La distribution Ex-Gauss pour l'analyse du temps de réaction : Initialisation plus robuste et traitement des données manquantes

Alandra Zakkour [1,2], Yousri Slaoui [1] & Cyril Perret [1,2]

[1] *Laboratoire de Mathématiques et Applications, Université de Poitiers, France.*
[2] *Centre de Recherches sur la Cognition et l'Apprentissage, Université de Poitiers, France.*
E-mail : alandra.zakkour@univ-poitiers.fr
E-mail : yousri.slaoui@univ-poitiers.fr
E-mail : cyril.perret@univ-poitiers.fr

Résumé. Le temps entre la présentation d'un stimulus et la réponse motrice d'un participant est la mesure la plus ancienne et la plus largement utilisée pour explorer le fonctionnement de l'esprit humain. Donders a théorisé cette durée, appelée temps de réaction (RT), comme impliquant trois ensembles d'activités : les mécanismes perceptifs, le traitement cognitif et la préparation motrice. Partant de l'hypothèse que les premier et dernier ensembles de traitements peuvent être considérés comme ayant des durées quasi identiques pour une même tâche, tout changement de RT entre deux conditions expérimentales est alors interprété comme indiquant un changement de la durée des traitements cognitifs. RTs sont alors considérées par les psychologues comme un outil pour explorer les mécanismes de traitement cognitif.
Pour analyser cette mesure, nous nous référons à la distribution Ex-Gauss, largement étudiée dans la littérature. Notre étude propose une méthode permettant d'obtenir des estimations moins biaisées pour les trois paramètres de cette distribution (μ, σ, et τ) en utilisant une approche bayésienne. Cette méthode consiste à adapter l'initialisation des paramètres en recourant au rééchantillon-nage de Bootstrap plutôt qu'à une sélection aléatoire de vecteurs de paramètres initiaux.
Un deuxième aspect essentiel de ce travail est la résolution du problème des données manquantes de type MAR, caractérisées par différents pourcentages de présence.
Mots-clés. RT; Distribution Ex-Gauss; Initialisation des paramètres; Données Manquantes; MAR

Abstract. The time between the presentation of a stimulus and a participant's motor response is the oldest and most widely used measure for exploring the functioning of the human mind. Donders theorized this duration, called reaction time (RT), as involving three sets of activities: perceptual mechanisms, cognitive processing, and motor preparation. Assuming that the first and last sets of processes can be considered to have nearly identical durations for the same task, any change in RT between two experimental conditions is then interpreted as indicating a change in the duration of cognitive processing. RTs are thus considered by psychologists as a tool for exploring cognitive processing mechanisms.
To analyze this measure, we refer to the Ex-Gaussian distribution, widely studied in the literature. Our study proposes a method to obtain less biased estimates for the three parameters of this distribution (μ, σ, and τ) using a Bayesian approach. This method involves adapting

the parameter initialization by resorting to Bootstrap resampling instead of randomly selecting initial parameter vectors.
A second essential aspect of this work is the resolution of the missing data problem of the MAR type, characterized by different percentages of presence.
Keywords. RT; Ex-Gaussian Distribution; Parameter Initialization; Missing Data; MAR

1 Introduction

Plusieurs distributions de probabilité ont été proposées pour prendre en compte le temps de réaction (`RT`). La distribution de probabilité qui semble être la plus proche de la distribution observée est obtenue en convoluant une distribution gaussienne et une distribution exponentielle, c'est-à-dire une distribution Ex-Gaussienne (Burbeck (1982) ; Hohle (1965) ; Luce (1986) ; El Haj et al., (2021)).
El Haj et al., (2021) ont proposé une méthode bayésienne pour estimer les trois paramètres Ex-Gaussiens, μ, σ et τ. L'objectif de cette méthode était d'obtenir des estimations non biaisées dans les conditions restrictives d'un petit échantillon ($n < 100$), fréquemment observées dans les études de psychologie scientifique. Cette méthodologie bayésienne utilise une initialisation aléatoire des paramètres.

Le premier objectif de notre article est d'adopter cette méthode en utilisant une initialisation appropriée dans l'approche bayésienne basée sur le rééchantillonnage et plus spécifiquement sur le Bootstrap.
Le bootstrap est une approche qui peut fournir des estimations précises, particulièrement lorsque les approximations habituelles sont invalides. Cette technique implique la création de multiples échantillons bootstrap en sélectionnant aléatoirement des observations à partir de l'ensemble de données original avec remplacement. Chaque échantillon a la même taille que les données initiales, permettant aux observations d'être incluses plusieurs fois ou pas du tout. Pour plus de détails sur cette méthode et leurs applications pratiques, voir le tutoriel de Wehrens (2000) ; Davison (1997).

Le deuxième objectif de ce travail est de traiter le problème des données manquantes de type Missing At Random (`MAR`), qui est le cas le plus classique défini par Rubin (1976) où la valeur manquante est prédite uniquement sur la base des données observées.
La méthode d'Imputation Multiple (`MI`) est devenue l'une des approches les plus avancées pour aborder le problème des données manquantes. C'est une technique statistique basée sur la création de plusieurs ensembles de valeurs possibles pour les données manquantes. L'imputation multiple est un cas général de l'Imputation Simple (`SI`) dans lequel les données manquantes sont remplacées par une valeur possible, puis les paramètres sont estimés. Pour obtenir la méthode d'imputation multiple, nous répétons la méthode simple plusieurs fois avec différentes valeurs prédites, puis les résultats sont combinés. `MI` aide à produire des inférences statistiques plus précises et fiables par rapport à la `SI`.
Dans notre article, nous avons suggéré de comparer la méthode `MI` avec trois autres méthodologies,

qui sont documentées dans la littérature.

Pour valider nos propositions, nous avons illustré son application en utilisant à la fois des données simulées et réelles. La comparaison est réalisée en utilisant l'Augmentation du Risque Absolu "Absolute Risk Increase" (`ARI`) pour les vecteurs de paramètres, obtenus à travers notre méthode proposée et l'approche précédemment utilisée.

2 Methodologie

2.1 La distribution Ex-Gauss

Supposons que la variable aléatoire Z suit la loi ex-gaussienne. Elle peut donc s'écrire sous la forme de la somme de deux variables aléatoires X et Y:

$$Z = X + Y$$

où $X \sim \mathrm{N}(\mu, \sigma)$ et $Y \sim \exp(\lambda)$. Cette distribution possède la fonction de densité suivante:

$$f(x; \mu, \sigma, \lambda) = \frac{\lambda}{2} \exp\left(\frac{\lambda}{2}(2\mu + \lambda\sigma^2 - 2x)\right)\phi\left(\frac{\mu + \lambda\sigma^2 - x}{\sqrt{2}\sigma}\right) \tag{1}$$

Avec ϕ est la fonction d'erreur complémentaire définit par:

$$\phi(x) = \frac{2}{\sqrt{\pi}} \int_x^{\infty} \exp(-t^2)dt$$

Pour estimer les trois paramètres μ, σ et λ, El Haj et al., (2021) ont proposé une méthode basée sur l'inférence bayésienne pour produire des prédictions avec de petits échantillons. Notre étude présente une version plus efficace de cette méthode, dans laquelle le choix du vecteur initial est basé sur l'algorithme Bootstrap plutôt que d'être effectué de manière aléatoire.
Dans la section suivante, nous présentons la méthode de rééchantillonnage utilisée.

2.2 Bootstrap

La méthode de bootstrap est une technique de rééchantillonnage largement utilisée en statistiques pour estimer la distribution d'un échantillon statistique en se basant sur les données disponibles. Elle consiste à générer de multiples échantillons bootstrap en tirant aléatoirement avec remplacement des observations à partir de l'échantillon original. Ces échantillons bootstrap sont de taille égale à l'échantillon original (Efron and Tibshirani (1993)).

Soit $X = \{X_1, \ldots, X_n\}$ notre échantillon initial de taille n avec une fonction de distribution $F(x)$. Pour générer un échantillon bootstrap, nous effectuons un tirage aléatoire de n observations de l'échantillon initial avec remplacement. Cette procédure est répétée B fois (où

B est le nombre d'itérations). L'échantillon bootstrap obtenu est noté $X^* = \{X_1^*, \ldots, X_n^*\}$.

Pour chaque échantillon bootstrap, nous calculons la fonction de distribution $F^*(x)$ et l'estimateur de l'intérêt $\hat{\theta}^* = \{\hat{\theta}_1^*, \ldots, \hat{\theta}_B^*\}$.

En résumé, la méthode Bootstrap permet d'approximer la distribution de l'estimateur statistique $\hat{\theta}$ en utilisant des échantillons bootstrap et de fournir des estimations robustes.

L'objectif est de sélectionner un vecteur de paramètres initiaux plus robuste qu'un vecteur arbitraire. Notre vecteur comprend trois paramètres à estimer, notés $\hat{\theta} = \{\hat{\mu}, \hat{\sigma}, \hat{\lambda}\}$. La méthode de rééchantillonnage est alors appliquée à trois fonctions d'intérêt statistique (voir l'algorithm 1).

Pour les applications numériques, nous avons comparé cette méthode proposée avec la

Algorithm 1 L'algorithm Bootstrap: X est le vecteur initial; B est le nombre d'échantillons Bootstrap; $\theta = \{\mu, \sigma^2, \lambda\}$ est le vecteur de paramètre.

Input: X, B, μ, σ^2 and λ.

1: **for** $b = 1, \ldots, B$ **do**

2: $X_b^* =$ échantillon de X avec remplacement et de taille n;
$\mu_b^* = \mathbb{E}(X_b^*) - \left(0.5 * \sqrt{\mathbb{V}(X_b^*)}\right)$;
$(\sigma^2)_b^* = \mathbb{V}(X_b^*) - \left(0.5 * \sqrt{\mathbb{V}(X_b^*)}\right)^2$;
$\lambda_b^* = \frac{1}{0.5 * \sqrt{\mathbb{V}(X_b^*)}}$.

3: **end for**

output: $\mu^* = \frac{1}{B}\sum_{b=1}^{B} \mu_b^*$, $(\sigma^2)^* = \frac{1}{B}\sum_{b=1}^{B} (\sigma^2)_b^*$ and $\lambda^* = \frac{1}{B}\sum_{b=1}^{B} \lambda_b^*$.

méthode arbitraire de El Haj et al, (2021) ainsi qu'avec la méthode du maximum de vraisemblance sur trois exemples de données simulées de taille n=50. Les résultats obtenus démontrent l'efficacité de l'approche Bootstrap en se basant sur le critère d'Augmentation absolue du risque (ARI). Dans ce document nous avons présenté un seul exemple dans le tableau 1.

Dans la figure 1, nous présentons la convergence des paramètres de la distribution exgaussienne vers les valeurs initiales lorsque la taille de l'échantillon augmente.

	Data	Aléa.	Max.V.	B
μ	**400**	72.2533	497.6375	**397.1385**
σ	**200**	42.77476	**203.7802**	181.9681
$\frac{1}{\lambda}$	**100**	420.4193	12.17656	**80.19481**
ARI	**0**	4.8096	1.1412	**0.2953**

Table 1: Prédiction des paramètres pour un jeu de données simulées de valeur $\mu = 400$, $\sigma = 200$, et $\frac{1}{\lambda} = 100$ en utilisant l'inférence bayésienne avec 5000 itérations. "Aléa" désigne la méthode d'initialisation aléatoire; "Max.V." désigne la méthode du maximum de vraisemblance; et "B" désigne la méthode de rééchantillonnage "Bootstrap".

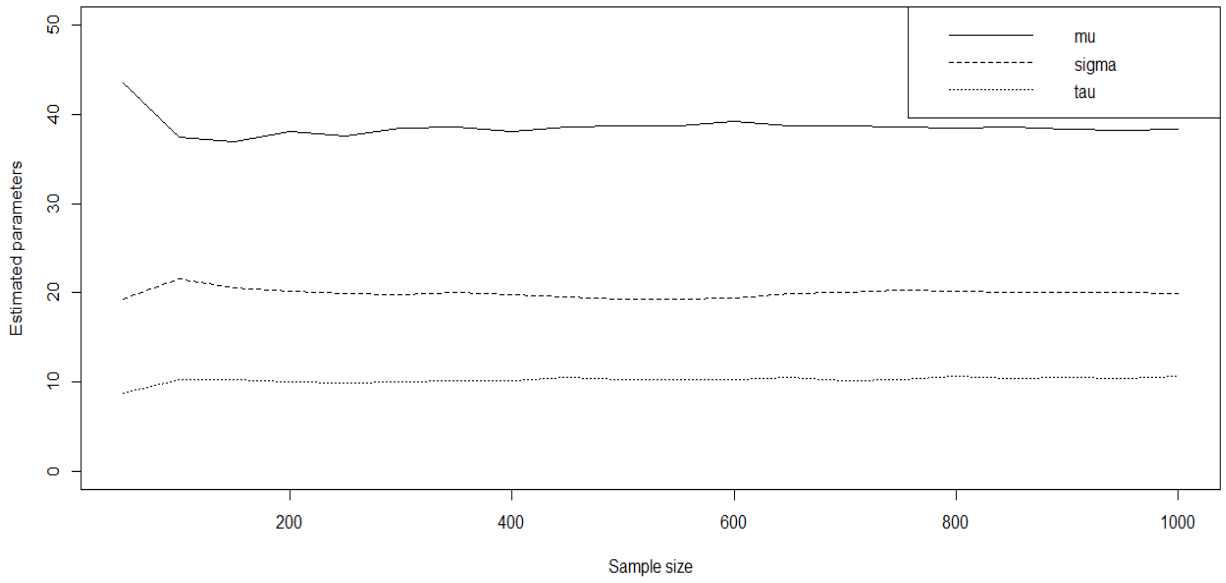

Figure 1: Exemple des paramètres estimés en utilisant la méthode Bootstrap pour différentes tailles d'échantillons à partir d'un jeu de données avec des valeurs initiales de $\mu = 40$, $\sigma = 20$ et $\frac{1}{\lambda} = 10$.

Notre deuxième objectif dans cette étude concerne le problème des données manquantes de type MAR, où la présence des données manquantes est associée à certaines caractéristiques observées. Dans la section suivante, nous abordons cette problématique ainsi que les méthodes que nous proposons pour y remédier.

2.3 Missing Data Imputation

Nous pouvons parler de données incomplètes lorsque les valeurs dans notre vecteur de réponse ne sont pas toutes observées pour de nombreuses raisons; on peut également dire qu'il s'agit d'une question sans réponse. Ces types de données sont un problème courant reconnu par les statisticiens.
En général, les données manquantes se produisent lorsqu'une valeur dans les données n'est pas représentée pour une variable donnée, pour de nombreuses raisons qui peuvent être liées à l'objectif de l'étude (par exemple, les participants ne répondent pas aux questions). Cela apparaît dans de nombreuses études de recherche, en particulier lors de la collecte de données et lorsque les participants sont étudiés sur une période de temps. Au début, les études étaient développées en supposant l'absence de valeurs manquantes. À la fin des années 1980, avec l'avancée de la technologie, ce problème a attiré l'attention de nombreux chercheurs qui souhaitaient étudier plusieurs techniques pour le gérer. En fonction des raisons de leur absence, ces valeurs pourraient être divisées en trois types : manquant complètement au hasard

		$\pi = 15\%$				$\pi = 20\%$			
	Data	Stand	K-NN	Bay	IM	Stand	K-NN	Bay	IM
μ	**4**	3.691	3.613	3.531	**3.739**	3.738	3.678	3.589	**3.818**
σ	**2**	1.447	1.436	**1.627**	1.548	1.440	1.469	**1.627**	1.584
τ	**1**	0.979	**0.997**	0.966	0.952	**0.970**	0.959	0.955	0.935
ARI	**0**	0.373	0.380	**0.336**	0.338	0.374	0.386	0.333	**0.317**

Table 2: Prédiction des paramètres pour un jeu de données incomplètes dans deux cas (π c'est le pourcentage de présence des données manquantes).

(MCAR), manquant de manière aléatoire (MAR) et manquant de manière non aléatoire (MNAR).

Pour comprendre chaque type et leur différence voir les références Mack et al. (2018); Heitjan and Basu (1996). Dans cette étude, nous nous intéressons aux données manquantes de type MAR (la probabilité qu'une valeur soit manquantes dépend des données observées) et avons exploré plusieurs scénarios. Nous avons considéré quatre exemples de pourcentages de données incomplètes (5%, 10%, 15% et 20%) dans trois ensembles de données simulées. Les résultats ont été comparés entre quatre méthodes différentes. La méthode standard ("Stand") consiste à remplacer les données manquantes par la moyenne des données observées. La méthode des k plus proches voisins ("K-NN") prédit les valeurs manquantes en se basant sur les valeurs des voisins les plus proches. La méthode bayésienne ("Bay") remplace les données incomplètes en utilisant le théorème de Bayes. Enfin, la méthode d'imputation multiple ("IM") remplace les données non observées par des valeurs numériques obtenues par imputation simple, cette opération étant répétée plusieurs fois pour obtenir une imputation multiple.

Après avoir fixé la méthode de Bootstrap pour le choix du vecteur initial dans la méthode bayésienne pour l'estimation des paramètres ($\hat{\mu}$, $\hat{\sigma}$, $\hat{\lambda}$), nous avons appliqué les quatre méthodes proposées pour résoudre le problème des données manquantes (Na) pour les quatre pourcentages et sur les trois exemples de données simulées. Ensuite, le même phénomène a été répété sur quatre exemples de données réelles en psycholinguistique. Les résultats obtenus montrent une compétition entre les méthodes, notamment entre la méthode bayésienne et l'imputation multiple. Cela est observé dans le tableau 2, où nous avons présenté un exemple de jeu de données ($\mu = 4$, $\sigma = 2$ et $\lambda = 1$) avec deux pourcentages de valeurs manquantes (15% et 20%). Pour un taux de 15% de données manquantes, la méthode bayésienne présente la plus petite valeur du critère ARI (0.336), mais cette valeur est très proche de celle obtenue pour l'imputation multiple (0.338). Tandis que pour un taux de 20% de valeurs manquantes, le meilleur résultat est obtenue avec la méthode "IM", avec une valeur de 0.317, tandis que la méthode "Bay" présente une valeur de 0.333.

En répétant l'exemple sur plusieurs scénarios, avec différents taux de valeurs manquantes et plusieurs jeux de données, les résultats obtenus nous permettent d'envisager une autre méthode plus robuste consistant à imputer les données manquantes dans le cadre de l'estimation bayésienne.

3 Conclusion

Notre projet aborde deux problématiques distinctes. Dans un premier temps, notre objectif initial était de développer une méthode d'initialisation des paramètres permettant d'obtenir des estimations plus robustes et moins biaisées dans le cadre bayésien.
Les résultats obtenus montrent que la méthode de rééchantillonnage proposée, nommée Bootstrap, s'avère plus efficace que la méthode d'initialisation arbitraire utilisée par El Haj et al,. (2021).
Ensuite, nous avons exploré la résolution du problème des données manquantes de type MAR en variant les pourcentages de présence des données. Cette partie de notre étude n'a pas permis d'identifier de manière concluante une méthode prédominante, car les méthodes bayésiennes et d'imputation multiple ont produit des résultats similaires. Face à cette compétition, nous ouvrons une nouvelle voie de recherche pour trouver une méthode permettant de résoudre ce problème de manière plus efficace.

References

Burbeck, S. and Luce, R. (1982), Evidence from auditory simple reaction times for both change and level detectors, *Perception & psychophysics*, 32, pp. 117–133.

Davison, A. and Hinkley, D. (1997). Bootstrap Methods and Their Applications. *Cambridge Univ. Press, Cambridge.*

Donders, F. C. (1869/1969), On the speed of mental processes, *Acta Psychologica*, 30, pp. 412-431.

Efron, B. and Tibshirani, R. (1993), An Introduction to the Bootstrap, *Chapman & Hall/CRC, New York.*

El Haj, A., Slaoui, Y., Solier, C., and Perret, C. (2021), Bayesian Estimation of The Ex-Gaussian Distribution, *Stat. Optim. Inf. Comput. 9*, pp 809-819.

Enders, C.K. (2010), Applied Missing Data Analysis, *Guilford Press: New York, NY, USA.*

Heitjan, D. and Basu, Srabashi. (1996), Distinguishing "Missing at Random" and "Missing Completely at Random". *Amer. Statist.*, 50, pp 207-213.

Hohle, R.H. (1965), Inferred components of reaction times as function of foreperiod duration, *Journal of Experimental Psychology*, 69, pp 382-386.

Little, R.J. and Rubin, D.B. Statistical Analysis with Missing Data, (2019), *John Wiley & Sons: Hoboken, NJ, USA*, 793.

Luce, R. D. (1986), Response times: Their role in inferring elementary mental organization, *Oxford: Oxford University Press.*

Mack, C., Su, Z. and Westreich, D. (2018), Managing missing data in patient registries: addendum to registries for evaluating patient outcomes: a user's guide. *Agency for Healthcare Research and Quality (US).*

Schafer, L.J. and Graham, W.J. (2002), Missing Data: Our View of the State of the Art, *Psychol. Methods*, 7, pp. 147–177.

Roelofs, A. (2018), One hundred fifty years after Donders : Insignts form unpublished data, a replication, and modeling of his reaction times, *Acta Psychologica*, 191, pp. 228-233.

Rubin, D. B. (1976). Inference and Missing Data. *Biometrika.*, 63, pp. 581-590.

Wehrens, R., Putter, H. and Buydens, L. (2000). The bootstrap: a tutorial. *Chemometr. Intell. Lab. Syst.*, 54, pp. 35-52.