# Utilisation de méthodes combinatoires en estimation de densité

Groupe de travail 2ème année

À partir de l'ouvrage « Combinatorial methods in density estimation »
par Luc Devroye et Gábor Lugosi, 2000

Disponible en 2 exemplaires à la bibliothèque.

En pratique : * Séances les jeudis après-midis, de 13h30 à 14h45
- exposé durant entre 30 minutes et 1 heure
- suivi de 15 minutes de questions / discussions
- horaire strict car de nombreux jeudis nous avons un séminaire à 15h à l'Ecole des Mines

DÉBUT LE JEUDI 7 OCTOBRE !

* Scannez vos notes manuscrits afin que les suivants aient une trace écrite

---

## Présentation du problème.

$X_1 \dots X_n$ observations réelles iid selon une loi $\nu$, de fonction de répartition $F$

(1) - Comment estimer $F$ si celle-ci est inconnue ? (problème statistique)

- On note $\hat{F}_n$ la fonction de répartition empirique, i.e., la fonction de répartition associée à la mesure empirique $\nu_n = \frac{1}{n} \sum_{t=1}^{n} \delta_{X_t}$ :

$$\forall x \in \mathbb{R}, \quad \hat{F}_n(x) = \frac{1}{n} \sum_{t=1}^{n} \mathbb{1}_{\{X_t \leq x\}}$$

C'est un bon estimateur : il est consistant, c'est-à-dire qu'il converge en probabilité vers la quantité à estimer $F$, pour la distance de la norme du supremum, et ce, quelle que soit $F$ :

$$d_\infty(\hat{F}_n, F) \overset{\text{not.}}{=} \| \hat{F}_n - F \|_\infty \xrightarrow{\mathbb{P}} 0$$

En fait, la convergence a même lieu p.s : c'est le théorème de

Glivenko – Cantelli.

(2) – La méthode précédente a donc permis d'estimer simultanément et de manière convergente toutes les quantités

$$\nu\left(]-\infty, x]\right), \qquad x \in \mathbb{R}$$

– Que faire si l'on veut désormais faire de même pour tous les

$$\nu(A), \qquad A \in \mathcal{B}(\mathbb{R})$$

où $\mathcal{B}(\mathbb{R})$ désigne la tribu des boréliens de $\mathbb{R}$?

(3) – On introduit à cet effet le critère suivant, dit distance en variation : soient $\mu_1$ et $\mu_2$ deux lois sur $\mathbb{R}$,

$$d_{VT}(\mu_1, \mu_2) = \sup_{A \in \mathcal{B}(\mathbb{R})} |\mu_1(A) - \mu_2(A)|$$

(C'est bien une distance.)

– Par exemple, lorsque $\nu$ est diffuse (n'a aucun atome) et désignant toujours par $\nu_n$ la mesure empirique (qui, elle, est p.s. totalement atomique):

$$d_{VT}(\nu_n, \nu) = 1.$$

(4) Simplification et formulation d'une stratégie plus raisonnable

– Simplification : Supposer que $\nu$ est absolument continue par rapport à la mesure de Lebesgue $d\lambda$, de densité notée $f$ :

$$\forall A \in \mathbb{R}, \qquad \nu(A) = \int_A f \, d\lambda$$

Par le théorème fondamental du calcul différentiel, cela est le cas si et seulement si $F$ est absolument continue, si et seulement si $F$ est presque partout dérivable, de dérivée $F' = f$ avec en outre

$$F(x) = \int_{-\infty}^{x} f \, d\lambda \quad \text{p.p.}$$

– Nouvelle stratégie :

Il suffit alors d'estimer la densité $f$, qui est la dérivée de $F$, plutôt que $F$ elle-même, cf. le théorème de Scheffé. On détaille cela :

## Le modèle de l'estimation de densité.

$X_1 \dots X_n$ iid selon une loi $\nu \ll \lambda$, de densité $f$ : $\frac{d\nu}{d\lambda} \overset{not.}{=} f$

$f$ est inconnue et on veut l'estimer ; tout ce qu'on sait, c'est que $f$ vit dans

$$\mathcal{P} = \{ g \in \mathbb{L}^1(\lambda) : \int_{\mathbb{R}} g \, d\lambda = 1 \text{ et } g \geq 0 \}$$

(et $\mathcal{P}$ est un très gros ensemble !).

**Th (de Scheffé) :** Soient $\mu_1$ et $\mu_2$ deux probabilités, avec $\mu_1 \ll \lambda$ et $\mu_2 \ll \lambda$, si l'on note $f_1 = \frac{d\mu_1}{d\lambda}$ et $f_2 = \frac{d\mu_2}{d\lambda}$. Alors :

$$d_{VT}(\mu_1, \mu_2) = \int_{\mathbb{R}} (f_1 - f_2)_+ \, d\lambda = \frac{1}{2} \int_{\mathbb{R}} |f_1 - f_2| \, d\lambda$$

**Preuve :** $B \overset{not.}{=} \{ x \in \mathbb{R} : f_1(x) \geq f_2(x) \} \overset{not.}{=} \{ f_1 \geq f_2 \}$

Pour tout $A \in \mathcal{B}(\mathbb{R})$,

$$\mu_1(A) - \mu_2(A) = \int_A (f_1 - f_2) \, d\lambda \leq \int_{A \cap B} (f_1 - f_2) \, d\lambda \quad \text{par déf. de } B$$

$$\leq \int_B (f_1 - f_2) \, d\lambda \quad \text{tjs par déf de } B$$

$$= \int (f_1 - f_2)_+ \, d\lambda$$

En passant à $\sup_{A \in \mathcal{B}(\mathbb{R})}$, on obtient la première égalité.

La seconde procède de ce que

$$0 = 1 - 1 = \int (f_1 - f_2) \, d\lambda = \int (f_1 - f_2)_+ \, d\lambda - \int (f_1 - f_2)_- \, d\lambda$$

soit $\int (f_1 - f_2)_+ \, d\lambda = \int (f_1 - f_2)_- \, d\lambda$ ;

or, $$\int |f_1 - f_2| d\lambda = \int (f_1 - f_2)_+ d\lambda + \int (f_1 - f_2)_- d\lambda = 2 \int (f_1 - f_2)_+ d\lambda,$$

ce qui conclut la preuve.

**Stratégie :** Estimer $f$ de manière convergente pour la norme $\|\cdot\|_1$ :

Construire $\hat{f}_n \in \mathcal{F}$ à partir de $X_1, \dots X_n$ telle que

$$\|\hat{f}_n - f\|_1 \overset{\text{not.}}{=} \int |\hat{f}_n - f| d\lambda \xrightarrow{\mathbb{P}} 0$$

**Remarque souvent utile :** Comme $\int |\hat{f}_n - f| d\lambda$ est toujours $\leq 2$ et $\geq 0$, quelles que soient $\hat{f}_n$ et $f$, il vient :

$$\|\hat{f}_n - f\|_1 \xrightarrow{\mathbb{P}} 0 \iff \mathbb{E}\left[\|\hat{f}_n - f\|_1\right] \longrightarrow 0.$$

(Ici, la convergence en probabilité et celle dans $\mathbb{L}^1(\mathbb{P})$ sont équivalentes pour notre critère.)

**Estimation par fenêtres glissantes.**

Son intuition repose sur le fait que $f$ est la dérivée de $F$, et que $F$ est bien approchée par la fonction de répartition empirique $F_n$ :

$$\hat{f}_n = \frac{F_n(\cdot + h_n) - F_n(\cdot - h_n)}{2h_n}$$

où $h_n > 0$ est un paramètre à régler

Il faut :

- $h_n \searrow 0$ pour respecter l'intuition
- mais pas trop vite (on verra que $nh_n \to +\infty$ est souhaitable) afin qu'il y ait suffisamment d'observations $X_t$ dans chaque intervalle $]x - h_n, x + h_n]$ :

$$\forall x \in \mathbb{R}, \quad \hat{f}_n(x) = \frac{1}{2nh_n} \sum_{t=1}^{n} \mathbb{1}_{\{X_t \in ]x - h_n, x + h_n]\}}$$

**Lem:** $\hat{f}_n \in \mathcal{P}$

**Preuve:** Par Fubini-Tonelli,

$$\int_{\mathbb{R}} \hat{f}_n(x)\,dx = \frac{1}{2nh_n} \sum_{t=1}^{n} \int \mathbb{1}_{\{X_t \in ]x-h_n,\, x+h_n[\}}\,dx$$

$$= \frac{1}{2nh_n} \sum_{t=1}^{n} \underbrace{d\big([X_t-h_n,\, X_t+h_n[\big)}_{=2h_n}$$

$$= 1$$

**Prop:** Lorsque $h_n \to 0$ et $nh_n \to +\infty$, on a en tout point $x$ tel que $F$ est dérivable en $x$ (i.e., pour presque tout $x$),

$$\mathbb{E}\left[ (f(x) - \hat{f}_n(x))^2 \right] \longrightarrow 0.$$

**Preuve:** On commence par une décomposition biais-variance :

$$(*) \qquad \mathbb{E}\left[ (\hat{f}_n(x) - f(x))^2 \right] = \Big( \mathbb{E}[\hat{f}_n(x)] - f(x) \Big)^2 + \mathbb{E}\left[ \big( \hat{f}_n(x) - \mathbb{E}[\hat{f}_n(x)] \big)^2 \right]$$

(cf. terme croisé nul)

$$= \Big( \underbrace{\mathbb{E}[\hat{f}_n(x)] - f(x)}_{\text{biais}} \Big)^2 + \underbrace{\operatorname{Var} \hat{f}_n(x)}_{\text{variance}}$$

Terme de biais : $\mathbb{E}[\hat{f}_n(x)] = \dfrac{F(x+h_n) - F(x-h_n)}{2h_n} \longrightarrow f(x)$ par hyp. ($F'(x) = f(x)$ et $h_n \to 0$)

(cf. $\forall y$, $\mathbb{E}[F_n(y)] = F(y)$)

Terme de variance : $2nh_n \hat{f}_n(x) = \sum_{t=1}^{n} \mathbb{1}_{\{X_t \in ]x-h_n,\, x+h_n]\}}$

suit une loi binomiale de paramètres $n$ et :

$$p_{nx} = \mathbb{P}\{ X_t \in ]x-h_n,\, x+h_n] \} = F(x+h_n) - F(x-h_n)$$

Donc $\operatorname{Var} \hat{f}_n(x) = \dfrac{1}{4n^2h_n^2} \underbrace{\operatorname{Var}(2nh_n \hat{f}_n(x))}_{n\, p_{nx}(1-p_{nx})} \leq \dfrac{1}{4nh_n} \dfrac{p_{nx}}{h_n} \sim \dfrac{1}{4nh_n} 2f(x)$

$\longrightarrow 0$

car $nh_n \to +\infty$

Donc $(*)$ est la somme de deux termes tendant vers 0.

**Th:** Lorsque $h_n \to 0$ et $nh_n \to +\infty$,

$$\| \hat{f}_n - f \|_1 \xrightarrow{\mathbb{P}} 0$$

**Preuve:** Comme une remarque précédente l'indiquait, il suffit de montrer que

$$\mathbb{E}[\, \| \hat{f}_n - f \|_1 ] \longrightarrow 0$$

$$= 2\, \mathbb{E}\Big[ \int (f - \hat{f}_n)_+ \, d\mu \Big] \overset{\text{Fubini-Tonelli}}{=} 2 \int \varphi_n \, d\mu,$$

où l'on a noté $\varphi_n(x) = \mathbb{E}[\, (f(x) - \hat{f}_n(x))_+ \,]$.

Or,

- $\varphi_n \leq f \in \mathbb{L}^1(\mu)$
- pour presque tout $x$,

$$\varphi_n(x) \underset{\text{Cauchy-Schwarz}}{\leq} \sqrt{\mathbb{E}[\, (f(x) - \hat{f}_n(x))^2 \,]} \longrightarrow 0 \quad \text{(par la proposition)}$$

D'où le résultat du théorème, par convergence dominée.

## Extensions et contenu du livre.

1 – Estimateurs à noyau : en notant $K = \frac{1}{2} \mathbb{1}_{[-1,1[}$, on a que l'estimateur par fenêtres glissants correspond à

$$\hat{f}_{n,h_n} = \frac{1}{n\, h_n} \sum_{t=1}^{n} K\left(\frac{\cdot - X_t}{h_n}\right)$$

↳ D'autres choix de $K$ sont possibles, qui conduisent à des estimés plus régulières (voir les illustrations graphiques)

↳ .... Mais il faudra prouver des théorèmes généraux de convergence !

OBJET DES PREMIÈRES SÉANCES

VS.

OBJET DES SÉANCES SUIVANTES

2 – Choix de $h_n$ : (qui est un paramètre d'utilisateur.)

– beaucoup de suites $(h_n)$ vérifient $h_n \to 0$ et $nh_n \to +\infty$ ↳ embarras du choix

– en pratique, $n$ est fixé (et non $n \to +\infty$ !) et on choisit $h_n$ en fonction de $X_1 \dots X_n$

↳ Choix aléatoire $H_n$ fondé sur les observations, qui conduit à un estimateur $f_{n,H_n}$, à comparer aux autres choix de $h$ possibles, qui auraient conduit aux estimateurs $f_{n,h}$ ; on veut faire presqu'aussi bien que le meilleur de ces derniers.

On va établir l'inégalité oracle suivante :

$$\frac{\mathbb{E}\left[\|\hat{f}_{n,H_n} - f\|_1\right]}{\inf_{h>0} \mathbb{E}\left[\|\hat{f}_{n,h} - f\|_1\right]} \leq 3 + o(1)$$

pour un choix à préciser de $H_n$ en fonction de $X_1 \dots X_n$ uniquement ; $\sup_{f \in \mathcal{F}}$

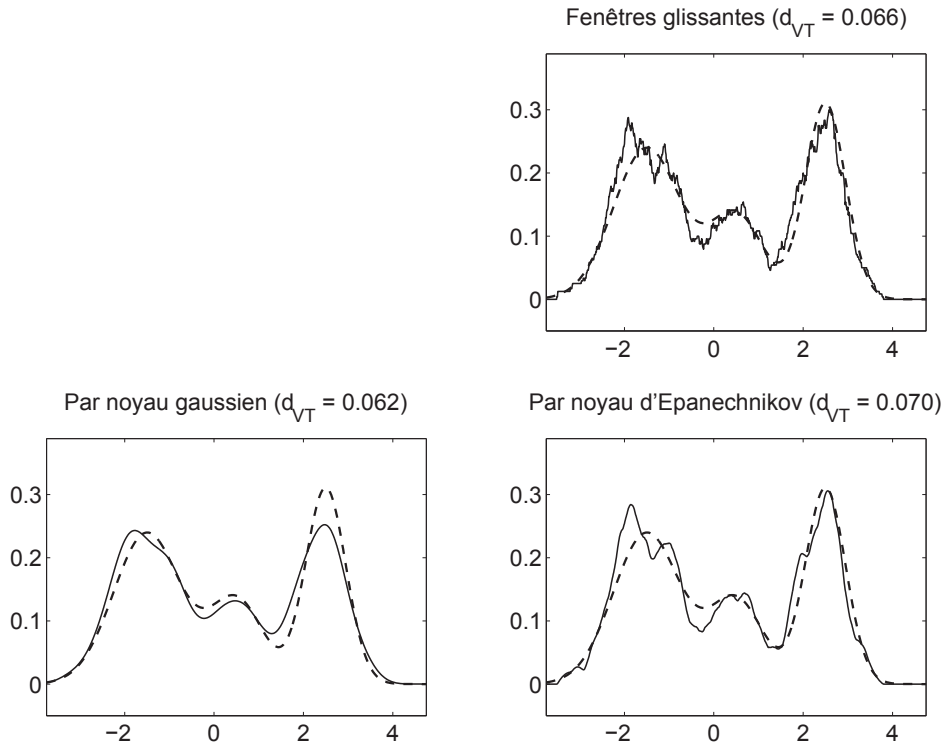


Figure 1. Une réalisation des estimations pour le mélange de lois gaussiennes, n = 400 et h = 0.30. On précise entre parenthèses les distances en variation totale réalisées.

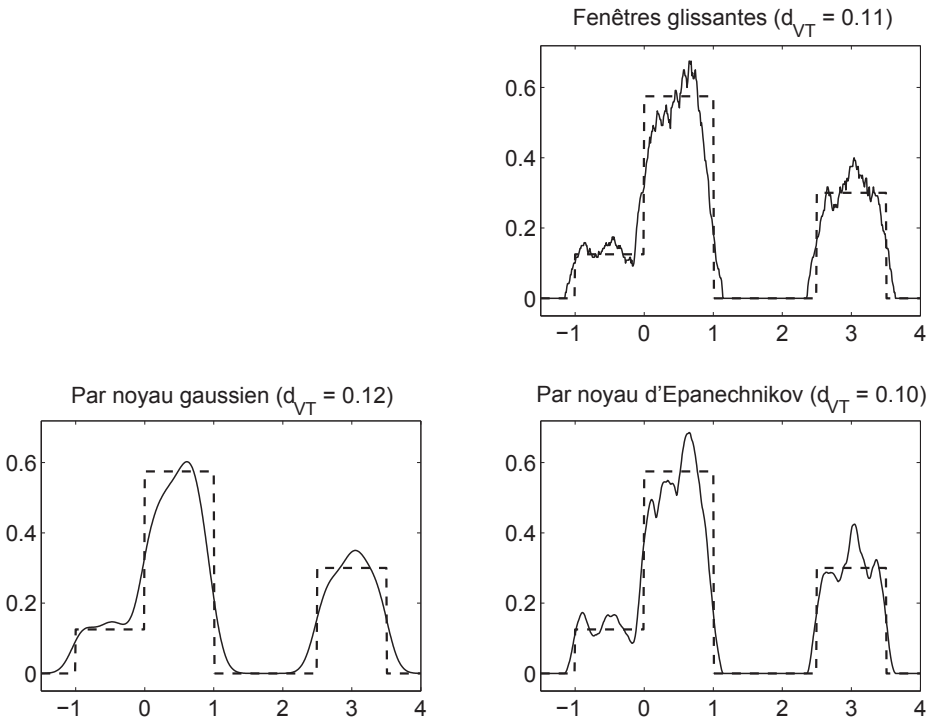


Figure 2. Une réalisation des estimations pour le mélange de lois uniformes, n = 400 et h = 0.15. On précise entre parenthèses les distances en variation totale réalisées.

3_ Ingrédients mis en jeu :

- des techniques d'analyse élémentaires (convolution dans $\mathbb{L}^1(d\lambda)$) pour la convergence des estimateurs par noyau
- de la concentration de sommes de variables aléatoires
- un peu de combinatoire (dimension de Vapnik, lemme de Sauer)

PLAN : DL00 sera notre référence principale

Deux références secondaires détaillant certains points délicats sont :
- Dev85 (Devroye : « A course in density estimation »)
- RS09 (Rivoirard et Stoltz : « Statistique en Action »)

(ces trois ouvrages sont disponibles à la bibliothèque)

Séance 1 : (7 octobre) Convergence des estimateurs à noyau à suite $(h_n)$ fixée (techniques d'analyse) — DL00, §9.1 – 9.3 ; RS09, § 24.6

Séance 2 : (14 octobre) Inégalités de concentration pour (essentiellement) des sommes de variables aléatoires iid — DL00, §2.1 – §2.3 ; RS09, § 26.2.2

Séance 3 : (21 octobre) Convergence des estimateurs à noyau : choix $(H_n)$ à partir des données $X_1, \dots X_n$ ; convergence ps (et pas seulement $\mathbb{L}^1$ ou en probabilité) — DL00, Th 9.3 ; plus : §2.4 et §9.4 ; Dev85, § 3.6

Séance 4 : (28 octobre) Un mot sur les vitesses de convergence : bornes supérieures uniquement (encore une séance avec pas mal d'analyse) — DL00 § 9.8

Il s'agit de faire simplement une mise en perspective ; voir aussi RS09, § 32.1.4 ← + éléments au choix dans le § 9.7 et les exercices (exo 9.24 p ex.)

Séance 5 : (4 novembre) — Méthodes et techniques combinatoires — DLOO §3.1 et §3.3

Séance 6 : (A définir) — Comment choisir entre un nombre fini de densités ? Vers une première inégalité oracle, à améliorer. — DLOO, Théorèmes : 6.3, 6.4 et 10.1

Séance 7 : (A définir) — Illustrations de l'inégalité oracle sur la classe des estimateurs de forme additive — DLOO, § 10.2

Séance 8 : (A définir) — L'inégalité oracle finale. — DLOO § 11.1 et § 11.2 et des éléments du § 11.3 si besoin.

En cas d'affluence, deux exposés supplémentaires sont disponibles :

Séance 5 bis : — Méthodes et techniques combinatoires, suite — DLOO § 4.1, § 4.2 et § 4.4

Séance 6 bis : — Autres illustrations de la première inégalité oracle — Chap 8 (choisir quelques exemples, ne pas tous les traiter)