Hearts: a mixture model for count data

Karim Chakroun
Maxime Gourceyraud
Charles Miranda
Vincent Sevestre

Mars 2022

1 Introduction

En 1987, Berry présente un jeu de données [2] sur l'effet d'un médicament utilisé pour traiter des patients souffrant de fréquentes contractions ventriculaires prématurées (CVP) du coeur. Dans les données qui nous sont fournies nous avons uniquement

- x_i le nombre de CVP par minute *avant* la prise du médicament
- y_i le nombre de CVP par minute *après* la prise du médicament
- $t_i = x_i + y_i$

Un modèle a été proposé par Farewell et Sprott en 1988 [4]. Les sections suivantes présentent le modèle et la méthode employée afin d'obtenir numériquement les coefficients du modèle [1].

2 Modèle

Farewell et Sprott proposent en 1988 un modèle du jeu de données [4] qui est un modèle de mélange de distributions de Poisson dans lequel certains patients sont "guéris" par le médicament, tandis que d'autres présentent des niveaux variables de réponse mais restent anormaux. Une valeur nulle après la prise du médicament peut indiquer une "guérison", ou peut représenter un zéro d'échantillonnage d'un patient avec un CVP normal. Le modèle suivant est donc proposé

[1] `https://github.com/viviseve/bayes-project-2`

$$
\begin{aligned}
x_i &\sim \mathcal{P}(\lambda_i) \text{ , pour tous les patients} \\
y_i &\sim \mathcal{P}(\beta\lambda_i) \text{ , pour tous les patients } \textit{non soignés} \\
\mathbb{P}(Cure) &= \theta
\end{aligned}
$$

Pour éliminer les paramètres de nuisance λ_i, Farewell et Sprott utilisent la distribution conditionnelle de y_i sachant $t_i = x_i + y_i$. En exploitant la remarque de [3] concernant la distribution conditionnelle pour des lois de Poisson, la loi jointe (X_i, Y_i) est

$$\mathbb{P}_{(X_i,Y_i)}(X_i = x_i, Y_i = y_i) = \frac{e^{-\lambda(1+\beta)}\lambda^{x_i+y_i}\beta^{y_i}}{x_i!y_i!}$$

Par conséquent, la distribution conditionnelle de (X_i, Y_i) sachant $T_i := X_i + Y_i = t_i$ est

$$\mathbb{P}_{(X_i,Y_i)|T_i}(X_i = x_i, Y_i = y_i|T_i = t_i) = \binom{t_i}{y_i}\frac{1}{(1+\beta)^{t_i}}\beta^{y_i} = \binom{t_i}{y_i}\frac{1}{(1+\beta)^{t_i-y_i}}\left(\frac{\beta}{1+\beta}\right)^{y_i}$$

Ainsi, le modèle de mélange final peut s'exprimer comme suit

$$
\begin{aligned}
\mathbb{P}(Y_i = 0|T_i) &= \theta + (1-\theta)(1-p)^{t_i} \\
\forall y_i > 0, \mathbb{P}(Y_i = y_i|T_i) &= (1-\theta)\binom{t_i}{y_i}p^{y_i}(1-p)^{t_i-y_i}
\end{aligned}
$$

3 Échantillonneur

3.1 Lois

À présent pour pouvoir déterminer les coefficients β et θ nous considérons les lois a priori $\alpha, \delta \sim \mathcal{N}(0, \sigma^2)$ telles que

$$
\begin{aligned}
\text{logit}(p) &= \alpha \\
\beta &= \exp(\alpha) \\
\text{logit}(\theta) &= \delta
\end{aligned}
$$

Afin de pouvoir échantillonner nous devons calculer les densités a posteriori $\pi(\alpha|\mathbf{y}, \delta)$ et $\pi(\delta|\mathbf{y}, \alpha)$.

$$
\begin{aligned}
\pi(\alpha|\mathbf{y}, \delta) &\propto \pi(\alpha)\prod_{i=1}^{n}\pi(y_i|\alpha, \delta) \\
&\propto \exp\left(-\frac{\alpha^2}{2\sigma^2}\right)\left(\prod_{\substack{i=1 \\ y_i=0}}^{n}(\theta + (1-\theta)(1-p)^{t_i})\right)\left(\prod_{\substack{i=1 \\ y_i>0}}^{n}(1-\theta)p^{y_i}(1-p)^{t_i-y_i}\right) \quad (1)
\end{aligned}
$$

Ainsi, par passage au log de (1) on obtient

$$\log \pi(\alpha|\mathbf{y},\delta) = cst - \frac{\alpha^2}{2\sigma^2} + \sum_{\substack{i=1 \\ y_i=0}}^{n} \log(\theta+(1-\theta)(1-p)^{t_i}) + \sum_{\substack{i=1 \\ y_i>0}}^{n} (y_i \log p + (t_i - y_i)\log(1-p)) \quad (2)$$

Par un calcul similaire à (1) et (2) on obtient pour δ

$$\log \pi(\delta|\mathbf{y},\alpha) = cst - \frac{\delta^2}{2\sigma^2} + \sum_{\substack{i=1 \\ y_i=0}}^{n} \log(\theta + (1-\theta)(1-p)^{t_i}) + \sum_{\substack{i=1 \\ y_i>0}}^{n} \log(1-\theta) \quad (3)$$

3.2 Algorithme

Étant donné que nous n'avons pas de lois explicites pour $\alpha|\mathbf{y},\delta$ et $\delta|\mathbf{y},\alpha$ nous allons employer l'algorithme *Metropolis-within-Gibbs* 1.

Algorithm 1: Metropolis-within-Gibbs

Input: N, σ, $\tilde{\sigma}^2$, α_0, δ_0
Output: $(\alpha_i, \delta_i)_{i=1}^N$
Data: $(t_i, x_i, y_i)_{i=1}^n$
for $i \leftarrow 0$ *à* $N-1$ **do**
  `// échantillonnage de` α
  $\tilde{\alpha} \sim \mathcal{N}(\alpha_i, \tilde{\sigma}^2)$
  $acc \leftarrow \frac{\pi(\tilde{\alpha}|\mathbf{y},\delta_i)}{\pi(\alpha_i|\mathbf{y},\delta_i)}$
  if $\mathcal{U}(0,1) < acc$ **then**
    $\alpha_{i+1} \leftarrow \tilde{\alpha}$
  else
    $\alpha_{i+1} \leftarrow \alpha_i$
  `// échantillonnage de` δ
  $\tilde{\delta} \sim \mathcal{N}(\delta_i, \tilde{\sigma}^2)$
  $acc \leftarrow \frac{\pi(\tilde{\delta}|\mathbf{y},\alpha_{i+1})}{\pi(\delta_i|\mathbf{y},\alpha_{i+1})}$
  if $\mathcal{U}(0,1) < acc$ **then**
    $\delta_{i+1} \leftarrow \tilde{\delta}$
  else
    $\delta_{i+1} \leftarrow \delta_i$
return $(\alpha_i, \delta_i)_{i=1}^N$

4 Résultats

Après avoir exécuté l'algorithme pour une chaîne de longueur $N = 10000$ et un *burnin* de 1000 nous obtenons les résultats suivants

	alpha	beta	delta	theta
mean	-4.57e-1	6.54e-1	2.93e-1	5.67e-1
std	2.59e-1	1.70e-1	6.24e-1	1.41e-1
min	-1.37e+0	2.54e-1	-2.28e+0	9.28e-2
2.5%	-9.83e-1	3.74e-1	-8.82e-1	2.93e-1
50%	-4.59e-1	6.32e-1	2.97e-1	5.74e-1
97.5%	2.76e-2	1.03e+0	1.53e+0	8.22e-1
max	2.92e-1	1.34e+0	2.44e+0	9.20e-1

Table 1: Résultats pour $N = 10000$, $\tilde{\sigma} \in \{0.05, 1\}$, $\alpha_0 = \delta_0 = 0$ et $burnin = 1000$

Les résultats obtenus sont très proches de ceux obtenus par [1], et par ailleurs les quantiles à 2.5% et 97.5% sont également presque les mêmes.

5 Conclusion

Pour conclure, malgré la dépendance en λ_i du modèle original proposé par [4] nous avons pu la contourner en passant par une distribution conditionnelle. Bien que nous n'ayons pas pu déterminer des lois explicites des distributions a posteriori nous avons pu tout de même échantilloner α et δ grâce à l'algorithme de *Metropolis-within-Gibbs* 1. Les résultats obtenus 1 sont très satisfaisants puisque les métriques comme les quantiles à 2.5% et 97.5% sont sensiblement les mêmes que ceux obtenus WinBugs [1].

References

[1] Hearts: a mixture model for count data. pages 14–15.

[2] D. A. Berry. Logarithmic transformations in ANOVA. *Biometrics*, 43(2):439, June 1987.

[3] D. Cox and D. Hinkley. Theoretical statistics. page 136, Sept. 1979.

[4] V. T. Farewell and D. A. Sprott. The use of a mixture model in the analysis of count data. *Biometrics*, 44(4):1191, Dec. 1988.