Année 2020-2021 **ENSAE 3A**

Bootstrap pour les séries temporelles

Introduction

Consignes

- Le sujet contient des exercices à traiter de manière séquentielle (mais je vous conseille de revenir aux questions ($\star\star$) après les simulations). Le dernier exercice est plus difficile, je le prendrai en compte en bonus.
- Le TP est à rendre sous la forme de deux notebooks (exécutés...) `R` ou `python` (aux formats html, pdf ou ipynb)
  - un premier le 5 mai
  - un second fini le 24/05
- Vous pouvez travailler en binôme, dans ce cas, envoyez moi un seul TP avec les deux noms.

Packages, simulations

Vous pouvez utiliser les fonctions suivantes en vérifiant bien la définition des coefficients qu'elles renvoient

- `ar` de `stat` dans `R` ou
- `AR` de `statsmodels.tsa.ar_model` dans `python`, voir https://www.statsmodels.org/stable/vector_ar.html

Dans les fichiers `TP_timeseries.Rmd` et `TP_timeseries.ipynb` vous trouvez des codes de simulations

- d'un processus AR(p) défini par $X_t = \sum_{j=1}^{p} \phi_j^\star X_{t-j} + \epsilon_t$
- d'un processus MA(q) défini par $X_t = \sum_{j=0}^{q} \psi_j^\star \epsilon_{t-j}$ avec $\psi_0^\star = 1$.

Dans la suite, on souhaite **comparer les différentes méthodes de ré-échantillonnage**, voir les exercices 5 à 7, sur des simulations de

- MA(1) de longueur 200 avec $\psi_1^\star = -0.6$ et (ϵ_t) i.i.d. de loi exponentielle décalée $(\mathcal{E}(1) - 1)$.
- AR(1) de longueur 200 avec $\phi_1^\star = 0.6$ et (ϵ_t) i.i.d. de loi exponentielle décalée $(\mathcal{E}(1) - 1)$.

i.i.d. bootstrap

Exercice 1

1. Coder une méthode de ré-échantillonage aléatoire avec remise dans $X_1, \ldots, X_n$.
2. Répresenter une réalisation de la série bootstrapées et de la série originale.
3. Quelles sont, à votre avis, les limites de cette méthode ?

Sieve bootstrap, in Kreiss 1992

On approxime la représentation $AR(\infty)$ par une répresentation AR(p_n) (avec $p_n \to \infty$ quand $n \to \infty$), puis on ré-échantillonne suivant la procédure suivante

1. Estimer l'ordre $\hat{p}_n$ et les coefficients associés $\hat{\phi}_{1,n}, \ldots, \hat{\phi}_{\hat{p}_n,n}$ (par la méthode de Yule-Walker ou le maximum de vraisemblance).
2. Définir les résidus
$$\widetilde{e}_t^{(\hat{p}_n)} = \hat{X}_t^{(\hat{p}_n)} - X_t = \sum_{j=1}^{\hat{p}_n} \hat{\phi}_{j,n} X_{t-j} - X_t \text{ pour } t = \hat{p}_n + 1, \ldots, n$$
les centrer en définissant
$$\widehat{e}_t^{(\hat{p}_n)} = \widetilde{e}_t^{(\hat{p}_n)} - \frac{1}{n - \hat{p}_n} \sum_{t=\hat{p}_n+1}^{n} \widetilde{e}_t^{(\hat{p}_n)}$$
3. Pour $b = 1, \ldots, B$, reconstruire
$$X_{b,\hat{p}_n+1}^{S,\star}, \ldots, X_{b,n}^{S,\star}.$$
comme un AR($\hat{p}_n$) avec les coefficients estimés $\hat{\phi}_{1,n}, \ldots, \hat{\phi}_{\hat{p}_n,n}$ et un tirage aléatoire avec remise dans les résidus $(\widehat{e}_t^{(\hat{p}_n)})_{t=p+1,\ldots,n}$.

Exercice 2

1. Coder cette méthode.
2. ($\star\star$) Quelles sont, à votre avis, ses limites ?

Block bootstrap, in Kunsch 1989

1. On commence par "circulariser" la série en posant
$$X_{n+1} = X_1, X_{n+2} = X_2, \ldots$$
2. On choisit un entier $1 < l < n$ et on pose $N = \lfloor n/l \rfloor$.
3. Pour $b = 1, \ldots, B$
   (a) on tire aléatoirement et avec remise N entiers $\nu_{b,k}$ ($k = 1, \ldots, N$) entre 1 et n
   (b) on crée les blocs $X_{\nu_{b,k}}, X_{\nu_{b,k}+1}, \ldots, X_{\nu_{b,k}+l-1}$ de longueur l
   (c) on les concatène pour créer la série bootstrapée
$$(X_{b,1}^{B,\star}, \ldots, X_{b,n}^{B,\star}) = (X_{\nu_{b,1}}, X_{\nu_{b,1}+1}, \ldots, X_{\nu_{b,1}+l-1}, \ldots, X_{\nu_{b,N}}, X_{\nu_{b,N}+1}, \ldots, X_{\nu_{b,N}+l-1}).$$

Exercice 3

1. Coder cette méthode.
2. Pourquoi exclut-on les valeurs $l = 1$ et $l = n$ a priori ?
3. ($\star\star$) Quelles sont, à votre avis, les limites de ce bootstrap ?

Stationnary bootstrap, in Politis and Romano 1994

On fait la même procédure que précédemment mais on choisit maintenant la longueur de chaque bloc de premier indice $\nu_{b,k}$ suivant une loi géométrique de moyenne l $(1 < l < n)$.

Exercice 4

1. Coder cette méthode.
2. Pour les deux dernières méthodes, à partir de la simulation du MA(1) calculer l'estimateur bootstrap de la variance de $\hat{\rho}(2)$ pour des tailles et des moyennes variant de $l = 1$ à $l = n$. Que constatez-vous ?

Maximum Entropy Bootstrap, in Vinod, López-de-Lacalle, et al. 2009

Cette méthode a été utilisée dans cet article récent Fenga 2020.

Exercice 5

($\star\star$) Coder la méthode de l'article.

Comparaison des méthodes

Exercice 6

Pour chaque méthode de bootstrap, à partir de la simulation du AR(1), créer 50 échantillons bootstrap. Estimer les ordres d'un ajustement AR de chaque échantillon, qu'observez vous ?

Exercice 7

Sur les deux simulations :

1. comparer les intervalles de confiance pour $\rho(2)$ calculés à partir des différentes méthodes de ré-échantillonnage et du TCL en terme de couverture et de longueur
2. comparer les niveaux et puissances des tests de $\mathcal{H}_0 : \rho(2) = 0$ associés.

References

[Fen20] Livio Fenga. "CoViD–19: An Automatic, Semiparametric Estimation Method for the Population Infected in Italy". In: *medRxiv* (2020).

[Kre92] Jens-Peter Kreiss. "Bootstrap procedures for AR (∞)—processes". In: *Bootstrapping and Related Techniques*. Springer, 1992, pp. 107–113.

[Kun89] Hans R Kunsch. "The jackknife and the bootstrap for general stationary observations". In: *The annals of Statistics* (1989), pp. 1217–1241.

[PR94] Dimitris N Politis and Joseph P Romano. “The stationary bootstrap”. In: *Journal of the American Statistical association* 89.428 (1994), pp. 1303–1313.

[V+09] Hrishikesh D Vinod, Javier López-de-Lacalle, et al. “Maximum entropy bootstrap for time series: the meboot R package”. In: *Journal of Statistical Software* 29.5 (2009), pp. 1–19.