

Bootstrap

1^{er} juin 2014

Ci-dessous, je comprends du sujet.

On considère un n-échantillon $\mathbf{X}^n = (X_1^n, \dots, X_n^n)$ indépendantes et identiquement distribuées suivant une loi dont on note F_X la fonction de répartition. On note le maximum $M_n = \max\{X_1, \dots, X_n\}$.

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

$$\frac{M_n - b_n}{a_n} \xrightarrow{L} G \quad (1)$$

où G est la loi de Gumbel dont la fonction de répartition F_G s'écrit $\forall x \in \mathbb{R}$

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

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

Dans cette situation, il a été montré par de Haan qu'on peut choisir la suite (a_n, b_n) comme suit

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

L'objectif de l'étude est de déterminer la distribution de M_n à partir du seul échantillon \mathbf{X}^n . En désignant par P une loi quelconque, le paramètre d'intérêt 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'écrit comme suit

$$T(P) = P(X_1 < x, \dots, X_n < x) = [F_X(x)]^n \quad (4)$$

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

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

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

La distribution bootstrap d'Efron de M_n s'écrit comme suit

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

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

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

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

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

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

où \hat{F}_X est la distribution empirique construite à partir de l'échantillon \mathbf{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 différence entre le bootstrap sous-échantillon et le m out of n. Sais-tu quelle est elle ?

A priori, la vitesse de convergence est donnée par a_n . On peut alors faire une régression pour trouver a_n sous la forme n^α . 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 ça fonctionne sous condition de choisir les points où on regarde l'évolution de la courbe lorsque m varie.

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