

Projet Bayes

Aurélien Mounier

Jules Gouy

Introduction

Nous étudions la croissance pondérale de jeunes rats au moyen de deux modèles hiérarchiques bayésiens inspirés de l'exemple classique de Gelfand *et al.* (1990). Les poids de $N = 30$ rats ont été mesurés hebdomadairement pendant $T = 5$ semaines. L'objectif est de modéliser ces trajectoires de croissance individuelles tout en estimant les paramètres au niveau population. Nous comparons (i) un modèle à effets aléatoires indépendants, appelé *Rats*, et (ii) un modèle hiérarchique bivarié, appelé *BiRats*, qui autorise une corrélation entre l'intercept initial et la pente de croissance de chaque rat. L'inférence bayésienne est réalisée via un échantillonneur MCMC (Gibbs) avec une phase de burn-in suivie de 1 000 itérations de collecte pour chaque modèle, assurant une convergence satisfaisante.

1 Modèle *Rats* : modèle hiérarchique linéaire

Le modèle *Rats* suppose que le poids y_{ij} du i -ème rat à l'âge x_j (en jours) suit une loi Normale dont l'espérance est une fonction linéaire de x_j . Plus précisément :

$$y_{ij} \mid \alpha_i, \beta_i, \sigma_c \sim \mathcal{N}(\alpha_i + \beta_i(x_j - \bar{x}), \sigma_c^2),$$

avec $\bar{x} = 22$ jours la moyenne des âges mesurés. Ici α_i et β_i représentent respectivement l'intercept (poids prévu au temps $x = \bar{x}$) et la pente de croissance pour le rat i . Au second niveau hiérarchique, on introduit des distributions *a priori* normales non informatives pour ces effets aléatoires :

$$\alpha_i \sim \mathcal{N}(\alpha_c, \sigma_\alpha^2), \quad \beta_i \sim \mathcal{N}(\beta_c, \sigma_\beta^2),$$

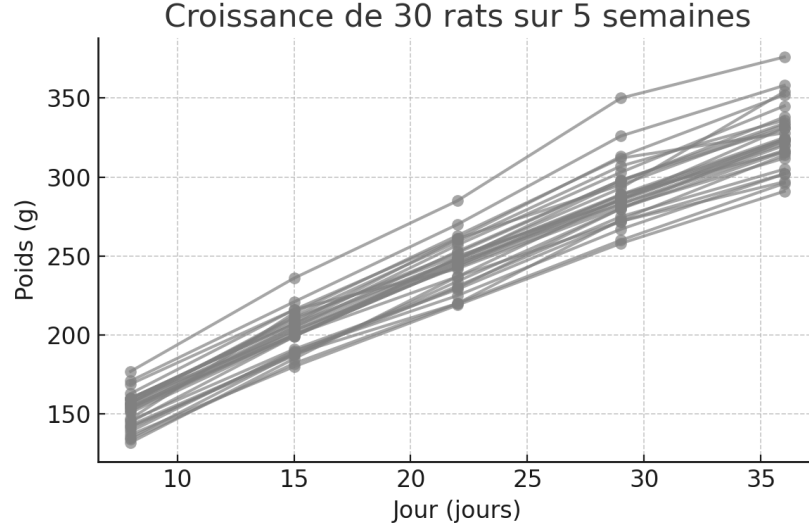


FIGURE 1 – Courbes de poids en fonction de l'âge pour les 30 rats (points : observations empiriques).

où α_c et β_c sont les moyennes populationnelles (intercept moyen et pente moyenne) et σ_α^2 , σ_β^2 les variances inter-individuelles. Des lois a priori larges et peu informatives sont choisies pour les hyperparamètres : par exemple α_c et β_c reçoivent des lois normales de variance élevée (ordre 10^6), et les précisions $\tau_\alpha = 1/\sigma_\alpha^2$, $\tau_\beta = 1/\sigma_\beta^2$ ainsi que $\tau_c = 1/\sigma_c^2$ reçoivent des lois Gamma(0.001, 0.001) (correspondant à des a priori quasi-plates sur les écarts-types).

A partir de ce modèle on déduit la loi à posteriori des α_i :

$$\pi(\alpha_i | \dots) \propto \left(\prod_j \pi(Y_{ij} | \dots) \pi(\alpha_i) \right) \propto \mathcal{N}(\mu_{\alpha_i}, \sigma_{\alpha_i}^2)$$

Avec,

$$\mu_{\alpha_i} = \frac{1}{\frac{1}{\sigma_\alpha^2} + \frac{n_j}{\sigma_c^2}} \left(\frac{\alpha_c}{\sigma_\alpha^2} + \frac{\sum_j e_{ij}}{\sigma_c^2} \right), \sigma_{\alpha_i}^2 = \left(\frac{1}{\sigma_\alpha^2} + \frac{n_j}{\sigma_c^2} \right)^{-1}$$

et

$$e_{ij} = Y_{ij} - \beta_i(x_j - \bar{x})$$

De même pour les β_i :

$$\pi(\beta_i | \dots) \propto \mathcal{N}(\mu_{\beta_i}, \sigma_{\beta_i}^2)$$

Avec,

$$\mu_{\beta_i} = \frac{C\sigma_{\beta}^2 + \beta_c\sigma_c^2}{S^2\sigma_{\beta}^2 + \sigma_c^2}, \sigma_{\beta_i}^2 = \frac{\sigma_c^2\sigma_{\beta}^2}{S^2\sigma_{\beta}^2 + \sigma_c^2}$$

et

$$S^2 = \sum_j (x_j - \bar{x})^2, C = \sum_j e_{ij}(x_j - \bar{x}), e_{ij} = Y_{ij} - \alpha_i$$

Les données observées (Figure 1) révèlent une croissance globalement linéaire des poids, avec une variabilité notable entre rats. Après ajustement du modèle *Rats* par MCMC, nous obtenons les estimations *a posteriori* suivantes pour les paramètres populationnels : le poids initial moyen extrapolé à la naissance (intercept moyen au temps 0) est d'environ $\hat{\alpha}_0 \approx 107$ g (intervalle de crédibilité 95% $\sim [99; 114]$ g). La pente de croissance moyenne est $\hat{\beta}_c \approx 6.2$ g/jour (IC_{95%} $\sim [6.0; 6.4]$ g/jour), ce qui correspond à un gain d'environ 43 g par semaine en moyenne. On note également une forte hétérogénéité entre individus : l'écart-type *a posteriori* des effets aléatoires d'intercept est de l'ordre de 10 g, tandis que celui de la pente est d'environ 0.5 g/jour. L'écart-type résiduel (variabilité intra-individu non expliquée) est estimé à $\sigma_c \approx 6.1$ g. Globalement, le modèle *Rats* s'ajuste bien aux données et reproduit correctement les tendances de croissance individuelles moyennant ces effets aléatoires.

2 Modèle *BiRats* : effets aléatoires corrélés

Le modèle *BiRats* reprend la même structure de base pour les données, à ceci près que les effets aléatoires (α_i, β_i) de chaque rat sont désormais supposés suivre conjointement une loi Normale multivariée. Autrement dit, on autorise une corrélation entre l'intercept spécifique α_i et la pente β_i . La formulation hiérarchique s'écrit :

$$y_{ij} \mid \alpha_i, \beta_i, \sigma_c \sim \mathcal{N}(\alpha_i + \beta_i x_j, \sigma_c^2),$$

$$(\alpha_i, \beta_i)^\top \mid \boldsymbol{\mu}_\beta, \Sigma \sim \mathcal{N}_2(\boldsymbol{\mu}_\beta, \Sigma),$$

où $\boldsymbol{\mu}_\beta = (\alpha_c, \beta_c)^\top$ désigne le vecteur des moyennes populationnelles et Σ la matrice de covariance des effets aléatoires. On peut paramétrer Σ par les écarts-types $\sigma_\alpha, \sigma_\beta$ (déviations inter-rat pour α_i et β_i) et la corrélation ρ entre α_i et β_i . Une corrélation $\rho > 0$ signifierait que les rats initialement plus lourds tendent à croître plus rapidement que les rats plus légers (pentes plus élevées), tandis qu’une corrélation négative impliquerait l’inverse. Pour les hyperparamètres, nous utilisons des a priori peu informatifs analogues à ceux du modèle précédent : $\boldsymbol{\mu}_\beta$ reçoit une loi normale vague (indépendante sur α_c et β_c), et la matrice de précision Σ^{-1} reçoit une loi de Wishart $\mathcal{W}(R, 2)$ avec 2 degrés de liberté (choix minimal pour une matrice 2×2) et une matrice d’échelle R fixée de façon à représenter une information faible (ordre de grandeur initial approximatif des variances, reprenant le choix de *Gelfand et al.*). Le paramètre de variance résiduelle σ_c^2 conserve la même a priori Gamma non informative que précédemment.

A partir de ce modèle on déduit la loi à posteriori des β_i :

$$\pi(\beta_i | \dots) \propto \mathcal{N}_2(\mu_i, \Sigma_i)$$

Avec,

$$\mu_i = A_i^{-1} B_i^T, \Sigma_i = A_i$$

où,

$$B = (1/\sigma_c^2) \sum_j Y_{ij} \tilde{X}_j^T + \mu_\beta^T \Sigma^{-1}, A = (1/\sigma_c^2) \sum_j \tilde{X}_j \tilde{X}_j^T + \Sigma^{-1}$$

Les résultats d’inférence du modèle *BiRats* montrent que les estimations des paramètres moyens α_c et β_c restent très proches de celles du modèle *Rats* (poids moyen de naissance ≈ 107 g, pente moyenne ≈ 6.2 g/jour), ce qui était attendu étant donné que la structure moyenne des données n’a pas changé. L’apport principal du modèle *BiRats* réside dans l’estimation de la corrélation *a posteriori* entre intercept et pente. Cette corrélation est estimée autour de $\rho \approx 0$ (IC_{95%} largement couvrant 0), suggérant qu’il n’y a pas de liaison significative entre le poids initial d’un rat et son rythme de croissance linéaire ultérieur dans ce jeu de données. Autrement dit, les rats ayant une masse initiale plus élevée ne présentent pas une vitesse de croissance sensiblement différente de celle des rats plus légers, une fois prises en compte les incertitudes.

3 Comparaison des modèles et discussion

En comparant les deux approches, on constate que le modèle *BiRats* n'améliore que marginalement la qualité d'ajustement par rapport au modèle *Rats*. Par exemple, les critères d'information (DIC) des deux modèles sont très proches, indiquant une performance prédictive comparable. Cela s'explique par le fait que la corrélation intercept-pente estimée est faible : le modèle plus simple, sans terme de corrélation, parvient déjà à capturer l'essentiel de la variabilité des courbes de croissance. Le modèle *BiRats* reste néanmoins utile pour vérifier l'existence d'une corrélation potentielle. Dans notre analyse, l'intervalle de crédibilité de ρ inclut 0, ce qui suggère qu'aucun effet de corrélation significatif n'est mis en évidence avec les données disponibles. En conclusion, les deux modèles hiérarchiques bayésiens décrivent adéquatement la croissance des rats et fournissent des inférences cohérentes. Le modèle *Rats*, plus parcimonieux, suffit ici à expliquer les données, tandis que le modèle *BiRats* apporte une flexibilité supplémentaire pour tester une corrélation intercept-pente, qui s'avère non significative dans ce cas.