Grandes déviations pour des L-statistiques

H.Boistard

http://www.eio.uva.es/∼helene

Definition des L-statistiques

Définition 1.

$$A_n = \sum_{i=1}^{n} a_{n,i} X_{(i)},$$

où $(X_i)_{i=1}^n$ *i.i.d.* $\sim F$, $(X_{(i)})_{i=1}^n$ *la statistique d'ordre,* $a_{n,i}$ *des coefficients dans* $\mathbb{R}^k$.

Souvent : $a_{n,i} \simeq \frac{1}{n} a(\frac{i}{n})$ où $a : [0,1] \to \mathbb{R}^k$.

Pour nous : $a_{n,i} = \int_{\frac{i-1}{n}}^{\frac{i}{n}} a(t)dt$. Dans ce cas :

$$A_n = \int_0^1 a(t) F_n^{-1}(t) dt.$$

Exemples de L-statistiques

- α-moyenne recoupée : $\frac{1}{n-2\lfloor \alpha n \rfloor} \sum_{i=\lfloor \alpha n \rfloor + 1}^{n-\lfloor \alpha n \rfloor} X_{(i)}$.

- Différence moyenne de Gini : estimateur du paramètre de dispersion $\theta = E(|X_1 - X_2|)$.

$$T_n = \frac{1}{C_n^2} \sum_{i<j} |X_i - X_j| = \frac{1}{C_n^2} \sum_{i=1}^{n} (-n + 2i - 1) X_{(i)}.$$

- Test d'ajustement de D'Agostino :

$$D = \frac{\sum_{i=1}^{n} (i - (n+1)2^{-1}) X_{(i)}}{n^2 S_n},$$

avec S_n^2 la variance empirique.

Lois limites

• LGN : via un raffinement du théorème de Glivenko-Cantelli (Wellner (1977, 1978), van Zwet (1980)).

• TCL : via méthode des projections de Hájek (Stigler (1974)), approximation par des U-statistiques (Helmers (1981)), ou processus empiriques (Mason et Shorack (1992)).

• Références complètes : Shorack et Wellner (1986) pour les lois limites, van der Vaart (1998) pour le TCL.

Rappels de grandes déviations

$\mathcal{X}$ un espace de Hausdorff, $\mathcal{B}(\mathcal{X})$ sa tribu de Borel.

Définition 2. *$I : \mathcal{X} \to \mathbb{R}_+ \cup \{+\infty\}$ est appelée fonction de taux (f.t.) si elle est s.c.i. Elle est dite bonne (b.f.t.) si $\forall \alpha$, l'ensemble de niveau*

$$\{x : I(x) \leq \alpha\}$$

est compact.

Définition 3. *Une suite (R_n) de mesures de probabilité sur $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ satisfait un PGD de f.t. I si $\forall A \in \mathcal{B}(\mathcal{X})$,*

$$-\inf_{x \in \mathring{A}} I(x) \leq \liminf_{n \to \infty} \frac{1}{n} \log R_n(A) \leq \limsup_{n \to \infty} \frac{1}{n} \log R_n(A) \leq -\inf_{x \in clo(A)} I(x).$$

Outils importants

Proposition 1. ***(Principe de contraction)*** *$\mathcal{X}$ et $\mathcal{Y}$ deux espaces de Hausdorff, $f : \mathcal{X} \to \mathcal{Y}$ une fonction continue. (R_n) satisfait un PGD sur $\mathcal{X}$ de b.f.t. I. Alors $\left(R_n \circ f^{-1}\right)$ satisfait un PGD sur $\mathcal{Y}$ de b.f.t. I' définie pour $y \in \mathcal{Y}$ par :*

$$I'(y) = \inf\{I(x) : x \in \mathcal{X}, f(x) = y\}.$$

Proposition 2. ***(Equivalence exponentielle)***
$(\mathcal{X}, d)$ espace métrique. ζ_n et ξ_n deux v.a. à valeurs dans $\mathcal{X}$. Elles sont exponentiellement équivalentes si $\forall \delta > 0$,

$$\limsup_{n\to\infty} \frac{1}{n} \log P\left(d\left(\zeta_n, \xi_n\right) > \delta\right) = -\infty.$$

Dans ce cas (ζ_n) satisfait un PGD de b.f.t. $\Rightarrow$ (ξ_n) satisfait le même PGD.

Théorème fonctionnel sous hypothèse forte de moment exponentiel

$\mathcal{M}_2(\mathbb{R})$ l'espace des fonctions quantile G^{-1} de lois P_G sur $\mathbb{R}$ telles que

$$\int x^2 dP_G = \int_0^1 \left(G^{-1}(t)\right)^2 dt < \infty.$$

Topologie : héritée de $L_2(0,1)$. Correspond à la topologie de la distance de Wasserstein sur les mesures de probabilités avec second moment fini :

$$\begin{aligned} \mathcal{W}^2(P_G, P_H) &= \inf\left\{E(X-Y)^2 : \mathcal{L}(X) = P_G, \mathcal{L}(Y) = P_H\right\} \\ &= \int_0^1 \left(G^{-1}(t) - H^{-1}(t)\right)^2 dt. \end{aligned}$$

Condition forte de moment exponentiel : $\exists \varphi : \mathbb{R} \to \mathbb{R}$, $\varphi(x) \to +\infty, |x| \to \infty$, et $t > 0$ tq

$$E\left(e^{tX_1^2 \varphi(X_1)}\right) < +\infty. \tag{1}$$

Théorème 1. *Sous (1), F_n^{-1} satisfait un PGD dans $\mathcal{M}_2(\mathbb{R})$ de bonne fonction de taux*

$$I_1(G^{-1}) = K(G, F).$$

K est la distance de Kullback.

Conséquence : PGD pour $A_n = \int_0^1 a(t) F_n^{-1}(t) dt$ lorsque $a \in L_2(0,1)$, de b.f.t. :

$$I_2(C) = \inf_{G \text{ f.r.}, \int a G^{-1} = C} K(G, F).$$

Plan de la preuve

- Hypothèse (1) $\Rightarrow$ tension exponentielle de F_n^{-1}.

- Identification de la fonction de taux par théorème de Sanov. (PGD pour la mesure empirique $\nu_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.)

Application au cas uniforme

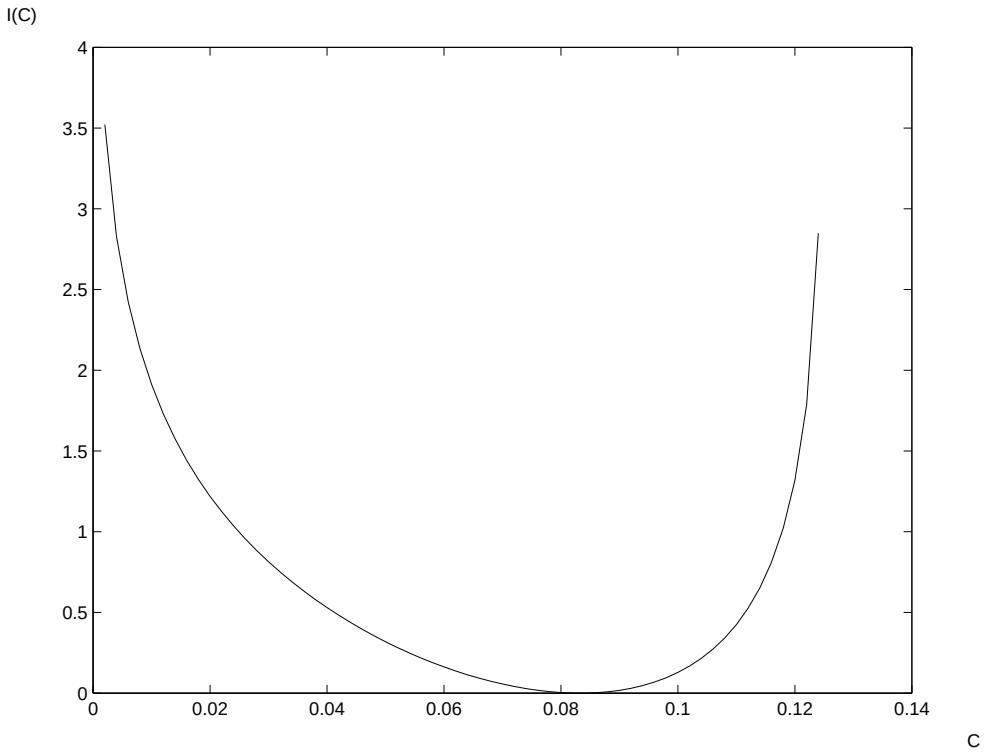
F la loi uniforme sur $[0,1]$, $a \in L_2(0,1)$ telle que $E\left(a(X_1)\right) = 0$.

PGD pour $\int_0^1 a(t) F_n^{-1}(t) dt$, fonction de taux exprimée grâce à outils de dualité.

Exemple : $a(t) = t - \frac{1}{2}$.

$$I(C) = \sup_{x,y \in \mathbb{R}} 1 + x + Cy + \int_0^1 \log\left(-x - \frac{y}{2} t(1-t)\right) dt.$$

Fonction de taux pour $a(t) = t - \frac{1}{2}$



PGD sous hypothèse de moment plus faible

Condition de moment plus faible :

$$E(e^{sX_1^2}) < \infty, \ \forall s \in \mathbb{R}. \tag{2}$$

Théorème 2. *Sous (2), F_n^{-1} satisfait un PGD, de b.f.t. :*

$$I_3(G^{-1}) = \sup_{\delta>0} \liminf_{T\to\infty} \inf_{\|H^{-1}-G^{-1}\|_2<\delta} K(H, F^T).$$

Preuve : argument de troncature.
$X_i^T = -T1_{X_i<-T} + X_i 1_{|X_i|\leq T} + T1_{X_i>T}$ satisfait (1).

PGD pour des L-statistiques normalisées

Sous (1),

$$F_n^{-1,N} = \sum_{i=1}^{n} 1_{\left(\frac{i-1}{n}\frac{i}{n}\right]} \frac{X_{(i)} - \bar{X}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}}$$

satisfait un PGD de b.f.t.

$$I_5(G^{-1}) = \begin{cases} \inf_{\mu\in\mathbb{R},\sigma>0} K\left(G\left(\frac{\cdot-\mu}{\sigma}\right), F\right) & \text{quand } G \text{ est } (0,1), \\ +\infty & \text{sinon.} \end{cases}$$

Preuve : la même que Théorème 1, avec la normalisation en plus.

Exemple : statistique D de D'Agostino.

PGD pour la loi exponentielle

PGD fonctionnel pour la mesure

$$\nu_n = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}\delta_{i/n}.$$

Espace $\mathcal{P}([0,1])$ des mesures régulières bornées sur $[0,1]$, muni de la convergence en loi, dual de $\mathcal{C}([0,1])$.

$$\frac{1}{n}\log E(\exp[\nu_n(na)]) \to \Lambda(a), \forall a \in \mathcal{C}([0,1])$$

où

$$\Lambda(a) = \begin{cases} -\int_0^1 \log\left[1 - \frac{\int_{1-t}^1 a(u)du}{t}\right] dt & \text{lorsque l'intégrale est définie} \\ +\infty & \text{sinon.} \end{cases}$$

Application du théorème de Gärtner-Ellis : ν_n satisfait un PGD de bonne fonction de taux

$$\Lambda^*(\mu) = \sup_{a \in C([0,1])} \left[\int_0^1 a(t) d\mu(t) - \Lambda(a) \right].$$

Expression de $\Lambda^*(\mu)$ par dualité : si μ s'écrit $\mu = l\lambda + \mu(\{1\})\delta_1$, où $l(u) = \int_0^u dm(s)$, m a pour décomposition de Lebesgue $m = \alpha\lambda + \chi$ et la mesure singulière $-td\chi(1-t) + \mu(\{1\})\delta_0$ est positive, alors

$$\Lambda^*(\mu) = \int_0^1 (t\alpha(1-t) - \log \alpha(1-t))\, dt - \int_0^1 td\chi(1-t) + \mu(\{1\}).$$

Sinon, $\Lambda^*(\mu) = +\infty$.