Statistiques en grande dimension

Cours 5

Previously...

- $X_1, ..., X_n$ iid sur $[0,1]$ de densité f
- $x_0 \in]0,1[$, on cherche à estimer $f(x_0)$
- Pour un $h>0$, on pose :

$$\hat{f}^{\mathrm{LC}}_{n,h}(x_0) = \frac{1}{2\,n\,h}\sum_{i=1}^{n} \mathbf{1}_{[x_0-h,x_0+h]}(X_i)$$

- Pour choisir h de façon adaptative dans une famille finie, $\{h_1, ..., h_M\}$, on utilise la méthode «Intersection of Confidence Intervals» $ICI \longrightarrow \hat{f}^{\mathrm{ICI}}(x_0)$

Théorème 1. *Si f est bornée par L et si $h_1 < h_2 < ... < h_M$, alors :*

$$\mathbb{P}\big(\big|\hat{f}^{\mathrm{ICI}}(x_0) - f(x_0)\big| \leqslant 6\,\rho_{j^*}\big) \geqslant 1-\varepsilon$$

...où :

$$\rho_j = \sqrt{\frac{3\,L}{2\,n\,h_j}\ln\left(\frac{2\,M}{\varepsilon}\right)}$$

Démonstration. On a noté $\xi_j(x_0) = \hat{f}^{\mathrm{LC}}_{n,h_j}(x_0) - \mathbb{E}\big[\,\hat{f}^{\mathrm{LC}}_{n,h_j}(x_0)\,\big]$ l'*erreur stochastique* de $\hat{f}^{\mathrm{LC}}_{n,h_j}$.

Soit $\Omega_0 = \{\forall\, j = 1, ..., M, |\xi_j(x_0)| < \rho_j\}$, on a prouvé que :

$$\mathbb{P}\big(\big|\hat{f}^{\mathrm{ICI}}(x_0) - f(x_0)\big| \leqslant 6\,\rho_{j^*}\big) \geqslant \mathbb{P}(\Omega_0)$$

Il reste à prouver que $\mathbb{P}(\Omega_0) \geqslant 1-\varepsilon \Leftrightarrow \mathbb{P}(\overline{\Omega_0}) < \varepsilon$. On a :

$$\begin{aligned}
\mathbb{P}(\overline{\Omega_0}) &= \mathbb{P}(\exists\, j \in [|1, M|] : |\xi_j(x_0)| > \rho_j) \\
&= \mathbb{P}\left(\bigcup_{j=1}^{M} \{|\xi_j(x_0)| > \rho_j\}\right) \\
&\leqslant \mathbb{P}(|\xi_0(x_0)| > \rho_0) + \cdots + \mathbb{P}(|\xi_M(x_0)| > \rho_M) \\
&\leqslant M \max_{1\leqslant j\leqslant M} \mathbb{P}(|\xi_j(x_0)| > \rho_j)
\end{aligned}$$

De plus, on peut écrire :

$$\xi_j(x_0) = \frac{1}{2\,n\,h_j}\sum_{i=1}^{n}\left[\underbrace{\mathbf{1}_{[x_0-h_j,x_0+h_j]}(X_i)}_{\in[0,1]} - \underbrace{\int_{x_0-h_j}^{x_0+h_j} f}_{\in[0,1]}\right]$$

Posons $Z_i = \frac{1}{2\,n\,h}\left(\mathbf{1}_{[x_0-h,x_0-h]}(X_i) - \int_{x_0-h_j}^{x_0+h_j} f\right)$: les Z_i sont iid, $\mathbb{E}[Z_i] = 0$ et $|Z_i| \leqslant \frac{1}{2\,n\,h}$.

On veut majorer $\mathbb{P}(|Z_1+\cdots+Z_n|>\rho)$, et on va pouvoir le faire grâce à l'inégalité de Bernstein :

Théorème 2. *(Inégalité de Bernstein)*

Si on dispose de variables indépendantes Z_i, de moyenne nulle, et bornées presque sûrement par une constante $c>0.$, alors :

$$\forall t>0, \mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i\right| \geqslant t\right) \leqslant 2\exp\left\{-\frac{t^2}{2\sum_{i=1}^{n}\mathrm{Var}(Z_i)+\frac{2}{3}c\,t}\right\}$$

Cette inégalité est plus fine que celle de Hoeffding (on remarquera que dans Hoeffding la variance n'intervient pas, uniquement la constante c). On utilise donc cette inégalité :

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} Z_i\right| > \rho\right) \leqslant 2\exp\left\{-\frac{\rho^2}{2\,n\,\mathrm{Var}(Z_1)+\frac{2}{3}\,\rho\,\frac{1}{2\,n\,h}}\right\}$$

Rappelons par ailleurs que :

$$\rho=\sqrt{\frac{3\,L}{2\,n\,h}\ln\left(\frac{2\,M}{\varepsilon}\right)}$$

$$\begin{aligned}
\mathrm{Var}(Z_1) &= \mathbb{E}[Z_1]^2 \\
&= \frac{1}{4\,n^2\,h^2}\,\mathrm{Var}[\mathbf{1}_{[x_0-h,x_0+h]}(X_1)] \\
&\leqslant \frac{1}{4\,n^2\,h^2}\,\mathbb{E}[\mathbf{1}_{[x_0-h,x_0+h]}(X_1)] \\
&= \frac{1}{4\,n^2\,h^2}\int_{x_0-h}^{x_0+h} f \\
&\leqslant \frac{2\,h\,L}{4\,n^2\,h^2}
\end{aligned}$$

D'où,

$$\begin{aligned}
\frac{\rho^2}{2\,n\,\mathrm{Var}(Z_1)+\frac{\rho}{3n\,h}} &\geqslant \frac{\frac{3\,L}{2\,n\,h}\ln\left(\frac{2\,M}{\varepsilon}\right)}{\frac{L}{n\,h}+\frac{1}{3\,n\,h}\sqrt{\frac{3\,L}{2\,n\,h}\ln\left(\frac{2\,M}{\varepsilon}\right)}} \\
&\geqslant \ln\left(\frac{2\,M}{\varepsilon}\right) \qquad (1)
\end{aligned}$$

...car $L=\max_{x\in[0,1]}f(x)\geqslant\int_0^1 f(x)\,\mathrm{d}x=1\Rightarrow\sqrt{L}\leqslant L$. En poursuivant, on obtient :

$$\frac{\rho^2}{2\,n\,\mathrm{Var}(Z_1)+\frac{\rho}{3n\,h}} \geqslant \frac{3}{2+\frac{2}{3}\sqrt{\frac{3}{2hn}\ln\left(\frac{2\,M}{\varepsilon}\right)}}\ln\left(\frac{2\,M}{\varepsilon}\right)$$

On peut choisir h tel que :

$$\frac{1}{n}\ln\left(\frac{2\,M}{\varepsilon}\right)\leqslant h$$

D'où,

$$\frac{\rho^2}{2\,n\,\mathrm{Var}(Z_1)+\frac{\rho}{3n\,h}} \geqslant \frac{3}{2+\sqrt{\frac{2}{3}}}\ln\left(\frac{2\,M}{\varepsilon}\right) \geqslant \ln\left(\frac{2\,M}{\varepsilon}\right)$$

On déduit de (1) que :

$$\begin{aligned}\mathbb{P}(|Z_1+\cdots+Z_n|>\rho) &\leqslant 2\exp\left\{-\ln\left(\frac{2\,M}{\varepsilon}\right)\right\}\\ &= \frac{\varepsilon}{M}\end{aligned}$$

$\mathbb{P}(\overline{\Omega_0}) \leqslant M \times \frac{\varepsilon}{M} = \varepsilon \Rightarrow \mathbb{P}(\Omega_0) \geqslant \varepsilon$. D'où le résultat. □

Régression non-paramétrique

On observe $(X_1, Y_1), ..., (X_n, Y_n)$, et $Y_i = f(X_i) + \xi_i$ pour $i = 1, ..., n$.

On note f la fonction de régression et ξ_i les erreurs (également appelées les résidus). Ces dernières vérifient les conditions : $\mathbb{E}[\xi_i|X_i] = 0$ et $\mathrm{Var}[\xi_i] = \sigma^2 < +\infty$ pour tout i.

Cas particulier : *Régression à design fixe* : $X_1, ..., X_n$ sont déterministes : cela correspond le plus souvent à des problèmes de traitement d'image ou de signal.

Les ξ_i sont iid, $\mathbb{E}[\xi_i] = 0$ et $\mathrm{Var}[\xi_i] = \sigma^2$.

- On cherche à estimer la fonction f
- On va mesurer la qualité d'un estimateur $\hat{f}$ de f
    - Soit en utilisant le *risque empirique* (typiquement pour le débruitage) :

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\left(\hat{f}(X_i)-f(X_i)\right)^2\right] = \mathbb{E}\left[\left\|\hat{f}-f\right\|_n^2\right]$$

    - Soit en utilisant le *risque intégré* MISE (typiquement pour la prédiction) :

$$\mathrm{MISE}_f\left[\hat{f}\right] = \mathbb{E}\left[\left\|\hat{f}-f\right\|_2^2\right]$$

    - On suppose que le design $\{X_i\}_i$ est unidimensionnel régulier sur $[0, 1]$:

$$X_i = \frac{i}{n}, i = 1, ..., n$$

Réduction au modèle de suites gaussiennes

Afin de présenter les fondements théoriques des méthodes basées sur les projections, il est commode de se ramener à un modèle inéalisé, celui de suites gaussiennes. Pour effectuer cette réduction, rappelons que, dans un modèle de régression, on observe $\{Y_i\}_i$ tels que :

$$Y_i = f\left(\frac{i}{n}\right) + \xi_i, i = 1, ..., n, \text{ où } \xi_1, ..., \xi_n \text{ sont iid et } \mathbb{E}[\xi_i] = 0, \mathrm{Var}[\xi_i] = \sigma^{2.}$$

On suppose que $f \in \mathcal{L}^2[0,1]$. Soit $\{\phi_m, m \in \mathcal{M}\}$ une base orthonormée de $\mathcal{L}^2[0,1]$:

$$\langle \phi_m, \phi_{m'} \rangle = \int_0^1 \phi_m(x)\, \phi_{m'}(x)\, \mathrm{d}x = \delta_{m,m'} = \begin{cases} 0 \text{ si } m \neq m' \\ 1 \text{ sinon} \end{cases}$$

On pose $\theta_m = \langle f, \phi_m \rangle, m \in \mathcal{M}$ — $\mathcal{M}$ étant isomorphe à $\mathbb{N}$, mais pas forcément égal, *cf.* cas d'une base d'ondelettes par exemple. D'après Parseval, on écrit :

$$\left\| \hat{f} - f \right\|_2^2 = \sum_{m \in \mathcal{M}} \left(\hat{\theta}_m - \theta_m \right)^2 \quad \text{où } \hat{\theta}_m = \left\langle \hat{f}, \phi_m \right\rangle$$

Conclusion. Estimer f équivaut à estimer la suite $\theta = \{\theta_m, m \in \mathcal{M}\}$. Si $\hat{\theta}$ est un estimateur de θ, alors $\hat{f} = \sum_{\mathcal{M}} \hat{\theta}_m \phi_m$ est un estimateur de f.

Remarque 3. En utilisant les sommes de Riemann, on peut écrire les approximations suivantes :

$$\theta_m = \int_0^1 f(x) \phi_m(x)\, \mathrm{d}x \approx \frac{1}{n} \sum_{i=1}^n f\left(\frac{i}{n}\right) \phi_m\left(\frac{i}{n}\right)$$

$$\theta_m \approx \underbrace{\frac{1}{n} \sum_{i=1}^n Y_i \phi_m\left(\frac{i}{n}\right)}_{Z_m} - \frac{\sigma}{\sqrt{n}} \underbrace{\frac{1}{\sqrt{n}\,\sigma} \sum_{i=1}^n \xi_i \phi_m\left(\frac{i}{n}\right)}_{\varepsilon_m}$$

Alors, on a, $\forall m \in \mathcal{M}$:

$$Z_m = \theta_m + \frac{\sigma}{\sqrt{n}} \varepsilon_m \quad / \ m \in \mathcal{M} \tag{2}$$

Le système (2) sous la condition que ε_m soit iid $\mathcal{N}(0,1)$ est appelé **modèle de suites gaussiennes**.

Important. Les Z_m sont observables. θ_m sont les valeurs à estimer.

Note 4. En utilisant la version de Lyapunov du Théorème central limite, on peut montrer que lorsque n tend vers l'infini, la famille des v.a. $\{\varepsilon_m, m \in \mathcal{M}\}$ converge en loi vers une famille des variables iid Gaussiennes centrées réduites.

Proposition 5. *Si $\xi_1, ..., \xi_n$ sont iid, $\mathbb{E}[\xi_i] = 0$ et $\mathbb{E}[\xi_i^2] = \sigma^2$, alors :*

i. $\mathbb{E}[\varepsilon_m[= 0$, $\forall m \in \mathcal{M}$

ii. On a :

$$\lim_{n \to \infty} \mathrm{Var}[\varepsilon_m] = 1$$

iii. On a également :

$$\lim_{n \to \infty} \mathrm{Cov}[\varepsilon_m, \varepsilon_{m'}] = 0 \text{ pour } m \neq m'$$

Démonstration.

i. $\mathbb{E}_m = \frac{1}{\sqrt{n}\,\sigma} \sum_{i=1}^n \mathbb{E}[\xi_i]\, \phi_m\left(\frac{i}{n}\right) = 0$

ii. On peut écrire :

$$\begin{aligned}
\mathrm{Var}[\varepsilon_m] &= \frac{1}{n\,\sigma^2}\,\mathrm{Var}\left[\sum_{i=1}^{n} \underbrace{\xi_i\,\phi_m\!\left(\frac{i}{n}\right)}_{\text{indépendants}}\right] \\
&= \frac{1}{n\,\sigma^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}\left[\,\xi_i\,\phi_m\!\left(\frac{i}{n}\right)\right]}_{=\sigma^2\,\phi_m\left(\frac{i}{n}\right)^2} \\
&= \frac{1}{n}\sum_{i=1}^{n} \phi_m\!\left(\frac{i}{n}\right)^2 \xrightarrow[n\to\infty]{} \int_0^1 \phi_m^2 = 1
\end{aligned}$$

iii. De la même manière,

$$\begin{aligned}
\mathrm{Cov}(\varepsilon_m, \varepsilon_{m'}) &= \mathbb{E}[\varepsilon_m\,\varepsilon_{m'}] \\
&= \frac{1}{n\,\sigma^2}\,\mathbb{E}\left[\sum_{i=1}^{n} \xi_i\,\phi_m\!\left(\frac{i}{n}\right)\sum_{j=1}^{n} \xi_i\,\phi_{m'}\!\left(\frac{j}{n}\right)\right] \\
&= \frac{1}{n\,\sigma^2}\sum_{i=1}^{n} \mathbb{E}\left[\,\xi_i^2\;\phi_m\!\left(\frac{i}{n}\right)\phi_{m'}\!\left(\frac{i}{n}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \phi_m\!\left(\frac{i}{n}\right)\phi_{m'}\!\left(\frac{i}{n}\right) \\
&\xrightarrow[n\to\infty]{} \int_0^1 \phi_m\,\phi_{m'} = 0
\end{aligned}$$

□

Estimateurs classiques dans le modèle (2)

Estimateurs par projection *(cut-off)*

$$\hat{\theta}_m = \begin{cases} Z_m & \text{si } \|m\| \leqslant \lambda \\ 0 & \text{si } \|m\| > \lambda \end{cases} = Z_m\,\mathbf{1}(\|m\| < \lambda)$$

Idée. Comme $f \in \mathcal{L}^2[0,1]$, on a :

$$\int f^2 < +\infty \Rightarrow \sum_{\mathcal{M}} \theta_m^2 < +\infty \Rightarrow \theta_m \xrightarrow[\|m\|\to+\infty]{} 0$$

L'estimateur par *cut-off* estime θ_m par 0 à partir d'un certain indice, ce qui n'induit pas beaucoup d'erreur car $\theta_m \to 0$.

Estimateurs par filtrage diagonal ($\mathcal{M} = \mathbb{N}$)

L'estimateur par *cut-off* étant un peu brutal, on peut définir $\hat{\theta}_m$ par :

$$\hat{\theta}_m = \alpha_m\,Z_m \quad \text{où } \alpha_m \in [0,1] \text{ est une suite donnée}$$

Il existe par exemple :

- *Tikhonov :* $\alpha_m = \dfrac{1}{1+\left(\dfrac{m}{\lambda}\right)^2}$ (années '70)

- *Pinsker :* $\alpha_m = \left(1 - \left|\frac{m}{\lambda}\right|^{\beta}\right)_+$ (année '80)

β et λ sont des paramètres de ces algorithmes.

Seuillage dur ***(hard thresholding)***

$$\hat{\theta}_m = Z_m \, \mathbf{1}(|Z_m| > \lambda)$$

Seuillage doux ***(soft thresholding)***

$$\hat{\theta}_m = |Z_m - \lambda| \, \mathrm{sign}(Z_m) \, \mathbf{1}(|Z_m| > \lambda)$$