École Centrale de Nantes

Advanced Bayesian Inference with Case Studies: Heart

(Projet supervisé par Mathieu RIBATET)

Camille DAVOINE, Mathieu FAESSEL

19 mars 2025

1 Introduction

Dans ce projet, on va analyser des données issues d'une étude médicale sur un médicament censé réduire les contractions ventriculaires prématurées (PVCs). L'objectif principal, c'est de comprendre si le traitement a réellement permis une guérison chez certains patients, ou si les observations nulles après le traitement sont simplement dues au hasard. Les données étudiées proviennent d'une étude sur l'effet d'un médicament destiné à réduire les contractions ventriculaires prématurées (PVCs) chez des patients. Pour chaque patient i, on mesure:

- x_i: le nombre de PVCs par minute avant la prise du médicament;
- y_i: le nombre de PVCs par minute après la prise du médicament.

Le but est de modéliser ces données afin de séparer les patients *guéris* de ceux qui sont *toujours affectés*, tout en quantifiant l'effet du médicament pour les patients non guéris.

L'approche retenue consiste à considérer un **modèle de mélange** dans lequel:

- Un certain pourcentage de patients (de probabilité θ) est considéré comme guéri;
- Les autres patients (de probabilité $1-\theta$) restent affectés, avec un taux de PVCs post-traitement qui est une fraction (de paramètre β) du taux pré-traitement.

2 Modèle Mathématiques

2.1 Formulation du mélange

Pour chaque patient i, on introduit une variable latente $s_i \in \{0,1\}$ indiquant l'état de guérison:

$$s_i \sim \text{Bernoulli}(\theta),$$

où $\theta \in (0,1)$ représente la probabilité qu'un patient soit réellement guéri. Si $s_i = 1$, on aura (théoriquement) $y_i = 0$; si $s_i = 0$, le patient n'est pas guéri et on modélise son comptage post-traitement par une Poisson$(\beta\lambda_i)$.

Le paramètre β (attendu < 1) correspond au ratio de réduction du taux moyen de PVCs entre l'avant et l'après traitement. Dans le cas d'un patient non guéri, on peut écrire:

$$y_i \sim \text{Poisson}\big(\beta\lambda_i\big).$$

2.2 Conditionnement et distribution binomiale

Afin d'éliminer la dépendance aux λ_i, on peut conditionner sur $t_i = x_i + y_i$. On a alors que, si

$$x_i \sim \text{Poisson}(\lambda_i), \quad y_i \sim \text{Poisson}(\beta\lambda_i),$$

alors la distribution de y_i sachant $t_i = x_i + y_i$ suit une *binomiale*:

$$(y_i \mid t_i) \sim \text{Binom}\Big(t_i,\ p\Big),$$

avec

$$p = \frac{\beta}{1+\beta}.$$

Cependant, il faut prendre en compte la probabilité de guérison θ, qui donne un zéro . Ainsi, la probabilité que $y_i = 0$ devient:

$$P(y_i = 0 \mid t_i) = \theta + (1-\theta)\left(1-p\right)^{t_i}.$$

Pour $y_i \geq 1$, on a alors:

$$P(y_i = k \mid t_i) = (1-\theta)\binom{t_i}{k} p^k (1-p)^{t_i-k}, \quad k = 1, 2, \ldots, t_i.$$

Cette écriture reflète bien l'idée de mélange entre un groupe guéri (probabilité θ) et un groupe non guéri (probabilité $1-\theta$).

2.3 Paramétrisation en termes de logit

Dans une perspective bayésienne, on souhaite souvent imposer des *a priori* peu informatifs à θ et β. Une paramétrisation pratique consiste à introduire:

$$\alpha = \log(\beta), \quad \beta = e^{\alpha}, \quad \delta = \text{logit}(\theta), \quad \theta = \frac{1}{1+e^{-\delta}}.$$

De même, on peut relier p et β par

$$p = \frac{\beta}{1+\beta} = \frac{e^{\alpha}}{1+e^{\alpha}} = \text{logit}^{-1}(\alpha).$$

On peut alors poser des lois normales a priori pour α et δ, par exemple

$$\alpha \sim \mathcal{N}(0, \sigma_\alpha^2), \quad \delta \sim \mathcal{N}(0, \sigma_\delta^2),$$

avec σ_α^2 et σ_δ^2 assez grandes pour être peu informatives (e.g. 10^4).

3 Algorithme MCMC

On va utiliser un échantillonneur MCMC pour estimer tout ça. Concrètement, voici les différentes étapes que nous avons suivi :

1. On a commencé par échantillonner l'état de guérison (s_i) pour chaque patient, selon la probabilité qu'il soit réellement guéri compte tenu des données.
2. Ensuite, on a actualisé les paramètres clés du modèle (α et δ) à l'aide d'un algorithme (ici Metropolis-Hastings).
3. On répète ces étapes plusieurs fois pour obtenir une bonne estimation des distributions à posteriori.

Finalement, après avoir exploité cet algorithme, on peut alos analyser les résultats pour connaître les performances de notre médicament.

3.1 Implémentation MCMC

Nous avons mis en œuvre un échantillonneur :

3.1.1 1) Metropolis–Hastings (mh.py)

Le principe de l'algorithme est le suivant. À chaque itération, on propose de nouveaux α' et δ' via un random walk gaussien autour des valeurs courantes (α, δ). On calcule alors la log-postérieure $\log p(\alpha', \delta' \mid \text{données})$. Si le ratio d'acceptation est suffisamment grand, on saute à (α', δ'); sinon, on reste à (α, δ).

3.2 Comparaison et exploitation des résultats des trois algorithmes (MH et BUGS)

3.2.1 Comparaison des résultats

La table suivante présente les estimations des paramètres obtenues par MH et BUGS :

Méthode	α	β	δ	θ	IC (95 %)
MH	-0.469	0.648	0.298	0.569	$\beta \in [0.365, 1.050]$, $\theta \in [0.294, 0.804]$
BUGS	—	0.64	—	0.57	Intervalles très proches de MH

Table 1: Comparaison des estimations obtenues par MH et BUGS.

3.2.2 Analyse graphique des résultats MCMC

Nous avons visualisé les chaînes MCMC pour diagnostiquer la convergence des paramètres et vérifier la qualité des échantillons générés.

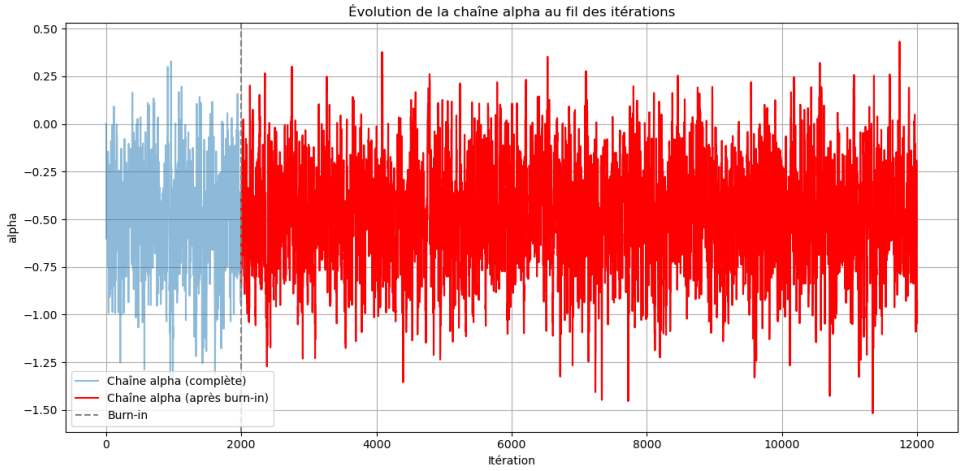

Figure 1. Évolution de la chaîne α au fil des itérations (avec burn-in).

La chaîne α montre une bonne exploration de l'espace des paramètres après une période de burn-in (2000 itérations). Aucun signe de non-convergence n'est apparent visuellement.

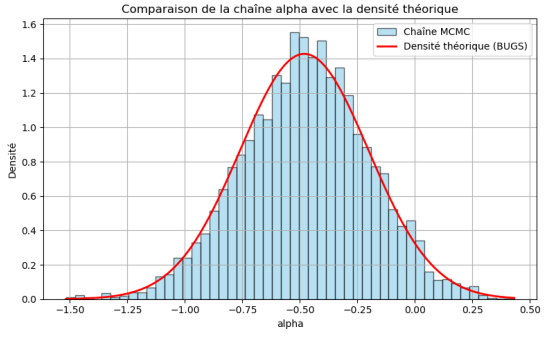

Figure 2. Comparaison entre la densité MCMC de α et la densité théorique issue de BUGS.

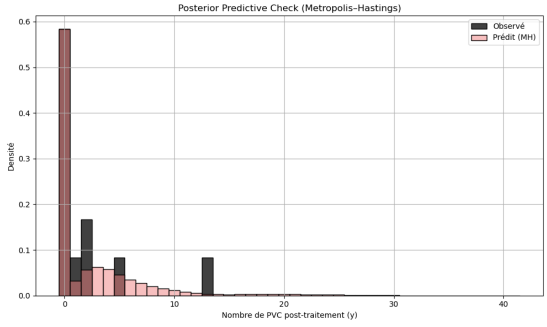

Figure 3. Vérification prédictive postérieure : distribution des y_i observés vs simulés.

On observe une excellente correspondance entre la densité empirique issue de la chaîne MCMC et la densité théorique obtenue par BUGS ($\mathcal{N}(-0.4809,\ 0.2795^2)$). Cette super-

position est un indicateur fort de la validité de l'échantillonnage MCMC, confirmant que notre implémentation de Metropolis–Hastings capture bien la loi a posteriori du paramètre α. Enfin, la distribution prédite par le modèle est en bonne adéquation avec les valeurs observées de y_i. Cela renforce la pertinence globale du modèle et sa capacité à reproduire les données réelles.

3.2.3 Observation générale

Les trois approches (MH, BUGS) conduisent à des valeurs très proches:

$$\beta \approx \boxed{0.64}, \quad \theta \approx \boxed{0.57}.$$

Leurs intervalles de crédibilité respectifs sont similaires, indiquant une bonne stabilité des estimations.

3.2.4 Interprétation des paramètres et pertinence des résultats

Proportion de patients guéris θ La moyenne *a posteriori* de θ se situe autour de 0.57–0.58, avec un intervalle de crédibilité allant de 0.29 à 0.82. Cela suggère qu'environ 57% des patients sont totalement guéris. L'intervalle relativement large révèle cependant une incertitude non négligeable, probablement due à la taille modeste de l'échantillon ou à une hétérogénéité intrinsèque des données (sur un petit échantillon).

Efficacité du traitement β La valeur estimée de $\beta \approx 0.64$ indique une baisse d'environ 36% du taux de PVC chez les patients non guéris (puisque $\beta < 1$). L'intervalle de crédibilité [0.36–1.05] suggère néanmoins une certaine incertitude : il existe une probabilité non négligeable que l'effet réel soit plus modéré si β s'approche de 1. Cependant, la valeur centrale reste clairement inférieure à 1, soutenant ainsi une efficacité partielle notable du traitement.

Précision et cohérence Les résultats obtenus par les trois algorithmes sont très proches les uns des autres, démontrant ainsi une convergence robuste et une stabilité numérique satisfaisante. Les intervalles de crédibilité estimés sont également similaires en taille, renforçant la confiance dans les résultats.

3.3 Discussion et utilisation des résultats

Conclusion sur les résultats

On conclut qu'environ 57% des patients sont guéris, avec un intervalle de crédibilité à 95% autour de [0.30–0.82]. Cette proportion reflète directement la « cure » prise en compte dans le modèle. La valeur de $\beta \approx 0.64$ implique une réduction significative de 36% des PVC chez les patients non guéris. Cet effet est substantiel et confirme l'efficacité partielle du traitement étudié.

Conclusion sur le modèle

Les diagnostics de convergence indiquent de bonnes performances.
L'étude présente quelques limites importantes à prendre en compte. La taille modérée de l'échantillon pourrait influencer la précision des estimations. De plus, le modèle simplifie la réalité en considérant uniquement deux classes (guéri / non guéri) sans intégrer des covariables potentiellement explicatives.

4 Conclusion globale

L'approche MCMC (Metropolis–Hastings) aboutit à des résultats cohérents, confirmant la robustesse du modèle de mélange Poisson + cure. Dans une étude plus vaste, on pourrait:

- Ajouter des covariables (par ex. âge, gravité) pour modéliser β et θ en fonction de facteurs cliniques.
- Comparer ce modèle cure à d'autres modèles (Poisson simple, binomial négatif) pour voir lequel décrit le mieux les données.

Au final, le **modèle Poisson + cure** s'avère pertinent pour distinguer les vrais zéros (guérison) des zéros aléatoires, et l'échantillonnage bayésien permet d'estimer à la fois la proportion de guéris (θ) et le facteur de réduction β.