

Un théorème limite conditionnel. Applications à l'inférence conditionnelle et aux méthodes d'Importance Sampling.

Soutenance de thèse.

Virgile Caron, LSTA-UPMC-Paris VI

Directeur de thèse : Michel Broniatowski

16 Octobre 2012

- 1 Introduction
- 2 Historique
- 3 Théorème Principal
- 4 Généralisation
- 5 Application aux méthodes d'Importance Sampling
- 6 Application à l'inférence conditionnelle
- 7 Perspectives

Contexte : Importance Sampling

- $\mathbf{X} \sim p$ sur \mathbb{R}
- $\mathbf{X}_1^n := (\mathbf{X}_1, \dots, \mathbf{X}_n)$ i.i.d.
- $A \subset \mathbb{R}$.
- $\mathbf{S}_{1,n} := \mathbf{X}_1 + \dots + \mathbf{X}_n$

Objectif

On veut estimer $P_n := P(\mathbf{S}_{1,n} \in nA)$.

- $(Y_1^n(1), \dots, Y_1^n(l)) \sim p$ i.i.d.

$$\frac{1}{L} \sum_{l=1}^L \mathbb{1}_{\mathcal{E}_n}(Y_1^n(l)) \rightarrow \int \mathbf{1}_{\mathcal{E}_n}(x_1^n) p(x_1^n) dx_1^n = P_n$$

avec

$$\mathcal{E}_n := \{(x_1, \dots, x_n) \in \mathbb{R}^n : (x_1 + \dots + x_n) \in nA\}.$$

Contexte : Importance Sampling

- $(Y_1^n(1), \dots, Y_1^n(L)) \sim g$ i.i.d. et les coordonnées de $Y_1^n(l)$ pas nécessairement i.i.d..
- Si $\text{supp}(p) \subset \text{supp}(g)$.

$$\frac{1}{L} \sum_{l=1}^L \frac{\prod_{i=1}^n p(Y_i(l))}{g(Y_1^n(l))} \mathbf{1}_{\mathcal{E}_n}(Y_1^n(l)) \rightarrow \int \frac{\prod_{i=1}^n p(x_i)}{g(x_1^n)} \mathbf{1}_{\mathcal{E}_n}(x_1^n) g(x_1^n) dx_1^n = P_n$$

- Choix optimal de g

$$g(x_1^k) = p(\mathbf{X}_1^n | \mathbf{S}_{1,n} \in nA) = \frac{p(\mathbf{X}_1^n)}{P_n} \mathbf{1}_{\mathcal{E}_n}(\mathbf{X}_1^n)$$

Objectif

Nous voulons approximer $p(\mathbf{X}_1^n = Y_1^n | \mathbf{U}_{1,n} \in nA)$.

Principe de Gibbs

- On définit la Distance de Kullback-Leibler :

$$K(Q, P) := \int \log \frac{dQ}{dP} dQ$$

$$K(\Omega, P) := \inf_{Q \in \Omega} K(Q, P)$$

- Notons :

$$\Pi = \arg \inf_{Q \in \Omega} K(Q, P)$$

où Ω est un ensemble de mesure.

Principe de Gibbs

- $A = \{Q \in \mathcal{P} : \int xQ(dx) \geq a\}$.
- Dans ce cas, Π existe et on l'appelle loi **tiltée** définie par

$$\frac{d\Pi^a}{dP}(x) = \frac{\exp tx}{\int \exp tx dP(x)}$$

- $\int v\Pi^a(dv) = a$ et t et a sont en bijection.
- Dans le cas où il existe une densité, elle s'écrit :

$$\pi^a(x) = \frac{\exp tx}{\Phi(t)} p(x)$$

où Φ est la fonction génératrice des moments.

Principe de Gibbs

- Csiszar (1984). On en déduit une version du principe de Gibbs pour a fixe,

$$KL\left(\mathcal{L}\left(\mathbf{x}_1 \mid \frac{1}{n} \sum \mathbf{x}_i \geq a\right), \Pi^a(\mathbf{x}_1)\right) \rightarrow 0$$

- Dembo et Zeitouni (1996). Pour k fixe ou $k/n \rightarrow c < 1$,

$$KL\left(\mathcal{L}\left(\mathbf{x}_1^k \mid \frac{1}{n} \sum \mathbf{x}_i \geq a\right), \prod_{i=1}^k \Pi^a(\mathbf{x}_i)\right) \rightarrow 0$$

Principe de Gibbs

On recherche une généralisation en densité de $\mathcal{L}(\mathbf{X}_1^k | \frac{1}{n} \sum \mathbf{X}_i \geq a_n)$ avec $k = k_n$ le plus grand possible, typiquement

$$0 \leq \limsup_{n \rightarrow \infty} k/n \leq 1$$

- pour des conditionnements de type $(\sum u(\mathbf{X}_i) \in nA)$
- avec des \mathbf{X}_i dans \mathbb{R}^d
- et des fonctions $u : \mathbb{R}^d \rightarrow \mathbb{R}^s$, $d, s \geq 1$.

Théorème Principal : Hypothèses et Notations.

- $\mathbf{X} \sim p_{\mathbf{X}}$ sur \mathbb{R}
- $\mathbf{X}_1, \dots, \mathbf{X}_n$ i.i.d.
- On suppose que \mathbf{X} est à densité à queue légère. Dans un voisinage non vide de 0 on a : $\Phi(t) := E \exp(t\mathbf{X}) < \infty$.
- $\mathbf{S}_{1,n} := \mathbf{X}_1 + \dots + \mathbf{X}_n$

$$p_{nA}(\mathbf{X}_1^k = Y_1^k) := p(\mathbf{X}_1^k = Y_1^k | \mathbf{S}_{1,n} \in nA)$$

$$p_{nA}(\mathbf{X}_1^k = Y_1^k) = \int_{nA} p(\mathbf{X}_1^k = Y_1^k | \mathbf{S}_{1,n} = nv) p(\mathbf{S}_{1,n} = nv | \mathbf{S}_{1,n} \in nA) dv$$

$$p_{nv}(\mathbf{X}_1^k = Y_1^k) := p(\mathbf{X}_1^k = Y_1^k | \mathbf{S}_{1,n} = nv)$$

Théorème Principal

Soit $a := a_n$ suite convergente.

Objectifs

On cherche g_{na} , une densité sur \mathbb{R}^k , proche de p_{na} pour k aussi proche de n que possible, facile à simuler tel que

$$p_{na}(\mathbf{X}_1^k = Y_1^k) \approx g_{na}(\mathbf{X}_1^k = Y_1^k)$$

- soit **sur les chemins** sous g_{na}
- soit **sur les chemins** sous p_{na}

Théorème Principal

$$\begin{aligned} p(\mathbf{x}_1^k = Y_1^k | \mathbf{s}_{1,n} = na) &= \prod_{i=0}^{k-1} p(\mathbf{x}_{i+1} = Y_{i+1} | \mathbf{x}_1^i = Y_1^i, \mathbf{s}_{1,n} = na) \\ &= \prod_{i=0}^{k-1} p(\mathbf{x}_{i+1} = Y_{i+1} | \mathbf{s}_{i+1,n} = na - S_{1,i}) \\ &= \prod_{i=0}^{k-1} \pi^{m_i}(\mathbf{x}_{i+1} = Y_{i+1} | \mathbf{s}_{i+1,n} = na - S_{1,i}) \\ &= \prod_{i=0}^{k-1} \pi^{m_i}(\mathbf{x}_{i+1} = Y_{i+1}) \frac{\pi^{m_i}(\mathbf{s}_{i+2,n} = na - S_{1,i+1})}{\pi^{m_i}(\mathbf{s}_{i+1,n} = na - S_{1,i})} \end{aligned}$$

où $S_{1,i} = Y_1 + \dots + Y_i$.

Théorème Principal

On définit

$$m(t) = \frac{d}{dt} \log \Phi(t).$$

Soit t_i l'unique solution de l'équation

$$m(t_i) := m_i = \frac{n}{n-i} \left(a - \frac{s_{1,i}}{n} \right)$$

où $s_{1,i} := y_1 + \dots + y_i$. On définit aussi

$$\pi^{m_i}(x) := \frac{\exp(t_i x)}{\phi(t_i)} p(x)$$

Lemmes permettant l'obtention du théorème

Lemme (sur les moments conditionnels)

$$\textcircled{1} E_{P_{na}}(\mathbf{X}_1) = a$$

$$\textcircled{2} E_{P_{na}}(\mathbf{X}_1 \mathbf{X}_2) = a^2 + o\left(\frac{1}{n}\right)$$

$$\textcircled{3} E_{P_{na}}(\mathbf{X}_1^2) = s^2(t) + a^2 + o\left(\frac{1}{n}\right)$$

où $m(t) = a$.

Lemme (sur le maximum des \mathbf{X}_1^n)

$$\max(|\mathbf{X}_1|, \dots, |\mathbf{X}_n|) = O_{P_{na}}(\log n).$$

Lemme (sur le maximum des m_i)

Soit ε_n une suite qui tends vers 0 telle que $\lim_{n \rightarrow \infty} \varepsilon_n \sqrt{n-k} = \infty$.

$$\max_{1 \leq i \leq k} |m_i| = a + o_{P_{na}}(\varepsilon_n).$$

Théorème Principal

On définit la densité g_{na} de façon itérative :

$$g_{na}(y_1^k) := \pi^a(y_1) \prod_{i=1}^{k-1} g(y_{i+1}|y_1^i)$$

où chacun des termes du produit approxime $p(\mathbf{X}_{i+1} = Y_{i+1} | \mathbf{X}_1^i = Y_1^i, \sum X_i = na)$ et s'écrit

$$g(y_{i+1}|y_1^i) = C_i p(y_{i+1}) n(\alpha\beta + a, \alpha, y_{i+1})$$

où $n(\mu, \tau, x)$ est la densité normale de moyenne μ et de variance τ en x . Les deux paramètres α et β s'expriment par rapport aux y_1^i précédent et aux cumulants de la loi de départ.

Théorème Principal

Théorème

On suppose satisfaites les conditions ci-dessus. Soit Y_1^k un échantillon de loi P_{na} . Alors

$$p_{na}(Y_1^k) := p(\mathbf{X}_1^k = Y_1^k | \mathbf{S}_{1,n} = na) = g_{na}(Y_1^k)(1 + O_{P_{na}}(\varepsilon_n(\log n)^2)).$$

Théorème

On suppose satisfaites les conditions ci-dessus. Soit Y_1^k un échantillon de loi G_{na} . Alors

$$p_{na}(Y_1^k) := p(\mathbf{X}_1^k = Y_1^k | \mathbf{S}_{1,n} = na) = g_{na}(Y_1^k)(1 + O_{G_{na}}(\varepsilon_n(\log n)^2)).$$

Théorème

Sous les hypothèses précédentes, la variation totale entre P_{na} et G_{na} tends vers 0 quand n tends vers l'infini et

$$\lim_{n \rightarrow \infty} \int |p_{na}(y_1^k) - g_{na}(y_1^k)| dy_1^k = 0.$$

- Théorème 1.6 dans Diaconis et Freedman (1988)
- Théorème 2.15 dans Dembo et Zeitouni (1996)

Théorème Principal

Notations

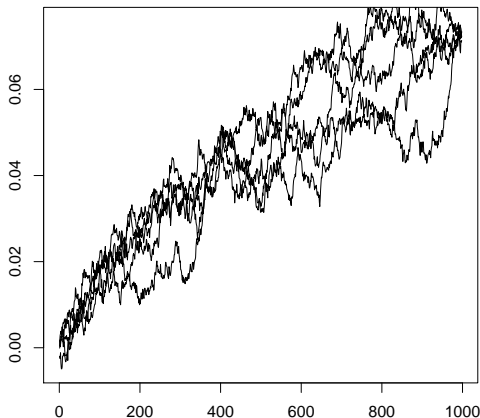
$$g(y_{i+1}|y_1^i) = C_i p(y_{i+1}) \exp \left(y_{i+1} t_i + y_{i+1} \frac{\mu_3(t_i)}{2s^4(t_i)(n-i-1)} - \frac{y_{i+1}^2 + \sigma^2}{2s^2(t_i)(n-i-1)} \right)$$

avec $s^2(t)$ et $\mu_3(t)$ les dérivées deuxième et troisième de la fonction génératrice des cumulants.

- g_{na} est une légère modification de π^{m_i} .
- Inversion de la fonction m .
- Simulation :
 - ▶ a_n est telle que $\lim_{n \rightarrow \infty} a_n = E[\mathbf{X}]$.
 - ▶ a_n est telle que $\lim_{n \rightarrow \infty} a_n \neq E[\mathbf{X}]$.

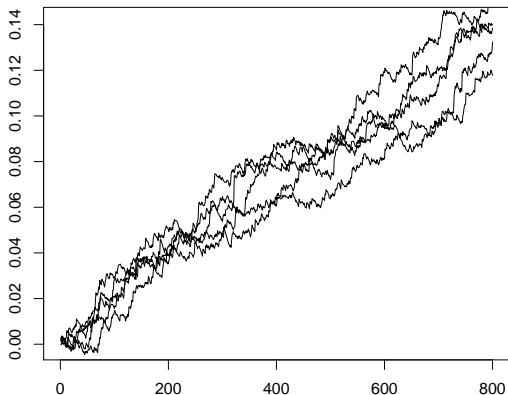
Trajectoires typiques

- $X_1, \dots, X_n \sim \mathcal{N}(0, 1)$.
- $n = 1000$, $k = n - 1$ et $P_n = 10^{-2}$.



Trajectoires typiques

- $X_1, \dots, X_n \sim \mathcal{E}(0, 1)$.
- $n = 1000$, $k = ???$ et $P_n = 10^{-8}$.



Choix de k

Choix de k

Le paramètre clé dans cette approximation est k qui représente la longueur maximale de la sous-trajectoire pour laquelle l'approximation est bonne à $\alpha\%$ près. Afin de déterminer la valeur maximal de k possible, on détermine un intervalle de confiance à deux sigma pour l'erreur relative :

$$CI(k) := [ERE(k) - 2\sqrt{VRE(k)}, ERE(k) + 2\sqrt{VRE(k)}]$$

avec

$$ERE(k) := E_{g_{na}} \left[\frac{p_{na}(Y_1^k) - g_{na}(Y_1^k)}{g_{na}(Y_1^k)} \right].$$

et

$$VRE(k) := Var_{g_{na}} \left[\frac{p_{na}(Y_1^k) - g_{na}(Y_1^k)}{g_{na}(Y_1^k)} \right].$$

Choix de k

Choix de k

On réécrit ces quantités sous la forme

$$\begin{aligned} VRE(k)^2 &= E_p \left(\frac{g_{na}^3(Y_1^k)}{p_{na}(Y_1^k)^2 p(Y_1^k)} \right) \\ &\quad - E_p \left(\frac{g_{na}^2(Y_1^k)}{p_{na}(Y_1^k) p(Y_1^k)} \right)^2 \end{aligned}$$

Lemme (Jensen (1995))

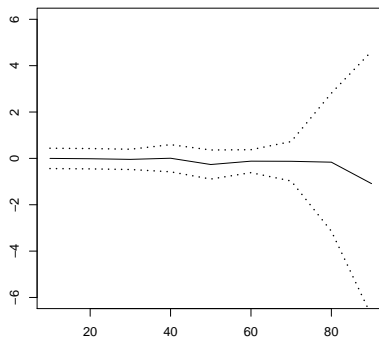
Soit $\mathbf{X}_1, \dots, \mathbf{X}_n$ un échantillon i.i.d. de densité p sur \mathbb{R} satisfaisant la condition de Cramer avec f.g.m. ϕ . Alors avec $m(t) = a$ et, quand $|u|$ est bornée, on a

$$p(\mathbf{S}_{1,n}/n = a) = \frac{\sqrt{n}\phi^n(t) \exp -nta}{s(t)\sqrt{2\pi}} (1 + o(1))$$

Et on utilise une méthode Monte-Carlo.

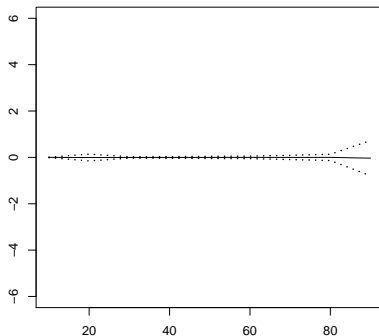
Graphe : CI Empirique

- $\mathbf{X}_1^n \sim \mathcal{E}(1)$
- $(\mathbf{S}_{1,n} = \mathbf{X}_1 + \dots + \mathbf{X}_n)$
- Trait pointillée : $\overline{CI}(k)$
- Trait plein : $\overline{ERE}(k)$



Exemple : CI Théorique

- $\mathbf{X}_1^n \sim \mathcal{E}(1)$
- $(\mathbf{S}_{1,n} = \mathbf{X}_1 + \dots + \mathbf{X}_n)$
- Trait pointillée : $CI(k)$
- Trait plein : $ERE(k)$



Généralisation du conditionnement

- $\mathbf{X} \in \mathbb{R}^d$ de densité $p_{\mathbf{X}}$
- $\mathbf{X}_1^n := (\mathbf{X}_1, \dots, \mathbf{X}_n)$ i.i.d.
- $u : \mathbb{R}^d \rightarrow \mathbb{R}^s$ mesurable avec $d, s \geq 1$.
- $\mathbf{U}_{1,n} = \sum_{i=1}^n u(\mathbf{X}_i)$.
- $\mathbf{U} := u(\mathbf{X})$ a une densité $p_{\mathbf{U}}$
- $u_{1,n}/n$ suite convergente.
- Fonction caractéristique de $\mathbf{U} = u(\mathbf{X}) \in L^r$ pour $r \geq 1$.

$$(\mathbf{U}_{1,n} := u_{1,n})$$

Théorème

Sous les hypothèses équivalentes aux théorèmes réels,

- *Soit Y_1^k un échantillon de loi $P_{u_{1,n}}$. Alors*

$$p(\mathbf{X}_1^k = Y_1^k | \mathbf{U}_{1,n} = u_{1,n}) = g_{u_{1,n}}(Y_1^k)(1 + o_{P_{u_{1,n}}}(1 + \varepsilon_n(\log n)^2))$$

- *Soit Y_1^k un échantillon de loi $G_{u_{1,n}}$. Alors*

$$p(\mathbf{X}_1^k = Y_1^k | \mathbf{U}_{1,n} = u_{1,n}) = g_{u_{1,n}}(Y_1^k)(1 + o_{G_{u_{1,n}}}(1 + \varepsilon_n(\log n)^2))$$

Trajectoires dans un hyperplan

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right).$$

- a, b et c constantes.
- $u(x, y) = ax + by$

$$\left(\sum_{i=1}^n (aX_i + bY_i) = nc \right)$$

Trajectoires dans un hyperplan

Loi Tiltée

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \frac{ac}{a^2+b^2} \\ \frac{bc}{a^2+b^2} \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right).$$

Approximation par g_{nc}

- Densité conditionnelle exacte et g_{nc} coïncident sans terme d'erreur.

Approximation de p_{nA} dans le cas $A = (a, +\infty)$

On rappelle

$$p_{nA}(x_1^k) = \int_a^\infty p_{nv}(\mathbf{x}_1^k = x_1^k) p(\mathbf{U}_{1,n}/n = v | \mathbf{U}_{1,n} > na) dv$$

On approxime la loi de $p(\mathbf{U}_{1,n}/n = v | a < \mathbf{U}_{1,n}/n < a + c)$ par

$$\frac{nm^{-1}(a) (\exp(-nm^{-1}(a)(v-a))) \mathbb{1}_{(a, a+c)}(v)}{1 - \exp(-nm^{-1}(a)c)}$$

où $c = c_n$ telle que $\lim_{n \rightarrow \infty} c_n = 0$.

Approximation de p_{nA} dans le cas $A = (a, +\infty)$

Théorème

Si Y_1^k est un vecteur aléatoire généré sous p_{nA} , sous les hypothèses mentionnées ci-dessus et sous d'autres conditions techniques, alors

$$p_{nA}(Y_1^k) = g_{nA}(Y_1^k)(1 + o_{p_{nA}}(\delta_n))$$

où

$$\delta_n := \max \left(\varepsilon_n (\log n)^2, (\exp(-ncm^{-1}(a)))^\delta \right).$$

pour tout $\delta < 1$.

Integration

- A union d'intervalles.
- Simulation de v avec Metropolis-Hastings selon

$$p\left(\frac{\mathbf{U}_{1,n}}{n} = v \mid \mathbf{U}_{1,n} \in nA\right)$$

car

$$r(v, v') := \frac{p(\mathbf{U}_{1,n}/n = v \mid \mathbf{U}_{1,n} \in nA)}{p(\mathbf{U}_{1,n}/n = v' \mid \mathbf{U}_{1,n} \in nA)}$$

est indépendante de $P(\mathbf{U}_{1,n} \in nA)$. Evidemment, la densité de proposition doit avoir comme support A .

Forme finale de la densité d'Importance Sampling

Densité d'IS

On propose comme densité d'échantillonnage

$$g_{nA}(Y_1^n) = g_{nA}(Y_1^k) \prod_{i=k+1}^n \pi^{m_k}(Y_i)$$

où m_k dépend de $S_{1,k}$.

Estimateur IS

On propose estimateur pour $P_n = P(\mathbf{U}_{1,n} \in nA)$

$$\hat{P}_n := \frac{1}{L} \sum_{l=1}^L \frac{\prod_{i=0}^{k-1} p(Y_{i+1}(l))}{g_{nA}(Y_1^k(l))} \prod_{i=0}^{k-1} \frac{p(Y_{i+1}(l))}{\pi^{m_k}(Y_{i+1}(l))} \mathbf{1}_{nA}(U_{1,n}(l))$$

avec $Y_1^n(l)$ simulé de manière i.i.d. sous g_{nA} . Les k premières coordonnées ne sont pas i.i.d..

Propriété de l'IS dans le cas $A = (a, +\infty)$

Erreur Relative dans le cas IS indépendant

L'IS classique est défini par L simulations d'un échantillon de taille n i.i.d. $X_1^n(j)$, $1 \leq j \leq L$, sous la densité tiltée π^{a_n} de façon non adaptative. L'erreur relative de l'estimateur \overline{P}_n est donnée par

$$RE(\overline{P}_n) := \frac{\text{Var} \overline{P}_n}{\overline{P}_n^2} = \frac{\sqrt{2\pi}\sqrt{n}}{L} a_n(1 + o(1))$$

Réduction de la Variance

L'erreur relative pour un estimateur fabriquée avec notre méthode :

$$RE(\widehat{P}_n) = \frac{\sqrt{2\pi}\sqrt{n-k-1}}{L} a_n(1 + o(1))$$

Exemple dissymétrique d -dimensionnelle

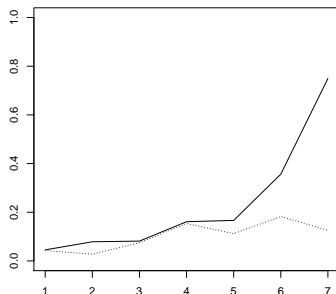
- Dupuis and Wang (2004) et Glasserman and Wang (1997).
- X_1, \dots, X_{100} échantillon i.i.d. dans \mathbb{R}^d de densité $\mathcal{N}_d(0.05, I_d)$.
- $B := (\mathcal{E}_{100})^d$ où

$$\mathcal{E}_{100} := \left\{ x_1^{100} : \frac{|x_1 + \dots + x_{100}|}{100} > 0.28 \right\}.$$

- $P_{100} = P[B]$.

Exemple dissymétrique d -dimensionnelle

- Point dominant : $(0.28, \dots, 0.28)$.
- $P_{100} = 10^{-2d}$.
- $L = 1000$.



Objectif

On veut estimer $P[h(\mathbf{X}_1^n) \in B_n]$ pour une fonction h mesurable et un ensemble mesurable B_n .

- On suppose qu'on peut trouver une fonction u et une suite a telles que

$$B_n \subset (u(\mathbf{X}_1) + \dots + u(\mathbf{X}_n) > na)$$

- On propose alors de prendre la loi sous optimal correspondant au conditionnement par $(u(\mathbf{X}_1) + \dots + u(\mathbf{X}_n) > na)$
- Puis d'utiliser notre théorème d'approximation, en simulant des échantillons \mathbf{X}_1^n de densité g_{nA} , pour fabriquer un estimateur d'Importance Sampling.

Cadre de l'étude.

Famille exponentielle de plein rang.

Une famille de lois $\{P_{\theta,\eta}, (\theta,\eta) \in \Theta\}$ est une famille exponentielle de dimension 2 si $P_{\theta,\eta}$ a une densité de la forme

$$p_{\theta,\eta}(x) := \frac{dP_{\theta,\eta}(x)}{dx} = \exp[\theta u(x) + \eta t(x) - K(\theta_1, \theta_2)] h(x)$$

Θ est un ensemble convexe de \mathbb{R}^2

Famille exponentielle courbe.

S'il existe $f : \mathbb{R} \rightarrow \mathbb{R}$ non linéaire telle que $f(\theta) = \eta$, la famille exponentielle est courbe.

Exhaustivité conservée

Objectif

Il faut vérifier si l'approximation qu'on propose garde l'exhaustivité.

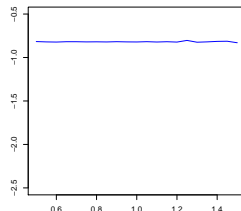
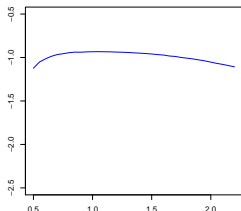
Cas Gamma

On considère la densité Gamma

$$f_{r,\theta}(x) := \frac{\theta^{-(\eta+1)}}{\Gamma(\eta+1)} x^\eta \exp -x/\theta \quad \text{for } x > 0. \quad (1)$$

- Paramètres canoniques : θ et η ,
- Statistiques exhaustives minimales : $u(x) := x$ et $t(x) := \log x$.
- $X_1^n := (X_1, \dots, X_n)$ i.i.d. de densité f_{θ_T, η_T} .
- $T_{1,n} := \log X_1 + \dots + \log X_n$
- $U_{1,n} := X_1 + \dots + X_n$.

Graphes



Evaluations des approximations des vraisemblances conditionnelles

- sur X_1^k en fonction de θ avec η_T connu (à gauche).
- sur X_1^k en fonction de η avec θ_T connu (à droite).

Bootstrap paramétrique

Définition (Efron (1979))

On suppose que les données X_1^n sont de loi p_θ . On estime par maximum de vraisemblance $\hat{\theta}$. Le bootstrap paramétrique permet la simulation d'échantillons Y_1^n sous la loi $p_{\hat{\theta}}$. On infère sur les quantités considérées en utilisant ces répliques.

Dans les familles de plein rang

Lockhart et O'Reilly (2005) propose de comparer la loi d'un sous-échantillon conditionnée à une statistique exhaustive à la loi induite par bootstrap paramétrique. On considère une famille exponentielle de plein rang de statistique exhaustive $U_{1,n}$

Théorème

On suppose que la vraie valeur du paramètre θ_0 est dans l'intérieur de Θ . On suppose qu'il existe un entier r et un voisinage \mathcal{N} de θ_0 tel que $U_{1,r}$ a une densité bornée par rapport à la mesure de Lebesgue pour tout $\theta \in \mathcal{N}$. Alors pour tout entier fixé m et pour $\delta > 0$,

$$\lim_{n \rightarrow \infty} n^{1-\delta} \sup_H |P_{\hat{\theta}}((X_1, \dots, X_m) \in H) - P((X_1, \dots, X_m) \in H | U_{1,n})| \rightarrow 0$$

presque surement. Le supremum est pris sur tous les ensembles boréliens H de \mathbb{R}^m .

Distance en variation totale

Théorème

La variation totale entre $P_{u_{1,n}}$ et $G_{u_{1,n}}$ tends vers 0 quand n tends vers l'infini et

$$\lim_{n \rightarrow \infty} \int |p_{u_{1,n}}(y_1^k) - g_{u_{1,n}}(y_1^k)| dy_1^k = 0.$$

Estimation par vraisemblance conditionnelle

Exemple : Sundberg(2010)

Soit X et Y deux r.v. telle que

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \mathcal{N} \left(\begin{pmatrix} \psi_T \\ \psi_T^2 \end{pmatrix}, \begin{pmatrix} \sigma_T^2 & 0 \\ 0 & \sigma_T^2 \end{pmatrix} \right)$$

Si σ_T^2 inconnu,

- Paramètres canoniques : $1/\sigma_T^2$, $(-2\psi_T/\sigma_T^2)$ et $(-2\psi_T^2/\sigma_T^2)$.
- Statistiques exhaustives minimales : $(X^2 + Y^2)$, X et Y .

Estimation par vraisemblance conditionnelle

Soit (X_i, Y_i) , $1 \leq i \leq n$ un échantillon i.i.d. de la distribution précédente avec $\psi_T = 2$ et $\sigma_T^2 = 1$. On suppose ψ_T paramètre de nuisance et σ_T^2 paramètre d'intérêt, tous les deux inconnus.

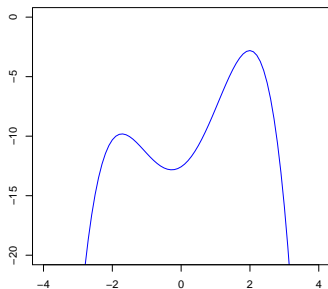
Vraisemblance multimodale

On calcule la vraisemblance par rapport à ψ sur l'échantillon précédent et on obtient

$$(U_{1,n} - \psi) + 2\psi(V_{1,n} - \psi^2) = 0$$

avec $U_{1,n} := X_1 + \dots + X_n$ and $V_{1,n} := Y_1 + \dots + Y_n$.

Estimation par vraisemblance conditionnelle

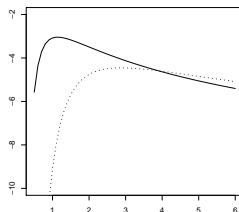
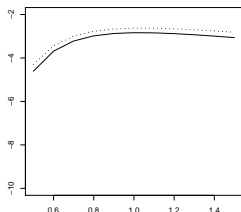


Estimation de ψ_T en utilisant la méthode de Newton-Raphson. Soit ψ_0 le point de départ.

- Si $\psi_0 > 1$, convergence vers $\psi_T = 2$.
- Si $\psi_0 < 1$, convergence vers -2 .

Vraisemblance par rapport à ψ .

Estimation par vraisemblance conditionnelle



Vraisemblances

- Trait pointillé : vraisemblance de X_1^n par rapport à σ^2 sous $\hat{\psi}$.
- Trait plein : approximation de la vraisemblance de X_1^n par rapport à σ^2 conditionnée à $(U_{1,n}, V_{1,n})$.

Théorème de Rao-Blackwell

Enoncé classique

Si δ est un estimateur sans biais et S une statistique exhaustive alors l'estimateur $E[\delta|S]$ a une variance plus faible que la variance de l'estimateur initial pour la même espérance.

Explication

Loi de l'espérance totale :

$$E[E[\delta|S]] = E[\delta] \quad (2)$$

Loi de la variance totale :

$$\text{Var}[\delta] = \text{Var}[E[\delta|S]] + E[\text{Var}[\delta|S]] \quad (3)$$

Théorème de Rao-Blackwell

On doit prouver que remplacer la loi conditionnelle par notre approximation est valide.

Théorème

Soit k fixé et $\lim_{n \rightarrow \infty} (u_{1,n}/n) = E[t(\mathbf{X})]$.

$$\lim_{n \rightarrow \infty} KL(p_{u_{1,n}}(x_1^k), \pi^{u_{1,n}/n}(x_1^k)) = 0$$

En utilisant le Lemme 3.1 de Csiszar (1973),

$$\lim_{n \rightarrow \infty} \left(\int t(x_1^k) p_{u_{1,n}}(x_1^k) dx_1^k - \int t(x_1^k) \pi^{u_{1,n}/n}(x_1^k) dx_1^k \right) = 0$$

pour une fonction $t(x_1^k)$ qui satisfait

$$\int \exp(rt(x_1^k)) \pi^{E[t(\mathbf{X})]}(x_1^k) dx_1^k < \infty$$

dans un voisinage non vide de 0.

Théorème de Rao-Blackwell

Conjecture

Sous les conditions du théorème principal et d'autres conditions additionnelles,

$$\lim_{n \rightarrow \infty} KL(p_{u_1,n}(x_1^k), g_{u_1,n}(x_1^k)) = 0$$

En utilisant le Lemme 3.1 de Csiszar (1973),

$$\lim_{n \rightarrow \infty} \left(\int t(x_1^k) p_{u_1,n}(x_1^k) dx_1^k - \int t(x_1^k) g_{u_1,n}(x_1^k) dx_1^k \right) = 0$$

pour une fonction $t(x_1^k)$ qui satisfait

$$\lim_{n \rightarrow \infty} \int \exp(rt(x_1^k)) p_{u_1,n}(x_1^k) dx_1^k < \infty$$

dans un voisinage non vide de 0.

Rao-Blackwellisation

Explications

$U_{1,n}$ est exhaustive pour le paramètre θ dans $g_{U_{1,n}}$, on peut utiliser cette densité pour améliorer les estimateurs de θ_T par Rao Blackwellization. Dans le modèle gamma précédent, on propose d'estimer θ .

Estimateur initial

Soit X_1^n un échantillon de densité f_{η_T, θ_T} . Un premier estimateur non biaisé est choisi

$$\hat{\theta}_2 := \frac{X_1 + X_2}{2\eta_T}.$$

Rao-Blackwellisation

Estimateur Rao-Blackwellisé

La version Rao-Blackwellisé de $\hat{\theta}_2$ est défini par

$$\theta_{RB,2} := E \left(\hat{\theta}_2 \middle| U_{1,n} \right)$$

dont la variance est plus petite que la variance de l'estimateur original.

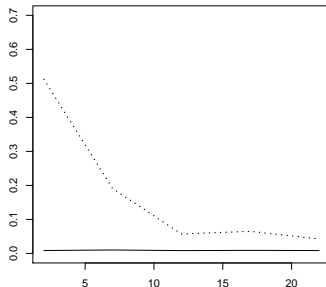
Graphes

On considère $k = 2$ dans $g_{u_{1,n}}(y_1^k)$. On simule $L (Y_1, Y_2)$ selon $g_{u_{1,n}}(y_1^2)$ pour $u_{1,n}$ fixé et on fabrique

$$\hat{\theta}_{RB,2} = \frac{1}{L} \sum_{l=1}^L \hat{\theta}_2(Y_1, Y_2)$$

On itère la procédure M fois pour estimer la variance.

Rao-Blackwellisation



Trait pointillé : Variance de l'estimateur initial. Trait plein : Variance de l'estimateur Raoblackwellisé.

Théorème de Rao-Blackwell : Perspective

Enoncé

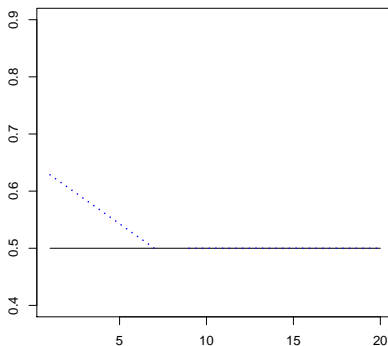
On réduit toujours la variance par conditionnement.

- Soit X_1^n un échantillon de densité p_{θ_T} .
- Soit $\hat{\theta} = f(X_1, \dots, X_k)$ un estimateur préliminaire de θ basé sur k termes.
- Soit $T_{1,n} = X_1 + \dots + X_n$
- On veut améliorer $\hat{\theta}$ en écrivant

$$\theta_{RB} = E_{\hat{\theta}}[f(Y_1, \dots, Y_k) | T_{1,n} = t_{1,n}]$$

- Soit $\tilde{\theta} = f(X_1, \dots, X_n)$ un estimateur de θ basé sur n termes.

Théorème de Rao-Blackwell : Perspective



- Trait plein : Variance de $\tilde{\theta}$ calculé sur n termes.
- Pointillé : Variance de θ_{RB} calculé sur k termes.

- Cas où \mathbf{X}_1^n sont indépendants et pas nécessairement de même loi.
 - ▶ Développements d'Edgeworth → résultats d'approximation similaires.
 - ▶ Applications aux modèles de regression linéaire ou logistique. Amélioration d'estimateurs par Rao-Blackwellisation.
 - ▶ Etude des déviations en moyenne mobile.
- Cas des chaines de Markov.

MERCI POUR VOTRE ATTENTION