

## Handout 16: Hierarchical Bayesian model <sup>a</sup>

Lecturer: Georgios P. Karagiannis

georgios.karagiannis@durham.ac.uk

### Aim

To be able to specify and analyze a Hierarchical Bayesian, as well as to extend previously introduced concepts in the Hierarchical Bayes framework.

### Basic reading list:

- Berger, J. O. (2013; Section 4.6). Statistical decision theory and Bayesian analysis. Springer.
- Robert, C. (2007, Sections 10.1-10.3). The Bayesian choice: from decision-theoretic foundations to computational implementation. Springer Science & Business Media.
- Robert, C. P., & Reber, A. (1998). Bayesian modelling of a pharmaceutical experiment with heterogeneous responses. Sankhyā: The Indian Journal of Statistics, Series B, 145-160. (<https://www.jstor.org/stable/pdf/25053027.pdf>)

### R-scripts:

- [https://github.com/georgios-stats/Bayesian\\_Statistics/blob/master/LectureHandouts/Rscripts/HierarchicalBayes/HierarchicalBayesPharmaceutical.R](https://github.com/georgios-stats/Bayesian_Statistics/blob/master/LectureHandouts/Rscripts/HierarchicalBayes/HierarchicalBayesPharmaceutical.R)

<sup>a</sup>Author: Georgios P. Karagiannis.

## 1 Hierarchical Bayesian Model

A Bayesian model can be hierarchical due to the sampling distribution modeling the observations or due to the decomposition of the prior information. A hierarchical Bayesian model involves several levels of conditional distributions.

**Definition 1.** A hierarchical Bayes model is a Bayesian statistical model with sampling distribution  $x \sim f(y|\theta)$  and prior  $\theta \sim \pi(\theta)$ , where the prior distribution  $\pi(\theta)$  is decomposed in conditional distributions. The Bayesian model is

$$\left\{ \begin{array}{l} y|\theta \sim f(y|\theta), \text{ sampling distribution} \\ \theta \sim \pi(\theta) \text{ marginal prior specified} \end{array} \right. \xrightarrow{\text{extend space}} \left\{ \begin{array}{ll} y|\theta \sim f(y|\theta) & \\ \theta \sim \pi_1(\theta|\phi_1) & \text{1st level prior} \\ \phi_1|\phi_2 \sim \pi_2(\phi_1|\phi_2) & \text{2nd level hyper-prior} \\ \vdots & \\ \phi_j|\phi_{j+1} \sim \pi_{j+1}(\phi_j|\phi_{j+1}) & j\text{th level hyper-prior} \\ \vdots & \\ \phi_{m-1}|\phi_m \sim \pi_m(\phi_{m-1}|\phi_m) & m\text{th level hyper-prior} \end{array} \right. \quad (1)$$

The joint distribution  $p(y, \theta, \phi_1, \dots, \phi_j, \dots, \phi_{m-1})$  has pdf

$$p(y, \theta, \phi_1, \dots, \phi_j, \dots, \phi_{m-1}) = f(y|\theta)\pi_1(\theta|\phi_1)\pi_2(\phi_1|\phi_2)\pi_3(\phi_2|\phi_3)\dots\pi(\phi_{m-1}|\phi_m)$$

The marginal prior distribution  $\pi(\theta)$  has pdf

$$\pi(\theta) = \int_{\Phi_1 \times \Phi_{m-1}} \pi_1(\theta|\phi_1)\pi_2(\phi_1|\phi_2)d\phi_1\pi_3(\phi_2|\phi_3)d\phi_2\dots\pi(\phi_{m-1}|\phi_m)d\phi_{m-1}.$$

The parameters  $\phi_j \in \Phi_j$  are called random hyper-parameters of level  $j$  for  $1 \leq j \leq m-1$ .

**Remark 2.** Hierarchical Bayesian model is simply a special type of Bayesian model, where

$$\begin{cases} y|\theta & \sim f(y|\theta) \\ \theta|\phi & \sim \pi(\theta|\phi) \\ \phi|\phi_m & \sim \pi(\phi|\phi_m) \end{cases} \quad (2)$$

for  $\phi = (\phi_1, \dots, \phi_{m-1})$ , and  $\phi_m$  fixed hyper-parameter.

**Note 3.** A hierarchical Bayesian model can be used as a mean to specify more diverse priors. This is achieved by setting  $\phi$  to be a random hyper-parameter with  $\phi|\phi_m \sim \pi_2(\phi|\phi_m)$  instead of setting  $\phi$  to have a fixed value.

**Note 4.** A hierarchical Bayesian model can be used when the sampling distribution or the prior distributions justify a certain structure.

**Example 5.** Regarding the fully hierarchical model (1), the full conditionals distributions of each element of  $\vartheta = (\theta, \phi_1, \dots, \phi_{m-1}) \in \Theta \times \Phi$  are given as:

$$\pi(\vartheta_j|y, \vartheta_{-j}) = \pi(\vartheta_j|y, \vartheta_{j-1}, \vartheta_{j+1})$$

with the convention

$$\vartheta_j = \begin{cases} \theta & , j = 1 \\ \phi_{j-1} & , j = 2, \dots, m \\ \phi_m & , j = m \end{cases}$$

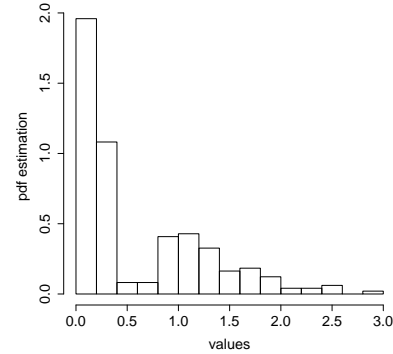
and  $\vartheta_{-j} = (\vartheta_1, \dots, \vartheta_{j-1}, \vartheta_{j+1}, \dots, \vartheta_m)$ .

**Proof.** Straightforward by using the Bayesian theorem. □

**Example 6.** Consider the following application where our concern is the distribution of enzymatic activity in the blood, for an enzyme involved in the metabolism of carcinogenic substances, among a group of  $n = 245$  unrelated individuals; aka cluster analysis.

Our observables are  $y = (y_1, \dots, y_n)$  with  $n = 245$ . In the Boxplot on the right, we can clear see that the distribution is multimodal, suggesting the existence of subpopulations/groups.

Interest lies on identifying subgroups of slow or fast metabolizers as a marker of genetic polymorphism in the general population. As we are interested in learning/identifying the sub-populations as the identity/label of the group from which each observation is drawn is unknown.



Questions of interest:

- How many groups exist?

Histogram of Enzyme dataset which is available from:  
<https://people.maths.bris.ac.uk/~mapjg/mixdata>

- To which group each observation belongs?.

As for the sampling model, we can assume that the  $i$ -th observation  $y_i$  is randomly drawn from the  $j$ -th group which has proportion  $\varpi_j$  in the population and which is distributed according to the sampling distribution  $y_i|\theta_j \sim f_j(y_i|\theta_j)$ .

For simplicity, let's assume that all groups are Normally distributed but with different parameter values  $\{\theta_j\}$ ; hence  $j$ -th group is  $y_i|\mu_j, \sigma_j^2 \sim N(y_i|\mu_j, \sigma_j^2)$  with  $\theta_j = (\mu_j, \sigma_j^2)$ .

35 It is natural to regard the group label  $z_i$  for the  $i$ th observation as a latent allocation variable: then  $z_i$  is supposed  
 36 to be distributed as  $z_i \sim f(z_i) = \varpi_{z_i}$  for  $z_i \in \{1, \dots, k\}$ , and  $y_i$  is supposed to be distributed as  $y_i|z_i, \theta_{z_i} \sim$   
 37  $f_{z_i}(y_i|z_i, \theta_{z_i}) := N(y_i|\mu_{z_i}, \sigma_{z_i}^2)$ , for  $i = 1, \dots, n$ ; i.e.

$$38 \quad \begin{cases} y_i|z_i, \mu_{z_i}, \sigma_{z_i}^2 & \sim f_{z_i}(y_i|\mu_{z_i}, \sigma_{z_i}^2) \\ z_i & \sim f(z_i) \end{cases} \implies \begin{cases} y_i|z_i, \mu_{z_i}, \sigma_{z_i}^2 & \sim N(y_i|\mu_{z_i}, \sigma_{z_i}^2) \\ z_i & \sim f(z_i) := \varpi_{z_i} \end{cases} \quad (3)$$

39 To complete the Bayesian model, we specify priors on the unknown quantities: Given there are  $k$  groups,  $\varpi_{1:k} \sim \text{Di}(\delta)$   
 40 for the group proportions,  $\mu_j \sim N(\xi_j, \sigma_j^2/\kappa)$  for the mean, and  $\sigma_j^2 \sim \text{Ga}(\alpha, \beta)$  for the variances. Assume we wish a  
 41 more spread prior for  $\sigma_j^2$  (for some reason), and hence we specify a hyper-prior on  $\beta$  as  $\beta \sim \text{Ga}(g, h)$ . As the number  
 42 of the groups is unknown, we assign prior  $k \sim \pi(k) \in \text{U}_{\text{discr}}(1, k_{\max})$ .

$$43 \quad \left\{ \begin{array}{ll} y_i|z_i, k, \varpi, \mu, \sigma^2 & \sim f_{z_i}(y_i|\varpi_{z_i}, \mu_{z_i}, \sigma_{z_i}^2) \\ & \text{for } i = 1, \dots, n \\ z_i|k & \sim f(z_i) := \varpi_{z_i} \\ & \text{for } i = 1, \dots, n \\ \varpi|k & \sim \text{Di}(\delta) \\ \mu_j|\sigma_j^2, k & \sim N(\xi, \sigma_j^2/\kappa) \\ & \text{for } j = 1, \dots, k \\ \sigma_j^2|\beta, k & \sim \text{Ga}(\alpha, \beta) \\ & \text{for } j = 1, \dots, k \\ \beta & \sim \text{Ga}(g, h) \\ k & \sim \text{U}_{\text{discr}}(1, k_{\max}) \end{array} \right. \quad (4)$$

A Directed Acyclic Graph representation of the hierarchical model (4). Circles correspond to random quantities. Squares correspond to fixed quantities.

44 The joint distribution has pdf

$$45 \quad p(y_{1:n}, z_{1:n}, k, \varpi_{1:k}, \mu_{1:k}, \sigma_{1:k}^2) = \underbrace{\prod_{i=1}^n N(y_i|\mu_{z_i}, \sigma_{z_i}^2)}_{f(y_{1:n}|z_{1:n}, \mu_{1:k}, \sigma_{1:k}^2)} \underbrace{\prod_{i=1}^n \varpi_{z_i}}_{f(z_{1:n}|k)} \underbrace{\prod_{j=1}^k N(\mu_j|\xi, \sigma_j^2/\kappa)}_{\pi(\mu_{1:k}|\sigma_{1:k}^2, k)} \underbrace{\prod_{j=1}^k \text{Ga}(\sigma_j^2|\alpha, \beta)}_{\pi(\sigma_{1:k}^2|\beta, k)} \underbrace{\text{Ga}(\beta|g, h)}_{\pi(\beta)} \underbrace{\frac{1}{|k_{\max}|}}_{\pi(k)}$$

46 The posterior  $\pi(k, \varpi, \mu, \sigma^2, z|y)$  can be computed with the Bayesian theorem, and factorized as

$$47 \quad \pi(k, \varpi, \mu, \sigma^2, \beta, z|y) = \frac{p(y, z, k, \varpi, \mu, \sigma^2, \beta)}{\int p(y, z, k, \varpi, \mu, \sigma^2, \beta) d(z, k, \varpi, \mu, \sigma^2, \beta)} \quad (5)$$

48 where to infer the number of groups from  $\pi(k|y)$ , the proportions in each group from  $\pi(\varpi_{1:k}|y, k)$ , the moments of  
 49 each group from  $\pi(\mu_{1:k}, \sigma_{1:k}^2|y, k)$ , and the allocation of each observation to each group with  $\pi(z|y, k, \varpi, \mu, \sigma^2)$ .

50 As the required integrals are intractable, we can resolve to numerical methods, etc... Monte Carlo e.g. via JAGS...

51 *Remark 7.* The Bayesian model with sampling distribution  $y \sim f(y|\theta)$  and prior  $\theta \sim \pi(\theta)$ , can be recovered from 2  
 52 by marginalizing the prior as

$$53 \quad \pi(\theta|\phi_m) = \int_{\Phi} \pi(\theta|\phi) \pi(\phi|\phi_m) d\phi = \int_{\Phi_1 \times \Phi_{m-1}} \pi(\theta|\phi_1) \pi(\phi_1|\phi_2) d\phi_1 \dots \pi(\phi_{m-1}|\phi_m) d\phi_{m-1}, \quad (6)$$

where  $\phi_m$  is just a fixed hyper-parameter. This reduction shows that hierarchical modelings are indeed included in the Bayesian paradigm.

*Note 8.* A particularly appealing aspect of hierarchical models is that they allow for conditioning on all levels, and this easy decomposition of the posterior. Consider the Bayesian hierarchical model (2) a parametric model  $f(y|\theta)$  with hierarchical prior  $\theta \sim \pi_1(\theta|\phi)$ , and  $\phi \sim \pi(\phi)$ . The posterior distribution of  $\theta$  is

$$\pi(\theta|y) = \int_{\Phi} \pi(\theta|y, \phi) \pi(\phi|y) d\phi \quad (7)$$

where

$$\begin{aligned} \pi(\theta|y, \phi) &= \frac{f(y|\theta) \pi_1(\theta|\phi)}{f_1(y|\phi)}; & \pi(\phi|y) &= \frac{f_1(y|\phi) \pi_2(\phi)}{f(y)}; \\ f_1(y|\phi) &= \int_{\Theta} f(y|\theta) \pi_1(\theta|\phi) d\theta; & f(y) &= \int_{\Theta} f_1(y|\phi) \pi_2(\phi) d\phi \end{aligned}$$

*Remark 9.* Note 8 has important consequences in terms of the computation of Bayes estimators, since it shows that  $\pi(\theta|y)$  can be simulated by generating, first,  $\phi$  from  $\pi(\phi|y)$  and then  $\theta$  from  $\pi(\theta|y, \phi)$ , if these two conditional distributions are easier to work with. (Snapshot from Term 2).

*Note 10.* Hierarchical decomposition (2) may facilitate the computation of intractable posterior moments. Let  $h$  be a function  $h : \Theta \rightarrow \mathbb{R}$ , then

$$E_{\pi}(h(\theta)|y) = E_{\pi}(E_{\pi}(h(\theta)|y, \phi) | y).$$

If  $E_{\pi}(h(\theta)|y) = \int h(\theta) \pi(\theta|y) d\theta$  is intractable and  $\theta$  has high dimensionality, one could possibly try to specify the prior decomposition  $\pi(\theta) = \int_{\Phi} \pi_1(\theta|\phi) \pi_2(\phi|\phi_m) d\phi$  in (6) such that  $E_{\pi}(h(\theta)|y, \phi)$  can be computed analytically, and  $\phi$  has low dimensionality. In that case one would have to compute the equivalent but lower dimensional (and hence easier) integral  $E_{\pi}(E_{\pi}(h(\theta)|y, \phi) | y) = \int E_{\pi}(h(\theta)|y, \phi) \pi(\phi|y) d\phi$ .

**Example 11.** Consider the 'Challenger O-ring' example from the Computer practicals. Let  $y_i$  denote the presence of a defective O-ring in the  $i$ th flight (0 for absence, and 1 for presence).

Assume that  $y_i$  can be modeled as observations generated independently from a Bernoulli distribution with parameter  $p_i$ . Here,  $p_i$  denotes the relative frequency of defective O-rings at flight  $i$ . We study if 'presence of a defective O-ring' ( $y$ ) depends on the 'temperature' ( $t$ ), or the 'pressure' ( $s$ ).

Let  $t_i$  denote the temperature (in F) in the platform, and let  $s_i$  denote the Leak check pressure (in PSI) before the  $i$ th flight. Here are some possible models of interest:

$$\begin{aligned} \mathcal{M}^I : p(t; \beta_{\mathcal{M}^I}, \mathcal{M}^I) &= \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} & ; \mathcal{M}^{IV} : p(t; \beta_{\mathcal{M}^{IV}}, \mathcal{M}^{IV}) &= \frac{\exp(\beta_0 + \beta_1 t + \beta_2 s)}{1 + \exp(\beta_0 + \beta_1 t + \beta_2 s)} \\ \mathcal{M}^{II} : p(t; \beta_{\mathcal{M}^{II}}, \mathcal{M}^{II}) &= \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} & ; \mathcal{M}^V : p(t; \beta_{\mathcal{M}^V}, \mathcal{M}^V) &= \frac{\exp(\beta_0 + \beta_1 t + \beta_2 s + \beta_3 ts)}{1 + \exp(\beta_0 + \beta_1 t + \beta_2 s + \beta_3 ts)} \\ \mathcal{M}^{III} : p(t; \beta_{\mathcal{M}^{III}}, \mathcal{M}^{III}) &= \frac{\exp(\beta_0 + \beta_2 s)}{1 + \exp(\beta_0 + \beta_2 s)} & \text{etc...} \end{aligned}$$

83 The Bayesian hierarchical model under consideration is:

$$84 \quad \begin{cases} y_i|\theta \sim f(y_i|\theta) :: & \left\{ y_i|\mathcal{M}, \beta_{\mathcal{M}} \sim \text{Br}\left(y_i \mid \frac{\exp(x_i^\top \beta_{\mathcal{M}})}{1 + \exp(x_i^\top \beta_{\mathcal{M}})}\right), \quad \text{for } i = 1, \dots, n \right. \\ \theta|\phi_1 \sim \pi(\theta|\phi_1) :: & \begin{cases} \beta_j|\mathcal{M} \sim (1 - \gamma_j)1_0(\beta_j) + \gamma_j\text{N}(\beta_j|\mu_0, \sigma_0^2) \quad j = 1, \dots, d \\ \mathcal{M} = (\gamma_1, \dots, \gamma_d) \\ \gamma_j|\varpi \sim \text{Br}(\varpi), \quad j = 1, \dots, d \end{cases} \\ \phi_1|\phi_2 \sim \pi(\phi_1|\phi_2) :: & \left\{ \varpi \sim \text{Be}(a_0, b_0) \right. \end{cases}$$

85 where  $\theta = (\mathcal{M}, \beta_{\mathcal{M}})$ ,  $\phi_1 = \varpi$ , and  $\phi_2 = (a_0, b_0)$ . Above, in the prior we considered an extra level of uncertainty by  
86 considering  $\varpi \sim \text{Be}(a_0, b_0)$ .

87 Here we added an additional level of uncertainty, and set  $\varpi \sim \text{Be}(a_0, b_0)$  which creates a more diverse prior model,  
88 compared to the computer practical handout example where we had set  $\varpi = 0.5$ .

89 Now the joint probability distribution has pdf

$$90 \quad p(y, \beta_{\mathcal{M}}, \mathcal{M}, \varpi) = \underbrace{\prod_{i=1}^n \text{Br}\left(y_i \mid \frac{\exp(x_i^\top \beta_{\mathcal{M}})}{1 + \exp(x_i^\top \beta_{\mathcal{M}})}\right)}_{f(y|\theta)} \underbrace{\prod_{i=1}^n ((1 - \gamma_j)1_0(\beta_j) + \gamma_j\text{N}(\beta_j|\mu_0, \sigma_0^2))}_{\pi(\theta|\phi_1)} \underbrace{\prod_{i=1}^n \text{Br}(\gamma_i|\varpi)\text{Be}(\varpi|a_0, b_0)}_{\pi(\phi_1|\phi_2)}$$

91 **Example 12.** Robert and Reber (1998) considers an experiment under which rats are intoxicated by a substance, then  
92 treated by either a placebo or a drug. (See: <https://www.jstor.org/stable/pdf/25053027.pdf>)

93 **Statistical model** ( $f(y|\theta)$ ): The model associated with this experiment is a linear additive model effect: given  $x_{ij}$ ,  
94  $y_{ij}$  and  $z_{ij}$ ,  $j$ th responses of the  $i$ th rat at the control, intoxication and treatment stages, respectively. The statistical  
95 model was specified such as that ( $1 \leq i \leq I$ )

$$\begin{aligned} 96 \quad x_{i,j} &\sim \text{N}(\theta_i, \sigma_c^2) & , 1 \leq j \leq J_i^c \\ 97 \quad y_{i,j} &\sim \text{N}(\theta_i + \delta_i, \sigma_a^2) & , 1 \leq j \leq J_i^a, \\ 98 \quad z_{i,j} &\sim \text{N}(\theta_i + \delta_i + \xi_i, \sigma_t^2) & , 1 \leq j \leq J_i^t, \end{aligned}$$

99 where  $\theta_i$  is the average control measurement,  $\delta_i$  the average intoxication effect and  $\xi_i$  the average treatment effect  
100 for the  $i$ th rat, the variances of these measurements being constant for the control, the intoxication and the treatment  
101 effects. An additional (observed) variable is  $w_i$ , which is equal to 1 if the rat is treated with the drug, and 0 otherwise.

102 **Prior model**  $\pi(\theta|\phi)$ : The different individual averages are related through a common (conjugate) prior distribution,

$$\begin{aligned} 103 \quad \theta_i &\sim \text{N}(\mu_\theta, \sigma_\theta^2), & \delta_i &\sim \text{N}(\mu_\delta, \sigma_\delta^2), & \xi_i|w_i &\sim \begin{cases} \text{N}(\mu_P, \sigma_P^2) & , w_i = 0 \\ \text{N}(\mu_D, \sigma_D^2) & , w_i = 1 \end{cases} \\ 104 \quad \sigma_c &\sim \pi(\sigma_c) \propto \frac{1}{\sigma_c}, & \sigma_a &\sim \pi(\sigma_a) \propto \frac{1}{\sigma_a}, & \sigma_t &\sim \pi(\sigma_t) \propto \frac{1}{\sigma_t}, \end{aligned} \quad (8)$$

105 This modeling seems to describe the natural phenomenon realistically enough, in the sense the responses  $x_{ij}$ ,  $y_{ij}$  and  
106  $z_{ij}$

**Hyper-priors**  $\pi(\phi|\phi_m)$ : For the higher levels of prior ( $\pi(\phi|\phi_m)$  in Eq 2), they considered improper (Jeffrey's) hyper-priors.

$$(\mu_\theta, \sigma_\theta) \sim \pi(\mu_\theta, \sigma_\theta) \propto \frac{1}{\sigma_\theta}, \quad (\mu_\delta, \sigma_\delta) \sim \pi(\mu_\delta, \sigma_\delta) \propto \frac{1}{\sigma_\delta}, \quad (\mu_P, \sigma_P) \sim \pi(\mu_P, \sigma_P) \propto \frac{1}{\sigma_P}, \quad (9)$$

$$(\mu_D, \sigma_D) \sim \pi(\mu_D, \sigma_D) \propto \frac{1}{\sigma_D}. \quad (10)$$

The priors in lines (8), (9) and (10) are improper non-informative priors. One could have specify proper priors, like Normal-Inverse Gamma which are conjugate, however in that case he/she should have to specify the values for the fixed hyper-parameters.

As improper priors are specified, one need to study under what conditions the above improper priors lead to a proper (well defined) posterior –we omit this step here...

I have an R script with a demo in [https://github.com/georgios-stats/Bayesian\\_Statistics/blob/master/LectureHandouts/Rscripts/HierarchicalBayes/HierarchicalBayesPharmaceutical.R](https://github.com/georgios-stats/Bayesian_Statistics/blob/master/LectureHandouts/Rscripts/HierarchicalBayes/HierarchicalBayesPharmaceutical.R)

**Example 13.** (Cont. Example 6) From another point of view, recall that the compound distribution  $f(y_i|k, \varpi_{1:k}\theta_{1:k})$  of (3) is mixture model of distribution

$$y_i|k, \theta_{1:k} \sim f(y_i|k, \varpi_{1:k}, \theta_{1:k}) = \sum_{j=1}^k \varpi_j f_j(y_i|\theta_j) = \sum_{j=1}^k \varpi_j \mathcal{N}(y_i|\mu_j, \sigma_j^2), \text{ for } i = 1, \dots, n \quad (11)$$

is a suitable sampling distribution for modeling heterogeneous populations. Then by marginalizing, we can get the equivalent model

$$\begin{cases} y_i|k, \varpi, \mu, \sigma^2 & \sim f(y_i|k, \varpi, \mu, \sigma^2) := \sum_{j=1}^k \varpi_j \mathcal{N}(y_i|\mu_j, \sigma_j^2), \text{ for } i = 1, \dots, n \\ \varpi|k & \sim \text{Di}(\delta) \\ \mu_j|\sigma_j^2, k & \sim \mathcal{N}(\mu_j|\xi, \sigma_j^2), \text{ for } j = 1, \dots, k \\ \sigma_j^2|k & \sim \text{Ga}(a, \beta), \text{ for } j = 1, \dots, k \\ \beta & \sim \text{Ga}(g, h) \\ k & \sim \text{U}_{\text{discr}}(1, k_{\max}) \end{cases} \quad (12)$$

The joint distribution that admits pdf

$$p(y, k, \varpi, \mu, \sigma^2, \beta) = f(y|k, \varpi, \mu, \sigma^2) \pi(\varpi|k) \pi(\mu|\sigma^2, k) \pi(\sigma^2|k, \beta) \pi(\beta) \pi(k).$$

The posterior  $\pi(k, \varpi, \mu, \sigma^2|y)$  can be computed with the Bayesian theorem, and factorized as

$$\pi(k, \varpi, \mu, \sigma^2, \beta|y) = \frac{p(y, \varpi, \mu, \sigma^2, \beta)}{\int p(y, k, \varpi, \mu, \sigma^2, \beta) d(k, \varpi, \mu, \sigma^2, \beta)} \quad (13)$$

Models (4) and (12) in the sense that posterior (13) is the marginal of the posterior (5).

## 2 Non-identifiability issue

A parametric model for which an element of the parametrisation is redundant is said to be non-identified. Let Bayesian model  $(f(y|\theta), \pi(\theta))$ , where  $\theta = (\theta_1, \theta_2) \in \Theta_1 \times \Theta_2$ , and assume that the parametric model does not depend on  $\theta_1$ ; i.e.  $f(y|\theta_1, \theta_2) = f(y|\theta_2)$ . The fact that the likelihood does not depend on  $\theta_1$  suggests that  $y$  does not provide information about  $\theta_1$  directly.

Bayesian analysis of a non-identified model is always possible if a suitable prior  $\Pi(\theta_1, \theta_2)$  on all the parameters is specified. For instance, if one specifies a priori that learning the value of  $\theta_2$  may change his belief about  $\theta_1$ , via  $\pi(\theta_1|\theta_2) \neq \pi(\theta_1)$ .

Factorize the prior distribution as  $\pi(\theta_1, \theta_2) = \pi(\theta_1|\theta_2)\pi(\theta_2)$ . Then, we have the following PDF/PMF

$$\begin{aligned}\pi(\theta_1, \theta_2|y) &\propto f(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2) = f(y|\theta_2)\pi(\theta_1|\theta_2)\pi(\theta_2) \implies \\ \pi(\theta_1, \theta_2|y) &= \pi(\theta_2|y)\pi(\theta_1|\theta_2) \implies \\ \pi(\theta_1|y, \theta_2) &= \pi(\theta_1|\theta_2)\end{aligned}\tag{14}$$

$$\begin{aligned}\pi(\theta_2|y) &= \frac{f(y|\theta_2)\pi(\theta_2)}{\int_{\Theta_2} f(y|\theta_2)\pi(\theta_2)d\theta_2} \cdot \\ \pi(\theta_1|y) &= \int_{\Theta_2} \pi(\theta_1|\theta_2)\pi(\theta_2)d\theta_2\end{aligned}\tag{15}$$

Here,  $\theta_1$  is said to be non-identifiable parameter from the data  $y$ , because  $y$  provides no direct information about  $\theta_1$ . Inference about  $\theta_1$  based on marginal posterior  $\pi(\theta_1|y)$  depends on  $y$  but the information provided about  $\theta_1$  comes indirectly through the marginal posterior of  $\theta_2$ , see (15). Equivalently, (15) implies that  $y$  provides no information about  $\theta_1$  given  $\theta_2$ .

If we a priori specify that learning the value of  $\theta_2$  does not change our belief about  $\theta_1$   $\pi(\theta_1|\theta_2) = \pi(\theta_1)$ , then (15) becomes  $\pi(\theta_1|y) = \pi(\theta_1)$  and hence data  $y$  provide no information about  $\theta_1$  at all.

**Example 14.** (Cont Example 13) It is not difficult to understand that the Bayesian model as defined in Example 6 is non-identifiable. For simplicity we focus on the Bayesian mixture of  $k = 2$  components with

$$\begin{aligned}y|\varpi, \mu, \sigma^2 &\sim f(y|\varpi, \mu, \sigma^2) := \varpi_1 N(y|\mu_1, \sigma_1^2) + \varpi_2 N(y|\mu_2, \sigma_2^2) \\ \pi(\varpi, \mu, \sigma^2) &= \underbrace{N(\mu_1|\xi, \sigma_1^2)N(\mu_2|\xi, \sigma_2^2)}_{\pi(\mu|\sigma^2)} \underbrace{\text{Ga}(\sigma_1^2|\alpha, \beta)\text{Ga}(\sigma_2^2|\alpha, \beta)}_{\pi(\sigma^2)} \text{Di}(\varpi|\delta)\end{aligned}\tag{16}$$

which leads to a posterior such as

$$\pi(\varpi, \mu, \sigma^2|y) \propto [\varpi_1 N(y|\mu_1, \sigma_1^2) + \varpi_2 N(y|\mu_2, \sigma_2^2)] N(\mu_1|\xi, \sigma_1^2) N(\mu_2|\xi, \sigma_2^2) \text{Ga}(\sigma_1^2|\alpha, \beta) \text{Ga}(\sigma_2^2|\alpha, \beta) \text{Di}(\varpi|\delta)$$

Here the parametrization is non-identifiable because the symmetry in the sampling distribution

$$\varpi_1 N(y|\mu_1, \sigma_1^2) + \varpi_2 N(y|\mu_2, \sigma_2^2) = \varpi_2 N(y|\mu_2, \sigma_2^2) + \varpi_1 N(y|\mu_1, \sigma_1^2)$$

and the naive prior in (16) produce a posterior such that

$$\pi\left(\varpi = \begin{pmatrix} \varpi_1 \\ \varpi_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \sigma^2 = \begin{pmatrix} \sigma_1^2 \\ \sigma_2^2 \end{pmatrix} | y\right) = \pi\left(\varpi = \begin{pmatrix} \varpi_2 \\ \varpi_1 \end{pmatrix}, \mu = \begin{pmatrix} \mu_2 \\ \mu_1 \end{pmatrix}, \sigma^2 = \begin{pmatrix} \sigma_2^2 \\ \sigma_1^2 \end{pmatrix} | y\right)$$

This parametrization is not meaningful for parametric inference. In Bayesian stats this can be resolved for instance by changing the prior and imposing an identifiability constrain as

$$\pi^*(\mu|\sigma^2) = \frac{N(\mu_1|\xi, \sigma_1^2)N(\mu_2|\xi, \sigma_2^2)1(\mu_1 \leq \mu_2)}{\int N(\mu_1|\xi, \sigma_1^2)N(\mu_2|\xi, \sigma_2^2)1(\mu_1 \leq \mu_2) d(\mu_1, \mu_2)} \propto \pi(\mu|\sigma^2)1(\mu_1 \leq \mu_2)$$

162 A schematic of the non-identifiability issue:

163 The non-identifiable model

$$\begin{cases} y_i | \varpi, \mu, \sigma^2 & \sim \varpi_1 N(y | \mu_1, \sigma_1^2) + \varpi_2 N(y | \mu_2, \sigma_2^2) \\ \varpi & \sim \text{Di}(\delta) \\ \mu | \sigma^2 & \sim N(\mu_1 | \xi, \sigma_1^2) N(\mu_2 | \xi, \sigma_2^2) \\ \sigma^2 & \sim \text{Ga}(\sigma_1^2 | \alpha, \beta) \text{Ga}(\sigma_2^2 | \alpha, \beta) \end{cases} \quad (17)$$

165 produces marginal posteriors

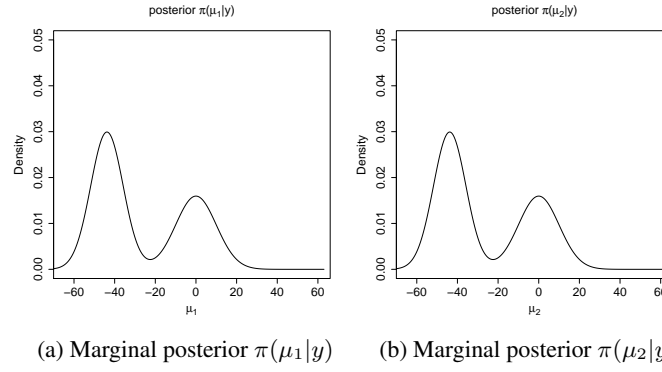


Figure 1: Some marginal posteriors of the non-identifiable model (17)

166 After non-identifiability is resolved, the identifiable model

$$\begin{cases} y_i | \varpi, \mu, \sigma^2 & \sim \varpi_1 N(y | \mu_1, \sigma_1^2) + \varpi_2 N(y | \mu_2, \sigma_2^2) \\ \varpi & \sim \text{Di}(\delta) \\ \mu | \sigma^2 & \sim N(\mu_1 | \xi, \sigma_1^2) N(\mu_2 | \xi, \sigma_2^2) 1(\mu_1 \leq \mu_2) \\ \sigma^2 & \sim \text{Ga}(\sigma_1^2 | \alpha, \beta) \text{Ga}(\sigma_2^2 | \alpha, \beta) \end{cases} \quad (18)$$

168 produces marginal posteriors

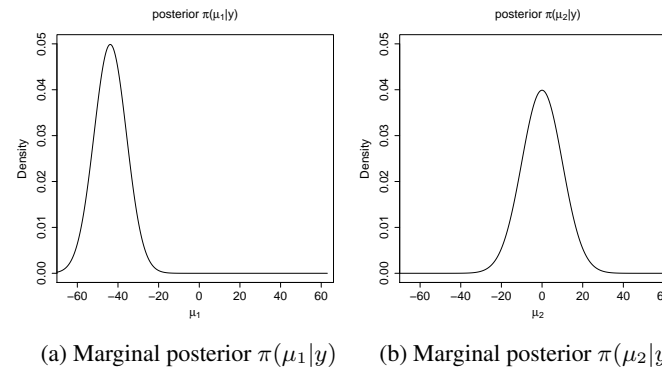


Figure 2: Some marginal posteriors of the non-identifiable model (18)



## A Appendix

**Example 15.** (Cont...) You may use

$$-\frac{1}{2} \sum_{i=1}^n \frac{(x - \mu_i)^2}{\sigma_i^2} = -\frac{1}{2} \frac{(x - \hat{\mu})^2}{\hat{\sigma}^2} + C; \quad \hat{\sigma}^2 = \left( \sum_{i=1}^n \frac{1}{\sigma_i^2} \right)^{-1}; \quad \hat{\mu} = \hat{\sigma}^2 \left( \sum_{i=1}^n \frac{\mu_i}{\sigma_i^2} \right); \quad C = \frac{1}{2} \frac{(\sum_{i=1}^n \frac{\mu_i}{\sigma_i^2})^2}{\sum_{i=1}^n \frac{1}{\sigma_i^2}} - \frac{1}{2} \sum_{i=1}^n \frac{\mu_i^2}{\sigma_i^2}$$

The joint posterior pdf of  $\vartheta = (\theta_{1:I}, \delta_{1:I}, \xi_{1:I}, \sigma_c^2, \sigma_a^2, \sigma_t^2, \sigma_\theta^2, \sigma_\delta^2, \sigma_P^2, \sigma_D^2, \mu_\theta, \mu_\delta, \mu_P, \mu_D)$  given obs.  $x, y, z$  is

$$\begin{aligned} \pi(\vartheta|x, y, z) &\propto \prod_{i=1}^I \left[ \exp \left( -\frac{(\theta_i - \mu_\theta)^2}{2\sigma_\theta^2} - \frac{(\delta_i - \mu_\delta)^2}{2\sigma_\delta^2} \right) \prod_{j=1}^{J_i^c} \exp \left( -\frac{(x_{i,j} - \theta_i)^2}{2\sigma_c^2} \right) \times \prod_{j=1}^{J_i^a} \exp \left( -\frac{(y_{i,j} - \theta_i - \delta_i)^2}{2\sigma_a^2} \right) \right. \\ &\quad \times \prod_{j=1}^{J_i^t} \exp \left( -\frac{(z_{i,j} - \theta_i - \delta_i - \xi_i)^2}{2\sigma_t^2} \right) \times \prod_{w_i=0} \exp \left( -\frac{(\xi_i - \mu_P)^2}{2\sigma_P^2} \right) \prod_{w_i=0} \exp \left( -\frac{(\xi_i - \mu_D)^2}{2\sigma_D^2} \right) \Big] \\ &\quad \times \sigma_c^{-\sum_i J_i^c - 1} \sigma_a^{-\sum_i J_i^a - 1} \sigma_t^{-\sum_i J_i^t - 1} \sigma_\theta^{I-1} \sigma_\delta^{I-1} \sigma_P^{I_D-1} \sigma_D^{I_P-1}. \end{aligned}$$

The joint posterior distributions is not of standard form, and its pdf is intractable. However the full conditionals are of standard form. For instance, the full conditional posterior distribution density

$$\begin{aligned} \pi(\delta_{1:I}|x_{\text{all}}, y_{\text{all}}, z_{\text{all}}, \theta_{1:I}, \xi_{1:I}, \sigma_c^2, \sigma_a^2, \sigma_t^2, \sigma_\theta^2, \sigma_\delta^2, \sigma_P^2, \sigma_D^2, \mu_\theta, \mu_\delta, \mu_P, \mu_D) \\ &\propto \prod_{i=1}^I \left[ \exp \left( -\frac{(\delta_i - \mu_\delta)^2}{2\sigma_\delta^2} \right) \times \prod_{j=1}^{J_i^a} \exp \left( -\frac{(y_{i,j} - \theta_i - \delta_i)^2}{2\sigma_a^2} \right) \times \prod_{j=1}^{J_i^t} \exp \left( -\frac{(z_{i,j} - \theta_i - \delta_i - \xi_i)^2}{2\sigma_t^2} \right) \right] \\ &\propto \prod_{i=1}^I \left[ \exp \left( -\frac{(\delta_i - \mu_\delta)^2}{2\sigma_\delta^2} - \sum_{j=1}^{J_i^a} \frac{(\delta_i - (y_{i,j} - \theta_i))^2}{2\sigma_a^2} - \sum_{j=1}^{J_i^t} \frac{(\delta_i - (z_{i,j} - \theta_i - \xi_i))^2}{2\sigma_t^2} \right) \right] \\ &\propto \prod_{i=1}^I \left[ \exp \left( -\frac{(\delta_i - \mu_{\delta,i}^*)^2}{2(\sigma_{\delta,i}^*)^2} + \text{const...} \right) \right] \propto \prod_{i=1}^I \left[ \exp \left( -\frac{(\delta_i - \mu_{\delta,i}^*)^2}{2(\sigma_{\delta,i}^*)^2} + \text{const...} \right) \right] \\ &\propto \prod_{i=1}^I \text{N}(\delta_i | \mu_{\delta,i}^*, (\sigma_{\delta,i}^*)^2) \end{aligned}$$

with

$$\delta_i | \text{rest}, \dots \stackrel{\text{ind}}{\sim} \text{N}(\mu_{\delta,i}^*, (\sigma_{\delta,i}^*)^2), \quad \forall i = 1, \dots, n$$

where

$$(\sigma_{\delta,i}^*)^2 = \left( \frac{1}{\sigma_\delta^2} + \frac{1}{\sigma_a^2} J_i^a + \frac{1}{\sigma_t^2} J_i^t \right)^{-1}; \quad \mu_{\delta,i}^* = (\sigma_{\delta,i}^*)^2 \left( \frac{\mu_\delta}{\sigma_\delta^2} + \frac{\sum_{j=1}^{J_i^a} y_{i,j} - J_i^a \theta_i}{\sigma_a^2} + \frac{\sum_{j=1}^{J_i^t} y_{i,j} - J_i^t \theta_i - J_i^t \xi_i}{\sigma_t^2} \right)$$

Notice that  $\delta_i$  are a postriori independent given all the resp unknown parameters  $(\theta_{1:I}, \xi_{1:I}, \sigma_c^2, \sigma_a^2, \sigma_t^2, \sigma_\theta^2, \sigma_\delta^2, \sigma_P^2, \sigma_D^2, \mu_\theta, \mu_\delta, \mu_P, \mu_D)$ . Notice that the prior  $\delta_i \sim \text{N}(\mu_\delta, \sigma_\delta^2)$  in Example 12 is conditional conjugate prior of  $\delta_i$ .

Try to compute the rest

$$\begin{aligned} \pi(\theta_{1:I} | \text{rest}, \dots) &\sim?; & \pi(\sigma_t^2 | \text{rest}, \dots) &\sim?; & \pi(\sigma_c^2 | \text{rest}, \dots) &\sim?; & \pi(\sigma_a^2 | \text{rest}, \dots) &\sim?; \\ \pi(\xi_{1:I} | \text{rest}, \dots) &\sim?; & \pi(\sigma_\theta^2 | \text{rest}, \dots) &\sim?; & \pi(\sigma_\delta^2 | \text{rest}, \dots) &\sim?; & \pi(\sigma_P^2 | \text{rest}, \dots) &\sim?; \text{ etc...} \end{aligned}$$

See the solutions in: Robert, C. P., & Reber, A. (1998) from the link (<https://www.jstor.org/stable/pdf/25053027.pdf>).

**Example 16.** (Cont. Example 6) As seen later, although equivalent in the sense that they produce the same inference, the expended hierarchical model (5) is computational convenient compared to hierarchical model (13) in the sense that its priors are conditional conjugate.

For simplicity assume that the number of groups is known and fixed to  $k$ . The full conditional posteriors in model (5) are:

$$\begin{aligned}
 w|y... &\sim \text{Di}(\delta + n_1, \dots, \delta + n_k); \text{ where } n_j = \sum_{i=1}^n 1(z_i = j) \\
 \mu_j|y... &\sim \text{N}\left(\frac{\sum_{i:z_i=j} y_i - \xi\kappa}{n_j + \kappa}, \frac{\sigma_j^2}{n_j + \kappa}\right), \text{ for } j = 1, \dots, k \\
 \sigma_j^2|y... &\sim \text{IG}\left(a + \frac{n_j}{2}, \beta + \frac{1}{2} \sum_{i:z_i=j} (y_i - \mu_j)^2\right), \text{ for } j = 1, \dots, k \\
 z_i|y... &\sim \pi(z_i = j|y...) = \frac{\frac{w_j}{\sigma_j} \exp\left(-\frac{1}{2} \frac{(y_i - \mu_j)^2}{\sigma_j^2}\right)}{\sum_{j'=1}^k \frac{w_{j'}}{\sigma_{j'}} \exp\left(-\frac{1}{2} \frac{(y_i - \mu_{j'})^2}{\sigma_{j'}^2}\right)}; \text{ for } i = 1, \dots, n \\
 \beta|y... &\sim \text{Ga}\left(g + k\alpha, h + \sum_{j=1}^k \sigma_j^2\right)
 \end{aligned}$$

which can be used in Monte Carlo integration.

Model (13) does not produce full conditional posteriors of standard form, due to the summation in the likelihood.