours réorganisée

Mars 2026

1 Introduction et Motivation

L'objectif de l'estimation de densité est de reconstruire une fonction de densité de probabilité f à partir d'un échantillon de variables aléatoires X_1, \dots, X_N indépendantes et identiquement distribuées (i.i.d.) selon f .

Définition 1.1 (Densité de probabilité). *Une fonction $f : \mathbb{R} \rightarrow \mathbb{R}$ est une densité de probabilité si elle vérifie :*

1. $f(x) \geq 0$ pour presque tout x .
2. $f \in L^1(\mathbb{R})$ et $\int_{\mathbb{R}} f(x)dx = 1$.

Remarque 1.1. *La valeur ponctuelle $f(x_0)$ n'est pas informative par elle-même car la mesure de Lebesgue d'un singleton $\{x_0\}$ est nulle. Pour estimer f , nous devons imposer une **contrainte de régularité** (smoothness). On suppose généralement que f appartient à un sous-espace de fonctions régulières $\mathcal{F} \subsetneq L^1$.*

2 L'approche "Naïve" : L'estimateur par intervalle

Considérons une réalisation i.i.d. x_1, \dots, x_N d'une loi de densité f . Pour un point x donné et un petit paramètre $h > 0$, la probabilité que X appartienne à l'intervalle $[x - h, x + h]$ est :

$$\mathbb{P}(X \in [x - h, x + h]) = \int_{x-h}^{x+h} f(u)du$$

Si f est continue en x et h est petit, alors $\int_{x-h}^{x+h} f(u)du \approx 2hf(x)$. On peut estimer cette probabilité par la proportion empirique de données tombant dans cet intervalle :

$$\hat{p} = \frac{1}{N} \sum_{n=1}^N \mathbb{1}_{[x-h, x+h]}(x_n)$$

D'où l'estimateur "naïf" de la densité :

$$\hat{f}(x) = \frac{1}{2Nh} \sum_{n=1}^N \mathbb{1}_{[x-h, x+h]}(x_n) = \frac{1}{Nh} \sum_{n=1}^N K\left(\frac{x - x_n}{h}\right)$$

où $K(u) = \frac{1}{2}\mathbb{1}_{[-1,1]}(u)$ est appelé le **noyau rectangulaire**.

3 L'approche par Projection (Espaces de Hilbert)

Cette méthode repose sur la décomposition de f dans une base orthonormée de l'espace de Hilbert $L^2([0, 1])$.

3.1 Espace de Hilbert et Bases Orthonormées

On munit $L^2([0, 1])$ du produit scalaire :

$$\langle f, g \rangle = \int_0^1 f(x) \overline{g(x)} dx$$

Soit $\{e_k\}_{k \in \mathbb{Z}}$ une base orthonormée de $L^2([0, 1])$ (par exemple la base de Fourier $e_k(x) = e^{i2\pi kx}$). Toute fonction $f \in L^2([0, 1])$ peut s'écrire :

$$f(x) = \sum_{k \in \mathbb{Z}} \alpha_k e_k(x) \quad \text{avec} \quad \alpha_k = \langle f, e_k \rangle$$

3.2 Régularité et Espaces de Sobolev

Pour quantifier la "douceur" de f , on utilise les espaces de Sobolev W^s .

Définition 3.1 (Espace de Sobolev). *Pour $s > 0$, on définit $W^s([0, 1])$ comme l'espace des fonctions $f \in L^2([0, 1])$ telles que :*

$$\sum_{k \in \mathbb{Z}} |\alpha_k|^2 (1 + |k|)^{2s} < +\infty$$

On définit l'ellipsoïde de Sobolev $B(s, R) = \{f \in W^s : \|f\|_{W^s}^2 \leq R^2\}$.

3.3 Construction de l'estimateur par projection

On approche d'abord f par une version tronquée à l'ordre M :

$$\tilde{f}(x) = \sum_{|k| \leq M} \alpha_k e_k(x)$$

Cependant, les coefficients α_k sont inconnus. On remarque que :

$$\alpha_k = \int_0^1 e_k(x) f(x) dx = \mathbb{E}[e_k(X)]$$

Par la méthode des moments, on estime α_k par la moyenne empirique :

$$\hat{\alpha}_k = \frac{1}{N} \sum_{n=1}^N e_k(X_n)$$

L'estimateur final est :

$$\hat{f}(x) = \sum_{|k| \leq M} \hat{\alpha}_k e_k(x)$$

4 Analyse des Performances : Risque Quadratique

Le risque est mesuré par l'Erreur Quadratique Moyenne Intégrée (MISE).

$$\text{Risk}(\hat{f}, f) = \mathbb{E} \left[\int_0^1 |\hat{f}(x) - f(x)|^2 dx \right]$$

4.1 Décomposition Biais-Variance

En utilisant l'orthogonalité de la base, le risque se décompose en :

$$\text{Risk} = \underbrace{\|\tilde{f} - f\|^2}_{\text{Biais}^2 \text{ (Troncature)}} + \underbrace{\mathbb{E}[\|\hat{f} - \tilde{f}\|^2]}_{\text{Variance (Estimation)}}$$

1. **Le Biais (Erreur d'approximation)** : Pour $f \in B(s, R)$, on montre que :

$$\text{Biais}^2 = \sum_{|k| > M} |\alpha_k|^2 = \sum_{|k| > M} |\alpha_k|^2 \frac{(1 + |k|)^{2s}}{(1 + |k|)^{2s}} \leq \frac{R^2}{M^{2s}} = O(M^{-2s})$$

2. **La Variance (Erreur statistique)** :

$$\text{Variance} = \sum_{|k| \leq M} \mathbb{E}[|\hat{\alpha}_k - \alpha_k|^2] = \sum_{|k| \leq M} \frac{\text{Var}(e_k(X))}{N}$$

Si les fonctions de base sont bornées ($|e_k| \leq C$), alors $\text{Var}(e_k(X)) \leq C^2$. D'où :
 $\text{Variance} = O\left(\frac{M}{N}\right)$.

Le risque total est donc de l'ordre de :

$$\text{Risk} \approx \frac{M}{N} + \frac{1}{M^{2s}}$$

5 Optimisation du paramètre de troncature

Pour minimiser le risque, nous cherchons le M optimal (M^*) qui équilibre le biais et la variance.

$$\frac{\partial}{\partial M} \left(\frac{M}{N} + M^{-2s} \right) = 0 \implies \frac{1}{N} - 2sM^{-(2s+1)} = 0$$

On obtient :

$$M^* \sim N^{\frac{1}{2s+1}}$$

Vitesse de convergence : En réinjectant M^* dans l'expression du risque, on trouve :

$$\text{Risk}^* \sim N^{-\frac{2s}{2s+1}}$$

- Si $s \rightarrow \infty$ (fonction très lisse), la vitesse approche N^{-1} (vitesse paramétrique).
- Si $s \rightarrow 0$ (fonction peu régulière), $M^* \sim N$ et la vitesse est très lente.

6 Introduction à l'Approche par Noyau

L'approche par noyau généralise l'idée de la section 2.

Définition 6.1 (Noyau). *Un noyau $K : \mathbb{R} \rightarrow \mathbb{R}$ est une fonction intégrable telle que $\int_{\mathbb{R}} K(u)du = 1$. On définit le noyau translaté et mis à l'échelle par la fenêtre (bandwidth) h :*

$$K_h(u) = \frac{1}{h} K\left(\frac{u}{h}\right)$$

L'estimateur à noyau est défini par le produit de convolution entre la mesure empirique et le noyau :

$$\hat{f}(x) = (K_h * f_{emp})(x) = \frac{1}{N} \sum_{n=1}^N K_h(x - x_n)$$

7 Analyse de l'Estimateur à Noyau

L'estimateur à noyau repose sur deux idées fondamentales (souvent appelées "Tricks" dans les notes) :

1. **L'approximation** : La convolution $(K_h * f)(x)$ converge vers $f(x)$ quand $h \rightarrow 0$.
2. **L'estimation** : L'espérance de l'estimateur $\hat{f}(x)$ est précisément cette convolution.

7.1 Cadre d'analyse : Espaces de Hölder

Pour garantir une certaine vitesse de convergence, nous supposons que la densité f appartient à un espace de régularité fonctionnelle.

Définition 7.1 (Espace de Hölder $\Lambda(s, L)$). *Soit $s > 0$. On écrit $s = k + \beta$ avec $k \in \mathbb{N}$ et $\beta \in]0, 1]$. Une fonction f appartient à l'espace de Hölder $\Lambda(s, L)$ si :*

1. f est k fois dérivable.
2. La k -ième dérivée est β -Höldérienne : $\forall x, y \in \mathbb{R}, |f^{(k)}(x) - f^{(k)}(y)| \leq L|x - y|^\beta$.

7.2 Hypothèses sur le noyau (Noyaux de Parzen-Rosenblatt)

Pour exploiter la régularité d'ordre s , le noyau K doit posséder des propriétés de moments :

1. $\int K(u)du = 1$.
2. $\int |u|^s |K(u)|du < +\infty$.
3. $\int u^l K(u)du = 0$ pour tout $l \in \{1, \dots, k\}$. On dit alors que K est un **noyau d'ordre k** .

7.3 Résultat 1 : Contrôle du Biais

Théorème 7.1 (Borne sur le biais). *Si $f \in \Lambda(s, L)$ et K est un noyau d'ordre k , alors le biais de l'estimateur vérifie :*

$$\sup_x |\mathbb{E}[\hat{f}(x)] - f(x)| = \sup_x |(K_h * f)(x) - f(x)| \leq \frac{L \cdot C}{k!} h^s = O(h^s)$$

où $C = \int |y|^s |K(y)|dy$.

Esquisse de preuve. Par changement de variable $y = \frac{x-u}{h}$:

$$(K_h * f)(x) - f(x) = \int K(y)[f(x - hy) - f(x)]dy$$

En utilisant un développement de Taylor-Young de f à l'ordre k en x :

$$f(x - hy) = f(x) - hyf'(x) + \dots + \frac{(-hy)^k}{k!}f^{(k)}(x - \epsilon hy)$$

Grâce aux propriétés d'annulation des moments du noyau (ordre k), les termes de dérivées s'annulent à l'intégration, ne laissant que le reste de Taylor qui est borné par la condition de Hölder en h^s . \square

7.4 Résultat 2 : Contrôle de la Variance

Théorème 7.2 (Borne sur la variance). *Si K est de carré intégrable ($K \in L^2$) et f est bornée, alors :*

$$\text{Var}(\hat{f}(x)) \leq \frac{C'}{Nh} = O\left(\frac{1}{Nh}\right)$$

où $C' = \|f\|_\infty \int K^2(u)du$.

Démonstration. Puisque les X_n sont i.i.d. :

$$\text{Var}(\hat{f}(x)) = \frac{1}{N} \text{Var}(K_h(x - X_1)) \leq \frac{1}{N} \mathbb{E}[K_h(x - X_1)^2]$$

$$\mathbb{E}[K_h(x - X_1)^2] = \int \frac{1}{h^2} K\left(\frac{x-u}{h}\right)^2 f(u)du$$

Par changement de variable $u' = \frac{x-u}{h}$, on obtient :

$$\frac{1}{h} \int K(u')^2 f(x - hu')du' \leq \frac{\|f\|_\infty}{h} \int K^2(u')du'$$

\square

8 Conclusion : Le compromis Biais-Variance

Le risque quadratique total se comporte comme :

$$\text{Risque}(h) \approx \underbrace{h^{2s}}_{\text{Biais}^2} + \underbrace{\frac{1}{Nh}}_{\text{Variance}}$$

8.1 Fenêtre optimale h^*

En minimisant cette expression par rapport à h , on trouve :

$$h^{2s+1} \sim \frac{1}{N} \implies h^* = N^{-\frac{1}{2s+1}}$$

Méthode	Paramètre de lissage	Rôle
Projection	M (nombre de modes)	Régularisation par troncature
Noyau	$1/h$ (inverse de la fenêtre)	Régularisation par lissage

8.2 Vitesse de convergence minimax

En remplaçant h^* dans l'expression du risque, on obtient la vitesse de convergence :

$$\text{Risque}^* \sim N^{-\frac{2s}{2s+1}}$$

Remarque 8.1 (Synthèse entre les deux approches). *Dans les deux cas, on retrouve la même vitesse de convergence. Le paramètre M en projection joue un rôle inversement proportionnel à h dans l'approche par noyau. Plus s (la régularité) est grand, plus la vitesse approche $1/N$, la vitesse "standard" de la statistique paramétrique.*