Détection bayésienne d'outliers et ses applications en archéologie

JM GALHARRET, A PHILIPPE, N MERCIER

Laboratoire Jean Leray (Univ. Nantes), CRP2A (Univ. Bordeaux)

June 4, 2019

Introduction

En archéologie, quelle que soit la méthode de datation utilisée (C14, OSL, ...), on est confronté au problème des outliers:

- Erreur de mesure (laboratoire),
- Erreur de prélèvement (fouilles archéologiques).

Logiciels de modélisation chronologique :

- OxCal : modèle de mélange Bronk Ramsey (2009),
- Chronomodel : Méthode d'estimation robuste Lanos and Philippe (2017, 2018).

Notre approche :

- Identification des outliers (via le modèle robuste),
- Ré-estimation du paramètre sur le sous-échantillon.

Event model [Lanos and Philippe (2017)]

Modèle hiérarchique pour estimer l'âge A d'un événement à partir de la datation de n objets le caractérisant.

- Pour le i-ème objet d'âge A_i, le laboratoire fournit la mesure X_i avec une erreur s_i.

$$X_i|A_i, s_i \sim \mathcal{N}(A_i, s_i^2)$$

- On suppose que $A_1, \ldots, A_n$ sont contemporains de A.

$$A_i|A, \sigma \sim \mathcal{N}(A, \sigma^2)$$

- Problème lié aux outliers

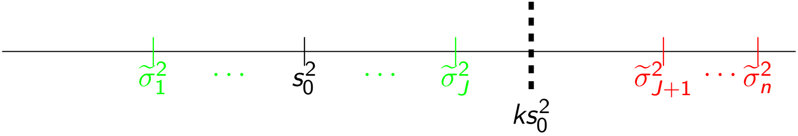
Event model :

$$A_i|A, \sigma_i \sim \mathcal{N}(A, \sigma_i^2)$$

où les σ_i^2 sont i.i.d de loi de shrinkage π_s telle que $\mathrm{med}(\sigma_i^2) = s_0^2$ où s_0^2 est la moyenne harmonique des s_i^2.

Test de détection des outliers [Galharret et al. (2019)]

Soit $\widetilde{\sigma}_i^2$ la médiane de la loi a posteriori de σ_i^2.



Règle de décision

X_i est un outlier si : $\widetilde{\sigma}_i^2 > ks_0^2$, où $k \geq 1$ est calibré pour que si l'échantillon ne contient pas d'outlier, on ne détecte qu'au plus 5% de faux positifs.

k est calibré numériquement par la méthode de Monte-Carlo :

- On simule B échantillons indépendants de taille n ne contenant aucun outlier.
- Pour tout $k \geq 1$, on calcule le pourcentage moyen $\overline{p}_{n,k}$ d'outliers détectés à tord.

$$\text{argmin}_{k \geq 1}\left(\bar{p}_{n,k} < 0.05\right).$$

Ré-estimation de A

- On ne conserve que le sous-échantillon $(X_j)_{j \in J}$ correspondant aux mesures qui n'ont pas été détectées comme outliers.
- Sur ce sous-échantillon on estime A avec (selon l'avis de l'expert):
  1. M2 : objets contemporains de l'événement

$$\begin{aligned} X_i | A_i, s_i &\sim \mathcal{N}(A_i, s_i^2) \\ A_i | A, \sigma &\sim \mathcal{N}(A, \sigma^2) \\ \sigma &\sim \pi_s \\ A &\sim \mathcal{U}[\, \underline{A}, \overline{A} \,] \end{aligned}$$

  2. M3 : âge identique.

$$\begin{aligned} X_i | A, s_i &\sim \mathcal{N}(A, s_i^2) \\ A &\sim \mathcal{U}[\, \underline{A}, \overline{A} \,] \end{aligned}$$

Application sur données simulées

On simule sous le modèle M_3 des échantillons de taille n et d'âge commun $A = 0$ (sans perte de généralité) contaminés par τ % d'outliers selon le modèle :

$$X_i \sim (1-\tau)\mathcal{N}(0, s_i^2) + \tau\mathcal{N}(15, s_i^2) \tag{1}$$

On va comparer les résultats numériques des modèles :

- Event model (M_1): $\forall\ i = 1, ..., n :\ \sigma_i^2$ individuels.
- M_2: $\forall\ i \in J :\ \sigma_i^2 = \sigma^2$
- M_3: $\forall\ i \in J :\ \sigma_i^2 = 0$

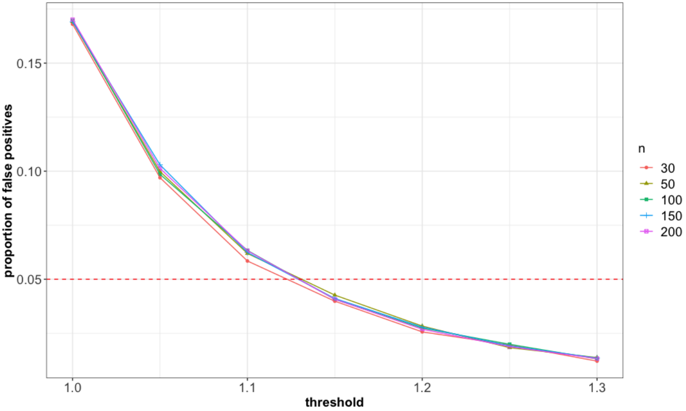


Figure: % moyen de faux positifs $\overline{p}_{n,k} = \sum_{b=1}^{B} p_{n,k}^{(b)}/B$ détectés pour $B = 1000$ réplications indépendantes sous $\mathcal{H}_0$ en fonction de la taille n de l'échantillon et de $k \geq 1$.

Comparaison des approches pour l'estimation de l'âge

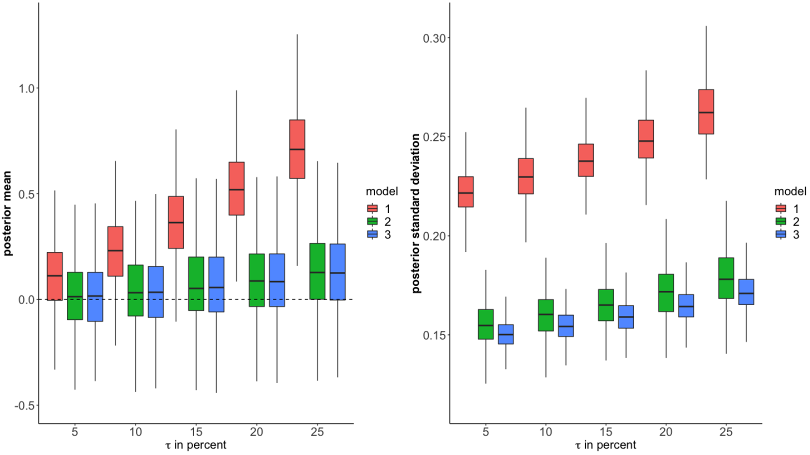


Figure: Boxplot des moyennes (gauche) et des écarts types (droite) a posteriori en fonction du taux de contamination τ.

Application à la datation par luminescence

Relation fondamentale

$$D \overset{\mathcal{L}}{=} A\dot{d} \qquad (2)$$

où $\dot{d}$ est le débit de dose, D est la dose équivalente absorbée et A l'âge cible.

- Le débit de dose $\dot{d}$ n'est pas observable,
- On sait simuler des échantillons $\dot{d}$ avec une erreur systématique $\dot{\varepsilon}$.
- Modélisation de $\dot{d}$:

$$\widetilde{\dot{d}} = \dot{d} + \dot{\varepsilon} \sim \mathcal{C}(\dot{\mu} + \dot{\varepsilon}, \dot{\sigma}^2), \qquad \dot{\varepsilon} \sim \mathcal{N}(0.10\dot{\mu}, 0.01\dot{\mu}^2)$$

- La relation (2) devient

$$D \sim \mathcal{C}\left(A(\dot{\mu} - \dot{\varepsilon}), A^2\dot{\sigma}^2\right)$$

Modèle d'âge

Les doses absorbées $(D_j)_{j\in\{1,\ldots,n\}}$ sont mesurées avec une erreur gaussienne de variance connue :

$$\widetilde{D}_j \sim \mathcal{N}(D_j, s_j^2),\ j \in \{1, \ldots, n\}$$

où $\widetilde{D}_j$ désigne la dose absorbée mesurée.

$$\begin{aligned} D_j &\sim \mathcal{C}\left(A_j(\dot{\mu} - \dot{\varepsilon}), A_j^2\dot{\sigma}^2\right) \\ A_j &\sim \mathcal{N}(A, \sigma_j^2) \\ \sigma_j^2 &\sim \mathcal{S}(3, S_0^2) \\ A &\sim \mathcal{U}[\underline{A}, \overline{A}] \end{aligned}$$

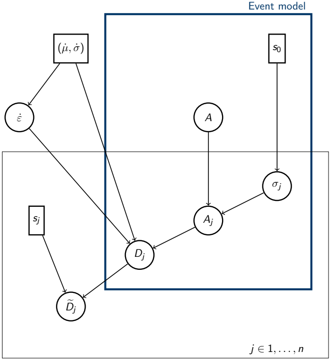


Figure: DAG of the model

Calibration du seuil

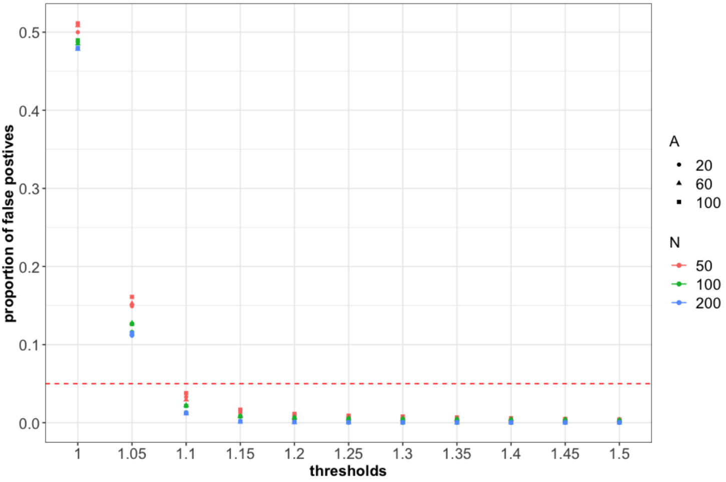
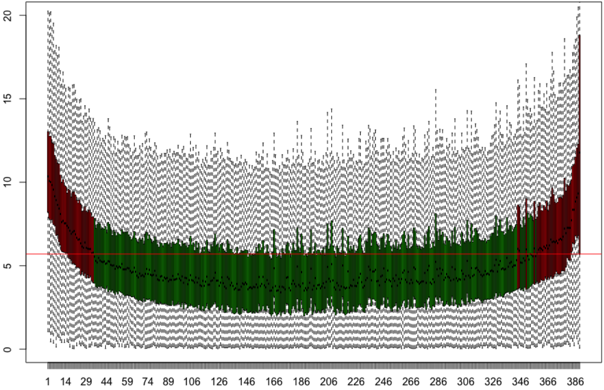


Figure: Calibration du seuil k en fonction de A, n.

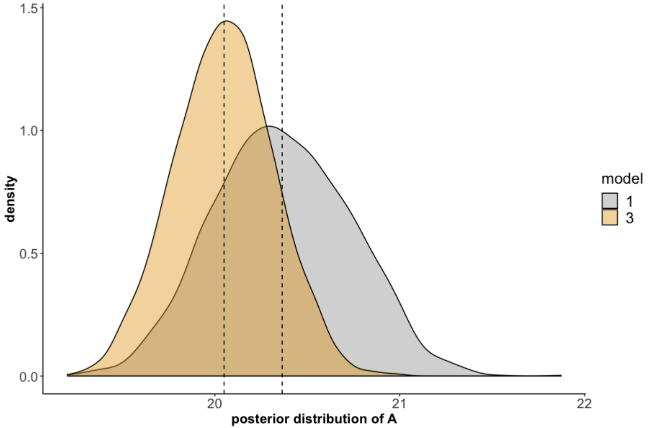
Analyse de données

On considère un échantillon de $n = 389$ doses absorbées. Boxplot des distributions a posteriori de σ_j ordonnées selon la médiane a posteriori des âges A_j. La couleur rouge indique les outliers détectés (18 %). La ligne rouge correspond à $1.1 s_0$.



Comparaison des estimations

Comparaison des lois a posteriori de A : en or l'estimation sans les outliers ($\widehat{A} = 20.08,\ 95\%\text{-CI} = [19.58, 20.64]$) et en gris celle avec le modèle robuste 20.54 and $\widehat{A} = 20.54,\ 95\%\text{-CI} = [19.77, 21.34]$)



Pour valider le choix du modèle on va comparer :

- La fonction de répartition empirique de $(D_j)_{j\in J}$ (notée F_D)
- La fonction de répartition de $A\dot{d}$ (notée $F_{A\dot{d}}$).

F_D et $F_{A\dot{d}}$ dépendent des paramètres inconnus $(D_j)_{j\in J}$ et $(A, \dot{\varepsilon})$.
On calcule leur estimateur de Bayes:

$$\mathbb{E}\left(F_D(t) \middle| (\widetilde{D}_j)_{j\in J}\right) = \frac{1}{|J|}\sum_{j\in J} F_{D_j|\widetilde{D}}(t).$$

$$\mathbb{E}(F_{A\dot{d}}(t)|(\widetilde{D}_j)_{j\in J}) = \mathbb{E}(\dot{G}\left(\frac{t}{A} - \dot{\varepsilon}\right)|(\widetilde{D}_j)_{j\in J})$$

où $\dot{G}$ est la fonction de répartition de la loi de Cauchy $(\dot{\mu}, \dot{\sigma})$.

Validation du modèle

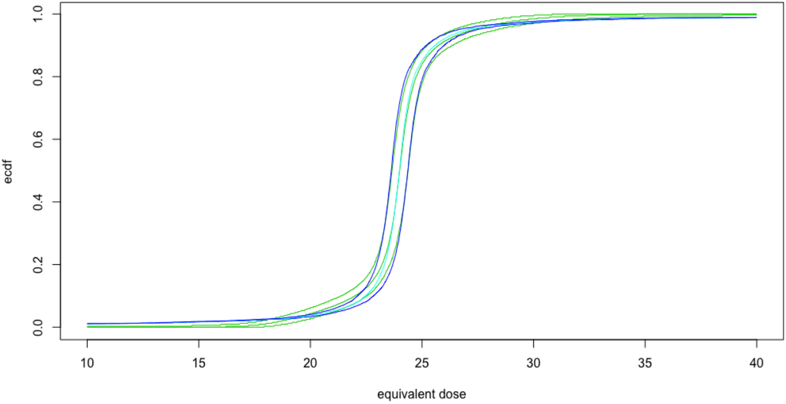


Figure: Représentation des intervalles de crédibilité à 95% de $F_D, F_{A\dot{d}}$, respectivement en bleu et en vert.

Bibliography

Bronk Ramsey, C. (2009). Dealing with outliers and offsets in radiocarbon dating. *Radiocarbon*, 51(3):1023–1045.

Lanos, P. and Philippe, A. (2017). Hierarchical Bayesian modeling for combining dates in archaeological context. *Journal de la Societe Française de Statistique*, 158(2):72–88.

Lanos, P. and Philippe, A. (2018). Event date model: a robust bayesian tool for chronology building. *Communications for Statistical Applications and Methods*, 25(2):131–157.