Bootstrap et méthodes de rééchantillonnage
P. Bertail

Dans chaque mémoire, Il s'agit d'étudier, dans un modèle donné avec une certaine loi sous-jacente, F paramétrique ou non paramétrique, les propriétés d'un estimateur T_n d'un paramêtre d'intérêt $\theta(F)$. Le but est essentiellement de construire la ou les distributions bootstrap adaptées au modèle et de comparer ses propriétés avec la distribution exacte, la distribution asymptotique, eventuellement le développement d'Edgeworth quand il est disponible etc... Nous suggérons dans la suite quelques modèles et quelques type de comparaison, mais les étudiants peuvent également choisir un modèle qui les intéresse (par exemple issu de leur mémoires de GT) ou une statistique dont la loi est particulièrement difficile à estimer (par exemple la loi du maximum, ou d'une étendue).

Dans chacun de ces modèles, il est demandé de :

1. Rappeler brièvement les propriétés théoriques (asymptotiques ou non) disponibles pour l'estimateur

proposé ;

- sa distribution d'échantillonnage, biais, variance quand c'est possible.

- étudier la possibilité de construire un intervalle de confiance pour θ par une méthode classique (exact, asymptotique...) ;

- imaginer un scénario de test et proposer une statistique de test avec évaluation de la p-valeur.

2. Proposer un algorithme de type bootstrap permettant de répondre par simulations aux questions

posées ci-dessus. Si possible, proposer plus plusieurs approches

- paramétrique, non paramétrique (bootstrap naif, smooth bootstrap, rééchantillonnage des résidus, sous-échantillonnage, etc...)

- percentile ou de base, t-percentile pour la construction d'intervalle de confiance

3. A partir d'un échantillon simulé de taille n = 30 et puis de taille n = 100 (voir plus pour certains modèles), illustrer numériquement les points précédents et comparer les résultats. Si vous avez des données réelles n'hesitez par à présenter les résultats pour une simulation et pour ces données. On pourra, lorsqu'elle n'est pas disponible analytiquement, simuler la distribution exacte de la statistique d'intérêt par Monte-Carlo (avec un grand nombre de tirages, L$\geq$100 00).

4. Etudier et comparer les performances des résultats théoriques et de certaines approximations de bootstrap proposées dans une expérience de Monte-Carlo (indiquer la précision des quantités estimées dans cette expérimentation). On pourra aussi

- évaluer en répétant le tirage de base, la couverture exacte des intervalles de confiance obtenus.

- évaluer la distance exacte entre la distribution bootstrap et la vraie distribution (pour une métrique de type kolmogorov, L2, L1 etc... au choix).

Votre rapport, clair et concis, doit contenir une présentation du problème et des procédures utilisées (expliquez par exemple pourquoi telle ou telle méthode n'est pas disponible ou pas adaptée pour votre sujet), ainsi qu'un commentaire sur les résultats obtenus. Le mémoire ne devra pas dépasser 20 pages (graphiques inclus). Le rapport (fichier pdf) doit être envoyé par e-mail à patrice.bertail@ensae.fr et en copie à Arbel@ensae.fr avant le 30 juin 2009, 18h00.

Suggestions de thèmes : vous pouvez vous interesser à des problèmes spécifiques (estimation, tests, prediction,. . .), dans un cadre purement i.i.d., en régression logistique, régression non-linéaire, régression non-paramétrique, modèle linéaire avec héteroscédasticité, estimation de densité, nombre de modes d'une densité, aux modèles de frontières, aux séries chronologiques, échantillons appariés, aux modèles ANOVA non-parametrique,aux modèles de survie, Value at Risk, aux extrèmes... L'originalité du thème proposé sera particulièrement appréciée. Dans la suite, quelques exemples très spécifiques sont évoqués : ils peuvent être choisis, modifiés, complétés et/ou adaptés...

Quelques exemples de problèmes spécifiques d'abord simples et de plus en plus complexes

1. Soit un échantillon i.i.d X_i de loi $Power(\beta, c)$, c.à.d. $F(x) = (x/c)^\beta$ pour $x \in [0, c]$,

$F(x) = 0$ pour $x < 0$ et $F(x) = 1$ pour $x > c$, où $c > 0$ est donné. On veut estimer

$\theta = Prob(X_1 > a)$ où $a \in]0, c[$ est donné, ou $\theta = \beta > 2$, la seconde quantité étant plus intéressante.

- proposer des estimateurs pour ces quantitées et éventuellement étudier leurs propriétés asymptotiques. Déterminer (par simulation) la distribution exacte de l'estimateur pour des tailles d'échantillon fixe données (n=30 et n=100). Voir éventuellement l'effet de c.

- proposer plusieurs type de méthode bootstrap selon que l'on suppose connue la forme de F ou non.

- construire les distributions bootstrap des estimateurs d'intérêt, les intervalles de confiances correspondants

- comparer ces distributions, les courvertures des intervalles de confiance etc...

2. Soit un échantillon i.i.d X_i de loi de Pareto (β, c), c.à.d. $F(x) = 1-(c/x)^\beta$ pour $x > c$ et

$F(x) = 0$ pour $x < c$, où $c > 0$ est donné. On veut estimer $\theta = Prob(X_1 > a)$ où

$a > c$ est donné. On souhaite également estimer α ou $\gamma = 1/\alpha$ (l'estimateur du maximum de vraisemblance de γ est appelé estimateur de Hill sur lequel il existe une très abondante littérature) . Même questions que précédemment.

3. Même problème qu'en 1 et 2 mais lorsqu'on s'intéresse non pas directement à β mais à un quantile d'ordre γ de la loi i. e. $\theta = F^{-1}(\gamma)$. On pourra proposer différents types d'estimateurs (selon que l'on connait ou pas la forme de la loi sous-jacente). Proposer la version bootstrap adaptée dans chaque cas. Etudier précisément ce qui se passe lorsque $\gamma = 75\%$, $\gamma = 90\%$, $\gamma = 1 - 1/n$ (la dernière valeur correspondant au maximum).

4. Même problème qu'en 1, 2 ou 3 mais en supposant que la loi est de la forme pareto généralisée $F(x) = 1 - (c/x)^{\beta} \log(x)^{\xi}$ selon diverses valeurs de ξ et de β. Etudier précisément l'impact de la fonction à variation sur l'estimation de β : proposer des méthodes de type bootstrap pour corriger le biais de l'estimateur de Hill, évaluer sa variance etc...

5. Etudier le bootstrap des estimateurs du maximum de vraisemblance dans un modèle de type probit

(avec résidus de la variable latente de type gaussien N(0,1)). On pourra générer les variables explicatives

uniforme sur [0,1]. Refaire les simulations en voyant ce qui se passe si la loi des résidus étaient $Student(p)$ $p > 2$.

6. Même type de questions mais sur des modèles en deux étapes type tobit ou tobit généralisés ou modèles double-hurdle.

7. Pour ceux qui ont suivis le cours de statistique non paramétrique de M. Delecroix ou de statistique semi-paramétrique de M. Bertail, choisir un des modèles étudiés en détail en cours et déterminer la distribution, exacte, asymptotique, bootstrap d'un estimateur efficace.

- par exemple dans une modèle de régression non-paramétrique avec "design" uniforme sur [0,1], étudier l'estimateur de Nadarya-Watson : proposer plusieurs type de méthodes bootstrap pour construire des intervalles de confiance ponctuel, uniforme sur [0,1]. Me demander eventuellement de la bibliographie pour les propriétés théoriques du bootstrap dans ce cadre.

- Même question mais pour un estimateur à noyau de la densité sur un compact.

- Etudier le modèle de symétrie, $X_1, ..., X_n$ i.i.d. de loi $\eta(x - \theta)$ où η est une densité symétrique. Proposer une méthode bootstrap adaptée à ce modèle et étudier la distribution bootstrap de l'estimateur efficace de θ dans ce modèle.

- Etudier un modèle contraint par des moments et étudier la distribution d'un estimateur efficace.

8. Même question qu'en 1,2...ou 7 mais avec des données eventuellement censurées par une certaine variable. Voir ce qui se passe selon que l'on suppose connue la loi de la censure ou pas. On pourra essayer de déterminer pour 1,2 une bande de confiance pour l'estimateur de Kaplan-Meier basée sur des techniques bootstrap.

9. Choisir un modèle de série temporelle simple stationnaire du type AR(p), MA(q), ARCH(p) GARCH(p,q), etc...et etudier les distributions exactes, asymptotiques, bootstrap des estimateurs des paramètres d'intéret.

Comparer differentes méthodes bootstrap disponibles en série temporelles (moving block, sive bootstrap, regenerative bootstrap quand c'est possible).

10. Même questions qu'en 9 mais avec des paramètres conduisant à des phénomènes non-stationnaires ou proche de la non-stationnarité. On pourra choisir dans ce cadre des tailles d'echantillon très grandes.

11. Combiner plusieurs modèles évoqués précédemment, par exemple modèles probit avec résidus autocorrellés, de loi de Student ou Pareto et variables explicatives retardées. Proposer une méthode bootstrap adaptée, comparer avec la distribution exacte (ne pas chercher à obtenir la distribution asymptotique dans ce cadre)...

12. Faire une étude précise de la loi du maximum dans le cas d'une loi dans le domaine d'attraction de la loi de Gumbel ou de la loi Paréto généralisée. Distribution exacte, asymptotique, bootstrap, sous-échantillonnage, m out of n bootstrap . Estimer la vitesse de convergence du maximum par des méthodes de type sous-echantillonnage. On pourra choisir dans ce cas des tailles d'échantillon grand $n > 100000$.

13. Même questions qu'en 12 mais pour des données dépendantes (par exemple AR(1)) et différentes type de méthodes bootstrap.