Empirical and Fully Bayesian Hierarchical Models : Two Applications in the Stellar Evolution

Shijing Si*, David van Dyk*, Ted von Hippel†

*Imperial College London

†Embry-Riddle Aeronautical University

February 10, 2015

Overview

Hierarchical Modelling

- Many data are of hierarchical structure, which motivates hierarchical models.
- Hierarchical modelling usually produces a shrinkage estimate, which is widely considered to be more reasonable.
- Also, it can reduce the mean squared errors.
- However, fitting a hierarchical model may be computationally intensive.

Hierarchical Models

- One commonly used and simple hierarchical model is:

$$Y_i = \mu_i + \epsilon_i;\ \ \epsilon_i \sim N(0, \sigma_i^2), \tag{1}$$

$$\mu_i \sim N(\gamma, \tau^2),\ \ i = 1, \cdots, I. \tag{2}$$

in which:

- ϵ_is are independent of μ_js and σ_i^2s are known.
- μ_is are group effects, γ pooling effects, and τ^2 between-group variation.
- What we want to learn from this model is μ_is and γ and τ^2.

Statistical Inference to Hierarchical Models

- Commonly used ways are Maximum likelihood, fully Bayesian and empirical Bayes.
- MLE: Find the MLE of hyper-parameters by maximising their likelihood function;
- Fully Bayesian: Find the fully joint posterior for all parameters and make inferences from this postrior, usually via MCMC;
- Empirical Bayesian: Find the MLE or MAP of hyper-parameters by maximising their likelihood function and then fit group-level parameters in a general Bayesian way.

Pros and Cons

- Fully Bayesian will yield a shrinkage estimator for all μ_is. However, it often brings computational challenges, eg. high-dimensional sampling.
- Empirical Bayes is the combination of the previous two. It will produce a shrinkage estimator but not much as fully Bayesian, also less intensive than it.

Fully Bayesian Approach

Shijing Si*, David van Dyk*, Ted von Hippel†

We fit a dataset with these two approaches and compare the result later.

Fully Bayesian: The joint posterior density

$$p(\gamma, \tau^2, \mu_1, \cdots, \mu_I | \mathbf{Y}) \propto p(\gamma, \tau^2) \prod_{i=1}^{I} N(Y_i | \mu_i, \sigma_i) N(\mu_i | \gamma, \tau^2).$$

Shrinkage Estimator

- For brevity, we take the hyper-prior $p(\gamma, \tau^2) \propto 1$. Given γ and τ^2, the conditional densities for μ_is are

$$p(\mu_i|\gamma, \tau^2, \mathbf{Y}) \sim N(\hat{\mu}_i, \hat{\sigma}_i^2), \tag{3}$$

with $\hat{\mu}_i = \frac{\tau^2 Y_i + \sigma_i^2 \gamma}{\sigma_i^2 + \tau^2}$ and $\hat{\sigma}_i^2 = \frac{\sigma_i^2 \tau^2}{\sigma_i^2 + \tau^2}$.

$$p(\gamma|\tau^2, \mu_1, \cdots, \mu_I, \mathbf{Y}) \sim N(\frac{1}{I}\sum_{i=1}^{I}\mu_i, \frac{\tau^2}{n}),$$

$$p(\tau^2|\gamma, \mu_1, \cdots, \mu_I, \mathbf{Y}) \sim inv - \Gamma(\frac{n-2}{2}, \frac{1}{2}\sum_{i=1}^{I}(\mu_i - \gamma)^2).$$

Empirical Bayes Estimation

- It proceeds by obtaining an estimation for hyper parameters γ and τ^2 first. Usually, a MLE or MAP estimator. Then, μ_is are fitted in a general Bayesian framework.
- EM-type algorithms can be employed to find the MLE of hyper-parameters. μ_is are treated as missing data.

The log complete data likelihood function:

$$L(\gamma, \tau^2 | \mathbf{Y}, \mu_1, \cdots, \mu_I) = \sum_{i=1}^{I} \Big[-\frac{1}{2}\log(2\pi\sigma_i^2) - \frac{1}{2\sigma_i^2}(Y_i - \mu_i)^2 - \frac{1}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}(\mu_i - \gamma)^2 \Big].$$

EM algorithm

- Given the t-th iteration $\gamma^{(t)}$ and $\tau^{2(t)}$, conditional distribution of μ_is are normal with mean $E(\mu_i|\mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}) = \frac{\tau^{2(t)} Y_i + \sigma_i^2 \gamma^{(t)}}{\sigma_i^2 + \tau^{2(t)}}$ and variance $Var(\mu_i|\mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}) = \frac{\tau^{2(t)} \sigma_i^2}{\sigma_i^2 + \tau^{2(t)}}$.
- The $t+1$-st update:

$$\gamma^{(t+1)} = \frac{1}{I} \sum_{i=1}^{I} E(\mu_i|\mathbf{Y}, \gamma^{(t)}, \tau^{2(t)});$$

$$\tau^{2(t+1)} = \frac{1}{I} \sum_{i=1}^{I} E[(\mu_i - \gamma^{(t+1)})^2|\mathbf{Y}, \gamma^{(t)}, \tau^{2(t)}].$$

One Example

- We take one dataset in Rstan (with modifications on σ). It is a survey conducted by ETS to check the effect of coaching on students' performance in eight states in US. The dataset is presented as follows:

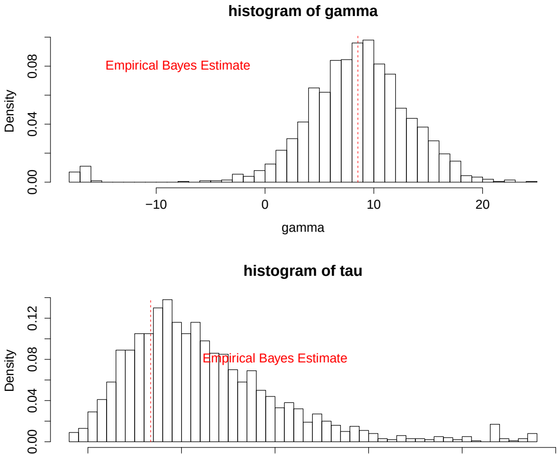
Y	28	8	-3	7	-1	1	18	12
σ	5	1	5	2	3	1	2	3

- The hierarchical model:

$$Y_i = \mu_i + \epsilon_i; \;\; \epsilon_i \sim N(0, \sigma_i^2),$$
$$\mu_i \sim N(\gamma, \tau^2), \;\; i = 1, \cdots, I.$$

Results

We use Rstan to fit this hierarchical model in a fully Bayesian way and code in R to find the MLE of hyper-parameters by EM algorithm.



Cont'd

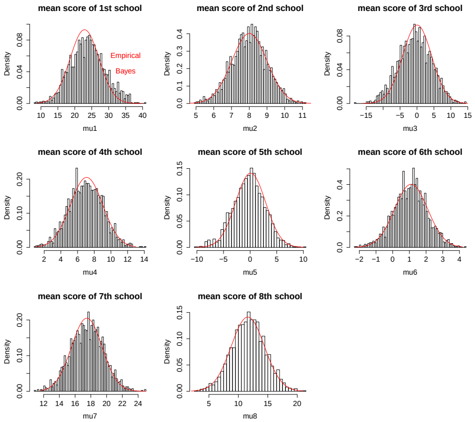


Figure 2 : Comparisons on indivual means from fully Bayesian (Histograms) and empirical Bayes (Red line)

Diagnostics about EM algorithm

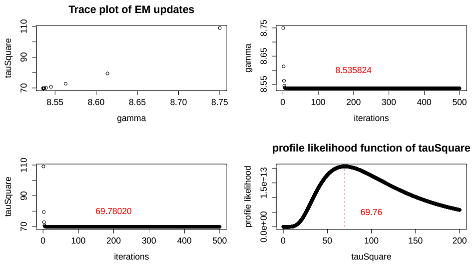


Figure 3 : Diagnostic pictures about EM algorithm

Contour plot of Marginal likelihood

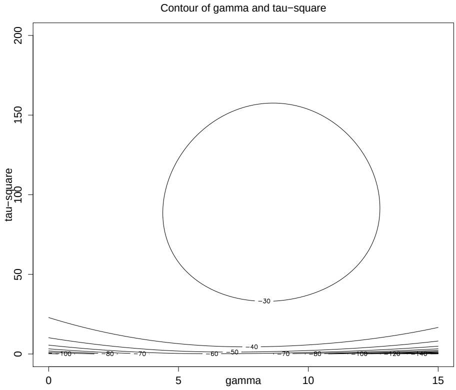
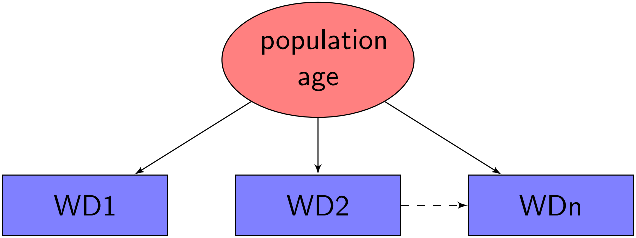


Figure 4 : Contour plot of marginal likelihood of γ and τ^2

Fitting the Distributions of Age of Halo WDs

- The goal here is the distribution of the population age of galactic halo WDs.
- μ the mean of the population logAge.
- Here is the simple hierarchical structure:



Mathematical Formula

- For the i-th star, denote its logAge, distance modulus, metallicity, mass as A_i, D_i, T_i, M_i.
- The data available is its photometry, $\mathbf{Y}_i$, the brightness on a range of wavelengths.
- Assume that the logAge of the Galactic halo is normal with mean γ and variance τ^2.

$$(\mathbf{Y}_i | A_i, D_i, T_i, M_i) \sim N(G(A_i, D_i, T_i, M_i), \Sigma_i);$$
$$A_i \sim N(\gamma, \tau^2), \ \ i = 1, \cdots, I,$$

with $G(\cdot)$ is a computer model and Σ_is are known from astronomers. Also, astronomers have very good priors on other parameters, $D_i, T_i, M_i, i = 1, \cdots, I$.

Photometry Data

Photometry is the intensity of an astronomical object's electromagnetic radiation, over a braod band of wavelengths.

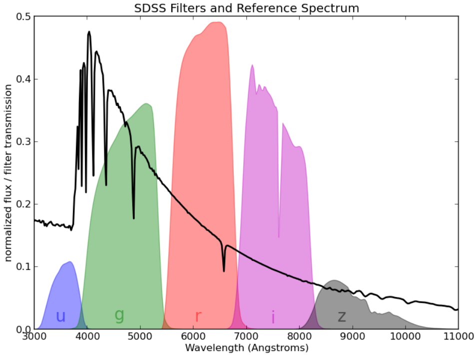
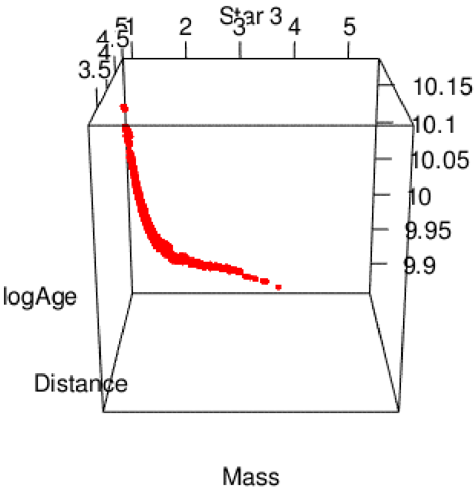


Figure 5 : The photometry over several filters.

Fitting one Star at a Time



Empirical Bayes Fitting

BASE9 can fit one star at a time, namely, it can draw a very good sample from the i-th star's posterior density for given γ and τ^2:

$$p(A_i, D_i, T_i, M_i|\mathbf{Y}_i) \sim p(\mathbf{Y}_i|A_i, D_i, T_i, M_i)\times p(A_i|\gamma, \tau^2)p(M_i)p(D_i)p(T_i),$$

where $p(A_i|\gamma, \tau^2), p(M_i), p(D_i), p(T_i)$ are prior densities.
We incorporate this fit-one-at-a-time algorithm in our fitting to a more complicated hierarchical model.

Monte Carlo EM algorithm

Monte Carlo EM (MCEM) algorithm is used to find the MLE of γ and τ^2.

- **MC-step:** Given the t-th update $\gamma^{(t)}$ and $\tau^{2(t)}$, draw a sample of size N, $(A_i^{[j]}, D_i^{[j]}, M_i^{[j]}, T_i^{[j]}),\ j = 1, 2, \cdots, N, i = 1, 2, \cdots, I$ from its posterior;
- **E-step:** Replace expectations with sample means;
- **M-step:**

$$\gamma^{(t+1)} = \frac{1}{I \cdot N} \sum_{i=1}^{I} \sum_{j=1}^{N} A_i^{[j]};$$

$$\tau^{2(t+1)} = \frac{1}{I \cdot N} \sum_{i=1}^{I} \sum_{j=1}^{N} (A_i^{[j]} - \gamma^{(t+1)})^2.$$

Primary Results

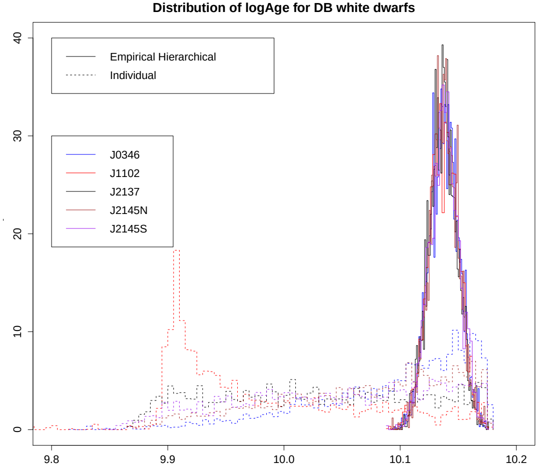


Figure 6 : Fit 5 stars in a hierarchical model.

IFMRs

- In stellar evolution, IFMR, initial-final mass relationship connects the mass of a white dwarf with the mass of its progenitor in the main-sequence.
- Commonly used IFMRs: William, Weidemann, Salaris I and Salaris II.
- Assume $M_{final} = f(M_{initial}, \beta)$, usually $f(\cdot)$ is taken as linear or piecewise linear.

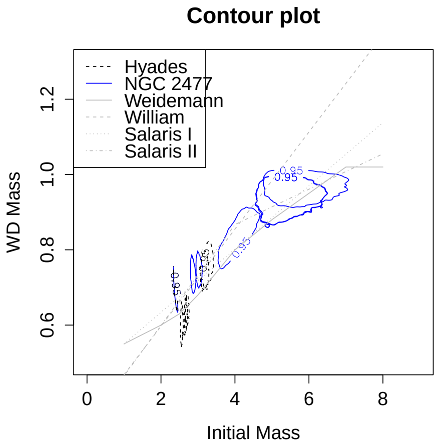


Figure 7 : Contour of the joint distribution of final mass and initial mass, which contains 95% of the distribution.

Hierarchical Modelling of IFMRs

$$(\mathbf{X}_i|\boldsymbol{\beta}_i) \sim N(G(A_i, \theta_i, \boldsymbol{\beta}_i), \boldsymbol{\Sigma}_i), i = 1, 2, \cdots, I;$$
$$(\boldsymbol{\beta}_i|\boldsymbol{\xi}, \boldsymbol{\Psi}) \sim N(\boldsymbol{\xi}, \boldsymbol{\Psi}),$$

- Σ_i are known;
- $\boldsymbol{\beta}_i$s are coefficients in IFMRs;
- We want to learn from some dataset about $\boldsymbol{\xi}$ and $\boldsymbol{\Psi}$.