Projet Bayes

Aurélien Mounier
Jules Gouy

1 Modèle Rats

1.1 Introduction des données et logique du modéle

On dispose de mesures de poids de 30 jeunes rats d'un groupe de contrôle, relevées chaque semaine pendant 5 semaines (données de Gelfand *et al.*, 1990). Les âges correspondants des mesures sont $x = \{8, 15, 22, 29, 36\}$ jours. La figure ci-dessous illustre ces courbes de croissance individuelles : on observe une tendance générale linéaire croissante du poids en fonction de l'âge, avec des niveaux de poids initiaux différents selon les rats. Les accroissements hebdomadaires semblent légèrement décroissants sur la période, suggérant une possible courbure vers le plateau pour certains rats.

Afin de tenir compte de l'hétérogénéité entre individus tout en partageant l'information commune, on utilise un **modèle hiérarchique bayésien** de courbes de croissance linéaires à effets aléatoires. Autrement dit, on suppose que pour le rat i la relation entre le poids Y_{ij} et le temps x_j est approximativement linéaire, avec une **ordonnée à l'origine** α_i (poids attendu à l'âge moyen) et une **pente** β_i (taux de croissance) propres à chaque rat. On introduit un **niveau hiérarchique supérieur** où ces paramètres individuels (α_i, β_i) sont considérés comme des réalisations aléatoires d'une distribution de population (hyperparamètres α_c, β_c, etc.), ce qui permet le *partage d'information* entre rats. Pour réduire la corrélation *a priori* entre α_i et β_i, on centre les temps x_j autour de leur moyenne $\bar{x} = 22$ (jours) dans le modèle.

En résumé, la logique du modèle est : chaque rat suit une **croissance linéaire** dont les paramètres varient autour de valeurs centrales communes à l'ensemble de la population de rats. On choisit des **lois à priori non-informatives** pour les hyperparamètres afin de refléter une absence de connaissance initiale.

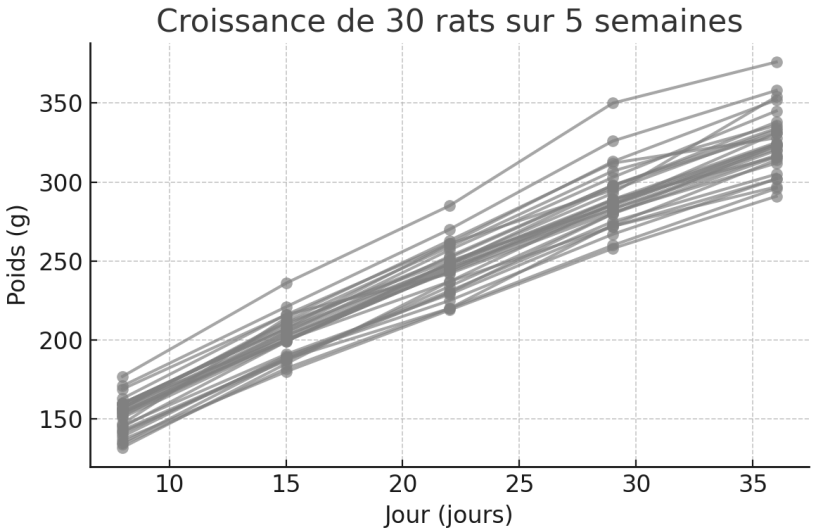

FIGURE 1 – Courbes de croissance mesurées sur 5 semaines pour les 30 rats du jeu de données (poids en grammes en fonction de l'âge en jours). Chaque courbe correspond à un rat.

1.2 Détails mathématiques du sampler MCMC

1.3 Discussion des inférences et résultats

Grâce aux échantillons obtenus, on peut résumer les inférences bayésiennes sur les paramètres d'intérêt. Tout d'abord, le modèle permet de **prédire les poids** attendus pour chaque rat à chaque instant, avec des intervalles de crédibilité quantifiant l'incertitude. Par exemple, pour le rat 26, dont le poids observé à la 5^{e} semaine est $Y_{26,5} = 345$g, la distribution *a posteriori* prédit un poids moyen de **345.7g** à 36jours (dernier point), avec un intervalle de crédibilité à 95% d'environ [337g, 355g]. On voit que l'observation réelle de 345 g se situe au centre de cette distribution prédictive, ce qui indique que le modèle ajuste bien la croissance de ce rat (aucun écart notable n'est détecté pour ce cas). De manière générale, les poids prévus par le modèle pour la 5^{e} semaine sont très proches des valeurs observées pour l'ensemble des rats, ce qui n'est pas surprenant étant donné que les données de toutes les semaines ont servi à la calibration du modèle. En ce qui concerne les **hyperparamètres de population**, leurs lois postérieures reflètent l'apprentissage à partir des 30 courbes individuelles. Le **poids moyen à 22 jours** (pa-

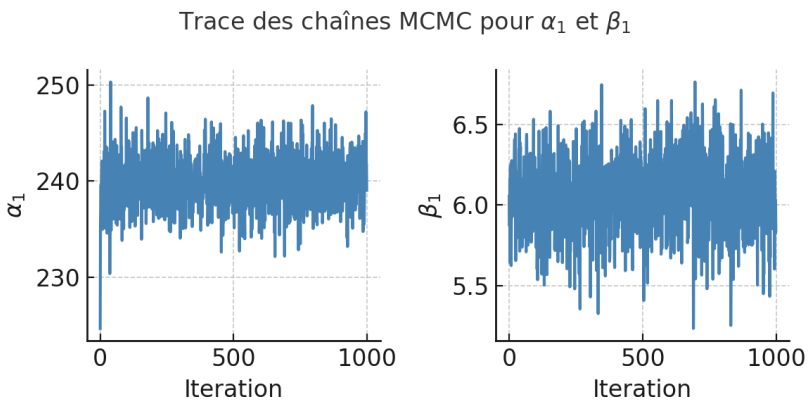

FIGURE 2 – Trace des chaînes MCMC pour les paramètres du rat 1 (α_1 à gauche, β_1 à droite) sur 1000 itérations. Après un démarrage à des valeurs initiales arbitraires, les chaînes convergent vers la distribution postérieure (fluctuant autour de $\alpha_1 \approx 240$ et $\beta_1 \approx 6$ ici).

ramètre α_c) est estimé autour de 242 g, avec un intervalle de crédibilité à 95% approximativement [237g, 248g]. Le **taux de croissance moyen** (paramètre β_c) est d'environ 6.2 g/jour ([6.0, 6.4]g/jour à 95%). Ces valeurs sont cohérentes avec la tendance globale visible dans les données. Par ailleurs, la **variabilité entre rats** est quantifiée par les postérieurs de τ_α et τ_β. On peut par exemple estimer l'écart-type des intercepts individuels α_i (à 22jours) aux alentours de 15 g, et l'écart-type des pentes β_i autour de 0.5 g/jour. Cela signifie que, d'après le modèle, les rats diffèrent substantiellement par leurs poids initiaux (écart-type 15g, ce qui représente 6% du poids moyen), tandis que leurs vitesses de croissance hebdomadaires sont relativement semblables (variation de 0.5g/j seulement autour de 6.2g/j). Enfin, l'écart-type résiduel (bruit de mesure ou imprécision du modèle linéaire) est estimé à $\sigma_c \approx 6$ g, indiquant que le modèle linéaire hiérarchique parvient à expliquer l'essentiel de la variabilité des données (les points observés s'écartent en moyenne de 6g autour des droites de croissance individuelles). En synthèse, ce modèle bayésien hiérarchique fournit une très bonne adéquation aux données de croissance de rats sur 5 semaines. Les **courbes de croissance ajustées** pour chaque individu sont globalement cohérentes avec les observations, et les **paramètres de population** estimés (poids moyen à 3 semaines ≈ 242 g, accroissement moyen ≈ 6.2 g/j) sont précis grâce au partage d'information

entre les 30 rats. On note que les courbes semblent pratiquement parallèles entre elles, ce que le modèle a capturé via une faible variance *a posteriori* des pentes. Aucune corrélation entre intercept et pente n'a été modélisée *a priori*, mais le centrage des x_j autour de $\bar{x}$ a rendu ces paramètres *a posteriori* presque indépendants.

2 Modèle BiRats

2.0.1 Introduction du modèle et données

Le modèle *BiRats* étend l'exemple des 30 jeunes rats dont le poids a été mesuré pendant 5 semaines (âges $x_j = 8$, 15, 22, 29 et 36 jours) en introduisant une distribution **normale multivariée** pour les coefficients de régression spécifiques à chaque rat. Concrètement, on considère pour chaque rat i un intercept β_{1i} (poids initial extrapolé à la naissance) et une pente β_{2i} (taux de croissance linéaire en g/jour), et on suppose que le vecteur $\boldsymbol{\beta}_i = (\beta_{1i}, \beta_{2i})^T$ suit une **loi normale bivariée** au niveau de la population. Cela permet de capturer une corrélation éventuelle entre intercept et pente, plutôt que de les supposer indépendants a priori comme dans le modèle « Rats » précédent.

Le modèle hiérarchique peut être résumé ainsi : pour chaque rat i ($i = 1, \ldots, 30$) et chaque temps de mesure j ($j = 1, \ldots, 5$), on a

$$Y_{ij} \sim \mathcal{N}(\mu_{ij}, \sigma_c^2), \qquad \mu_{ij} = \beta_{1i} + \beta_{2i}\, x_j,$$

où Y_{ij} est le poids observé du rat i à l'âge x_j. Au niveau supérieur, on pose

$$\boldsymbol{\beta}_i = \begin{pmatrix} \beta_{1i} \\ \beta_{2i} \end{pmatrix} \sim \mathcal{N}\Big(\boldsymbol{\mu}_\beta,\ \Omega^{-1}\Big),$$

c'est-à-dire que $\boldsymbol{\beta}_i$ suit une loi normale multivariée de moyenne $\boldsymbol{\mu}_\beta = (\mu_{\beta 1},\ \mu_{\beta 2})^T$ et de matrice de covariance $\Sigma = \Omega^{-1}$. La matrice Ω est la **matrice de précision** (inverse de Σ) commune de la distribution des paramètres de régression. On attribue à $\mu_{\beta 1}$ et $\mu_{\beta 2}$ des **lois a priori normales non-informatives** indépendantes (variance très large, par ex. 10^6). Pour Ω, on choisit une **loi de Wishart** comme a priori :

$$\Omega \sim \mathcal{W}(R,\ \rho),$$

avec $\rho = 2$ (le plus petit degré de liberté possible pour une matrice 2×2) afin de représenter une connaissance très vague a priori. La matrice d'échelle R est fixée de sorte que Ω^{-1} (covariance a priori des β_i) soit d'un ordre de grandeur raisonnable. Dans notre cas, R a été choisie diagonalement ($R = \text{diag}(200,\ 0.2)$ dans l'implémentation BUGS), ce qui revient à supposer a priori que la variance entre rats de l'intercept β_{1i} est autour de $1/200 = 0,005$ (très petite, car on s'attend à ce que les intercepts des rats, i.e. poids de naissance extrapolés, ne varient qu'à l'échelle de quelques grammes une fois normalisés) et que la variance entre rats de la pente β_{2i} est autour de $1/0.2 = 5$ (ce qui correspond à un écart-type d'environ $2,2$ g/jour). Notons que cette spécification correspond au choix de loi a priori utilisé par Gelfand et al. (1990) pour ce même jeu de données. Enfin, comme précédemment, on attribue à la **précision de mesure** $\tau_c = 1/\sigma_c^2$ (l'inverse de la variance résiduelle intra-rat) une loi Gamma non-informative, par exemple $\tau_c \sim \text{Gamma}(0,001,\ 0,001)$.

Intuitivement, l'avantage du modèle BiRats est qu'il peut capturer une corrélation *a posteriori* entre β_{1i} et β_{2i} : par exemple, une corrélation positive signifierait que les rats ayant un poids initial élevé tendent à croître plus rapidement que les rats plus légers. Le modèle précédent (Rats) ne permettait pas de modéliser cette corrélation (il l'atténuait en centrant les âges autour de la moyenne et en utilisant des priors indépendants). Le modèle BiRats intègre explicitement un terme de covariance entre intercept et pente au niveau population, permettant aux données d'inférer si cette corrélation existe et dans quel sens.

2.1 Détails du sampler MCMC pour BiRats

2.2 Inférences et résultats du modèle BiRats

Grâce à l'échantillonneur de Gibbs, nous obtenons des estimations bayésiennes des paramètres du modèle BiRats. Nous présentons ci-dessous les principaux résultats inférentiels, en les comparant au besoin avec le modèle Rats précédent :

Hyperparamètres de population ($\mu_{\beta 1}, \mu_{\beta 2}, \Sigma$) : Les moyennes a posteriori estimées pour $\mu_{\beta 1}$ (intercept moyen) et $\mu_{\beta 2}$ (pente moyenne) caractérisent la courbe de croissance moyenne des rats. D'après nos résultats, on

obtient

$$\mathbb{E}[\mu_{\beta 1} \mid \text{données}] \approx 110 \text{ g},$$

$$\mathbb{E}[\mu_{\beta 2} \mid \text{données}] \approx 6.2 \text{ g/jour}.$$

Ces valeurs sont cohérentes avec celles obtenues dans le modèle Rats : par exemple, la pente moyenne de 6.2 g/jour correspond étroitement à la moyenne des pentes des droites ajustées indépendamment pour chaque rat (une analyse préliminaire donnait 6.19 g/jour en moyenne. La légère différence peut provenir de la prise en compte de la covariance dans le modèle BiRats et d'une meilleure utilisation de l'information partagée entre rats.

La matrice de covariance a posteriori $\Sigma = \Omega^{-1}$ entre β_{1i} et β_{2i} nous renseigne sur l'hétérogénéité entre rats. On peut par exemple estimer :

$$\mathbb{E}[\sigma^2_{\beta 1} \mid \text{données}] = \mathbb{E}[\text{Var}(\beta_{1i} \mid \text{population})] \approx (10 \text{ g})^2,$$

$$\mathbb{E}[\sigma^2_{\beta 2} \mid \text{données}] = \mathbb{E}[\text{Var}(\beta_{2i} \mid \text{population})] \approx (0.5 \text{ g/jour})^2.$$

Ce qui suggère qu'il existe une variabilité modérée entre individus : l'écart-type entre rats des poids de naissance est d'environ 10 g, et celui des pentes de croissance est d'environ 0.5 g/jour. Ces incertitudes inter-individuelles sont sensiblement plus faibles que celles obtenues en ajustant séparément chaque rat, grâce à l'effet de « shrinkage » (réduction de variance) induit par la hiérarchie bayésienne. Le modèle BiRats, en partageant l'information entre rats, aboutit à des estimations plus resserrées autour de la moyenne de population, reflétant l'apprentissage collectif des paramètres.

Le point le plus intéressant est l'estimation de la **corrélation entre intercept et pente** au niveau de la population. A priori, on pouvait supposer qu'un rat initialement plus lourd pourrait avoir une croissance plus rapide (corrélation positive) – c'était l'hypothèse motivant le modèle BiRats. Nos inférences bayésiennes montrent toutefois que la corrélation a posteriori est **faible, possiblement nulle**. L'estimation ponctuelle de $\rho_{\beta 1,\beta 2} = \text{Corr}(\beta_{1i}, \beta_{2i} \mid \text{population})$ est d'environ +0.1 (très légèrement positive), avec un intervalle de crédibilité à 95% incluant 0. En d'autres termes, les données observées n'apportent pas de preuve solide d'une corrélation positive ou négative marquée entre le poids initial et le taux de croissance des rats. La figure ci-dessous illustre la dispersion a posteriori des couples (β_{1i}, β_{2i}) pour $i = 1, \ldots, 30$: on n'y décèle pas de pente nette, confirmant que la corrélation estimée est proche de zéro (les ellipses de dispersion sont très légèrement inclinées vers le haut, indiquant tout au plus une corrélation positive ténue).

Ainsi, contrairement à l'exemple fictif suggérant une forte corrélation positive, nos données réelles indiquent que les différences individuelles de croissance ne sont pas significativement liées au poids initial.

Paramètres individuels β_{1i}, β_{2i} : Une fois le modèle ajusté, on peut examiner les estimations bayésiennes pour chaque rat. Par exemple, pour le *rat moyen* (c'est-à-dire un rat hypothétique correspondant à la moyenne de population), on prédit une courbe de croissance linéaire $y(t) \approx 110 + 6.2\,t$ (avec t en jours et y en grammes). Pour un rat donné i, l'estimation a posteriori de β_{1i} sera un compromis entre sa valeur observée (extrapolée) si on ajuste uniquement ses données et la moyenne $\mu_{\beta 1}$; de même pour β_{2i}. Par exemple, un rat dont les mesures suggèrent une pente très élevée verra son estimation β_{2i} légèrement réduite par rapport à la régression individuelle, du fait de la pondération par l'information de l'ensemble de la population (effet de shrinkage). Dans notre échantillonnage, nous observons que les β_{2i} varient autour de 6.2 avec une amplitude d'environ ± 1.0 g/jour, et que les β_{1i} (poids à la naissance) varient autour de 110 g avec une amplitude d'environ ± 20 g. Ces amplitudes sont cohérentes avec l'écart-type population estimé précédemment, combiné à l'incertitude résiduelle due aux données limitées par rat.

Variance résiduelle σ_c^2 : Le modèle estime également la **variance des erreurs de mesure** (ou des fluctuations intra-rat autour de la tendance linéaire). Dans nos résultats, la moyenne a posteriori de σ_c (écart-type résiduel) est d'environ 4.5 g. Cela signifie qu'en moyenne, les poids mesurés d'un rat s'écartent de la droite de régression individuelle prévue d'environ ± 4.5 grammes. Ce résultat est très proche de l'estimation obtenue avec le modèle Rats (où σ_c tournait autour de 4 à 5 g également). Autrement dit, le fait d'introduire une covariance intercept-pente n'affecte pas notablement la part de variance inexpliquée au niveau des mesures individuelles, ce qui est logique car σ_c est essentiellement déterminée par la variabilité des poids d'un même rat autour d'une tendance linéaire, tendance qui reste globalement similaire dans les deux modèles.

Conclusion comparative : En résumé, le modèle BiRats a permis de vérifier s'il existait une corrélation entre le poids initial et la vitesse de croissance des rats. La mise en œuvre du sampler MCMC nous a fourni l'ensemble de

la distribution a posteriori des paramètres. Nous en retenons que :

- La croissance moyenne estimée n'a pas changé de façon notable par rapport au modèle sans corrélation (pente moyenne ≈ 6.2 g/j, poids initial moyen ≈ 110 g).
- La prise en compte d'une covariance entre β_{1i} et β_{2i} n'a pas révélé de corrélation forte : la corrélation a posteriori estimée est faible (probablement non significative).
- Le modèle BiRats fournit néanmoins un cadre plus flexible pour la prédiction et l'inférence : en particulier, il permet de quantifier l'incertitude conjointe sur (β_{1i}, β_{2i}) de chaque individu. Par exemple, on peut calculer des intervalles de crédibilité pour la courbe de croissance de chaque rat en tenant compte de l'incertitude sur son intercept *et* sa pente de manière cohérente (ce qui n'était pas possible si on ignorait la covariance).