Projet STA211 - Sujet 2 au choix

"Méthodes de simulation numérique statistique"

Sophie Ancelet et Merlin Keller

02 Mai 2023

Ce devoir maison peut être réalisé seul ou en binôme. Sa réalisation nécessite un ordinateur. Vous rédigerez :

— soit un fichier `RMarkdown` intégrant simultanément un rappel des questions, vos réponses écrites à ces questions et vos codes `R`

— soit un document word/pdf intégrant un rappel des questions et vos réponses écrites à ces questions ainsi qu'un fichier `R` contenant vos codes.

Attention ! Les fichiers de code transmis doivent être directement exécutables sous `R`. Vos fichiers seront à envoyer **au plus tard le vendredi 19 mai 2023** aux deux adresses suivantes : **sophie.ancelet@irsn.fr et merlin.keller@edf.fr** avec pour objet DMSTA211 suivi de votre nom (ou de vos deux noms si vous travaillez en binôme). Vos réponses doivent être systématiquement justifiées.

Ajustement d'une loi de Weibull sur des données de durée de vie d'un composant industriel, avec censures à droite

On cherche à estimer la distribution $\mathcal{P}$ de la durée de vie T d'un composant industriel (batterie de portable ou de voiture, turbine d'une centrale à énergie renouvelable, ...). On dispose pour cela d'un jeu de données de n temps de fonctionnement observés $t_1, \ldots, t_n$, où :

— pour $i = 1, \ldots, p$ t_i est un temps à défaillance, c'est-à-dire que la durée de vie T_i du i-ème composant est exactement t_i : $T_i = t_i$;

— pour $i = p+1, \ldots, n$, t_i est une censure à droite, c'est-à-dire que la durée de vie T_i du i-ème composant est au moins $T_i \geq t_i$;

On suppose de plus que les durées de vie suivent la loi de Weibull $\mathcal{W}_{\alpha,\kappa}$, de fonction de répartition :

$$\mathcal{P}[T_i \leq t|\alpha,\kappa] = F(t|\alpha,\kappa) = 1 - \exp\left(-\alpha\ t^{\kappa}\right).$$

Le but de cet exercice est donc de proposer plusieurs méthodes, fréquentistes et bayésiennes, pour estimer les paramètres α, κ à l'aide du jeu de données $t_{1:n}$, et de comparer leurs résultats.

Les données sont fournies dans le fichier `donnees_Weibull_censuree`. Le nombre de temps à défaillance observé est $p = 6$.

Simulation

1 Calculer la fonction de répartition inverse $F^{-1}(x|\alpha,\kappa)$ pour tout $x \in [0, 1]$ de la loi de Weibull, et en déduire un algorithme de simulation de cette loi basée sur l'inversion générique.

2 Coder cet algorithme au sein d'une fonction `R` qui simule `n` observations de la loi de Weibull de paramètres `alpha` et `kappa` donnés, censurées au-dessus d'un niveau `t0` donné (concrètement, on censure au-dessus de t_0

en remplaçant chaque valeur simulée T par $\min(T, t_0)$). Le programme renvoie le vecteur de données simulées, ordonné de telle sorte que les p premières observations ne soient pas censurées, ainsi que le nombre p.

Calcul de la vraisemblance

3 Montrer que la log-vraisemblance $\ell(t_{1:n}|\alpha, \kappa, p)$, dans le modèle de Weibull dont les $n - p$ dernières données sont censurées à droite, s'écrit :

$$\ell(t_{1:n}|\alpha, \kappa, p) = p(\log \alpha + \log \kappa) + (\kappa - 1) \sum_{i=1}^{p} \log t_i - \alpha \sum_{i=1}^{n} t_i^{\kappa}$$

4 Donner l'expression exacte du gradient et de la matrice hessienne de ℓ

Approche fréquentiste

5 Déterminer le maximum de ℓ en α **à** κ **fixé**, et en déduire l'expression exacte du maximum de vraisemblance conditionnel $\widehat{\alpha}_{MLE}(\kappa) := \arg\max_{\alpha} \ell(t_{1:n}|\alpha, \kappa, p)$. En déduire l'expression de la vraisemblance profilée en κ, $\ell_{prof}(t_{1:n}|\kappa, p) := \max_{\alpha} \ell(t_{1:n}|\alpha, \kappa, p)$

6 À l'aide de la fonction `optimize`, écrire un programme `R` qui prend en entrée le vecteur `t` $= (t_1, \ldots t_n)$ des durées de fonctionnement observées et le nombre `p` de données non censurées, et renvoie les estimateurs du maximum de vraisemblance de (α, κ).

Utiliser ce programme pour estimer (α, κ) par maximum de vraisemblance, à partir du jeu de données fourni.

7 À l'aide de la fonction écrite en première partie (cf. question 2), simuler 10 000 jeux de données $(t_{1:n}^{(k)}, p^{(k)})$, de même taille n que le jeu de données initial, pour la même valeur seuil t_0 et en prenant (α, κ) égaux à leurs valeurs estimées sur le jeu de données fourni (cf. question 6). Puis, pour chaque jeu de données simulé, ré-estimer :

— les paramètres (α, κ) ;

— le quantile à 60% $q_{60\%} := F^{-1}(0.6|\alpha, \kappa)$ de la loi de Weibull estimée.

On fera attention au fait que le nombre p de données non censurées varie à chaque simulation. On obtient ainsi des échantillons dit "bootstrap" $(\widehat{\alpha}_{MLE}^{(k)}, \widehat{\kappa}_{MLE}^{(k)}, \widehat{q}_{60\%})_{1 \leq k \leq 10\,000}$ représentatifs de la loi des estimateurs du maximum de vraisemblance.

8 Construire des intervalles de confiance à 95% pour α, κ et $q_{60\%}$, obtenus en considérant les quantiles empiriques d'ordre 2.5% et 97.5% des échantillons bootstrap ci-dessus.

Approche bayésienne

9 Montrer que, si la loi *a priori* sur α est la loi Gamma $\mathcal{G}(a, b)$, alors la loi *a posteriori* conditionnelle de α sachant κ, encore appelée loi conditionnelle complète et notée $\pi(\alpha|\kappa, t_{1:n}, p)$, est encore une loi Gamma, dont on précisera les hyperparamètres.

Dans la suite, on prendra de même une loi *a priori* de type Gamma pour κ, de paramètres c et d, et on utilisera le choix "faiblement informatif" suivant : $a = b = c = d = 10^{-3}$.

Calculer l'expression à une constante multiplicative près de la densité marginale *a posteriori* de β. Celle-ci correspond-elle à une loi connue ?

10 Implémenter un algorithme de Metropolis within Gibbs sous la forme d'une fonction `R` nommée `MCMC` qui va permettre d'échantillonner dans la loi jointe *a posteriori* du couple (α, κ) sachant les données $(t_{1:n}, p)$ en mettant à jour :

— le paramètre α en tirant dans sa loi conditionnelle complète trouvée à la question 9

— le paramètre κ avec l'algorithme de Metropolis-Hastings (MH), en utilisant comme loi de proposition une loi uniforme (discrète) sur $\{\kappa^{curr} - k, \kappa^{curr} + k\}$ où κ^{curr} désigne la valeur courante du paramètre N à une itération donnée et k est un paramètre de saut.

11 **Choix du saut k :** Utiliser la fonction `MCMC` précédemment implémentée pour calculer puis tracer l'évolution du taux d'acceptation associé à la mise à jour de N en fonction de différentes valeurs du paramètre k (par exemple, allant de 1 à 301 par pas de 10). Pour chaque valeur de k, on pourra faire tourner l'algorithme de Metropolis within Gibbs pendant 10 000 itérations et avec une unique chaîne de Markov pour cette étape de calibration. Quelle valeur de k vous semble la meilleure (rappel : viser un taux d'acceptation d'environ 40%) ? Vous conserverez cette valeur pour la suite.

12 Lancer à présent 3 chaînes de Markov à partir de positions initiales différentes en fixant k à la valeur précedemment choisie afin de générer 3 échantillons $((\alpha^{(1)}, \kappa^{(1)}), \ldots, (\alpha^{(G)}, \kappa^{(G)}))$ de taille $G = 20\,000$. Faites un examen visuel des chaînes de Markov obtenues et calculer la statistique de Gelman-Rubin. Identifiez-vous un problème de convergence de l'algorithme MCMC implémenté vers sa loi stationnaire ? Si oui, comment proposez-vous d'y remédier ? Combien d'itérations G_0 vous semblent a minima nécessaires pour espérer avoir atteint l'état stationnaire ?

13 Analyser les autocorrélations intra-chaînes. Qu'en pensez-vous ?

14 Supprimer les G_0 premières itérations correspondant à votre temps-de-chauffe "estimé" de l'algorithme afin de constituer votre échantillon *a posteriori*. Calculer la taille d'échantillon effective (ESS) de l'échantillon *a posteriori* constitué. Qu'en pensez-vous ? Si l'ESS vous semble trop petit, refaites tourner l'algorithme en augmentant le nombre d'itérations G jusqu'à obtenir un ESS "satisfaisant" pour bien estimer α et κ.

15 Donner les statistiques *a posteriori* et représenter les lois *a posteriori* approchées pour les paramètres inconnus de votre modèle.

16 Construire les intervalles de crédibilité à 95% pour α, κ et $q_{60\%}$ (tel que défini à la question 7), obtenu en considérant les quantiles empiriques d'ordre 2.5% et 97.5% des échantillons *a posteriori* correspondants.

17 Comparer les résultats obtenus par les deux approches, à la fois en terme de résultats numériques, et de mise en œuvre pratique. Quels avantages et inconvénients voyez-vous à chaque approche ?