ENSAE
3ème année

Estimation non paramétrique

Professeur : Cristina Butucea

Prise de notes : Camille Charreaux, Guillaume Lachaud et Corentin Odic

Table des matières

1 Estimation non paramétrique de fonctions

(voir livre Estimation non paramétrique de A. Tsybakov)

1.1 Modèles statistiques

1.1.1 Densité de probabilité

On observe n variables aléatoires $X_1, \dots, X_n$ i.i.d., ayant une loi commune absolument continue par rapport à la mesure de Lebesgue, donc admettant une densité de probabilité $p : \mathbb{R} \to [0, +\infty)$ telle que :

$$p \geq 0 \quad \text{et} \qquad \int_{\mathbb{R}} p(x)\, dx = 1$$

On veut estimer p.

Définition 1.1 (Estimateur)**.** Un estimateur $\hat{p}_n(x) = \hat{p}_n(x, X_1, \dots, X_n)$ est une fonction mesurable des observations.

Définition 1.2 (Problème paramétrique)**.** Si on suppose que $p \in \{p(x,\theta), \theta \in \Theta\}$ avec $\Theta \subset \mathbb{R}^k$ ou un autre espace vectoriel de dimension k finie, alors estimer p revient à estimer θ.

Définition 1.3 (Problème non paramétrique)**.** On suppose que $p \in \mathcal{F}$ où $\mathcal{F}$ ne peut être en bijection avec aucun espace de dimension finie.

Exemple 1.1. On prend $\mathcal{F} = \Sigma(1, L)$, la classe de fonctions 1-Lipschitz, c'est-à-dire que si $p \in \mathcal{F}$, pour tout x, y dans T un intervalle de $\mathbb{R}$, $|p(x) - p(y)| \leq L|x - y|$. $\mathcal{F}$ est alors une boule de rayon L dans un espace de fonctions Lipschitziennes avec pour norme

$$\|p\|_{\mathrm{L}ip} = \sup_{x \neq y} \frac{|p(x) - p(y)|}{|x - y|}.$$

Exemple 1.2 (Autres exemples)**.** On peut aussi considérer les classes de fonctions monotones ou convexes.

1.1.2 La régression non-paramétrique

On observe des couples de variables aléatoires réelles $(X_1, Y_1), \dots, (X_n, Y_n)$ i.i.d. tels que $Y_i = f(X_i) + \xi_i$, où les ξ_i sont indépendants des X_i et centrés (pour tout i, $E(\xi_i) = 0$).

Supposons que $X_i \in [0, 1]$. On veut estimer $f : [0, 1] \to \mathbb{R}$.

Définition 1.4 (Modèle paramétrique linéaire)**.** On appelle *modèle paramétrique linéaire* un modèle de régression où f peut s'écrire $f_{a,b}(x) = ax + b$.

1.1.3 Modèle de bruit blanc Gaussien

Considérons l'équation différentielle stochastique suivante :

$$dY(t) = f(t)\, dt + \frac{1}{\sqrt{n}}\, dW(t), \quad t \in [0, 1]$$

W représente le mouvement Brownien sur $[0, 1]$. Le but est d'estimer $f : [0, 1] \to \mathbb{R}$.

1.1.4 Modèle de densité spectrale

Définition 1.5 (Densité spectrale)**.** On observe $X_1, \dots, X_n$ issus d'un processus $\{X_t\}_{t \in \mathbb{Z}}$ stationnaire au second ordre. Si on note $\gamma_j = \mathrm{cov}(X_t, X_{t+j})$ pour tout $t \in \mathbb{Z}$ et si $\sum_j |\gamma_j| < \infty$, on peut définir la *densité spetrale* par :

$$f(x) = \frac{\gamma_0}{2\pi} + \frac{1}{\pi} \sum_{j=1} \gamma_j \cos(jx), \quad x \in [-\pi, \pi]$$

1.2 Estimation de la densité de probabilité

Soient $X_1, \dots, X_n$ n variables aléatoires i.i.d. ayant la même densité p.

Définition 1.6 (Estimateur empirique)**.** L'estimateur empirique de la fonction de répartition $F(x) = \mathbb{P}(X \leq x)$ est donné par

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}(X_i \leq x)$$

On a en particulier les propriétés suivantes :
- pour tout x dans $\mathbb{R}$, $F_n(x) \xrightarrow[n\to\infty]{p.s.} F(x)$
- (Glivenko-Cantelli) $\|F_n - F\|_\infty = \sup_{x\in\mathbb{R}} |F_n(x) - F(x)| \xrightarrow[n\to\infty]{p.s.} 0$,
- pour tout $\varepsilon > 0$, $P(\|F_n - F\|_\infty \geq \varepsilon) \leq 2\exp(-2n\varepsilon^2)$.

Pour $h > 0$ suffisamment petit, on peut faire l'approximation

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

En remplaçant F par F_n dans la formule, on arrive à la définition suivante.

Définition 1.7 (Estimateur de Rosenblatt, 1956)**.** On appelle *estimateur de Rosenblatt* la fonction

$$\hat{p}_n^R(x) = \frac{F_n(x+h) - F_n(x-h)}{2h}.$$

Il peut se réécrire sous la forme suivante

$$\hat{p}_n^R(x) = \frac{1}{2nh} \sum_{i=1}^{n} \mathbf{1}(x - h < X \leq x + h) = \frac{1}{nh} \sum_{i=1}^{n} K_0\left(\frac{X_i - x}{h}\right),$$

où $K_0(u) = \frac{1}{2}\mathbf{1}(-1 < u \leq 1)$.

Cet estimateur peut être généralisé en modifiant le choix de K_0 et mène à la définition suivante.

Définition 1.8 (Estimateur à noyau ou estimateur de Parzen-Rosenblatt, 1962)**.** Soit $K : \mathbb{R} \to \mathbb{R}$ une fonction intégrable telle que $\int K(u)\, du = 1$. Soit $h > 0$ la *fenêtre (bandwidth)*. Alors K est appelé *noyau* et on désigne par *estimateur à noyau* ou *estimateur de Parzen-Rosenblatt* la fonction qui à x associe

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

Exemple 1.3 (Noyau rectangulaire)**.** $K(u) = \mathbf{1}(|u| \leq 1)$

Exemple 1.4 (Noyau triangulaire)**.** $K(u) = (1 - |x|)_+$

Exemple 1.5 (Noyau Gaussien)**.** $K(u) = \frac{1}{\sqrt{2\pi}} \exp(-u^2/2)$

Exemple 1.6 (Noyau de Silverman)**.** $K(u) = \frac{1}{2} \exp(-|u|/\sqrt{2}) \sin(|u|/\sqrt{2} + \pi/4)$

Exemple 1.7 (Noyau d'Epanechnikov)**.** $K(u) = \frac{3}{4}(1 - u^2)\mathbf{1}(|u| \leq 1)$

1.2.1 Erreur quadratique moyenne des estimateurs à noyaux

Un moyen de mesurer la précision d'un estimateur est de calculer son *erreur quadratique moyenne* (appelée **MSE** pour ***M**ean **S**quared **E**rror*) pour un point $x_0 \in \mathbb{R}$:

$$\mathrm{MSE} = \mathrm{MSE}(x_0) \triangleq \mathbb{E}_p\left[\left(\hat{p}_n(x_0) - p(x_0)\right)^2\right]$$

où
$$\mathbb{E}_p\left[\left(\hat{p}_n(x_0) - p(x_0)\right)^2\right] \triangleq \int \dots \int \left(\hat{p}_n(x_0, x_1, \dots, x_n) - p(x_0)\right)^2 \prod_{i=1}^{n} \left[p(x_i)\, dx_i\right].$$

L'espérance sous p de $\hat{p}_n(x_0)$ peut s'écrire

$$\begin{aligned}
\mathbb{E}_p\left[\hat{p}_n(x_0)\right] &= \mathbb{E}_p\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}K\left(\frac{X_i - x_0}{h}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h}\int K\left(\frac{z - x_0}{h}\right)p(z)\, dz \\
&= \frac{1}{h}\int K\left(\frac{z - x_0}{h}\right)p(z)\, dz
\end{aligned}$$

On peut réécrire MSE sous la forme

$$\begin{aligned}
\mathrm{MSE}(x_0) &= \mathbb{E}_p\left[\left(\left(\hat{p}_n(x_0) - \mathbb{E}_p\left[\hat{p}_n(x_0)\right]\right) + \left(\mathbb{E}_p\left[\hat{p}_n(x_0)\right] - p(x_0)\right)\right)^2\right] \\
&= \mathbb{E}_p\left[\left(\hat{p}_n(x_0) - \mathbb{E}_p\left[\hat{p}_n(x_0)\right]\right)^2\right] + \mathbb{E}_p\left[\left(\mathbb{E}_p\left[\hat{p}_n(x_0)\right] - p(x_0)\right)^2\right] \\
&\quad + \mathbb{E}_p\left[\left(\hat{p}_n(x_0) - \mathbb{E}_p\left[\hat{p}_n(x_0)\right]\right)\left(\mathbb{E}_p\left[\hat{p}_n(x_0)\right] - p(x_0)\right)\right]
\end{aligned}$$

Le deuxième terme ne dépendant pas de p, le troisième étant nul et en notant $b(x_0) = \mathbb{E}_p\left[\hat{p}_n(x_0)\right] - p(x_0)$ le biais de l'estimateur $\hat{p}_n(x_0)$ et $\sigma^2(x_0) = \mathbb{E}_p\left[\left(\hat{p}_n(x_0) - \mathbb{E}_p\left[\hat{p}_n(x_0)\right]\right)^2\right]$ sa variance, on obtient

$$\mathrm{MSE}(x_0) = b^2(x_0) + \sigma^2(x_0) \tag{2}$$

Définition 1.9 (Biais et variance)**.** On appelle la quantité $b(x_0)$ le *biais* de l'estimateur $\hat{p}_n$ au point x_0. La quantité $\sigma^2(x_0)$ est sa *variance*.

1.2.2 Étude la variance

$\hat{p}_n(x_0) - \mathbb{E}_p\left[\hat{p}_n(x_0)\right]$ peut se réécrire sous la forme

$$\begin{aligned}
\hat{p}_n(x_0) - \mathbb{E}_p\left[\hat{p}_n(x_0)\right] &= \frac{1}{nh}\sum_{i=1}^{n}\left[K\left(\frac{X_i - x_0}{h}\right) - \mathbb{E}_p\left[K\left(\frac{X_i - x_0}{h}\right)\right]\right] \\
&= \frac{1}{nh}\sum_{i=1}^{n} Z_i
\end{aligned}$$

en posant $Z_i = K\big((X_i - x_0)/h\big) - \mathbb{E}_p\left[K\big((X_i - x_0)/h\big)\right]$

Proposition 1. *Supposons que pour tout z, $p(z) \leq p_{\max} < \infty$. Soit $K : \mathbb{R} \to \mathbb{R}$ telle que*

$$\int K^2(u)\, du < \infty. \tag{3}$$

Alors, pour tout $x_0 \in \mathbb{R}, h > 0$ et $n \geq 1$ on a

$$\sigma^2(x_0) \leq \frac{C_1}{nh}$$

avec $C_1 = p_{\max} \int K^2(u)\, du$.

Démonstration. Les Z_i sont des variables aléatoires i.i.d. centrées de variance

$$\begin{aligned}
\mathbb{E}_p\left[Z_i^2\right] &\leq \mathbb{E}_p\left[K^2\left(\frac{X_i - x_0}{h}\right)\right] \\
&\leq \int K^2\left(\frac{z - x_0}{h}\right) p(z)\, dz \\
&\leq p_{\max} \int K^2\left(\frac{z - x_0}{h}\right)\, dz \\
\left({\scriptstyle u=(z-x_0)/h \atop du=dz/h}\right) &\leq p_{\max} h \int K^2(u)\, du
\end{aligned}$$

Ainsi,

$$\begin{aligned}
\sigma^2(x_0) &= \mathbb{E}_p\left[\left(\frac{1}{nh}\sum_{i=1}^{n} Z_i\right)^2\right] \\
&= \frac{1}{n^2h^2}\sum_{i=1}^{n} \mathbb{E}_p\left[Z_i^2\right] \\
&= \frac{1}{nh^2}\sum_{i=1}^{n} \mathbb{E}_p\left[Z_1^2\right] \\
&\leq \frac{p_{\max}}{nh}\int K^2(u)\, du
\end{aligned}$$

□

Remarque. Si $h = h_n$ (i.e. une fonction de n) de telle sorte que $nh_n \to 0$ quand $n \to \infty$, alors la variance $\sigma^2(x_0)$ tend vers 0 quand $n \to \infty$.

1.2.3 Étude du biais

On rappelle que le biais d'un estimateur à noyau à la forme

$$b(x_0) = \mathbb{E}_p\left[\hat{p}_n(x_0)\right] - p(x_0) = \frac{1}{h}\int K\left(\frac{z - x_0}{h}\right) p(z)\, dz - p(x_0)$$

Définition 1.10 (Classe de Hölder)**.** Soit T un intervalle dans $\mathbb{R}$, et soit β et L deux nombres positifs et $\ell = \lfloor \beta \rfloor$. La *classe de Hölder* $\Sigma(\beta, L)$ sur T est définie comme l'ensemble des fonctions $f : T \to \mathbb{R}$ telles que leur dérivée $f^{(\ell)}$ vérifie

$$\left|f^{(\ell)}(x) - f^{(\ell)}(x')\right| \leq L|x - x'|^{\beta - \ell}, \quad \forall x, x' \in T.$$

Remarque. $f^{(l)}$ est $(\beta - \ell, L)$- Lipschitz.

Remarque. $\lfloor 3.2 \rfloor = 3$, $\lfloor 4 \rfloor = 3$.

Définition 1.11 (Noyau d'ordre ℓ)**.** Soit $\ell \geq 1$ un entier. On dit que $K : \mathbb{R} \to \mathbb{R}$ est un *noyau d'ordre* ℓ si les fonctions $u \to u^j K(u)$, $j = 0, 1, \ldots, \ell$ sont intégrables et vérifient

$$\int K(u)\, du = 1, \qquad \int u^j K(u)\, du = 0, \quad j = 1, \ldots, \ell$$

Remarque. D'autres définitions supposent $\int u^{\ell+1} K(u)\, du \neq 0$ en plus.

Proposition 2. *On note*

$$\mathcal{P}(\beta, L) = \left\{ p \;\middle|\; p \geq 0, \int p(x)\, dx = 1, \text{ and } p \in \Sigma(\beta, L) \text{ on } \mathbb{R} \right\}$$

Si $p \in \mathcal{P}(\beta, L)$ et soit K un noyau d'ordre $\ell = \lfloor \beta \rfloor$ tel que

$$\int |u|^\beta |K(u)|\, du < \infty.$$

Alors, pour tout x_0 dans $\mathbb{R}, h > 0$ et $n \geq 1$, on a

$$\left| b(x_0) \right| \leq C_2 h^\beta$$

avec

$$C_2 = \frac{L}{\ell\,!} \int |u|^\beta |K(u)|\, du.$$

Démonstration. On a

$$\begin{aligned} b(x_0) &= \frac{1}{h} \int K\left(\frac{z - x_0}{h}\right) p(z)\, dz - p(x_0) \\ &= \int K(u) \big[p(x_0 + uh) - p(x_0) \big]\, du. \end{aligned}$$

Pour un certain $0 \leq \tau \leq 1$, on a

$$p(x_0 + uh) = p(x_0) + p'(x_0) uh + \ldots + \frac{(uh)^\ell}{\ell\,!} p^{(\ell)}(x_0 + \tau uh)$$

Comme K est d'ordre $l = \lfloor \beta \rfloor$, les $\int u^j K(u)\, du$ pour $j = 1, \ldots, l$ sont nuls, donc

$$\begin{aligned} b(x_0) &= \int K(u) p(x_0)\, du - p(x_0) + \int K(u) \frac{(uh)^\ell}{\ell\,!} p^{(\ell)}(x_0 + \tau uh)\, du \\ &= \int K(u) \frac{(uh)^\ell}{\ell\,!} p^{(\ell)}(x_0 + \tau uh)\, du \\ &= \int K(u) \frac{(uh)^\ell}{\ell\,!} \left(p^\ell(x_0 + \tau uh) - p^{(\ell)}(x_0) \right) du \end{aligned}$$

En prenant la valeur absolue et en utilisant le fait que $p \in \mathcal{P}(\beta, L)$, on a alors

$$\begin{aligned} |b(x_0)| &\leq \int |K(u)| \frac{|uh|^\ell}{\ell\,!} \left| p^{(\ell)}(x_0 + \tau uh) - p^{(\ell)}(x_0) \right| du \\ &\leq L \int |K(u)| \frac{|uh|^\ell}{\ell\,!} |x_0 + \tau uh - x_0|^{\beta - L}\, du \\ &\leq \frac{L}{\ell\,!} h^\beta \int |u|^\beta |K(u)|\, du \end{aligned}$$

□

Lorsque p et K satisfont les hypothèses des propositions 1 et 2, on a

$$\mathrm{MSE}(x_0) \leq \frac{C_1}{nh} + C_2^2 h^{2\beta} \tag{4}$$

Notons h_n^* la valeur qui minimise le terme de droite. h_n^* vérifie

$$-\frac{C_1}{(h_n^*)^2 n} + 2\beta C_2^2 (h^*)^{2\beta - 1} = 0$$

Soit encore

$$h_n^* = \left(\frac{C_1}{2\beta C_2^2} \right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+1}}$$

Ainsi, en réinjectant h_n^* dans l'inégalité, on obtient

$$\mathrm{MSE}(x_0) = O\left(n^{-\frac{2\beta}{2\beta+1}} \right), \quad n \to \infty$$

Théorème 3. *Sous les hypothèses des propositions 1 et 2, en prenant* $h_n^* = \alpha n^{-1/(2\beta+1)}, \alpha > 0$, *on a*

$$\sup_{p\in\mathcal{P}(\beta,\mathcal{L})} \mathrm{MSE}\big[\hat{p}_n(x_0)\big] \leq C n^{-2\beta/(2\beta+1)}, \quad \textit{pour tout } x_0 \in \mathbb{R},$$

où $C = C(\beta, L, K, \alpha)$.

Démonstration. Il reste à montrer qu'il existe $p_{\max} < \infty$ tel que

$$\sup_{z\in\mathbb{R}} \sup_{p\in\mathcal{P}(\beta,L)} p(z) \leq p_{\max}.$$

Considérons K^* un noyau borné d'ordre ℓ (pas nécessairement égal à K). En appliquant la proposition 2 à K^* pour $h = 1$, on a, pour tout x_0 dans $\mathbb{R}$ et pour tout p dans $\mathcal{P}(\beta, L)$,

$$\left| \int K^*(z - x_0)p(z)\,dz - p(x_0) \right| \leq C_2^* \triangleq \frac{L}{\ell\,!} \int |u|^\beta \big|K^*(u)\big|\,du.$$

Ainsi, pour tout x dans $\mathbb{R}$ et pour tout p dans $\mathcal{P}(\beta, L)$,

$$p(x) \leq C_2^* + \int \big|K^*(z - x)\big|p(z)\,dz \leq C_2^* + K^*_{\max}$$

avec $K^*_{\max} = \sup_{u\in\mathbb{R}}\big|K^*(u)\big|$. On obtient le résultat souhaité en prenant $p_{\max} = C_2^* + K^*_{\max}$. □

On note $\psi^2_{n,\beta} = n^{-2\beta/(2\beta+1)}$. On dit que "$\hat{p}_n(x_0)$ atteint la vitesse $\psi^2_{n,\beta}$ sur la classe $\mathcal{P}(\beta, L)$."

1.2.4 Résultat d'optimalité

On considère l'inégalité suivante

$$\inf_{\tilde{p}_n} \sup_{p\in\mathcal{P}(\beta,L)} \mathrm{MSE}\big(\tilde{p}_n(x_0)\big) \geq c\psi^2_{n,\beta} \tag{5}$$

où l'inf est pris sur tous les estimateurs $\hat{p}_n$ de p.

Définition 1.12 (Vitesse minimax)**.** On dit que $\psi^2_{n,\beta}$ est la *vitesse minimax* du MSE sur la classe $\mathcal{P}(\beta, L)$ si il existe un estimateur qui atteint les bornes supérieures du Théorème 3 et les bornes inférieures 5 sont vérifiées.

Remarque. Si $\beta = 1/2$, $\psi^2_{n,\beta} = n^{-1/2}$. La vitesse diminue si β augmente.

Remarque (Choix de h). Si l'on prend h trop petit, on risque d'être dans une situation de *sous-lissage*. À l'inverse, un h trop grand peut entraîner du *sur-lissage*.

1.3 Estimation confidentielle de la densité

Si on note $X = (X_1, ..., X_n)$ l'échantillon, X suit la loi jointe $P^{\otimes n}$ (si les coordonnées sont i.i.d.) pour P une loi de probabilité sur $(\chi^n, \mathcal{A}^n)$ dont on veut estimer une fonctionnelle $\theta = \theta(P)$ à valeurs réelles, e.g. $E_P(X)$, $p(x_0)$ pour un x_0 réel fixé.

Pour des raisons de confidentialité, le statisticien doit randomiser X pour obtenir Z sur un espace de probabilité $(\mathcal{Z}, \mathcal{B})$ (souvent différent de l'espace d'origine). La loi conditionnelle

$$P(Z \in A|X = x) = Q(A|x), \quad A \in \mathcal{B},\ x \in \chi^n,$$

est décrite par le noyau de Markov Q. Ainsi, Z est distribué selon la loi $QP^{\otimes n}$:

$$QP^n(dz) = \int Q(dz|x)P^n(dx).$$

Dans un 2ème temps on doit utiliser Z pour estimer le paramètre $\theta(P)$ d'intérêt.

Définition 1.13. Si $\alpha > 0$, un noyau de Markov $Q : \mathcal{B} \times \chi^n \to [0,1]$ est $\alpha-$ DP (*differentially private*) si pour tout A dans $\mathcal{B}$, pour tous x, x' dans χ^n tels que $H(x,x') = 1$ (la perte de Hamming : nombre de coordonnées différentes entre x et x') on a

$$Q(A|x) \leq e^\alpha Q(A|x').$$

Conséquences : si x et x' sont tels que $H(x,x') = 1$, les mesures $Q(\cdot|x)$ et $Q(\cdot|x')$ sont équivalentes et

$$e^{-\alpha} \leq \frac{Q(A|x)}{Q(A|x')} \leq e^\alpha.$$

Si α décroit vers 0, on plus de confidentialité.

Estimation paramétrique de la moyenne : On suppose que la loi P est définie sur $[-M,M]$ et que l'on estime la moyenne $\theta = E_P(X)$.

Il suffit de rendre publique la donnée confidentielle :

$$Z = \overline{X_n} + \frac{2M}{n\alpha} W, \quad p^W(x) = \frac{1}{2}\exp(-|x|)$$

c-à-d W suit une loi de Laplace. Alors, cet algorithme est $\alpha-$DP.

Preuve : En effet, le noyau de Markov admet une densité de probabilité conditionnelle :

$$q(z|x) = \frac{1}{2}\frac{n\alpha}{2M}\exp\left(-\frac{n\alpha}{2M}|z - \overline{x_n}|\right).$$

On a, si x et x' sont différenets en la coordonnée i :

$$\begin{aligned}
\frac{q(z|x)}{q(z|x')} &= \exp\left(-\frac{n\alpha}{2M}(|z - \overline{x_n}| - |z - \overline{x'_n}|)\right) \\
&\leq \exp\left(\frac{n\alpha}{2M}(|\overline{x_n} - \overline{x'_n}|)\right) \\
&\leq \exp\left(\frac{\alpha}{2M}(|x_i - x'_i|)\right) \leq e^\alpha.
\end{aligned}$$

Remarquons que $E(Z) = E(\overline{X_n}) = E(X)$ et

$$Var(Z) = Var(\overline{X_n}) + \frac{4M^2}{n^2\alpha^2} \cdot Var(W).$$

Cette approche est globale car, pour produire Z, on a besoin de connaitre tout l'échantillon X. Une approche locale est encore plus confidentielle car elle n'utilise pour chaque Z_i qu'une valeur de X_i et, éventuellement, les autres $Z_1, ..., Z_{i-1}$ déjà produits.

Définition 1.14. Un noyau $Q : \mathcal{B}^{\otimes n} \times \chi^n \to [0,1]$ est non-interactif si il existe $Q_i : \mathcal{B} \times \chi \to [0,1]$ noyaux tels que :

$$Q(dz|x) = \bigotimes_{i=1}^n Q_i(dz_i|x_i).$$

Remarque : Q non-interactif est $\alpha-$DP si et seulement si Q_i est $\alpha-$DP pour tout i de 1 à n.

Dans l'exemple précédent, un noyau interactif produirait

$$Z_i = X_i + \frac{2M}{\alpha} W_i, \quad W_i \text{ i.i.d. de loi de Laplace.}$$

Alors, le noyau Q aura la densité conditionnelle

$$q(z|x) = \prod_{i=1}^n \frac{1}{2}\frac{\alpha}{2M}\exp\left(-\frac{\alpha}{2M}|z_i - x_i|\right)$$

et est $\alpha-$DP.

L'estimateur $\overline{Z_n}$ de la moyenne θ sera sans biais : $E(\overline{Z_n}) = E(\overline{X_n}) = E(X)$ et de variance

$$Var(\overline{Z_n}) = Var(\overline{X_n}) + \frac{4M^2}{n\alpha^2} \cdot Var(W).$$

Définition 1.15. Un noyau $Q : \mathcal{B}^{\otimes n} \times \chi^n \to [0,1]$ est séquentiellement interactif si il existe $Q_i : \mathcal{B} \times (\chi \times \mathcal{Z}^{i-1}) \to [0,1]$ noyaux tels que :

$$Q(dz|x) = Q_1(z_1|x_1) \times \bigotimes_{i=2}^{n} Q_i(dz_i|x_i, z_1, ..., z_{i-1}).$$

Remarque : Q non-interactif est $\alpha-$DP si et seulement si Q_i est $\alpha-$DP pour tout i de 1 à n.

Smith (2008) Estimation paramétrique efficace en globale.

Estimation non paramétrique de la densité de probabilité : Histogramme Wasserman, Zhou (2010) estimateur de la densité Hölder de régularité $\beta \in (0,1)$ qui atteint la vitesse non paramétrique plus lente, qu'avec les observations X.

$X_1, ..., X_n$ i.i.d. de densité de probabilité $p : [0,1] \to \mathbb{R}$, p dans $\mathcal{P}(\beta, L)$, β dans (0,1).

Soit K classes égales qui partitionnent [0,1] : $C_j = \left[\frac{j-1}{K}, \frac{j}{K}\right)$, $j = 1, ..., K$. L'histogramme est défini par

$$\begin{aligned} \hat{p}(x) &= \frac{K}{n} \sum_{j=1}^{K} I(x \in C_j) \cdot \sharp\{i : X_i \in C_j\} \\ &= \frac{K}{n} \sum_{j=1}^{K} I(x \in C_j) \sum_{i=1}^{n} I(X_i \in C_j). \end{aligned}$$

Le noyau de Markov non-interactif produit $Z_{i\cdot}$ vecteur de dimension K comme suit :

$$Z_{i,j} = I(X_i \in C_j) + \frac{2}{\alpha} W_{i,j}, \quad i = 1, ..., n, \ j = 1, ..., K,$$

où $W_{i,j}$ sont i.i.d. de loi de Laplace.

Exercice : Montrer que ce noyau est $\alpha-$DP.

Estimateur confidentiel de la densité p au point x dans (0,1) :

$$\hat{p}^{NI}(x) = K \sum_{j=1}^{K} I(x \in C_j) \cdot \frac{1}{n} \sum_{i=1}^{n} Z_{i,j}.$$

Cet estimateur s'écrit aussi

$$\hat{p}^{NI}(x) = \hat{p}(x) + K \sum_{j=1}^{K} I(x \in C_j) \cdot \overline{W_{\cdot,j}}.$$

On a

$$E(\hat{p}^{NI}(x)) = E(\hat{p}(x)) = K \sum_{j=1}^{K} I(x \in C_j) \int_{C_j} p.$$

Le biais de cet estimateur s'écrit, si x appartient à C_j :

$$\begin{aligned} |E(\hat{p}^{NI}(x)) - p(x)| &= |K \int_{C_j} p - p(x)| = K| \int_{C_j} [p(u) - p(x)] du| \\ &\leq \int_{C_j} |p(u) - p(x)| du \leq KL \int_{C_j} |u - x|^{\beta} du \leq LK^{-\beta}. \end{aligned}$$

La variance est majorée, si x appartient à C_j, par

$$\begin{aligned} Var(\hat{p}^{NI}(x)) &= Var(\hat{p}(x)) + K^2 \frac{4}{n\alpha^2} Var(W_{1,1}) \\ &= \frac{K}{n} p(x)(1+o(1)) + \frac{8K^2}{n\alpha^2}. \end{aligned}$$

Vu que la densité p est bornée, le terme dominant de la variance est de l'ordre $K^2/(n\alpha^2)$. Pour un choix de $K = c(n\alpha^2)^{\frac{1}{2\beta+2}}$, on obtient

$$MSE(\hat{p}^{NI}(x), p(x)) \leq C(n\alpha^2)^{-\frac{2\beta}{2\beta+2}},$$

qui est plus lente que $n^{-\frac{2\beta}{2\beta+1}}$ si les X_i sont observés. Cette vitesse est néanmoins optimale (Duchi et Ruan, 2018).

1.4 Validation croisée

La validation croisée est une méthode pour choisir automatiquement la fenêtre h à partir des données. On considère la MISE (*Mean Integrated Squared Error*, ou erreur quadratique intégrée) :

$$\text{MISE} \triangleq \mathbb{E}_p\left[\int \left(\hat{p}_n(x) - p(x)\right)^2 dx\right]$$

On peut refaire des calculs de vitesse (ex) et on peut atteindre les mêmes vitesses de convergence, ce qui nous amène au même problème qui est la dépendance en β. On veut retrouver $h_{id} = \arg\min_{h>0} \text{MISE}(\hat{p}_n^h, p)$ avec $\hat{p}_n^h$ un estimateur à noyau K fixé et $h > 0$.

Précedemment, on a procédé de la façon suivante :

Par le théorème de Tonelli-Fubini, on a

$$\text{MISE} = \int \text{MSE}(x)\, dx = \int b^2(x)\, dx + \int \sigma^2(x)\, dx. \tag{6}$$

On donne des majorations des deux termes, fonction de n et de h et on minimise en h cette borne supérieure. On obtient $h = \alpha n^{-1/(2\beta+1)}$.

Remarque. $h = \alpha n^{-1/(2\beta+1)}$ dépend de β.

Ici, on cherche un estimateur sans biais du MISE et on minimise en h cet estimateur. On développe

$$\text{MISE}(\hat{p}_n^h, p) = \mathbb{E}_p\left[\int \left(\hat{p}_n^h(x)\right)^2 dx\right] - 2\,\mathbb{E}_p\left[\int \hat{p}_n^h(x) p(x)\, dx\right] + \int p^2(x)\, dx$$

Remarque. On peut oublier le dernier terme car il ne dépend pas de h.

On note $\mathcal{J}(h) = \mathbb{E}_p\left[\int \left(\hat{p}_n^h(x)\right)^2 dx\right] - 2\,\mathbb{E}_p\left[\int \hat{p}_n^h(x) p(x)\, dx\right]$ et on veut estimer sans biais $\mathcal{J}(h)$.

— $\int \left(\hat{p}_n^h(x)\right)^2$ est un estimateur sans biais (ESB) du premier terme de $\mathcal{J}(h)$.

— le deuxième terme se calcule de la manière suivante :

$$\begin{aligned} \mathbb{E}_p\left[\int \hat{p}_n^h(x) p(x)\, dx\right] &= \mathbb{E}_p\left[\int \left(\frac{1}{n}\sum_{i=1}^n \frac{1}{h} K\left(\frac{X_i - x}{h}\right)\right) p(x)\, dx\right] \\ &= \int \mathbb{E}_p\left[\frac{1}{h} K\left(\frac{X_i - x}{h}\right)\right] p(x)\, dx \\ &= \int\int \frac{1}{h} K\left(\frac{y - x}{h}\right) p(y) p(x)\, dy\, dx \end{aligned}$$

Définition 1.16 (Estimateur "leave-one-out" de p). On appelle *estimateur "leave-one"out* de p en x la valeur suivante :

$$\hat{p}_{n,(-i)}(x) = \frac{1}{n-1} \sum_{\substack{k=1 \\ k \neq i}}^{n} \frac{1}{h} K\left(\frac{X_k - X_i}{h}\right)$$

On pose

$$\hat{G} = \frac{1}{n} \sum_{i=1}^{n} \hat{p}_{n,(-i)}(X_i) = \frac{1}{n(n-1)} \sum_{\substack{k=1 \\ k \neq i}} \frac{1}{h} K\left(\frac{X_k - X_i}{h}\right)$$

On a

$$\mathbb{E}_p(\hat{G}) = \frac{1}{n(n-1)} \sum_{k \neq i} \frac{1}{h} \mathbb{E}_p\left[K\left(\frac{X_k - X_i}{h}\right)\right].$$

La loi jointe de (X_k, X_i) est $p(y)p(x)$ (variables indépendantes). Donc

$$\mathbb{E}_p(\hat{G}) = \frac{1}{h} \int \int K\left(\frac{y-x}{h}\right) p(y)p(x)\, dy\, dx$$

Ce qui est le deuxième terme de $\mathcal{J}(h)$ à un facteur près. On propose donc $\mathrm{CV}(h) = \int (\hat{p}_n^h)^2 - 2\hat{G}$ qui est un estimateur sans biais de $\mathcal{J}(h)$ si $\int |K\big((y-x)/h\big)| p(y)p(x)\, dx\, dy < \infty$.

Pour choisir h par validation croisée, on pose

$$h^{\mathrm{CV}} \in \arg\min_{h>0} \mathrm{CV}(h)$$

Remarque. De par le caractère aléatoire du critère $\mathrm{CV}(h)$, h^{CV} est une fenêtre aléatoire.

En 1984, Stone a montré que lorsque $n \to \infty$,

$$\mathrm{MISE}(\hat{p}_n^{h^{\mathrm{CV}}}, p) = \mathrm{MISE}(\hat{p}_n^{h^{id}})\big(1 + o(1)\big)$$

1.4.1 Estimation d'une densité multivariée

On observe $X = (X_1^1, \ldots, X_1^d)^\top, \ldots, X_n = (X_n^1, \ldots, X_n^d)^\top \in \mathbb{R}^d$, i.i.d. de loi $p : \mathbb{R}^d \to \mathbb{R}^+$, intégrable d'intégrale 1.

Un estimateur à noyau $K : \mathbb{R} \to \mathbb{R}$:

$$\hat{p}_n(x_1, \ldots, x_d) = \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{X_i^1 - x_1}{h}\right) \times \cdots \times K\left(\frac{X_i^d - x_d}{h}\right)$$

Soit $\beta \in (0,1]$, on suppose que $|p(x) - p(y)| \leq L\|x - y\|^\beta$ où $\|x\| = \sqrt{x_1^2 + \cdots + x_d^2}$. Alors

$$\mathbb{E}_p\big[\hat{p}_n(x) - p(x)\big] \leq C_B h^\beta$$

et

$$\mathrm{Var}_P\big(\hat{p}_n(x)\big) \leq \frac{C\mu}{nh^d} \quad \text{sous des hypothèses sur } K.$$

On a

$$h_{\mathrm{opt}} = \arg\min_h \left[C_B^2 h^{2\beta} + \frac{C\mu}{nh^d}\right] \sim n^{-1/(2\beta+d)},$$

donc l'estimateur $\hat{p}_n(x)$ atteint la vitesse $n^{-2\beta/(2\beta+d)}$ qui est exponentiellement lente en la dimension d. C'est ce que l'on appelle le "fléau de la dimension". En pratique, il faut chercher un espace de petite dimension qui approxime sufisamment bien le support des données et estimer la densité dans cet espace là, plutôt que dans l'espace de grande dimension. Il faut réduire la dimension dès qu'elle est supérieure à 4 ou 5.

2 Régression non paramétrique

On observe $(X_1, Y_1), \ldots, (X_n, Y_n)$ i.i.d. tels que chaque Y_i peut s'écrire

$$Y_i = f(X_i) + \xi_i, i = 1, \ldots, n,$$

où les ξ_i sont indépendants des X_i, centrés et de variance finie (ou de moment absolu fini $\mathbb{E}\,|\xi| < \infty$).

On veut estimer f. On note $\hat{f}_n(x, (X_1, Y_1), \ldots, (X_n, Y_n))$ l'estimateur.

2.1 Méthodes d'estimation de la régression

On distingue

— la régression à effets aléatoires si X_i ont une densité de proba $p : T$, T compact et $p(x) \geq c > 0$ pour tout x dans T.

— la régression à effets fixes (les X_i design) $x_1, \ldots, x_n$ et équiréparties.

Ici, on prend les $X_i = i/n$ pour $i = 1, \ldots, n$. On a $Y_i = f(i/n) + \xi_i$ avec $\{\xi_i\}_i$ indépendants, $\mathbb{E}(\xi_i) = 0$ et soit $\mathbb{E}\,|\xi_i| < \infty$, soit $\mathbb{E}\,\xi^2 < \infty$

2.1.1 Estimateurs à noyau ou de Nadaraya-Watson, régression à effets aléatoires

$$f(x) = \mathbb{E}\big[Y \mid X = x\big] = \frac{\int y f^{(X,Y)}(x, y)\, dy}{f^{(X)}(x)} = \int y f^{Y|X=x}(y)\, dy$$

On choisit K un noyau ($\int K = 1$) et on estime f^X par

$$\hat{f}_n^X(y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - y}{h}\right), \qquad h > 0 \text{ petit, la } \textit{fenêtre}$$

et $f^{(X,Y)}$ par

$$\hat{f}_n^{(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^2} K\left(\frac{X_i - x}{h}\right) K\left(\frac{Y_i - y}{h}\right)$$

On suppose que $\int uK(u)\, du = 0$ (K est une fonction paire).

$$\int y \hat{f}_n^{(X,Y)}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^2} K\left(\frac{X_i - x}{h}\right) \int yK\left(\frac{Y_i - y}{h}\right)\, dy$$

On pose $Z = (Y_i - y)/h$.

$$\begin{aligned}
\int y \hat{f}_n^{(X,Y)}(x, y) &= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^2} K\left(\frac{X_i - x}{h}\right) \int (Y_i - hz) K(z)\, dz \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h^2} K\left(\frac{X_i - x}{h}\right) \left[Y_i h \int K - h \int zK(z)\, dz\right] \\
&= \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{X_i - x}{h}\right) Y_i
\end{aligned}$$

L'estimateur de Nadaraya-Watson est alors défini par

$$
\begin{aligned}
\hat{f}_n(x) &= \begin{cases} \dfrac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} Y_i \frac{1}{h} K\left(\frac{X_i - x}{h}\right)}{\hat{f}_n^X(x)}, & \text{si } \hat{f}_n^X(x) \neq 0 \\ 0, & \text{sinon} \end{cases} \\
&= \begin{cases} \dfrac{\displaystyle\sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right)}{\displaystyle\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}, & \text{si } \sum_{i=1}^{n} K\big((X_i - x)/h\big) \neq 0 \\ 0, & \text{sinon} \end{cases} \\
&= \begin{cases} \sum_{i=1}^{n} Y_i W_{n,i}(x), & \text{où } W_{n,i} = \frac{K\left((X_i - x)/h\right)}{\sum_{j=1}^{n} K\left((X_j - x)/h\right)} \text{ et } \sum_{i=1}^{n} W_{n,i}(\cdot) = 1 \\ 0, & \text{si } \sum_{j=1}^{n} K\big((X_j - x)/h\big) = 0 \end{cases}
\end{aligned}
$$

Définition 2.1 (Estimateur linéaire)**.** Un estimateur de la régression tel que $\hat{f}_n(x) = \sum_{i=1}^{n} Y_i W_{n,i}(x)$, où $W_{n,i}(x)$ ne dépend pas de Y_i, et $\sum_{i=1}^{n} W_{n,i}(x) = 1$, est appelé *estimateur linéaire.*

Remarque. Si on connaît la densité f^X des X_i, alors

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

2.1.2 Estimateur par polynômes locaux

Si $K \geq 0$ est un noyau symétrique, $\hat{f}_n(x)$, l'estimateur à noyau de $f(x)$, vérifie

$$
\begin{aligned}
\hat{f}_n(x) &= \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (Y_i - \theta)^2 K\left(\frac{X_i - x}{h}\right) \\
&= \arg\min_{\theta \in \mathbb{R}} \sum_{i=1}^{n} (-2\theta Y_i + \theta^2) K\left(\frac{X_i - x}{h}\right) \\
&= \arg\min_{\theta \in \mathbb{R}} -2\theta \sum_{i=1}^{n} Y_i K\left(\frac{X_i - x}{h}\right) + \theta^2 \underbrace{\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right)}_{\text{on suppose } \neq 0}
\end{aligned}
$$

Remarque. La régression $f(z) \approx \theta$ est une approximation locale en x par un polynôme de degré 0 (constant).

De manière plus générale, on peut étendre le développement limité de f. On pose ℓ dans $\mathbb{N}$. On a

$$
\begin{aligned}
f(z) &\approx f(x) + hf'(x)\frac{z - x}{h} + \cdots + h^{\ell} \frac{f^{(\ell)}(x)}{\ell\,!} \frac{(z - x)^{\ell}}{h^{\ell}} \\
&= \theta_0 + \theta_1 \frac{z - x}{h} + \cdots + \theta_{\ell} \frac{(z - x)^{\ell}}{\ell\,!\, h^{\ell}} \\
&= \theta^T U\left(\frac{z - x}{h}\right)
\end{aligned}
$$

où $\theta = \begin{pmatrix} \theta_0 \\ \vdots \\ \theta_{\ell} \end{pmatrix}$ et $U(u) = \begin{pmatrix} 1 \\ u \\ \vdots \\ u^{\ell}/\ell\,! \end{pmatrix}$.

Définition 2.2 (Estimateur localement polynomial de degré ℓ). Soit K un noyau, alors $\hat{\theta}_n(x) \in \mathbb{R}^{\ell+1}$ est appelé *estimateur localement polynomial de degré* ℓ de $\theta(x)$ si

$$\hat{\theta}_n(x) = \arg \min_{\theta \in \mathbb{R}^{\ell+1}} \sum_{i=1}^{n} \left(Y_i - \theta^\top U \left(\frac{X_i - x}{h} \right) \right)^2 K \left(\frac{X_i - x}{h} \right)$$

La statistique $f_n^{\mathrm{LP}} = \hat{\theta}_n(x)^\top \times U(0)$ est l'estimateur localement polynomial de degré ℓ de $f(x)$.

On peut réécrire $\hat{\theta}_n(x)$ de la manière suivante

$$\begin{aligned}
\hat{\theta}_n(x) &= \arg \min_{\theta \in \mathbb{R}^{\ell+1}} -2\theta^\top \underbrace{\sum_{i=1}^{n} Y_i U \left(\frac{X_i - x}{h} \right) K \left(\frac{X_i - x}{h} \right)}_{\mathfrak{A}_\mathfrak{n}} + \sum_{i=1}^{n} \left(\theta^\top U \left(\frac{X_i - x}{h} \right) \right)^2 K \left(\frac{X_i - x}{h} \right) \\
&= \arg \min_{\theta \in \mathbb{R}^{\ell+1}} -2\theta^\top \mathfrak{A}_\mathfrak{n} + \theta^\top \underbrace{\sum_{i=1}^{n} U \left(\frac{X_i - x}{h} \right) U^\top \left(\frac{X_i - x}{h} \right) K \left(\frac{X_i - x}{h} \right)}_{\mathfrak{B}_\mathfrak{n}} \theta \\
&= \arg \min\nolimits_{\theta \in \mathbb{R}^{\ell+1}} -2\mathfrak{A}_\mathfrak{n} + \theta^\top \mathfrak{B}_\mathfrak{n} \theta
\end{aligned}$$

Théorème 4. *Si $\mathfrak{B}_n(x) > 0$, alors il existe un unique $\hat{\theta}_n(x)$, estimateur par polynômes locaux de degré ℓ et $\hat{\theta}_n(x) = \mathfrak{B}_\mathfrak{n}(x)^{-1} \mathfrak{A}_\mathfrak{n}(x)$.*

Démonstration. $\hat{\theta}_n$ est solution de $-2\mathfrak{A}_\mathfrak{n}(x) + 2\mathfrak{B}_\mathfrak{n}(x)\theta = 0$ $\square$

Remarque. $\hat{f}_n^{\mathrm{PL}}$ ne souffre pas d'effets de bords, alors que $\hat{f}^{\mathrm{NW}}(x)$ si.

Remarque. L'estimateur $\hat{f}_n^{\mathrm{LP}}(x)$ est linéaire :

$$\begin{aligned}
\hat{f}_n^{\mathrm{PL}}(x) &= \hat{\theta}_n^\top(x) U(0) \\
&= \Big((\mathfrak{B}_\mathfrak{n}(x))^{-1} \mathfrak{A}_\mathfrak{n}(x) \Big)^\top U(0) \\
&= \sum_{i=1}^{n} Y_i \underbrace{U^\top \left(\frac{X_i - x}{h} \right) K \left(\frac{X_i - x}{h} \right) \mathfrak{B}_\mathfrak{n}{}^{-1}(x) U(0)}_{W_{ni}(x)} \\
&= \sum_{i=1}^{n} Y_i W_{ni}(x)
\end{aligned}$$

Donc $\hat{f}_n^{\mathrm{LP}}(x)$ est linéaire.

Théorème 5 (Les polynômes locaux reproduisent exactement les polynômes de même degré). *Si $Y_i = P(X_i)$ avec $P \in \mathbb{R}^{[\ell]}(X)$ et si $\mathfrak{B}_\mathfrak{n}(x) > 0$, alors pour tout x, $\hat{f}_n^{\mathrm{LP}}(x) = P(x) \Leftrightarrow \sum_{i=1}^n P(X_i) W_{ni}(x) = P(x)$.*

En particulier, $P(x) \equiv c \Rightarrow \sum_{i=1}^n W_{ni}(x) = 1$ et si $P(u) = (u - x)^k$ alors $\sum_{i=1}^n (X_i - x)^k W_{ni}(x) = 0$, quelque soit $k = 1, ..., \ell$.

Démonstration. Si $P \in \mathbb{R}^{[\ell]}(X)$, alors

$$\begin{aligned}
P(X_i) &= P(x) + P'(x) h \frac{X_i - x}{h} + \cdots + \frac{P^{(\ell)}(x) h^\ell}{\ell\,!} \left(\frac{X_i - x}{h} \right)^\ell \\
&= p^\top(x) U \left(\frac{X_i - x}{h} \right)
\end{aligned}$$

avec $p(x) = \begin{pmatrix} P(x) \\ \vdots \\ P^{(\ell)}(x)h^{\ell} \end{pmatrix}$

$$\begin{aligned} \sum_{i=1}^{n}\left(Y_i - \theta^\top U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right) &= \sum_{i=1}^{n}\left((p(x)-\theta)^\top U\left(\frac{X_i - x}{h}\right)\right)^2 K\left(\frac{X_i - x}{h}\right) \\ &= (p(x)-\theta)^\top \mathfrak{B}_n (p(x)-\theta). \end{aligned}$$

Ceci est minimisé en $\hat{\theta}_n = p$. Donc $\hat{f}_n^{\mathrm{PL}}(x) = P(x)$. □

2.1.3 Estimateur "spline" de la régression

On suppose $0 < X_1 < \ldots < X_n < 1$ et $Y_i = f(X_i) + \xi_i$ (design fixe).

En 1964, Schoenberg introduit les **splines d'interpolation** : le but est de trouver $g : [0,1] \to \mathbb{R}$ tel que $g(X_i) = Y_i$ et qui minimise $\int_0^1 \left(g''(x)\right)^2 dx$. On peut réécrire le problème sous la forme suivante

$$\min_{\substack{g \in \mathcal{C}^2([0,1]) \\ g(X_i)=Y_i,\, i=1,\ldots,n}} \int \left(g''(x)\right)^2 dx$$

Dans le cas des **splines d'approximation**, on cherche à résoudre le problème suivant

$$\min_{\substack{g \in \mathcal{C}^2([0,1]) \\ \sum_{i=1}^n (g(X_i)-Y_i)^2 \leq \varepsilon}} \int \left(g''(x)\right)^2 dx$$

Cette formulation est équivalente à la formulation suivante

$$\min_{g \in \mathcal{C}^2([0,1])} \sum_{i=1}^{n} \left(g(X_i) - Y_i\right)^2 + \lambda \int \left(g''(x)\right)^2 dx$$

c-à-d le critère des moindres carrés pénalisé.

Définition 2.3 (Estimateur spline de degré $2\ell - 1$)**.** Soit ℓ dans $\mathbb{N}^*$. On appelle *estimateur spline de degré* $2\ell - 1$ la solution du problème

$$\min_{\substack{g:\ \ell \text{ fois dérivable} \\ \text{sur } [0,1]}} \sum_{i=1}^{n} \left(g(X_i) - Y_i\right)^2 + \lambda \int \left(g^{(l)}(x)\right)^2 dx$$

Remarque. Si $\ell = 1$, on parle de *spline de degré 1*

Remarque. Si $\ell = 2$, on parle de *spline cubique*

Théorème 6. *Si $0 < X_1 < \ldots < X_n < 1$ et $n \geq \ell$, alors il existe un unique $\hat{f}_n^S(x)$, estimateur spline de degré $2\ell - 1$ de $f(x)$.*

Définition 2.4. Soit $0 < X_1 < \ldots < X_n < 1$ (les X_i sont appelés les *nœuds*) et $n \geq \ell$. S_n, espace de splines de degré $2\ell - 1$, est l'espace des fonctions $s : [0,1] \to \mathbb{R}$ telles que

1. $s^{(2\ell-2)}$ est continue sur [0,1].
2. $s^{(2\ell-1)}$ est continue par morceaux $(0, X_1), (X_1, X_2), \ldots, (X_n, 1)$.
3. $s^{(j)}(0) = s^{(j)}(1) = 0$, $j = \ell, \ldots, 2\ell - 1$.

Proposition 7. *$\hat{f}_n^S(x)$ est un estimateur linéaire de $f(x)$ qui reproduit exactement les polynômes de degré $\mathbb{R}^{[\ell-1]}(X)$. Autrement dit, si $Y_i = P(X_i)$ avec P un polynôme de degré inférieur ou égal à $\ell - 1$, $\hat{f}_n^S = P$.*

2.1.4 Estimateurs par projection de la régression f

On se place dans le cadre d'un design fixe sur $[0,1]$.

Si $f \in L_2[0,1]$, on choisit une base orthonormée $\{\varphi_j\}_{j\geq 1}$ de $L_2[0,1]$ et on projette $f(x) = \sum_{j\geq 1} \theta_j \varphi_j(x)$ où $\theta_j = \int_0^1 f\varphi_j$.

On approxime f par $f_M(x) = \sum_{j=1}^M \theta_j \varphi_j(x)$ et on estime θ_j par $\hat{\theta}_j$, $j = 1, \ldots, M$ pour obtenir $\hat{f}_M(x) = \sum_{j=1}^M \hat{\theta}_j \varphi_j(x)$.

Exemple 2.1 (Base trigonométrique)**.** On prend la famille des $\{\varphi_j\}_{j\geq 1}$ définie par

$$\begin{cases} \varphi_1(x) \equiv 1 \\ \varphi_{2k}(x) = \sqrt{2}\cos(2\pi kx), k \in \mathbb{N}^* \\ \varphi_{2k+1}(x) = \sqrt{2}\sin(2\pi kx) \end{cases}$$

Exemple 2.2 (Base d'ondelettes)**.** On peut prendre une base d'ondelettes comme par exemple la base de Haar.

On travaille avec le modèle approximé :

$$Y_i = f_M(X_i) + \xi_i = \sum_{j=1}^M \varphi_j(X_i)\theta_j + \xi_i \tag{7}$$

avec les X_i fixés.

On pose

$$y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \xi = \begin{pmatrix} \xi_1 \\ \vdots \\ \xi_n \end{pmatrix} \quad \mathbb{X} = \begin{pmatrix} \varphi_1(X_1) & \ldots & \varphi_M(X_1) \\ \vdots & \ddots & \vdots \\ \varphi_1(X_n) & \ldots & \varphi_M(X_n) \end{pmatrix} \quad \theta = \begin{pmatrix} \theta_1 \\ \vdots \\ \theta_M \end{pmatrix}$$

Le modèle (7) s'écrit donc

$$y = \mathbb{X}\theta + \xi.$$

L'estimateur des moindres carrés de θ et de $f(x)$ s'écrivent respectivement

$$\hat{\theta}_{\mathrm{MC}} = (\mathbb{X}^\top \mathbb{X})^{-1}\mathbb{X}^\top y \quad \text{si } \mathbb{X}^\top\mathbb{X} \text{ est inversible}$$

et

$$\hat{f}_n = \big(\varphi_1(x), \ldots, \varphi_M(x)\big) \cdot \hat{\theta}_{\mathrm{MC}} = \varphi^\top(x)\hat{\theta}_{\mathrm{MC}}$$

Proposition 8. *$\hat{f}_n = \hat{f}_n^{\mathrm{MC}}$ est un estimateur linéaire.*

$$\hat{f}_n^{\mathrm{MC}}(x) = \varphi^\top(x) \times (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top y$$

Remarque. M est aussi un paramètre de lissage. Il joue le même rôle que $1/h$.

Définition 2.5 (Risque en moyenne quadratique)**.** Le risque en moyenne quadratique (MISE) de $\hat{f}_n$ (en

MC) de $f(x)$ est

$$
\begin{aligned}
R(\hat{f}_n, f) &= \mathbb{E}\Big[\int \Big(\hat{f}_n(x) - f(x)\Big)^2 dx\Big] \\
&= \underbrace{\mathbb{E}\Big[\int \Big(\hat{f}_n(x) - f_M(x)\Big)^2 dx\Big]}_{\mathbb{E}\left[\int \left(\sum_{j=1}^{M} (\hat{\theta}_j - \theta_j)\varphi_j(x)\right)^2 dx\right]} + \underbrace{\int \Big(f_M(x) - f(x)\Big)^2 dx}_{\int \left(\sum_{j\geq M} \theta_j \varphi_j(x)\right)^2 dx} \\
&= \mathbb{E}\Big[\int \Big(\sum_{j=1}^{M} (\hat{\theta}_j - \theta_j)\varphi_j(x)\Big)^2 dx\Big] + \sum_{j\geq M} \theta_j^2 \\
&= \mathbb{E}\Big[\sum_{j=1}^{M} \big(\hat{\theta}_j - \theta_j\big)^2\Big] + \sum_{j\geq M} \theta_j^2 \\
&= \mathbb{E}\Big[\big(\hat{\theta} - \theta\big)^\top \big(\hat{\theta} - \theta\big)\Big] + \sum_{j\geq M} \theta_j^2 \\
&= \mathbb{E}\Big[\mathrm{tr}\Big(\big(\hat{\theta} - \theta\big)^\top \big(\hat{\theta} - \theta\big)\Big)\Big] + \sum_{j\geq M} \theta_j^2 \\
&= \mathbb{E}\Big[\mathrm{tr}\Big(\big(\hat{\theta} - \theta\big)\big(\hat{\theta} - \theta\big)^\top\Big)\Big] + \sum_{j\geq M} \theta_j^2 \\
&= \mathrm{tr}\big(\mathrm{Var}(\hat{\theta})\big) + \sum_{j\geq M} \theta_j^2
\end{aligned}
$$

Remarque. $\mathbb{E}\big[\hat{f}_n(x)\big] = \varphi^\top(x)\,\mathbb{E}[\hat{\theta}_{\mathrm{MC}}] = \varphi^\top(x)\theta = f_M(x)$.

2.2 Propriétés statistiques des estimateurs linéaires

Si, plus généralement, $\hat{\theta} = Sy$ est un estimateur linéaire, alors $\mathrm{Var}\,\hat{\theta} = \mathrm{Var}(Sy) = Var(S\xi)$. Ainsi

$$
\begin{aligned}
\mathrm{Var}(\hat{\theta}) &= \mathrm{tr}\Big(\mathbb{E}\big[S\xi\xi^\top S^\top\big]\Big) \\
&= \mathrm{tr}\Big(S\,\mathbb{E}(\xi\xi^\top)S^\top\Big) \\
&= \mathrm{tr}\left(S\begin{pmatrix}\sigma^2 & & \\ & \ddots & \\ & & \sigma^2\end{pmatrix}S^\top\right) \qquad \text{si les } \xi_i \text{ sont indépendants et de variance } \mathrm{Var}(\xi_i) = \sigma^2 > 0 \\
&= \mathrm{tr}(\sigma^2 S S^\top) \\
&= \sigma^2\,\mathrm{tr}(S^\top S)
\end{aligned}
$$

En particulier, pour $\hat{\theta}_{\mathrm{MC}}$, $S = (\mathbb{X}^\top\mathbb{X})^{-1}\mathbb{X}^\top$.

$$
\begin{aligned}
[\mathbb{X}^\top\mathbb{X}]_{k,j} &= \sum_{i=1}^{n} \varphi_j(X_i)\varphi_k(X_i) \\
\underset{X_i = i/n}{} &= n\frac{1}{n}\sum_{i=1}^{n} \varphi_j\left(\frac{i}{n}\right)\varphi_k\left(\frac{i}{n}\right)
\end{aligned}
$$

En prenant la base trigonométrique, qui vérifie la propriété

$$
\frac{1}{n}\sum_{i=1}^{n} \varphi_j\left(\frac{i}{n}\right)\varphi_k\left(\frac{i}{n}\right) = \delta_{jk}
$$

D'où

$$\mathbb{X}^\top \mathbb{X} = nI_n$$

Puis

$$\sigma^2 \operatorname{tr}(S^\top S) = M\frac{\sigma^2}{n}$$

Théorème 9. *Si $y = \mathbb{X}\theta + \xi$ et $\hat{f} = Sy$ est un estimateur linéaire de f, alors*

$$\mathbb{E}\left[\frac{\|\hat{f} - f\|_2^2}{n}\right] = \frac{1}{n}\|Sf - f\|_2^2 + \frac{\sigma^2}{n}\operatorname{tr}(S^\top S)$$

Démonstration.

$$\|\hat{f} - f\|_2^2 = \|Sy - f\|_2^2 = \|Sf + S\xi - f\|_2^2$$

$$\mathbb{E}\big[\|\hat{f} - f\|_2^2\big] = \|Sf - f\|_2^2 + \mathbb{E}\big[\|S\xi\|_2^2\big] = \|Sf - f\|_2^2 + \sigma^2 \operatorname{tr}(S^\top S)$$

□

Remarque. Si $X_i = i/n$, alors $(1/n)\|\hat{f} - f\|_2^2 = (1/n)\sum_{i=1}^n \big(\hat{f} - f\big)^2(i/n) \approx \int \big(\hat{f} - f\big)^2(x)\,dx$.

Théorème

$$\frac{1}{n}E||\hat{f} - f||^2 \leq \min_{\theta\in\mathbb{R}^M} \frac{1}{n}||f_\theta - f||^2 + \frac{\sigma^2}{n}Tr(S^TS)$$

avec $Tr(S^TS) = \min(M, n)$

Application :

Estimation non paramétrique de la régression, à design régulier : $X_i = \dfrac{i}{n}$, $\{\xi_i\}_i$ indépendants, $(0, \sigma^2)$

$$Y_i = f(\frac{i}{n}) + \xi_i \Leftrightarrow y = f + \xi$$

On suppose $f \in \mathbb{L}_2[0,1]$. On fixe $\{\varphi_j\}_{j\geq 1}$ la base orthonormée trigonométrique. Pour évaluer le biais, on suppose $f \in W(\beta, L), \beta \in \mathbb{N}^*, L > 0$ classe de Sobolev périodique, donc $f \in \mathcal{C}^{(\beta)}$, $||f^{(\beta)}||_2 \leq L$ tq $f^{(j)}(0) = f^{(j)}(1)$ pour tout $j = 0, ..., \beta - 1$.

On note $\theta_j = \int_0^1 f(x)\varphi_j(x)dx,\ f(x) = \sum_{j\geq 1}\theta_j\varphi_j(x)$.

$$f \in \mathbb{L}_2[0,1] \Leftrightarrow \{\theta_j\}_j \in l_2(\mathbb{N}^*)\ (\sum_{j\geq 1}\theta_j^2 < \infty)$$

Propriété La fonction $f \in W(\beta, L)$ pour $\beta \in \mathbb{N}^*, L > 0$ ssi $\{\theta_j\}_{j\geq 1} \in w(\beta, L)$ défini par $\{\theta_j\}_{j\geq 1} \in l_2(\mathbb{N}^*$ tq

$$\sum_{j\geq 1}\theta_j^2 a_j^2 \leq \frac{L^2}{\pi^{2\beta}},\ \text{où } a_j = \begin{cases} j^\beta,\ j \text{ pair} \\ (j-1)^\beta,\ j \text{ impair}\end{cases}$$

Si β est grand, θ_j tend vers 0 plus vite quand j tend vers l'infini.

Propriété

$$\frac{1}{n}\sum_{i=1}^n \varphi_j(\frac{i}{n}).\varphi_k(\frac{i}{n}) = \begin{cases} 1,\ j = k \\ 0,\ j \neq k\end{cases} = \delta_{jk}$$

($\{\varphi_1, ..., \varphi_{n-1}\}$ sont orthonormées pour $||\cdot||_n = \frac{1}{n}\|\cdot\|_2$)

Remarque : $w(\beta, L)$ peut être défini pour $\beta > 0$, donc aussi $W(\beta, L)$, mais si $\beta \leq \dfrac{1}{2}$ alors $W(\beta, L)$ contient des fonctions pas continues.

Propriété Si $f \in W(\beta, L)$, $\beta \in \mathbb{N}^*$, $L > 0, X_i = \dfrac{i}{n}$, et si $\{\varphi_j\}_{j\geq 1}$ base trigonométrique, alors pour tout $n \geq 1$ et $M \geq 1$:

$$\inf_{\theta\in\mathbb{R}^M} ||f_\theta - f||_n^2 \leq C(\beta, L)(M^{-2\beta} + \frac{1}{n})$$

avec $C(\beta, L) > 0$ et $f_\theta(x) = \sum_j^M \theta_j\varphi_j(x)$.

Preuve

$$f_\theta(x) - f(x) = -\sum_{j\geq M+1} \theta_j\varphi_j(x)$$

On a $\sum_{j\geq 1} |\theta_j| < \infty$ car pour $2\beta < 1$:

$$\sum_{j\geq 1} |\theta_j| a_j.\frac{1}{a_j} = \sum_{j\geq 1} A_j.B_j \leq (\sum_{j\geq 1} \theta_j^2 a_j^2)^{\frac{1}{2}}.(\sum_{j\geq 1} \frac{1}{a_j^2})^{\frac{1}{2}} \leq \frac{L}{\pi^\beta} \times C_\beta$$

$$|\sum_{j\geq M+1} \theta_j\varphi_j(x)|$$

$$\begin{aligned}
||f_\theta - f||_n^2 &= \frac{1}{n}\sum_{i=1}^n (\sum_{j\geq M+1} \theta_j\varphi_j(\frac{i}{n}))(\sum_{k\geq M+1} \theta_k\varphi_k(\frac{i}{n})) \\
&= \frac{1}{n}\sum_{i=1}^n [(\sum_{j=M+1}^{n-1}\sum_{k=M+1}^{n-1} \theta_j\theta_k\varphi_j(\frac{i}{n})\varphi_k(\frac{i}{n}) \\
&+ \sum_{j\geq n}\sum_{k=M+1}^{n-1} \theta_j\theta_k\varphi_j(\frac{i}{n})\varphi_k(\frac{i}{n}) \\
&+ \sum_{j=M+1}^{n-1}\sum_{k\geq n} \theta_j\theta_k\varphi_j(\frac{i}{n})\varphi_k(\frac{i}{n}) \\
&+ \sum_{j\geq n}\sum_{k\geq n} \theta_j\theta_k\varphi_j(\frac{i}{n})\varphi_k(\frac{i}{n})] \\
&= \sum_{M+1}^{n-1} |\theta_j|^2 + T_1 + T_2 + T_3
\end{aligned}$$

Pour T_3, $|\varphi_j \leq \sqrt{2}$, $\forall x$:

$$\begin{aligned}
|T_3| &\leq \sum_{j\geq n}\sum_{k\geq n} |\theta_j||\theta_k| 2 = 2(\sum_{j\geq n} |\theta_j|)^2 \\
&\leq 2\sum_{j\geq n} \theta_j^2 a_j^2 \times \sum_{j\geq n} \frac{1}{a_j^2} \leq C_1.n^{-2\beta+1} \leq \frac{C_1}{n}
\end{aligned}$$

avec $\sum_{j\geq n} \theta_j^2 a_j^2 < \infty$, $a_j \sim j^\beta$ or $\beta \geq 1$, donc $\dfrac{1}{n^{2\beta-1}} \leq \dfrac{1}{n}$.

Or $M^{-2\beta}.\sum_{j\geq M+1}^{n-1} |\theta_j|^2.M^{2\beta} \leq M^{-2\beta}\sum_{j\geq M+1} \theta_j^2 a_j^2 \leq \dfrac{L^2}{\pi^{2\beta}} M^{-2\beta}$ et $a_j^2 \geq (a_M)^2 \approx M^{2\beta}$. La démonstration pour T_1 et T_2 est laissée en exercice.

Théorème Si $f \in W(\beta, L), \beta \in \mathbb{N}^*, L > 0$, soit $M = [c.n^{\frac{1}{2\beta+1}}]$ avec $c > 0$ une constante, alors :

$$\sup_{f \in W(\beta,L)} E(||\hat{f}^{MC} - f||_n^2) \leq C(\beta, L) \times n^{-\frac{2\beta}{2\beta+1}}$$

Remarque : cette vitesse est optimale.

Remarque : l'estimateur $\hat{f}^{MC}$ dépend de β, par le choix de M. On voudra un estimateur adaptatif en β, qui atteint la même vitesse.

Preuve Nous avons déjà vu que :

$$E||\hat{f}^{MC} - f||_n^2 \leq \inf_{\theta \in \mathbb{R}^M} ||f_\theta - f||_n^2 + \frac{\sigma^2}{n}\{M \wedge n\}.$$

Par la propriété on obtient

$$\leq C_1(\beta, L)(M^{-2\beta} + \frac{1}{n}) + \frac{\sigma^2 M}{n} \leq C(\beta, L) n^{-\dfrac{2\beta}{2\beta+1}}$$

pour $M \sim n^{\frac{1}{2\beta+1}}$.

2.3 Estimation adaptative (à la régularité de β) de la régression par estimation sous biais du risque

Définition 1. L'**ERM** est l'estimateur qui minimise le risque empirique. Pour déterminer cet estimateur, on définit la **norme empirique** $||.||_n$ telle que $||f||_n = \frac{1}{n} f^t f$. Pour un estimateur $\hat{f}$, qui se définit comme la projection de Y sur un certain espace issu des données ($\hat{f} = Sy$)

$$\mathbf{E}(||\hat{f} - f||_n^2) = \mathbf{E}(||\hat{f}||_n^2) - \frac{2}{n} f^t \mathbf{E}(\hat{f}) + ||f||_n^2 \tag{8}$$

On écrit $J \hat{=} \mathbf{E}(||\hat{f}||_n^2) - \frac{2}{n} f^t \mathbf{E}(\hat{f})$ la partie de l'erreur qui dépend du choix de l'estimateur. Pour un estimateur lié à un espace projecteur S, on note $\hat{J}(S)$ l'estimation de l'erreur empirique. On définit la règle de sélection dans un ensemble de modèle $\mathcal{S}$:

$$\hat{S} \in \arg\min_{S \in \mathcal{S}} \quad \hat{J}(S) \tag{9}$$

— Première partie de $\hat{J}$: on estime $\mathbf{E}(||\hat{f}||_n^2)$ par $||\hat{f}||_n^2$
— Deuxième partie de $\hat{J}$:

$$\begin{aligned} f^t \mathbf{E}(\hat{f}) &= f^t \mathbf{E}(S(f + \xi)) \\ &= f^t(Sf + S\mathbf{E}(\xi)) \\ &= f^t Sf \text{car } \xi \text{est centrée} \end{aligned}$$

On se pose la question.. Est-ce que $y^t Sy$ est un estimateur sans biais de $f^t Sf$? (Non.)

$$\begin{aligned} \mathbf{E}(y^t Sy) &= \mathbf{E}[(f + \xi)S(f + \xi)] \\ &= f^t Sf + \mathbf{E}(\xi^t Sf + f^t S\xi + \xi^t S\xi) \\ &= f^t Sf + \mathbf{E}(\xi^t S\xi) \text{ ce terme quadratique lié à l'erreur va créer un biais} \\ &= f^t Sf + \mathbf{E}(tr(\xi^t S\xi))) \text{ car pour un nombre réel } x,\ tr(x) = x \\ &= f^t Sf + \mathbf{E}(Str(\xi\xi^t) \text{ car } tr(AB) = tr(BA) \\ &= f^t Sf + \sigma^2 tr(S) \end{aligned}$$

En conclusion, $y^t Sy - \sigma^2 tr(S)$ est un estimateur sans biais de la deuxième partie de $\hat{J}(S)$

$$\begin{aligned}
\hat{J}(S) &= ||\hat{f}||_n^2 - \frac{2}{n}(y^t S y - \sigma^2 tr(S)) \\
&= \frac{1}{n}[||Sy||^2 - 2y^t S y] + \frac{2\sigma^2}{n} tr(S) \text{ on rappelle encore une fois que } \hat{f} = Sy \\
&= ||\hat{f} - y||_n^2 - ||y||_n^2 + \frac{2\sigma^2}{n} tr(S)
\end{aligned}$$

Notre problème (2) se ramène donc à :

$$\underset{S\in\mathcal{S}}{\arg\min}||\hat{f}-y||_n^2+\frac{2\sigma^2}{n}tr(S) \tag{10}$$

En particulier, si S correspond à un projecteur sur un espace de dimension p, $tr(S)=p$ et on retrouve le C_p-Mallows (qui fait référence à Colin Lingwood Mallows).

Théorème 1. (Kneip 1994) Inégalité Oracle pour la méthode C_p-Mallows. Sous les hypothèses du modèle de régression linéaire gaussien (c'est-à-dire $\xi_i \overset{iid}{\hookrightarrow} \mathcal{N}(0,\sigma^2)$), si classe $\mathcal{S}$ vérifie :

— $\forall S\in\mathcal{S},\quad 0\preccurlyeq \mathcal{S}\preccurlyeq I_n$. Id est $\forall v\in\mathbf{R}^n,\quad 0\leq v^tSv\leq v^tv$.

— $\forall(S,S')\in\mathcal{S}^2,\quad SS'=S'S$.

— $\forall(S,S')\in\mathcal{S}$, soit $S\preccurlyeq S'$ soit $S'\preccurlyeq S$.

Alors $\exists C<+\infty$, une constante, telle que quelque ce soit $\epsilon>0$, l'estimateur sélectionné $\hat{f}_{\hat{S}}$ vérifie :

$$\mathbf{E}(||\hat{f}_S-f||^2)\leq(1+\epsilon)\min_{S\in\mathcal{S}}\mathbf{E}(||\hat{f}_S-f||^2)+\frac{C}{\epsilon n} \tag{11}$$

Remarque 1. Une telle inégalité s'appelle une **inégalité d'oracle**. Elle lie le risque d'un estimateur sélectionné à celui du meilleur possible (celui que choisit "l'oracle" qui peut voir le signal non bruité) à une constante multiplicative près $(1+\epsilon)$ et un terme d'erreur $\frac{C}{\epsilon n}$

Remarque 2. ϵ joue un rôle inverse dans les deux termes de la majoration. D'une part, il tend à donner une inégalité d'oracle exacte, si $\epsilon=0$, d'autre part il rajoute une erreur infinie si $\epsilon\longrightarrow 0$).

Remarque 3. A titre de comparaison, pour un estimateur Lasso, ce risque est borné par $\log(M)\frac{\sigma^2 S}{M}$ où S est la sparcité du problème (c'est-à-dire le nombre de dimensions pertinentes dans le signal f) et M le nombre de paramètres considérés. Pour l'estimateur SLOPE, l'erreur est bornée par $\frac{\sigma^2 S}{M}\log(\frac{eM}{S})$.

Application du théorème de Kneip à l'estimation des moindres carrés.

On prend $X=X_M=((\phi_j(X_i))_{1\leq i\leq n,1\leq j\leq m})$ comme matrice de données. On sait que $\hat{\theta}^{MC}=(X^tX)^{-1}X^tY$. On pose alors la classe de modèles $\mathcal{S}=\{S_M=X_M(X_M^tX_M)^{-1}X_M^t|M\in 1,...,n\}$, c'est-à-dire qu'on considère la décomposition dans une base orthogonale fonctionnelle, et on sélectionne à quel rang on tronque la décomposition.

Proposition La classe $\mathcal{S}$ vérifie les trois hypothèses du théorème de Kneip. En effet,

1. $\forall M\in\{1,...n\}, S_M$ est une projection orthogonale. Donc, $S_M^2=S_M$ et les valeurs propres de S_M sont donc des 0 et des 1.
2. $\forall M_1,M_2\in\{1,...,n\}^2$, $S_{M_1}S_{M_2}=S_{\min(M_1,M_2)}=S_{M_2}S_{M_1}$.
3. Si $M_1\geq M_2$ (resp. $M_2\geq M_1$) , alors $S_{M_1}-S_{M_2}\preccurlyeq 0$ (resp. $S_{M_2}-S_{M_1}\preccurlyeq 0$).

Preuve 3. Si $M_1\geq M_2\Rightarrow S_{M_1}\geq S_{M_2}$ (id est valeurs propres de $S_{M_1}-S_{M_2}$ positives)
$||S_{M_1}u-u||_2^2=$ (par déf) $\min_{v\in V_{M_1}}||v-u||_2^2\leq\min_{v\in V_{M_2}}||v-u||_2^2=||S_{M_2}u-u||_2^2=||I-S_{M_2}u||_2^2$
$\Rightarrow I-S_{M_1}\leq I-S_{M_2}\Rightarrow S_{M_1}\geq S_{M_2}\quad\square$

Donc, si on se place dans le cas où $X^tX=I_n$ ($X_i=\frac{i}{n}\quad\forall i\in\{1,...,n\}$ par exemple).

En posant $\hat{f}_M(x)=\hat{\theta}_M^t\phi(x)$ avec $[\hat{\theta}_M]_j=[\mathbf{X}^ty]_j=\frac{1}{n}\sum_{i=1}^n y_i\phi_j(X_i)$,

$$\underset{f\in W(\beta,L)}{sup}\mathbf{E}(||\hat{f}_{\hat{S}}-f||^2)\leq C(\beta,L)n^{\frac{-2\beta}{2\beta+1}}\quad\forall\beta\geq 1,\quad\forall L>0 \tag{12}$$

On rappelle que $\hat{S}=\underset{S\in\mathcal{S}}{\arg\inf}[\frac{1}{n}\sum_{i=1}^n(Y_i-\hat{f}_S)^2+\frac{2\sigma^2}{n}tr(S)]$.

Corollaire Dans le modèle de régression $y=f+\xi$, avec les ξ iid, de loi $\mathcal{N}(0,\sigma^2)$, l'estimateur $\tilde{f}=\hat{S}\cdot y$ est tel que

$$\sup_{f\in W_{per}(\beta,L)}E_f(||\tilde{f}-f||_n^2)\leq C(\beta,L)n^{-\frac{2\beta}{2\beta+1}},\beta\geq 1,L>0.$$

On applique le théorème de Kneip avec $\epsilon = 1$, en utilisant la décomposition biais/variance de l'erreur vue précédemment dans le cours.

$$\begin{aligned}\mathbf{E}(||\hat{f}_{\hat{S}} - f||^2) &\leq 2 \min_{M \in \{1,..,n\}} (\overline{C}(\beta, L) M^{-2\beta} + \frac{\tilde{C}(\beta, L) M}{n}) + \frac{C}{2n} \\ &\leq C(\beta, L) n^{\frac{-2\beta}{-2\beta+1}}\end{aligned}$$

Pour $M* = c_0 n^{\frac{1}{2\beta+1}} \leq n$, car $\frac{C}{2n} << n^{\frac{-2\beta}{2\beta+1}}$.

Remarque 4. $\hat{f}_{\hat{S}}$ est un estimateur **adaptatif optimal**. (C'est-à-dire qu'il atteint la vitesse optimale sans hypothèse sur la sparcité). Mais, si on ne s'intéresse plus au MISE mais au $MSE(x)$ (le risque évalué en x), et si $f \in W(\beta, L)$ alors il n'existe pas d'estimateur f adaptatif optimal, mais des estimateurs adaptatifs à la vitesse $(\frac{log(n)}{n})^{\frac{2\beta}{2\beta+1}}$

2.4 Estimation de fonctionnelles de $f \in \mathcal{F}$ (estimation de $\Phi(f)$ où $\Phi : \mathcal{F} \longrightarrow \mathbf{R}$)

Exemples :

— $D = \int(f^2)$. Cela peut correspondre au cas où on veut tester dans le modèle de régression sur $\mathcal{F}(\beta, L)$ avec H_0 :"$f = 0$ sur $[0, 1]$" contre H_1 : " $||f||_2^2 \geq \phi^2$. On a alors besoin d'un estimateur de $||f||_2^2$. On pourrait bien sûr choisir $||\hat{f}||_2^2$ mais ce choix est en réalité sous-optimal.

— Excess-mass : pour f une densité de probabilité et $\lambda > 0$,

$$EM(\lambda) = \int_{u|f(u)\geq\lambda} (f(u) - \lambda) du$$

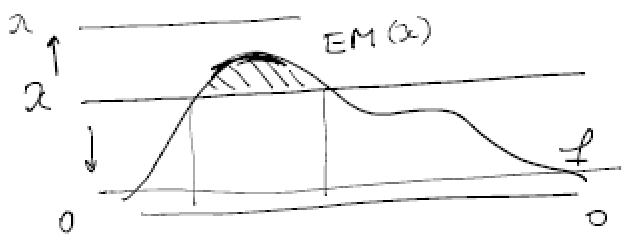

FIGURE 1 – Excess-Mass

— $||f||_p^p$ pour $p \geq 1$

Estimation de D pour un modèle de bruit blanc gaussien ou de suite gaussienne

On se place dans le modèle de bruit blanc gaussien,

$$dY_t = f(t)dt + \epsilon dW_t, \quad t \in [0, 1], \quad (\epsilon = \frac{1}{\sqrt{n}} \longrightarrow 0) \tag{13}$$

$f \in \mathbf{L}_2([0, 1])$ est le drift , W_t est un mouvement Brownien sur $[0, 1]$. Si $(\phi_j)_{j\geq1}$ est une b.o.n de $\mathbf{L}_2([0, 1])$, on a, pour $j, k \in \{1, ..., n\}$,

— $y_j \hat{=} \int \phi_j(t) dY_t$

— $\theta_j \hat{=} \int \phi_j(t) f(t) dt$

— $\eta_j \hat{=} \int \phi_j(t) dW_t \hookrightarrow \mathcal{N}(0, 1)$ et $Cov(\eta_j, \eta_k) = \delta_{j,k}$

Ou dans le modèle de suite gaussienne,

$$y_j = \theta_j + \epsilon\eta_j, \quad \eta_j \overset{\text{iid}}{\hookrightarrow} \mathcal{N}(0, 1). \tag{14}$$

Dans ce cas, $D = \int f^2 = \sum_{j\geq 1} \theta_j^2$.

Ces deux modèles sont équivalent et s'étudient de manière similaire, mais nous allons nous concentrer sur le modèle de suite gaussienne.

On suppose que $f \in W(\beta, \overline{L})$ avec $\beta > 0$. De manière équivalente, $\theta = (\theta_j)_{j\geq 1} \in w(\beta, L)$ c'est-à-dire, $\sum_{j\geq 1} j^{2\beta}\theta_j^2 \leq L^2$.

Un estimateur de D est :

$$\hat{D}_N = \sum_{j=1}^{N}(y_j^2 - \epsilon^2)$$

On détermine son risque quadratique en séparant le calcul du biais et de la variance.

— Biais :

$$\begin{aligned}
\mathbf{E}_\theta(\hat{D}_N) - D &= \sum_{j=1}^{N} \mathbf{E}_\theta(y_j^2 - \epsilon^2) - \sum_{j\geq 1}\theta_j^2 \\
&= \sum_{j=1}^{N}\theta_j^2 - \sum_{j\geq 1}\theta_j^2 \\
&= -\sum_{j\geq N+1}\theta_j^2
\end{aligned}$$

En utilisant les hypothèses sur θ, on conclut donc que :

$$\begin{aligned}
|b_\theta(\hat{D}_N)| &= \sum_{j\geq N+1}\theta_j^2 \times 1 \\
&\leq \sum_{j\geq N+1}\theta_j^2 \frac{j^2\beta}{N^{2\beta}} \\
&\leq L^2 N^{-2\beta}
\end{aligned}$$

— Variance :

$$
\begin{aligned}
\mathbf{V}_\theta(\hat{D}_N) &= \sum_{j=1}^{N} \mathbf{V}_{\theta_j}(y_j^2) \\
&= \sum_{j=1}^{N} \mathbf{E}(Y_j^4) - \mathbf{E}(Y_j^2)^2 \\
&= \sum_{j=1}^{N} \mathbf{E}[(\theta_j + \epsilon\eta_j)^4] - (\theta_j^2 + \epsilon^2)^2 \\
&= \sum_{j=1}^{N} \theta_j^4 + 6\theta_j^2\epsilon^2\mathbf{E}(\eta_j^2) + \epsilon^4\mathbf{E}(\eta_j^4) - \theta_j^4 - 2\epsilon^2\theta_j^2 - \epsilon^4 \\
&= \sum_{j=1}^{N} 4\epsilon^2\theta_j^2 + 2\epsilon^4 \\
&\leq 4\epsilon^2 D + 2N\epsilon^4
\end{aligned}
$$

Où on a utilisé que pour $\eta \hookrightarrow \mathcal{N}(0,1)$ alors $\mathbf{E}(\eta^2) = 1, \mathbf{E}(\eta^4) = 3$ et $\mathbf{E}(\eta^1) = \mathbf{E}(\eta^3) = 0$
En conclusion,

$$\mathbf{E}_\theta[(\hat{D}_N - D)^2] \leq L^4 N^{-4\beta} + 4\epsilon^2 D + 2\epsilon^4 N \tag{15}$$

On distingue deux cas selon le terme qui prévaut dans le terme de variance : $T_1 = 4\epsilon^2 D$ et $T_2 = 2\epsilon^4 N$:

1. Le régime paramétrique, quand $T_1 >> T_2$, c'est-à-dire le régime $\epsilon \longrightarrow 0, N \underset{\epsilon \to 0}{\longrightarrow} +\infty$.

   On chosisit N tel que $N^{-4\beta} \leq \epsilon^2$ et $N\epsilon^4 \leq \epsilon^2 D$, c'est-à-dire $N \geq (\epsilon^2)^{\frac{-1}{4\beta}}$ et $N\epsilon^2 \lesssim 1$ (c'est-à-dire à une constante près).

   Un tel rang N existe si

   $$
   \begin{aligned}
   (\epsilon^2)^{\frac{-1}{4\beta}} \times \epsilon^2 &\lesssim 1 \\
   (\epsilon^2)^{1-\frac{1}{4\beta}} &\lesssim 1
   \end{aligned}
   $$

   Comme on est dans le régime $\epsilon \to 0$, ceci est possible quand $\beta \geq \frac{1}{4}$. Et alors,

   $$
   \begin{aligned}
   \mathbf{E}_\theta[(\hat{D}_N - D)^2] &\lesssim \epsilon^2 \text{ (Vitesse paramétrique d'estimation).} \\
   &\lesssim 4D\epsilon^2 \text{ (Vitesse efficace d'estimation).}
   \end{aligned}
   $$

2. Le régime non-paramétrique, quand $T_1 << T_2$. N doit minimiser $L^4 N^{-4\beta} + 2\epsilon^4 N$, ce qui donne $N \simeq (\frac{\epsilon^4}{2\beta L^4})^{\frac{-1}{4\beta+1}}$.
   Alors, quand $\beta \in]0, \frac{1}{4}$,

   $$\sup_{\theta \in w(\beta, L)} \mathbf{E}_\theta[(\hat{D}_N - D)^2] \lesssim (\epsilon^2)^{\frac{8\beta}{4\beta+1}} \tag{16}$$

On peut résumer ce résultat avec le théorème suivant :

Théorème : Dans le modèle de suite gaussienne, pour toute suite $\theta \in w(\beta, L), \beta > 0, L > 0$, on a

$$\sup_{\theta \in w(\beta, L)} \mathbf{E}_\theta[(\hat{D}_N - D)^2] \leq \begin{cases} 4\epsilon^2 D & \text{si } \beta \geq \frac{1}{4} \text{ pour } N >> (\frac{L^2}{4\epsilon^2})^{\frac{1}{4\beta}} \\ C(\epsilon^2)^{\frac{8\beta}{4\beta+1}} & \text{si } 0 < \beta < \frac{1}{4} \text{ pour } N \simeq (\epsilon^2)^{\frac{-2}{4\beta+1}} \end{cases} \tag{17}$$

Remarque 2. Ces vitesses sont plus rapides que $n^{-\frac{4\beta}{2\beta+1}} = \epsilon^{\frac{8\beta}{4\beta+1}}$. ($\epsilon$ joue ici le rôle de $\frac{1}{n}$). Autrement dit, estimer D se fait à une vitesse plus rapide que d'estimer $\{\theta_j\}_{j\geq 1}$ en risque l_2.

Remarque 3. Dans le modèle de densite $X_1, ..., X_n$ iid de loi f, si $f \in W(\beta, L)$, la vitesse minimax se trouve en séparant les deux régimes est :

$$\frac{4}{n}(\int f^3 - (\int f^2)^2)\mathbf{1}_{\beta > \frac{1}{4}} + C(\beta, L) n^{\frac{-8\beta}{4\beta+1}} \mathbf{1}_{0<\beta\leq\frac{1}{4}} \tag{18}$$

Remarque 4. Dans le cas $f : \mathbf{R}^d \to \mathbf{R}$, le régime paramétrique correspond à $\beta \leq \frac{d}{4}$ et le non paramétrique $0 < \beta < \frac{d}{4}$ et donne une vitesse $n^{\frac{8\beta}{4\beta+d}}$

Estimation de fonctionnelles intégrables régulières.

Soit f une densité de probabilité et $D(f) = \int \Phi(x, f(x))dx$, avec $\Phi \in \mathcal{C}^4$ donnée.

Avec la moitié des données, on estime $\hat{f}_n$, pour $p > 1$:

$$\sup_{f\in\mathcal{F}_\beta} \mathbf{E}_f(||\hat{f}_n - f||_p^p) \leq C n^{\frac{-p\beta}{2\beta+1}} \tag{19}$$

Un Développement Limité de $D(f)$ en $\hat{f}_n$ donne :

$$\begin{aligned} D(f) = D(\hat{f}^n &+ \int (f - \hat{f}_n)\Phi^{'}(\hat{f}_n) + \frac{1}{2}\int (f - \hat{f}_n)^2 \Phi^{''}(\hat{f}_n) \\ &+ \frac{1}{3!}\int (f-\hat{f}_n)^3 \Phi^{'''} + C_4 ||\hat{f}_n - f||_4^4 \\ &\hat{=} J_0(\hat{f}_n) + \int f J_1(\hat{f}_n) + \int f^2 J_2(\hat{f}_n) + \int f^3 J_3(\hat{f}_n) + R_{n,4} \\ &\hat{=} J_0(\hat{f}_n) + G_1 + G_2 + G_3 + R_{n4} \end{aligned}$$

Avec l'autre moitié des données, on estime G_1, G_2, G_3, avec des moyennes empiriques :

$$\hat{G}_1 = \frac{1}{n/2}\sum_{i=1}^{n/2} J_i(\hat{f}_n, X_i) \tag{20}$$

On estime finalement D par $\hat{D} = J_0(\hat{f}_n) + \sum_{k=1}^{3} \hat{G}_k$, et le risque quadratique est $\mathbf{E}_f[(\hat{D}_N - D)^2] \lesssim n^{-\frac{4\beta}{2\beta+1}} + n^{\frac{-4\beta}{2\beta+1}}$, si $0 < \beta \leq \frac{1}{4}$.

Remarque 5. Dans le cas de l'entropie, $H(f) = \int f log(\frac{1}{f})$, la dérivée première n'est pas continue en 0. Si on se restreint à l'entropie sur un support compact, du type $\mathcal{F}_c = \{x, f(x) \geq c > 0\}$, ou bien dans les cas où $f(x)$ tend assez lentement vers 0, on peut atteindre une vitesse $\frac{1}{n}$.

3 Tests non paramétriques

3.1 Introduction

Soient $X_1, ..., X_n$ iid de densité commune $f \in \mathcal{F}$. Une **hypothèse** est un énoncé sur le paramètre du modèle f : "$f \in \mathcal{F}_0$".

On dit que l'hypothèse est **simple** si $\mathcal{F}_0$ détermine la loi $\mathbf{P}_f$. Sinon, l'hypothèse est dite **composite**.

Exemple 1. Pour le modèle $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2, \quad \mu \in \mathbf{R}, \sigma^2$ fixée). L'hypothèse - $\mathcal{H}_0$: "$\mu = 3$" c'est-à-dire $\mathcal{F}_0 = \{\mathcal{N}(3, \sigma^2)\}$ - est simple.
Exemple 2. Pour le modèle $\mathcal{F} = \{\mathcal{N}(\mu, \sigma^2), \quad \mu \in \mathbf{R}, \quad \sigma > 0\}$, la même hypothèse est associée à $\mathcal{F}_0 = \{\mathcal{N}(3, \sigma^2), \sigma > 0\}$, et il s'agit donc d'une hypothèse composite.

Définition : Un **test** est une fonction mesurable de $\chi^n vers \{0,1\}$ (ou parfois $[0,1]$ pour un test dit randomisé).

Définition : Un **problème de test** se formule H_0 vs H_1, avec ici $H_0 : f \in \mathcal{F}_0$ vs $H_1 : f \in \mathcal{F}_1$, où $\mathcal{F}_0 \cap \mathcal{F}_1 = \emptyset$.

On veut construire un test $\Delta_n = \Delta_n(X_1, ..., X_n)$, qui décide de répondre H_0 s'il renvoie 0 et H_1 s'il renvoie 1.

Tests à alternative locale

$$\begin{cases} H_0 : f \equiv f_0, \ f_0 \text{connu} \\ H_1 : f(x) = f_0(x) + \varphi_n g(x), \ g \text{ connu } et \, \varphi_n \underset{n \to \infty}{\to} 0 \end{cases}$$

Tests non paramétriques
Un **test d'adéquation** ou (goodness-of-fit)

$$\begin{cases} H_0 : f \equiv f_0, \ f_0 \text{ connu} \\ H_1(f_0, \varphi_n) : f \in W(\beta, L), \beta \geq 1, L > 0, \ \text{tel que } ||f - f_0||^2 \geq C \cdot \varphi_{n,\beta}^2 \end{cases}$$

Pour avoir des erreurs de test qui tendent vers 0, on est obligé d'exclure un voisinage de f_0 d'où la forme de H_1.
Remarque : on peut choisir d'autres normes $||.||_p$, $||.||_\infty$, ou semi-distance $|f(x_0) - f_0(x_0)|$, x_0 fixé.

Si H_0 composite :

$$\begin{cases} H_0 : & f \in \mathcal{F}_0 \subset \mathcal{F} \\ H_1(\mathcal{F}_0, \varphi_{n,\beta}) : & f \in \mathcal{F}, \ \text{tel que } \inf_{f_0 \in \mathcal{F}_0} ||f - f_0||^2 \geq C \cdot \varphi_{n,\beta}^2 \end{cases}$$

Erreurs de test

— erreur de première espècé : $P_{f_0}[\Delta_n = 1] = E_{f_0}[\Delta_n]$
— erreur maximale de première espèce : $\sup_{f_0 \in \mathcal{F}_0} E_{f_0}[\Delta_n]$
— erreur maximale de seconde espèce :

$$\sup_{f \in H_1(f_0, \varphi_{n,\beta})} P_f[\Delta_n = 0] = \sup_{f \in H_1(f_0, \varphi_{n,\beta})} E_f[1 - \Delta_n]$$

Risque d'un test Δ_n pour (3.1)

$$R(\Delta_n, \varphi_{n,\beta}) = E_{f_0}[\Delta_n] + \sup_{f \in H_1(f_0, \varphi_{n,\beta})} E_f[1 - \Delta_n]$$

On veut le plus petit $\varphi_{n,\beta}$ tq un test Δ_n : $R(\Delta_n, \varphi_{n,\beta}) \underset{n\to\infty}{\to} 0$ ou $R(\Delta_n, \varphi_{n,\beta}) \leq \gamma, \forall n$ et $0 < \gamma < 1$.

Définition On appelle $\{\varphi_{n,\beta}\}_{n\geq 1}$ vitesse minimax pour tester (1) si :

1. (borne sup) il existe un test Δ_n tq, pour $\gamma \in]0,1[$, $R(\Delta_n, \varphi_{n,\beta}) \leq \gamma, \forall n$
2. (borne inf) si $r_n = o(\varphi_{n,\beta})$, $(i.e. \, r_n/\varphi_{n,\beta} \underset{n\to\infty}{\to} 0)$ alors

$$\liminf_{n\to\infty} \inf_{\tilde{\Delta}_n} R(\tilde{\Delta}_n, \varphi_{n,\beta}) > 0$$

   où l'inf est pris sur tous les tests $\tilde{\Delta_n}$

Remarque : si $\lim_{n\to\infty} \inf_{\tilde{\Delta}_n} R(\tilde{\Delta}_n, \varphi_{n,\beta}) = 1$, on dit que H_0 et H_1 sont asymptotiquement indistinguables.

Remarque :

1. bornes sup : si $r_n/\varphi_{n,\beta} \underset{n\to\infty}{\to} \infty$ alors

$$R(\Delta_n, r_n) = E_{f_0}[\Delta_n] + \sup_{f_1\in H_1(f_0,r_n)} E_f[1-\Delta_n] \leq R(\Delta_n, \varphi_{n,\beta})$$

   car $H_1(f_0, r_n) \subseteq H_1(f_0, \varphi_{n,\beta})$

3.2 Tests de Kolomogorov-Smirnov

$X_1, ..., X_n$ iid, de f.d.r F.

$$\begin{cases} H_0 : F \equiv F_0, \ F_0 \text{ f.d.r connue} \\ H_1(F_0, \dfrac{1}{\sqrt{n}}) : \ F \text{ tq } ||F - F_0||_\infty \geq \dfrac{C}{\sqrt{n}} \end{cases}$$

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leq x)$$

Sous H_0 : $\sqrt{n}||\hat{F}_n - F_0||_\infty \xrightarrow{\mathcal{L}} V$, V de loi tabulée qui ne dépend pas de F_0. Soit t_α tq $P(V > t_\alpha) = \alpha$ et on pose le test :

$$\Delta_n = I[\sqrt{n}||\hat{F}_n - F_0||_\infty > t_\alpha]$$

On a : $E_{F_0}[\Delta_n] = P_{F_0}(\sqrt{n}||\hat{F}_n - F_0||_\infty > t_\alpha) \underset{n\to\infty}{\to} P(V > t_\alpha) = \alpha$

et

$$\begin{aligned} E_F(1-\Delta_n) &= P_F(\sqrt{n}||\hat{F}_n - F_0||_\infty \leq t_\alpha) \\ &\leq P_F(C - \sqrt{n}||\hat{F}_n - F||_\infty \leq t_\alpha) \end{aligned}$$

car $||\hat{F}_n - F||_\infty \geq ||F - F_0||_\infty - ||\hat{F}_n - F_0||_\infty$.

Finalement

$$\begin{aligned} \sup_{F\in H_1(F_0,\frac{1}{\sqrt{n}})} P_F(\sqrt{n}||\hat{F}_n - F_0||_\infty \leq t_\alpha) &\leq \sup_{F\in H_1(F_0,\frac{1}{\sqrt{n}})} P_F(\sqrt{n}||\hat{F}_n - F||_\infty \geq C - t_\alpha) \\ &\leq P(V > C - t_\alpha) + \delta, \quad \text{pour un } \delta > 0. \end{aligned}$$

Attention : on a besoin que $C > t_\alpha$ pour que $P(V \geq C - t_\alpha) < 1$. Donc $\gamma = \alpha + P(V \geq C - t_\alpha) + \delta$, qui est < 1 si C est assez grand, si n est assez grand.

3.3 Bornes inférieures

Pbm (1) : $\begin{cases} H_0 : f \equiv f_0 \\ H_1(f_0, \varphi_n), |||f||_2 \leq C_2,\ d(f, f_0)^2 \geq C.\varphi_n^2 \end{cases}$ où d peut être une norme, une distance, une semie-distance.

On choisit $r_n = o(\varphi_n)$ et un test Δ_n arbitraire. On réduit dans un premier temps l'ensemble sous H_1 à un nombre fini (fixe ou croissant avec n) de fonctions $f_1, ..., f_M \in H_1(f_0, r_n)$. Nous minorons le max par la moyenne et changeons de mesure dans chaque intégrale :

$$\begin{aligned}
\sup_{f \in H_1(f_0, r_n)} E_f[1 - \Delta_n] &\geq \max_{f \in \{f_1, .., f_M\}} E_f[1 - \Delta_n] \\
&\geq \frac{1}{M} \sum_{j=1}^{M} E_{f_j}[1 - \Delta_n] \\
&\geq \frac{1}{M} \sum_{j=1}^{M} E_{f_0}[(1 - \Delta_n) \frac{dPf_j}{dP_{f_0}}] \\
&\geq E_{f_0}[(1 - \Delta_n) \frac{1}{M} \sum_{j=1}^{M} \frac{dPf_j}{dP_{f_0}}]
\end{aligned}$$

On pose $Z = \frac{1}{M} \sum_{j=1}^{M} \frac{dPf_j}{dP_{f_0}}$. Si $0 < \eta < 1$:

$$\begin{aligned}
R(\Delta_n, r_n) &\geq E_{f_0}[\Delta_n] + E_{f_0}[(1 - \Delta_n).Z] \\
&\geq E_{f_0}[\Delta_n(1 - Z)] + E_{f_0}[Z] \\
&\geq (1 - \eta) P_{f_0}(Z \geq 1 - \eta) \quad (*)
\end{aligned}$$

en considérant successivement les cas $Z < 1 - \eta$ et $Z \geq 1 - \eta$.

On a deux minorations différentes (inégalités de Markov ou Chebyshev) :

1. $(*) \geq 1 - \frac{1}{\eta} E_{f_0}(|Z - 1|)$
2. $(*) \geq 1 - \frac{1}{\eta} E_{f_0}[(Z - 1)^2]$

Pour résumer : on construit $f_1, ...f_M$ dominées par f_0, tq $d(f_j, f_0)^2 \geq C \cdot r_n^2$ mais tq les vraisemblances soient assez proches, c-à-d $Z = \frac{1}{M} \sum_{j=1}^{M} \frac{dPf_j}{dP_{f_0}}$ vérifie $E_{f_0}(|Z - 1|) \underset{n \to \infty}{\rightarrow} 0$ ou bien $E_{f_0}[(Z - 1)^2] \underset{n \to \infty}{\rightarrow} 0$ (distance du χ^2 entre f_0 et $\frac{1}{M} \sum_{j=1}^{M} dPf_j$), alors $R(\Delta_n, r_n) \underset{n \to \infty}{\rightarrow} 1$.

Preuve de 1) :

$$\begin{aligned}
P_{f_0}(Z \geq 1 - \eta) &= 1 - P_{f_0}(\eta < 1 - Z) \\
&\geq 1 - \frac{1}{\eta} E_{f_0}(|1 - Z|) \\
&\geq 1 - \frac{1}{\eta} \sqrt{E_{f_0}(1 - Z)^2} \\
&= 1 \text{ si } Z > 1
\end{aligned}$$

et pour 2) : $E_{f_0}[(1 - Z)^2] \geq E_{f_0}^2(|1 - Z|)$

Application au test de K-S :

$$\begin{cases} H_0 : \quad F \equiv \mathbf{1}_{[0,1]} \text{ (loi uniforme)} \\ H_1(F_0, \dfrac{1}{\sqrt{n}}) : \quad ||F - F_0||_\infty \geq \dfrac{C}{\sqrt{n}} \end{cases}$$

Soit $r_n = o(\dfrac{1}{\sqrt{n}})$ (i.e. $r_n\sqrt{n} \to 0$) et $M = 1$. On choisit $F_1 \in H_1(F_0, \dfrac{1}{\sqrt{n}})$ tq F_1 admet la densité $f_1(x) = f_0(x) + r_n\Psi(x)$, où Ψ est tq $\int_0^1 \Psi = 0$ et borné, ce qui implique $f_1 \geq 0$ pour n assez grand, $||\Psi||_2^2 = 1$, $||\Psi||_1 < ||\Psi||_2 = 1$.

On a $F_1 \in H_1$, car $f_1 \geq 0$, $\int_0^1 f_1 = 1$, et $||F_1 - F_0||_\infty = \sup_{x\in[0,1]} |r_n \int_0^x \Psi(u)du| = r_n \sup_x |\int_0^x \Psi(u)du|$ où le sup est fini !

Ici $Z = \dfrac{f_1(X_1)...f_1(X_n)}{f_0(X_1)...f_0(X_n)}$ soit :

$$Z = \prod_{i=1}^n \frac{f_1}{f_0}(X_i) = \prod_{i=1}^n (1 + r_n\Psi(X_i))$$

Ainsi,

$$\begin{aligned} E_{f_0}[(Z-1)^2] &= E_{f_0}[Z^2] - 1 = E_{f_0}[\prod_{i=1}^n (1 + r_n\Psi(X_i))^2] - 1 \\ &= \prod_{i=1}^n E_{f_0}[(1 + r_n\Psi(X_i))^2] - 1 = \prod_{i=1}^n \int_0^1 (1 + r_n\Psi(x))^2 dx - 1 \end{aligned}$$

Or :

$$\begin{aligned} \int_0^1 (1 + r_n\Psi(x))^2 dx &= 1 + 2\int_0^1 r_n\Psi(x)dx + r_n^2 \int_0^1 \Psi^2(x)dx \\ &= 1 + r_n^2 \end{aligned}$$

D'où :

$$E_{f_0}[(Z-1)^2] = (1 + r_n^2)^n - 1 \leq e^{n\cdot r_n^2} - 1 \to 0$$

Donc finalement :

$$\lim_{n\to\infty} \inf_{\Delta_n} R(\Delta_n, r_n) \to 1$$

3.4 Modèle de densité

$f \in \Sigma(\beta, L)$, classe de Hölder, $\beta > 0$, $L > 0$

$$\begin{cases} H_0 : f \equiv f_0 \text{ connue} \\ H_1(f_0, \varphi_{n,\beta}) : \; f \in \Sigma(\beta, L), |f(x_0) - f_0(x_0)|^2 \geq C \cdot \varphi_{n,\beta}^2 \end{cases}$$

Ici $\varphi_{n,\beta}^2 = n^{-\frac{2\beta}{2\beta+1}}$, et on atteint les bornes sup par un test plug-in :

$$\Delta_n = I(|\hat{f}_{n,h}(x_0) - f_0(x_0)|^2 \geq \frac{2}{\gamma}\tilde{C}\varphi_{n,\beta}^2)$$

où $\hat{f}_{n,h}(x) = \dfrac{1}{nh}\sum_{i=1}^n K(\dfrac{x - X_i}{h})$, $h = \alpha n^{-\frac{1}{2\beta+1}}$

d	Estimation	Test
en x_0 fixé	$n^{-\frac{\beta}{2\beta+1}}$	$n^{-\frac{\beta}{2\beta+1}}$
$\|\|.\|\|_\infty$	$(\log(n)/n)^{\frac{\beta}{2\beta+1}}$	$(\log(n)/n)^{\frac{\beta}{2\beta+1}}$
$\|\|.\|\|_2$	$n^{-\frac{\beta}{2\beta+1}}$	$n^{-\frac{2\beta}{4\beta+1}}$

TABLE 1 – Vitesses minimax $\varphi_{n,\beta}$ sur $\Sigma(\beta, L)$

Rappel $\{\varphi_n\}_{n\geq 1}$ est dite vitesse minimax de test si :

1. $\forall \alpha > 0, \exists \Delta_n$ un test tq
$$E_{p_0}[\Delta_n] + \sup_{p\in H_1(p_0,\varphi_n)} E_p[1-\Delta_n] \leq \alpha$$
2. pour toute suite $r_n = \theta(\varphi_n)$
$$\lim_{n\to\infty} \inf_{\Delta_n} \{E_{p_0}[\Delta_n] + \sup_{p\in H_1(p_0,\varphi_n)} E_p[1-\Delta_n]\} > 0$$

— si d est la distance $|p(x_0) - p_0(x_0)|$ en x_0 fixé ou $\|p - p_0\|_\infty = \sum_x |p(x) - p_0(x)|$ alors on construit un test minimax Δ_n à partir de l'estimateur $\hat{p}_n$ qui atteint la vitesse minimax d'estimation de p.
/ ! / en norme l_2, la vitesse de test est plus rapide que la vitesse d'estimation.

Comparaison entre K.S et les tests non paramétriques basés sur la densité Soit $\Psi : [0,1] \to [-1,1]$ (cf. figure) et le test $H_0 : p = \mathbf{1}_{[0,1]}$. On choisit $p(x) = 1 + \epsilon . n^{-\frac{1}{4}} \Psi(x \cdot n^{\frac{1}{4}})$ pour un $\epsilon > 0$ fixé.

$$\begin{aligned} F_p(u) &= \int_0^u 1dx + \int_0^u \epsilon \cdot n^{-\frac{1}{4}} \Psi(x \cdot n^{\frac{1}{4}}) dx \\ &= F_{p_0}(u) + \epsilon n^{-\frac{1}{2}} \int_0^{u\cdot n^{1/4}} \Psi(u) du \end{aligned}$$

K-S :

$$\begin{aligned} \|\|F_p - F_{p_0}\|\|_\infty &= \sup_u \epsilon \cdot n^{-\frac{1}{2}} |\int_0^{u\cdot n^{1/4}} \Psi(u)du| \\ &\leq \epsilon \cdot n^{-\frac{1}{2}} n^{\frac{1}{4}} = \epsilon \cdot n^{-\frac{1}{4}} = o(\frac{1}{\sqrt{n}}) \end{aligned}$$

Donc p n'est pas détecté par K.S, il n'arrive pas à rejeter H_0, parce que la norme dans cette exemple n'accentue pas bien les différences là où il faut. On va voir que les tests non paramétriques dans cet exemple fonctionnent mieux.

Tests non paramétriques : $p \in \Sigma(1, L)$

— $\|\|p - p_0\|\|_\infty = \sup_x |\epsilon \cdot n^{\frac{-1}{4}} \Psi(x \cdot n^{\frac{1}{4}})| = \epsilon \cdot n^{-\frac{1}{4}}$

à comparer à $(\frac{\log(n)}{n})^{1/3}$: $(\frac{1}{n})^{1/4} >> (\frac{\log(n)}{n})^{1/3}$.

Donc $p \in H_1(p_0, (\frac{\log(n)}{n})^{1/3})$ en $\|\| \cdot \|\|_\infty$ et sera détectée par le test minimax.

— $||p-p_0||_2^2 = \epsilon^2 n^{-\frac{1}{2}} \int_0^1 \Psi^2(x \cdot n^{\frac{1}{4}})dx = \epsilon^2 n^{-\frac{3}{4}} \int_0^{n^{1/4}} \Psi^2(y)dy$ d'où :

$$||p-p_0||_2 = \epsilon\sqrt{C_0}n^{-\frac{3}{8}}$$

est à comparer à : $n^{-\frac{2}{5}}$ la vitesse de test quand $\beta = 1$: $n^{-\frac{3}{8}} >> n^{-\frac{2}{5}}$, donc p sera détecté avec le test minimax en norme l_2.

Modèle de bruit blanc Gaussien : tests en $||\cdot||_2$

$$dY_t = f(t)dt + \frac{1}{\sqrt{n}}dW_t,\ t \in [0,1]$$

$\{W_t\}$ mouvement brownien, $f \in W_per(\beta, L)$

$$\begin{cases} H_0 : f \equiv 0, \quad (f_0 = 0) \\ H_1(f_0, \varphi_{n,\beta}) : f \in W_{per}(\beta, L), ||f||_2^2 \geq C \cdot \varphi_{n,\beta}^2 \end{cases}$$

Soit $\{\Phi_j\}_{j\geq 1}$ la base trigonométrique de $\mathbb{L}_2[0,1]$.

$$y_j = \int \Phi_j(t)dY_t = \int \Phi_j(t)f(t)dt + \frac{1}{\sqrt{n}}\int \Phi_j(t)dW_t = \theta_j + \frac{1}{\sqrt{n}}\xi_j$$

$$\xi_j \sim \mathcal{N}(0,1), E\xi_j = \int \Phi_j = 0, E(\xi_j\xi_k) = \int \Phi_j\Phi_k = \delta_{jk}$$

donc les ξ_j sont indépendants. Pour construire Δ_n optimal, on estime $||f||_2^2 = \sum_{j\geq 1}\theta_j^2$. On doit tronquer parce qu'on ne peut pas estimer un nombre infini de paramètres. On l'estime donc par $S_n = \sum_{j=1}^M y_j^2 - \dfrac{M}{n}$.

Théorème Le test $\Delta_n = I(S_n \geq \tilde{C}\cdot\varphi_{n,\beta}^2)$, où $M = n^{\frac{2}{4\beta+1}}$ et $\tilde{C}$ tq $\tilde{C} > \dfrac{2}{\sqrt{\alpha}}$ et $C > \tilde{C} + L + \dfrac{2}{\sqrt{\alpha}}$, $\alpha > 0$ atteint la vitesse de test $\varphi_{n,\beta}^2 = n^{-\frac{4\beta}{4\beta+1}}$ qui est minimax.

Rq : faire un test c'est estimer l'écart entre deux ensembles, et selon la norme que l'on choisit on obtient des écarts plus ou moins importants.

3.5 Tests adaptatifs à la régularité

Plus facile que de faire de l'estimation adaptative.

Modèle de densité :

$$\begin{cases} H_0 : p \equiv p_0 \\ H_1(p_0, \{\Psi_{n,\beta}\}_{\beta\in(0,B]}) : \bigcup_{\beta\in(0,B]}\{p \in \Sigma(\beta,L) : d^2(p,p_0) \geq C\cdot\Psi_{n,\beta}^2\} \end{cases}$$

But : construire un Δ_n^* qui ne dépend pas de β (mais peut-être de B) et qui est minimax sur ce H_1 qui est beaucoup plus grand qu'auparavant.

Si d est $|p(x_0) - p_0(x_0)|$ ou $||p-p_0||_\infty$, alors on construit $\hat{p}_n^*$ (libre de β) tq

$$\sup_{\beta\in(0,B]}\ \sup_{p\in\Sigma(\beta,L)} E_p d^2(\hat{p}_n^*, p) \leq C_1 \cdot \Psi_{n,\beta}^2$$

alors Δ_n^* sera basé sur $\hat{p}_n^*$.

d	Estimation de f	Test
en x_0 fixé	$(\log(n)/n)^{-\frac{\beta}{2\beta+1}}$	$(\log(n)/n)^{\frac{\beta}{2\beta+1}}$
$\|\|.\|\|_\infty$	$(\log(n)/n)^{\frac{\beta}{2\beta+1}}$	$(\log(n)/n)^{\frac{\beta}{2\beta+1}}$
$\|\|.\|\|_2$	$(\log(n)/n)^{\frac{2\beta}{4\beta+1}}$	$(\sqrt{\log\log(n)}/n)^{\frac{2\beta}{4\beta+1}}$

TABLE 2 – Vitesses adaptatives $\Psi_{n,\beta}$

Estimation adaptative de $||f||_2^2 = \int f^2$ à la vitesse $(\frac{\log(n)}{n})^{\frac{2\beta}{4\beta+1}}$.

Modèle de bruit blanc Gaussien

$$\begin{cases} H_0 : f \equiv 0 (i.e.\ f_0 = 0) \\ H_1(f_0, \{\Psi_{n,\beta}\}_{0<\beta\leq B}) : \bigcup_{\beta\in(0,B]} H_1(f_0, \Psi_{n,\beta}) \end{cases}$$

où $H_1(f_0, \Psi_{n,\beta}) : f \in W_{per}(\beta, L), ||f||_2^2 \geq C \cdot \Psi_{n,\beta}^2$. Ici, $\Psi_{n,\beta}^2 = (\frac{\sqrt{\log\log(n)}}{n})^{\frac{4\beta}{4\beta+1}}$.

Construction de Δ_n^* : on se donne une grille de plus en plus fine (plus j'ai de données, plus j'ai de valeurs) $0 < \beta_1 < \beta_2 < ... < \beta_N \leq B$, $N \sim \log n$.

Pour chaque β_j on pose $\Delta_{n,j} = I(S_{n,j} \geq \tilde{C}\Psi_{n,\beta_j}^2)$ où $S_{n,j} = \sum_{k=1}^M y_k^2 - \frac{M_j}{n}$ et $M_j = \left(\frac{n}{\sqrt{\log\log n}}\right)^{\frac{2}{4\beta_j+1}}$. Au final $\Delta_n^* = \max_j \Delta_{n,j}$. On accepte H_0 ssi tous les N tests $\{\Delta_{n,j}\}$ acceptent H_0.

Erreur de première espèce :

$$\begin{aligned} E_0[\Delta_n^*] &= E_0[\max_{j\in 1,\dots,N} \Delta_{n,j}] \\ &= P_0(\bigcup_{j=1}^n (S_{n,j} \geq \tilde{C} \cdot \Psi_{n,\beta_j}^2)) \\ &\leq \sum_{j=1}^N P_0(S_{n,j} \geq \tilde{C} \cdot \Psi_{n,\beta_j}^2) \\ &= \sum_{j=1}^n P_0(\frac{1}{n}\sum_{k=1}^{M_j} \xi_k^2 \geq \tilde{C}\psi_{n,\beta_j}^2 + \frac{M_j}{n}) \end{aligned}$$

car

$$S_{n,j} = \sum y_k^2 - \frac{M_j}{n}, \text{ or sous } H_0 :\ y_j = \frac{1}{\sqrt{n}}\xi_j$$

et donc on obtient ensuite

$$= \sum_{j=1}^N P_0(\frac{\sum_{k=1}^{M_j} \xi_k^2 - M_j}{\sqrt{2M_j}} \geq \frac{n}{2M_j}\tilde{C}\Psi_{n,\beta_j}^2)$$

Or

$$\frac{\sum_{k=1}^{M_j}\xi_k^2 - M_j}{\sqrt{2M_j}} \to \mathcal{N}(0,1)$$

Inégalité de Berry-Esseen : Si $U_1, U_2, ...$ iid, $(0,\sigma^2)$, et $\rho = E(|X|^3) < \infty$, on a :

$$|P(\sqrt{n}\frac{\bar{U}_n}{\sigma} \leq x) - \Phi(x)| \leq \frac{C\cdot\rho}{\sigma^3\sqrt{n}}, \forall x$$

D'où :

$$\frac{n}{\sqrt{2M_j}}\Psi_{n,\beta_j}^2 = \frac{n}{\sqrt{2}(n/\sqrt{\log\log n})^{\frac{1}{4\beta_j+1}}} \times (\frac{\sqrt{\log\log n}}{n})^{\frac{4\beta_j}{4\beta_j+1}} = \sqrt{\frac{\log\log n}{2}}$$

Et donc :

$$\begin{aligned}
E_0[\Delta_n^*] &\leq \sum_{j=1}^{N}(\Phi(\frac{\tilde{C}}{\sqrt{2}}\sqrt{\log\log n}) + \frac{C'.\rho}{\sqrt{n}}) \\
&\leq N.e^{-\frac{1}{2}\frac{\tilde{C}^2}{2}\log\log n} + \frac{C'\rho N}{\sqrt{n}} \\
&\leq e^{(1-\frac{\tilde{C}^2}{4})\log\log n + o(1)}
\end{aligned}$$

Le tout tend vers 0 si $\tilde{C} > 2$, assez grand.

Erreur de 2ème espèce :

$$\begin{aligned}
E_f[1-\Delta_n^*] &= E_f[1 - \max_j \Delta_{n,j}] \\
&= E_f[\min_j(1-\Delta_{n,j}] \\
&= P_f[\bigcap_{j=1}^{N}\{S_{n,j} < \tilde{C}\Psi_{n,\beta_j}^2\}]
\end{aligned}$$

Or $f \in W_{per}(\beta, L)$, a une vraie régularité β que je ne connais pas, on choisit l entre 1 et N tq $\beta_l \leq \beta < \beta_{l+1}$, le plus proche inférieurement du vrai β. Et donc :

$$E_f[1-\Delta_n^*] \leq P_f(S_{n,l} < \tilde{C}\Psi_{n,\beta_l}^2)$$

Remarque :

$$\begin{aligned}
\frac{\Psi_{n,\beta_l}}{\Psi_{n,\beta}} &= (\frac{n}{\sqrt{\log\log n}})^{-\frac{2\beta_l}{4\beta_l+1}+\frac{2\beta}{4\beta+1}} \\
&= exp((\log\frac{n}{\sqrt{\log\log n}}) \times \frac{2(\beta-\beta_l)}{(4\beta_l+1)(4\beta+1)}) \underset{n\to\infty}{\asymp} 1
\end{aligned}$$

car $\beta - \beta_l \leq \frac{B}{N} = \frac{B}{\log n}$ car grille régulière

.