Schools: ranking school examination results using multivariate hierarcical models

Charles Miranda
Armand Rouyrre
Aymard Sahnoune
Vincent Sevestre

Mars 2022

1 Introduction

Le but de ce projet est de classer la performance de certaines écoles Londoniennes [1].

1.1 Jeu de données

On dispose d'un jeu de données comprenant les notes de 1978 élèves (une note par élève) dans 38 écoles différentes. On considère que la note de l'élève dépend des autres données que nous disposons, qui sont :

- Le genre de l'élève.
- L'école de l'élève.
- Une note de lecture de l'élève (LRT : London Reading Test).
- Une catégorie de raisonnement de l'élève (allant de 1 à 3) (VR : Verbal Reasoning).
- Le genre de l'école (Ecole pour fille, pour garçon, ou mixte).
- Le type du bâtiment de l'école (Church of England, Roman Catholic, State school ou autres).

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

1.2 Modèle mathématique

On utilise un modèle hierarchique. Notons Y_{ij} la note du l'élève i dans l'école j. Dans notre modèle. $Y_{ij} \sim \mathcal{N}(\mu_{ij}, \tau_{ij})$.

- Le logarithme de τ_{ij} est modélisé comme suivant une fonction linéaire de la note du test LRT. On note $\log(\tau_{ij}) = \theta + \phi LRT_{ij}$
- μ_{ij} est définie comme égale à

  $\alpha_{1j} + \alpha_{2j} \cdot LRT_{ij} + \alpha_{3j} \cdot VR_{1ij} + \beta_1 \cdot LRT_{ij}^2 + \beta_2 \cdot VR_{2ij} + \beta_3 \cdot Girl_{ij} + \beta_4 \cdot GirlSchool_j + \beta_5 \cdot BoySchool_j + \beta_6 \cdot CESchool_j + \beta_7 \cdot RCSchool_j + \beta_8 \cdot OtherSchool_j$

  Les distributions à priori sont les suivantes :

- Les variables aléatoires β_k où $k \in [1, 8]$, ainsi que θ et ϕ suivent des lois normales centrées et indépendantes, de précision 0.0001.
- Le vecteur aléatoire $(\alpha_{kj})_k$ pour $k \in [1, 3]$, suit une loi normal multivariée $\alpha_{\cdot j} \sim \mathcal{N}(\gamma, \Sigma)$. Où γ suit une loi normal multivariée non informative, et $T = \Sigma^{-1}$ suit une distribution de Wishart.

1.3 Échantillonneur

On utilise un échantillonneur de Gibbs avec un schéma de mise à jour Metropolis-Hasting. Les lois utilisés sont détaillées dans le paragraphe suivant.

2 Justifications mathématiques

2.1 Lois a posteriori reconnues

- $\pi(\beta_k|...) = \propto \pi(\beta_k) \prod_{i,j} \pi(Y_{ij}|...)$

  $\sim \mathcal{N}((\sum_{i,j} \nu_{ij}(Y_{ij} - \delta_{ij})\tau_{ij})(\tau + \sum_{i,j} \nu_{i,j}^2 \tau_{ij})^{-1}, (\tau + \sum_{i,j} \nu_{i,j}^2 \tau_{ij})^{-1})$
- $\pi(\gamma|...) \propto \pi(\gamma) \prod_j \pi(\alpha_j|...)$

  $\sim \mathcal{N}_3((T_0 + \sum_j T_j)^{-1}(T_0\mu_0 + \sum_j T\alpha_j), (T_0 + \sum_j T_j)^{-1})$
- $\pi(T|...) \propto \pi(T) \prod_j \pi(\alpha_j|...)$

  $\sim \mathcal{W}_3((R + \sum_j (\alpha_j - \gamma)(\alpha_j - \gamma)^T)^{-1}, n + 3)$

2.2 Lois a posteriori non reconnues

- $\log \pi(\theta|...) = cst - \frac{\theta^2}{2}\tau - \frac{1}{2}\sum_{i,j}(-\theta + (Y_{ij} - \mu_{ij})^2\tau_{ij})$
- $\log \pi(\phi|...) = cst - \frac{\phi^2}{2}\tau - \frac{1}{2}\sum_{i,j}(-\phi LRT_{ij} + (Y_{ij} - \mu_{ij})^2\tau_{ij})$
- $\log \pi(\alpha_j|...) = cst - \frac{(\alpha_j-\gamma)^T T(\alpha_j-\gamma)}{2} - \frac{1}{2}\sum_i \frac{(\alpha_j^T X_{ij})^2 - 2\alpha_j^T X_{ij}(Y_{ij}-\delta_{ij})}{\sigma_{ij}^2}$

3 Conclusion

3.1 Résultats

Pour obtenir une estimation des paramètres, on exécute la commande suivante :
`python main.py data\data.json data\init1.json`
Les valeurs obtenues pour les paramètres sont présentes dans le tableau 1.

coefficient	mean	std	q2.5pc	median	q97.5pc
beta[1]	1.08e-03	6.87e-04	-5.86e-04	1.08e-3	2.03e-3
beta[2]	3.69e-01	1.34e-01	1.22e-01	3.69e-01	6.43e-01
beta[3]	2.73e-01	8.28e-02	1.03e-01	2.73e-01	4.29e-01
beta[4]	7.09e+00	3.05e+00	2.28e+00	7.09e+00	1.51e+01
beta[5]	3.64e-01	3.81e-01	-2.75e-01	3.64e-01	1.36e+00
beta[6]	-1.96e-01	6.68e-01	-2.21e+00	-1.96e-01	8.42e-01
beta[7]	-1.09e+00	1.56e+00	-5.15e+00	-1.09e+00	4.92e-01
beta[8]	-5.51e-01	3.35e-01	-1.09e+00	-5.51e-01	1.66e-01
gamma[1]	6.74e-03	1.01e+00	-1.99e+00	6.74e-03	1.99e+00
gamma[2]	1.93e-03	9.92e-01	-1.94e+00	1.93e-03	1.96e+00
gamma[3]	-3.05e-02	9.91e-01	-1.93e+00	-3.05e-2	1.88e+00
phi	-8.19e-03	7.02e-03	-1.57e-02	-8.19e-03	9.37e-03
theta	-3.85e-02	2.49e-01	-8.28e-01	-3.85e-02	2.21e-01

Table 1: Résultats pour $N = 10000$ et $burnin = 1000$

Comparons ces résultats aux valeurs annoncées dans l'énoncé [1] :

1. β_2, β_3, β_5, β_6, β_8 et ϕ sont du même ordre de grandeur que leur valeurs données respectives.
2. β_1, γ_2 ne sont pas de même ordre de grandeur, mais la différence est faible (une puissance de 10).

3. β_4, β_7, γ_1, γ_3 et θ ont des valeurs bien différentes (de signe opposés et/ou d'ordre de grandeur différents).

Notons cependant que les valeurs d'écarts-type obtenus sont de même ordre de grandeur, voire plus faibles que les écarts-types données.

3.2 Conclusion

En conclusion, les résultats obtenus nous donne une estimation des paramètres. Il faut toutefois avoir conscience des variations par rapport aux résultats attendus.

References

[1] Schools: ranking school examination results using multivariate hierarcical models. pages 26–28. `https://www.mrc-bsu.cam.ac.uk/wp-content/uploads/WinBUGS_Vol2.pdf`.