Introduction à la T-estimation

Eddie Aamari 1
22/11/16

Référence : Lucien Birgé - Model selection via testing : an alternative to (penalized) maximum likelihood estimators

But : Developper une méthode générale d'estimation, basée sur des tests, robuste aux erreurs de modèle et adaptée à la selection de modèle.

I Estimateur du maximum de vraisemblance

II Construction des T-estimateurs

III Application en regression bornée

I Estimateur du maximum de vraisemblance (E.M.V.)

1) Défauts de l'E.M.V.

Au delà du fait que l'É.M.V. n'est pas intrinsèque, ce n'est pas un estimateur universel, au sens où il peut échouer au moins pour deux raisons :

* Comportement pénible de la vraisemblance :

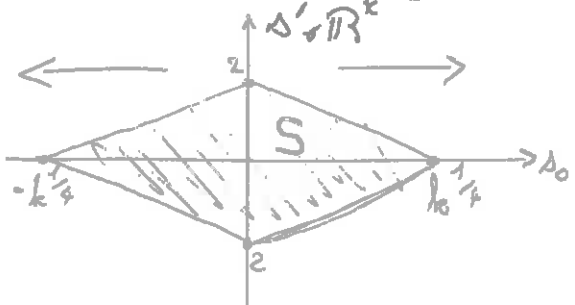
Pour le modèle de translation $f_\Delta(x) = f(x-\Delta)$, $\Delta \in \mathbb{R}$, où $f : \mathbb{R} \to \mathbb{R}_{>0}$ est une densité telle que $\lim_{x\to 0} f(x) = +\infty$. Si l'on observe un n-échantillon $X_1, \dots, X_n$, l'EMV associé n'existe pas car la vraisemblance est non-bornée.

Même en restreignant $\Delta \in K$ compact, le problème est identique.

* Massivité de l'espace des paramètres :

Soit $X = (X_0, \dots, X_k) \sim \mathcal{N}(\Delta, I_{k+1})$, avec $\Delta = (\Delta_0, \Delta_1, \dots, \Delta_k) = (\Delta_0, \Delta')$

On se place dans le modèle $S = \left\{ \Delta \in \mathbb{R}^{k+1} \mid |\Delta_0| \le k^{\frac{1}{2}}, \|\Delta\|_2 \le 2\left(1 - \frac{|\Delta_0|}{k^{\frac{1}{4}}}\right) \right\}$



Pour $k \ge 128$,

a.g.p. $\hat{\Delta}_{EMV} = \left(0, \frac{2X'}{\|X'\|}\right)$, et $\sup_{\Delta \in S} \mathbb{E}_\Delta \|\Delta - \hat{\Delta}_{EMV}\|^2 \ge \frac{3}{4}\sqrt{k} + 3$.

Pourtant l'estimateur $\tilde{\Delta} = (X_0, 0)$ donne le risque minimax

$$\inf_{\hat{\Delta}} \sup_{\Delta \in S} \mathbb{E}_\Delta \|\Delta - \hat{\Delta}\|^2 \leq 5 .$$

*Erreur dans le modèle :

De plus, même dans des modèles paramétriques compacts, une erreur dans le modèle peut conduire à une non-consistance

Modèle : les lois uniformes $\mathcal{U}_\theta$ $0 < \theta \leq 1$. On observe $X_1, \dots, X_n$ iid de loi $\bar{P}$ ayant pour densité $10\left[\left(1-\frac{2}{n}\right)\mathbb{1}_{[0,\frac{1}{10}]} + \frac{2}{n}\mathbb{1}_{[\frac{9}{10},1]}\right]$ par rapport à la mesure de Lebesgue sur $[0,1]$.

Le modèle étant faux, on mesure les distances de manière intrinsèque avec Hellinger $\boxed{h^2(P,Q) = \frac{1}{2}\int (\sqrt{dP} - \sqrt{dQ})^2}$

$$\rho(P,Q) = \int \sqrt{dP\,dQ} = 1 - h^2(P,Q)$$

L'E.M.V. est $\hat{\theta} = \max_{1 \leq i \leq n} X_i = X_{(n)}$. Avec proba $\geq 1-\left(1-\frac{2}{n}\right)^n \geq 1-e^{-2}$, $X_{(n)} \geq \frac{9}{10}$, d'où $\mathbb{E}_{\bar{P}}\, h^2(\bar{P}, \mathcal{U}_{X_{(n)}}) > 0.38$.

Or, un estimateur robuste devrait s'éloigner de la vitesse sans erreur de modèle $\left(\frac{1}{n}\right)$ du même ordre $O\left(\frac{1}{n}\right) = O\left(h^2\left(\bar{P}, \mathcal{U}_{[0,\frac{1}{10}]}\right)\right)$.

2) Une ré. interprétation du maximum de vraisemblance

On observe un n-échantillon $X = (X_1, \dots, X_n)$ d'une distribution $\overline{P}_s$, où $s \in \mathcal{S}$. On prend une paramétrisation $S \ni s \mapsto \overline{P}_s$ bijective et on munit $\mathcal{S}$ de la métrique $h(s,t) = h(\overline{P}_s, \overline{P}_t)$. On confondra allegrement distributions et paramètres dans la suite.

Supposons que $\mathcal{S}$ est compact. La famille $\{\overline{P}_s\}_{s \in S}$ est alors dominée, et on note $d\overline{P}_s$ la densité de $\overline{P}_s$.

$$\Lambda_n(t, X) = -\sum_{i=1}^{n} \log\left(d\overline{P}_t(X_i)\right)$$

L'E.M.V. est $\hat{s} = \underset{s \in S}{\operatorname{Argmin}}\, \Lambda_n(s, X)$ (supposé existant et unique)

↑ $S \subset \mathcal{S}$ fini par exemple.

Pour tous $t, u \in \mathcal{S}$,

$$\mathbb{P}_t\left(\Lambda_n(u,X) \le \Lambda_n(t,X)\right) = \mathbb{P}_t\left(\sum_{i=1}^{n} \log\left(\sqrt{\frac{d\overline{P}_u}{d\overline{P}_t}}(X_i)\right) \ge 0\right)$$

$$\overset{\text{(Chernoff)}}{\le} \mathbb{E}_t \exp\left(\sum_{i=1}^{n} \log\left(\sqrt{\frac{d\overline{P}_u}{d\overline{P}_t}}(X_i)\right)\right)$$

$$= \exp\left(n \log\left(\rho(\overline{P}_u, \overline{P}_t)\right)\right)$$

$$= \log\Big[\underbrace{\int \sqrt{d\overline{P}_u d\overline{P}_t} - 1}_{-h^2(d\overline{P}_u, d\overline{P}_t)} + 1\Big] \le -h^2(\overline{P}_u, \overline{P}_t)$$

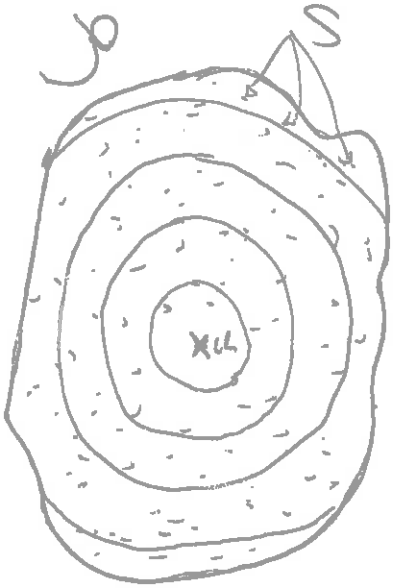
$$\le \exp\left(-n h^2(u,t)\right)$$

Pour donner une borne uniforme, on somme par couronnes :

Pour $K, \eta > 0$, $S_k = \{u \in S \mid 2^{k/\epsilon} K\eta \leq h(s,u) < 2^{(k+1)} K\eta\}$

$\forall s \in S$,

$$P_s[h(s,\hat{s}) \geq K\eta] \leq P_s[\exists u \in S,\ h(s,u) \geq K\eta,\ \Lambda_n(u,X) \leq \Lambda_n(s,X)]$$



$$\leq \sum_{k \geq 0} P_s[\exists u \in S_k,\ \Lambda_n(u,X) \leq \Lambda_n(s,X)]$$

$$\leq \sum_{k \geq 0} |S_k| \sup_{u \in S_k} P_s[\Lambda_n(u,X) \leq \Lambda_n(s,X)]$$

$$\leq \sum_{k \geq 0} |S_k| \exp[-2^k n K^2 \eta^2].$$

Pour revenir au modèle complet $\mathcal{S}$, on peut réquerir que S est un η-réseau. Il existe alors $s' \in S$ avec $h(s,s') \leq \eta$, et donc $P_s[h(s,\hat{s}) \geq (K+1)\eta]$

$$\leq P_s[h(s',\hat{s}) \geq K\eta]$$

Mais borner cette quantité est impossible en toute généralité. Autrement dit, la robustesse est parfois impossible, car le test de rapport de vraisemblance entre deux paramètres, même éloignés en Hellinger, fait des erreurs trop importantes.

Changement de point de vue :

Pour $t \in S$, on note $\mathcal{R}_t = \{u \in S,\ \Lambda_n(u,X) \leq \Lambda_n(t,X)\}$

$= \{u \in S,\ u \text{ est plus vraisemblable que } t\}$,

et $D_X(t) = \sup_{u \in \mathcal{R}_t} h(t,u)$. $D_X(t)$ s'interprète comme un indice de plausibilité : un $D_X(t)$ grand doit laisser penser que le vrai paramètre est loin de t.

Comme pour l'ETIV, $\hat{\Delta} = \operatorname{argmin}_{t \in S} \Lambda_n(t,X) \iff D_X(\hat{\Delta}) = 0$,

on obtient

$$\boxed{\hat{\Delta} = \operatorname{argmin}_{t \in S} D_X(t).}$$

Comme on a toujours $u \in \mathcal{R}_t$ ou $t \in \mathcal{R}_u$, $h(t,u) \leq D_X(t) \vee D_X(u)$.

En particulier, $h(t,\hat{\Delta}) \leq D_X(t) \vee D_X(\hat{\Delta}) \leq D_X(\hat{\Delta})$. Comme précédemment

$$\mathbb{P}_\Delta\left(h(\Delta,\hat{\Delta}) \geq (K+1)\eta\right) \leq \mathbb{P}_\Delta\left[h(\Delta',\hat{\Delta}) \geq K\eta\right]$$

$$\leq \mathbb{P}_\Delta\left[D_X(\Delta') \geq K\eta\right]$$

$$\leq \mathbb{P}_\Delta\left[\exists u \in \mathcal{R}_{\Delta'} \text{ tel que } h(\Delta',u) \geq K\eta\right],$$

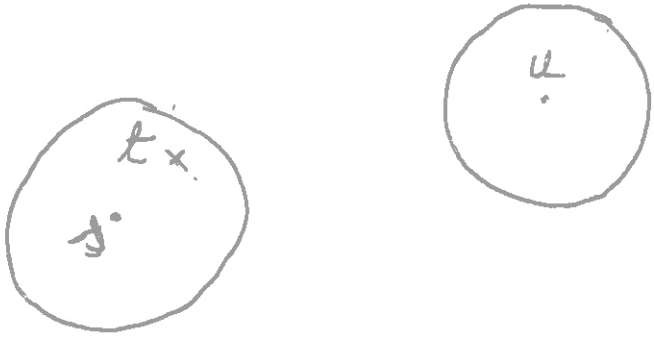
qui est sensiblement la même expression que plus haut.

Qu'a-t-on gagné ?

↳ On peut désormais remplacer le test de rapport de vraisemblance (non robuste) entre t et u par une version robuste, au sens où elle vérifierait

$$\mathbb{P}_s\big[\text{le test entre } t \text{ et } u \text{ choisit } u\big] \leq \exp\big[-c\,m\,h^2(t,u)\big]$$

si $h(t,u) \geq \kappa\, h(s,t)$



Notations pour la suite :

- On observe une variable aléatoire $X \in \Xi$ de loi P_X.
- (Π, d) espace des paramètres, métrique (ou presque).
- $s = F(P_X)$ est le paramètre d'intérêt
- On identifie s et P_s : $d(s,t) =: d(P_s, P_t)$
- Par "modèle", on entend un sous-ensemble $(S, S', \overline{S}, \mathcal{S})$ de M, qui contient s ou non.

II Construction des T-estimateurs

Def: Pour $X \in \square$, $t \neq u \in \Pi$, un test entre t et u est un couple de fonctions $\psi(t,u,X) = 1 - \psi(u,t,X) \in \{0,1\}$.

Convention : $\psi(t,u,X) = 1 \iff$ Accepter u
(H_0 « H_1»)

Vue d'ensemble : pour construire un T-estimateur, il faut :

(i) Un ensemble $S \subset M$ approximant ;

(ii) Un $\varepsilon \geq 0$ et une fonction de poids $\eta : S \to \mathbb{R}_{>0}$;

(iii) Une famille de tests entre les points de S.

Def: En notant $\forall u \neq t \in S$ $\psi(t,u,X)$ le test entre t et u, et

$$R_t = \{u \in S,\ \psi(t,u,X) = 1\},$$

puis $$D_X(t) = \begin{cases} \sup_{u \in R_t} d(t,u) & \text{si } R_t \neq \emptyset \\ 0 & \text{si } R_t = \emptyset \end{cases},$$

on appelle T-estimateur (ou T_ε-estimateur) tout $\hat{s}(X) \in S$ tel que

$$D_X(\hat{s}(X)) \vee \varepsilon \eta(\hat{s}(X)) = \inf_{t \in S} \{D_X(t) \vee \varepsilon \eta(t)\}$$

Lien avec les M-estimateurs : Si $\gamma'(\cdot, X)$ est un contraste sur S, et η une fonction poids, soit $\gamma(t, X) = \gamma'(t, X) + \tau \eta^2(t)$.

Les M-tests issus de γ' et de pénalité $\tau\eta$ sont

$$\varphi(t, u, X) = \begin{cases} 0 & \text{si } \gamma(t, X) > \gamma(u, X) \\ 1 & \text{si } \gamma(t, X) < \gamma(u, X) \end{cases}.$$

Remarque : Par définition, $\forall t \in S$, $d(t, \hat{S}(X)) \leq D_X(t) \vee \varepsilon \eta(t)$.

En particulier, si $\varepsilon = 0$, $\forall s \in M$,

$$d(s, \hat{S}) \leq \inf_{t \in S} \{ d(s, t) + D_X(t) \},$$

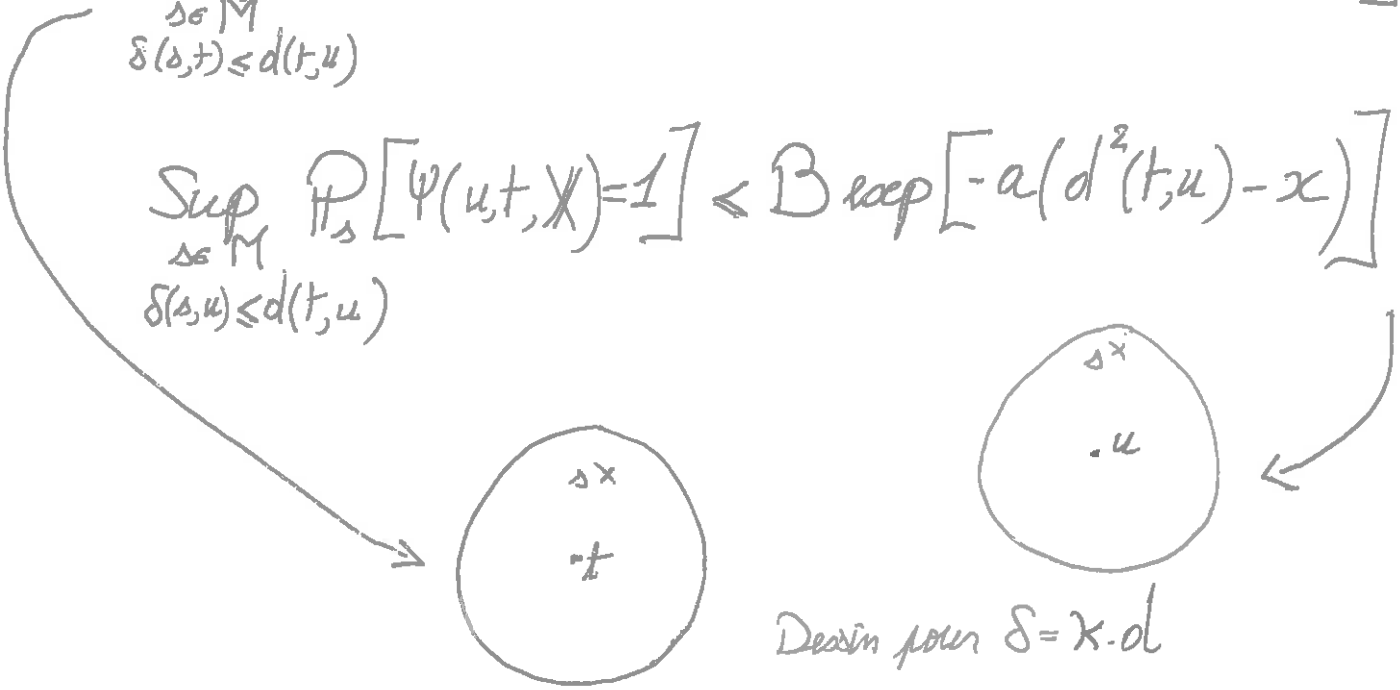
c'est-à-dire que $\hat{S}$ réalise le meilleur compromis entre distance à S et plausibilité.

Hypothèses sur la famille de tests :

(A1) Il existe $M_T \subset M$ $\delta : M \times M_T \to [0,\infty]$, $a, B > 0$ tels que pour tout $u \neq t \in M_T$ et $x \in \mathbb{R}$, il existe un test $\psi(t,u,X)$ tel que

$$\sup_{\substack{s \in M \\ \delta(s,t) \leq d(t,u)}} P_s[\psi(t,u,X)=1] \leq B \exp\left[-a(d^2(t,u)+x)\right]$$

$$\sup_{\substack{s \in M \\ \delta(s,u) \leq d(t,u)}} P_s[\psi(u,t,X)=1] \leq B \exp\left[-a(d^2(t,u)-x)\right]$$



Dessin pour $\delta = \chi \cdot d$

(A2) (Version contraste) $\psi(t,u,X) = 1 \iff \gamma'(t,X) - \gamma'(u,X) > \tau x$

↳ Hypothèse similaire

On prendra ensuite $x = \eta^2(u) - \eta^2(t)$

Hypothèses sur les modèles :

On mesurera la massivité des modèles avec la notion de D-modèle ("D" pour Discret et Dimension)

Def (D-modèle) Soient $\eta, D, B' > 0$ et $S' \subset M$. On dit que S' est un D-modèle pour les paramètres η, D, B' lorsque

$$|S \cap B_d(t, x\eta)| \leq B' \exp[Dx^2] \quad \forall x \geq 2, t \in M.$$

De manière équivalente,

$$|S \cap B_d(t, r)| \leq B' \exp\left[D\left(\frac{r}{\eta} \vee 2\right)^2\right] \quad \forall r > 0, t \in M.$$

Remarque :
- Le "2" est fixé par convention. L'idée est de regarder S à l'échelle $\gtrsim \eta \times 2\eta$.
- Voir section 6 pour le lien avec la dimension métrique et la dimension metrique intérieure. Et autres...
- Le carré "x^2" peut paraître contre-intuitif si l'on se réfère à $\mathbb{R}^D$ où il est inutile. Malgré tout, des "$\log x$" apparaissent naturellement dans la construction de modèles approximants. Ne serait-ce que pour des raisons techniques liées à l'entropie quand elle est utilisée. ◊ section 6 : Il semble être plus intéressant d'être optimal en "D" et de perdre sur le rayon des boules.

T-estimateurs basés sur un seul D-modèle :

Pas de pénalité ici : $\varepsilon = 0$.

Théorème (3) (Simplifié) : Sous l'hypothèse (A1), soit $S \subset M_T$ un D-modèle tel que $D \geq 1/2$ et $2a\eta^2 \geq 3D$

Pour $s \in \mathcal{M}$, si $\delta = \underset{>0}{\kappa} d$ et $1 \leq q \leq \frac{16a\eta^2}{3b} \wedge 17$ (Artificiel), on a

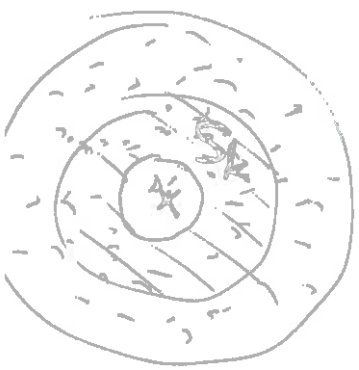
$$\mathbb{E}_s\, d^q(s,\hat{s}) \leq C_{q,B,B',\kappa} \left(d(s,S) \vee \frac{4\eta}{\kappa} \right)^q$$

Dem : Si $s' = \pi_S(s)$, $d(s,\hat{s}) \leq d(s,s') + \mathcal{D}_\kappa(s')$

$$= d(s,S) + \mathcal{D}_\kappa(s').$$

On borne $\mathbb{P}_s[\mathcal{D}_\kappa(s') > y]$ en sommant sur des couronnes centrées en s', notées S_k :

$$\mathbb{P}_s[\mathcal{D}_\kappa(s') > y] = \mathbb{P}_s\left[\exists t \in S,\ d(t,s') > y \text{ et } \psi(s',t,X) = 1\right]$$



$$\leq \sum_{k \geq 0} \mathbb{P}_s\left[\exists t \in S_k,\ \psi(s',t,X) = 1\right]$$

$$\leq \sum_{k \geq 0} |S_k| \sup_{s \in S_k} \mathbb{P}_s\left[\psi(s',t,X) = 1\right]$$

Borné via D-modèle — Borné via Hypothèse (A1)

□

Remarque : le pendant "contraste" fonctionne tout pareil.

T-estimateurs basés sur plusieurs D-modèles :

Ici, S est l'union de différents D-modèles.

(A3) $S = \bigcup_{m \in \mathcal{M}} S_m$ est l'union finie ou dénombrable de D-modèles S_m de paramètres η_m, D_m, B', avec $D_m \geq \frac{1}{2}$.

Pour $t \in S$, on note $\eta(t) = \inf\{\eta_m \mid m \in \mathcal{M} \text{ et } t \in S_m\}$

= le meilleur η_m disponible pour t.

On suppose que $\sum_{m \in \mathcal{M}} \exp(-a\eta_m^2/21) = \Sigma < \infty$

Théorème (5) (Simplifié) On suppose (A1) et (A3) avec $S \subset \mathcal{M}_T$,

$$a\eta_m^2 \geq 21 D_m / 5 \quad \forall m \in \mathcal{M}$$

Si $\varepsilon > 0$, $\delta = \chi d$ et $1 \leq q \leq 7.9$ (Artificiel), $\forall s \in \mathcal{M}$,

$$\mathbb{E}_s\left[d_2^q(s, \hat{s})\right] \leq C_{q,B,B',\Sigma,\chi} \times \inf_{m \in \mathcal{M}} \left\{ d(s, S_m) \vee \frac{4\eta_m}{\chi} \right\}^q$$

Remarque : le pendant "contraste pénalisé" est analogue.

III Application en regression bornée

On observe un n-echantillon $\underbrace{(X_1,Y_1),\dots,(X_n,Y_n)}_{=\mathbb{X}}$ où

$$\begin{cases} \cdot\ X \in \mathcal{X} \\ \cdot\ Y_i = s(X_i) + \xi_i \\ \cdot\ Y, s(X) \in [0,1] \end{cases}, \quad \begin{cases} X \sim \mu \text{ inconnue} \\ \mathbb{E}_s[\xi \mid X] = 0 \end{cases}$$

On note, pour $s,t \in \{\mathcal{X} \to [0,1] \text{ mesurables}\} = \mathcal{M}$

$$d_2^2(t,s) = \mathbb{E}_\mu \|s(X) - t(X)\|^2 .$$

On rappelle que l'on a besoin $\begin{cases} \text{d'une famille de tests} \\ \text{de D-modèle(s)} \end{cases}$

Proposition (5) Pour tout $s,t,u \in \mathcal{M}$, $z \in \mathbb{R}$, si $\gamma = 4\|s-t\|^2 - \|t-u\|^2/4$,

$$\mathbb{P}_s\left[\gamma'(t,X) - \gamma'(u,X) \geq nz\right] \leq \exp\left[-\frac{3n}{100}\left(\|t-u\|^2 + \frac{98(z-\gamma)}{25}\right)\right],$$

où $\gamma'(t,X) = \sum_{i=1}^{n} (Y_i - t(X_i))^2$

Dem Bernstein à $Z_i = (Y_i - t(X_i))^2 - (Y_i - u(X_i))^2$ qui vérifie

$$\begin{cases} |Z_i| \leq 1 \\ |Z_i| \leq 2|t(X_i) - u(X_i)| \end{cases}$$

À propos de la construction de D-modèles

Comme la loi de la covariable est inconnue, l'espace métrique à discrétiser est inconnu : $d_2^2(s,t) = E_{\mu=?}|t(X)-u(X)|^2$.

Par conséquent, la façon sûre de discrétiser est d'utiliser la métrique plus forte L^∞

Pour $\left\{ m=(j,k)\in\mathbb{N}^2 \;\middle|\; 3\le|m|\le\frac{n}{10}\right\} = \mathcal{M} \ni m$

où $|m| = j(k+1)$,

considérons l'espace $\mathcal{P}_m$ des polynômes par morceaux sur $[0,\frac{1}{j}], \dots [\frac{j-1}{j},1]$ de degrés $\le k$. $\mathcal{P}_m$ est un espace vectoriel de dimension $|m|$: on construit un D-modèle $S_m \subset \mathcal{P}_m$ de paramètres :

$$\eta_m = \sqrt{\frac{18|m|}{n}\log\left(\frac{n}{|m|}\right)} \;;\; D_m = \frac{|m|}{4}\log\left(1+\frac{2}{\eta_m}\right) \;;\; B' = 1$$

Troncature à $[0,1]$ et existence d'un réseau en dimension finie sur les compacts $\longrightarrow d_2(s,S_m) \le d_\infty(s,\mathcal{P}_m) + \eta_m$

Corollaire (8) On rappelle que si $\eta_m \ge \frac{40 D_m}{n}$ et $\sum \exp(-n\eta_m^2/200) = \Sigma < \infty$,

L'unique minimiseur de $\sum_{i=1}^{n}(Y_i - t(X_i))^2 + \frac{25}{24} n\eta^2(t)$ vérifie

$$E_s\|s-\hat{s}\|^q \le C_{q,B',\Sigma}\left\{\left(\inf_{m\in\mathcal{M}}\|s-t\|\vee\eta_m\right)^q\right\} \quad \text{pour } 1\le q\le 79.$$

Dans notre cas, les polynômes locaux fournissent :

$$\mathbb{E}_s \|s - \hat{s}\|^q \leq c_q \inf_{m \in \mathcal{M}} \left\{ d_\infty(s, \mathcal{P}_m) + \sqrt{\frac{B|m|}{n} \log\left(\frac{n}{|m|}\right)} \right\}^q .$$

Si $s \in \mathcal{L}^\beta(R)$ est dans une classe de Hölder, on a

$$d_\infty(s, \mathcal{P}_m) \leq c_\beta R\, j^{-\beta},$$

donc après optimisation en $m = (j,k) \in \mathcal{M}$, on obtient :

$$\mathbb{E}_s \|s - \hat{s}\|^2 \leq c_\beta\, R^{\frac{2}{2\beta+1}} \left(\frac{\log n}{n}\right)^{\frac{2\beta}{2\beta+1}} .$$

On perd le "log n" ici, mais $\hat{s}$ est robuste