Schools: ranking school examination results using multivariate hierarcical models

Charles Miranda
Armand Rouyrre
Aymard Sahnoune
Vincent Sevestre

March 2022

1 Introduction

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

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).

1.2 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

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

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 $\alpha_{*j} \sim \mathcal{N}(\gamma, \Sigma)$. Où γ suit une loi normal multivariée non informative, et $T = \Sigma^{-1}$ suit une distribution de Wishart.

2 Justifications mathématiques

2.1 Lois a posteriori reconnues

- $\pi(\beta_k) \sim \mathcal{N}(\frac{\sum_{i,j} c_k \tau_{ij} - a_{ij}}{\tau + \sum_{ij} c_k \tau_{ij}}, \frac{1}{\tau + \sum_{ij} c_k \tau_{ij}})$, avec
  - c_k vérifie $\mu = c_1\beta_1 + ... + c_k\beta_k + ...$
  - $a_{ij} = Y_{ij} - \mu_{ij} + c_k\beta_k$
- $\pi(\gamma) \sim \mathcal{N}_3((P + nT)^{-1}(P\mu_0 + T\sum_j \alpha_j), (P + nT))$, avec
  - $\gamma \sim \mathcal{N}(\mu_0, P)$ où P est la matrice de précision
- $\pi(T) \sim Wishart(n + 3, (R^{-1} + \sum(\alpha_j - \gamma)(\alpha_j - \gamma)^T)^{-1})$, avec
  - $T \sim Wishart(3, R)$

2.2 Lois a posteriori non reconnues

- $\pi(\theta) \propto \exp(-\theta^2\tau/2) \exp(-\sum_{i,j}(Y_{ij} - \mu_{ij})^2\theta \exp(\theta)/2)$
- $\pi(\phi) \propto \exp(-\phi^2\tau/2) \exp(-\sum_{i,j}(Y_{ij} - \mu_{ij})^2\phi LRT_{ij} \exp(\phi LRT_{ij})/2)$
- $\pi(\alpha_j) \propto \exp(-(\alpha_j - \gamma)^t T(\alpha_j - \gamma)/2) \exp(-\sum_{i,j}((\alpha_j A_{ij})^2 - 2b_{ij}(\alpha_j A_{ij}))\tau_{ij}/2)$, avec
  - $\alpha_j = (\alpha_{1j}, \alpha_{2j}, \alpha_{3j})$
  - $A_{ij} = (1, LRT_{ij}, VR_{1ij})$
  - $b_{ij} = \mu_{ij} - \alpha_j A_{ij} - Y_{ij}$