# Bootstrap

1er juin 2014

Ci-dessous, je comprends du sujet.

On considre un n-chantillon $\boldsymbol{X}^n = (X_1^n, ..., X_n^n)$ indpendantes et identiquement distribues suivant une loi dont on note $F_X$ la fonction de rpartition. On note le maximum $M_n = \max\{X_1, ..., X_n\}$.

Cette loi appartient au domaine d'attraction de Gumbel. Autrement dit, il existe une suite $(a_n, b_n)$ telle que

$$\frac{M_n - b_n}{a_n} \xrightarrow{L} G \tag{1}$$

o $G$ est la loi de Gumbel dont la fonction de rpartition $F_G$ s'crit $\forall x \in \mathbb{R}$

$$F_G(x) = 1 - e^{-e^{-x}} \tag{2}$$

Ce domaine est trs grand et de nombreuses lois classiques y appartiennent : exponentiel, gamma, logistique, log-normale, normale, ...

Dans cette situation, il a t montr par de Haan qu'on peut choisir la suite $(a_n, b_n)$ comme suit

$$a_n = F_X^{-1}\left(1 - \frac{1}{en}\right) - F_X^{-1}\left(1 - \frac{1}{n}\right) \qquad b_n = F_X^{-1}\left(1 - \frac{1}{n}\right) \tag{3}$$

L'objectif de l'tude est de dterminer la distribution de $M_n$  partir du seul chantillon $\boldsymbol{X}^n$. En dsignant par $P$ une loi quelconque, le paramtre d'intrt est donc $TP = P(M_n < x) = E_P[1(M_n < x)]$ $\forall x \in \mathbb{R}$.

La distribution (exacte) de $M_n$ s'crit comme suit

$$T(P) = P(X_1 < x, ..., X_n < x) = [F_X(x)]^n \tag{4}$$

Asymptotiquement, comme la loi appartient au domaine d'attraction de Gumbel

$$P\left(\frac{M_n - b_n}{a_n} < x\right) \simeq F_G(x) \quad \Longrightarrow \quad TP \simeq F_G(a_n x + b_n) \tag{5}$$

Par commodit, on considre donc aussi $\tilde{T}P = P\left(\frac{M_n - b_n}{a_n} < x\right)$.

La distribution bootstrap d'Efron de $M_n$ s'crit comme suit

$$P_n^* = \left[1 - \left(\frac{n-1}{n}\right)^n\right]\delta_{X_{(n)}} + \ldots + \left[\left(\frac{k}{n}\right)^n - \left(\frac{k-1}{n}\right)^n\right]\delta_{X_{(k)}} + \ldots + \left(\frac{1}{n}\right)^n \delta_{X_{(1)}} \tag{6}$$

Fukuchi a montr qu'asymptotiquement cette loi convergeait vers un processus stochastique fonction du tirage effectu. Faut-il dvelopper ce point ou utiliser un autre argument ?

Si on tire seulement $m$ avec $m < n$, on a

$$P_{m|n}^* = \left[1 - \left(\frac{n-1}{n}\right)^m\right]\delta_{X_{(n)}} + \ldots + \left[\left(\frac{k}{n}\right)^m - \left(\frac{k-1}{n}\right)^m\right]\delta_{X_{(k)}} + \ldots + \left(\frac{1}{n}\right)^m \delta_{X_{(1)}} \tag{7}$$

Fukuchi montre que $m = o(n)$, $\tilde{T}P_{m|n}^*$ converge vers $F_G$ avec la distance de Kolmogorov. Comme on ne connat pas $a_n$ et $b_n$, on peut l'estimer comme suit et travailler par analogie.

$$\hat{a}_n = \hat{F}_X^{-1}\left(1 - \frac{1}{en}\right) - \hat{F}_X^{-1}\left(1 - \frac{1}{n}\right) \qquad \hat{b}_n = \hat{F}_X^{-1}\left(1 - \frac{1}{n}\right) \tag{8}$$

o $\hat{F}_X$ est la distribution empirique construite  partir de l'chantillon $\boldsymbol{X}^n$. C'est la version m out of n qui a priori fonctionne pour la distribution de $M_n$.

Je ne vois pas la diffrence entre le bootstrap sous-chantillonn et le m out of n. Sais-tu quelle est elle ?

A priori, la vitesse de convergence est donne par $a_n$. On peut alors faire une rgression pour trouver $a_n$ sous la forme $n^\alpha$. Il y a un article de Bertail sur le sujet : on subsampling estimators with unknown rate of convergence. En regardant rapidement, l'article donne les preuves que a fonctionne sous condition de choisir les points o on regarde l'volution de la courbe lorsque $m$ varie.

La fonction de rpartition $\tilde{T}P_{m|n}^*$ est en escalier. On pourrait donc acclrer la convergence en la lissant par exemple  l'aide d'un noyau.