

Handout 16: Hierarchical Bayesian model ^a

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:

- 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

^aAuthor: Georgios P. Karagiannis.

1 Hierarchical Bayesian Model

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

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

$$\text{and we write } \left\{ \begin{array}{ll} y|\theta & \sim f(y|\theta) \\ \theta|\phi_1 & \sim \pi_1(\theta|\phi_1) \\ \phi_1|\phi_2 & \sim \pi_2(\phi_1|\phi_2) \\ \vdots & \\ \phi_j|\phi_{j+1} & \sim \pi_{j+1}(\phi_j|\phi_{j+1}) \\ \vdots & \\ \phi_{m-1} & \sim \pi_m(\phi_{m-1}|\phi_m) \end{array} \right. \quad (1)$$

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 ϕ_m fixed hyper-parameter.

Remark 3. The Bayesian model with sampling distribution $y \sim f(y|\theta)$ and prior $\theta \sim \pi(\theta)$, can be recovered from 2 by marginalizing the prior as

$$\pi(\theta) = \int_{\Phi} \pi(\theta|\phi)\pi(\phi|\phi_m)d\phi = \int_{\Phi_1 \times \Phi_{m-1}} \pi(\theta|\phi_1)\pi(\phi_1|\phi_2)d\phi_1\dots\pi(\phi_{m-1}|\phi_m)d\phi_{m-1}, \quad (3)$$

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

Note 4. A hierarchical Bayesian model can be used as a mean to specify more diverse priors. This is achieved by setting ϕ to be a random hyper-parameter with $\phi|\phi_m \sim \pi_2(\phi|\phi_m)$ instead of setting ϕ to have a fixed value. See Example 5.

Example 5. 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:

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

36 The Bayesian hierarchical model under consideration is:

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

38 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
 39 considering $\varpi \sim \text{Be}(a_0, b_0)$.

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

42 Now the joint probability distribution has pdf

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

44 *Note 6.* A hierarchical Bayesian model can be used when the sampling distribution or the prior distributions justify a
 45 certain structure. See Example 7.

46 **Example 7.** Robert and Reber (1998) considers an experiment under which rats are intoxicated by a substance, then
 47 treated by either a placebo or a drug. (See: <https://www.jstor.org/stable/pdf/25053027.pdf>)

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

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

54 where θ_i is the average control measurement, δ_i the average intoxication effect and ξ_i the average treatment
 55 effect for the i th rat, the variances of these measurements being constant for the control, the intoxication and
 56 the treatment effects. An additional (observed) variable is w_i , which is equal to 1 if the rat is treated with the
 57 drug, and 0 otherwise.

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

$$\begin{aligned}
 \theta_i &\sim \text{N}(\mu_\theta, \sigma_\theta^2), & \delta_i &\sim \text{N}(\mu_\delta, \sigma_\delta^2), & \xi_i|w_i &\sim \begin{cases} \text{N}(\mu_P, \sigma_P^2) & , w_i = 0 \\ \text{N}(\mu_D, \sigma_D^2) & , w_i = 1 \end{cases} \\
 \sigma_c &\sim \pi(\sigma_c) \propto \frac{1}{\sigma_c}, & \sigma_a &\sim \pi(\sigma_a) \propto \frac{1}{\sigma_a}, & \sigma_t &\sim \pi(\sigma_t) \propto \frac{1}{\sigma_t},
 \end{aligned} \tag{4}$$

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

$$(\mu_\theta, \sigma_\theta) \sim \pi(\mu_\theta, \sigma_\theta) \propto \frac{1}{\sigma_\theta}, \quad (\mu_\delta, \sigma_\delta) \sim \pi(\mu_\delta, \sigma_\delta) \propto \frac{1}{\sigma_\delta}, \quad (\mu_P, \sigma_P) \sim \pi(\mu_P, \sigma_P) \propto \frac{1}{\sigma_P}, \quad (5)$$

$$(\mu_D, \sigma_D) \sim \pi(\mu_D, \sigma_D) \propto \frac{1}{\sigma_D}. \quad (6)$$

The priors in lines (4), (5) and (6) 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...

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 a hierarchical prior $\theta \sim \pi_1(\theta|\phi)$, and $\phi \sim \pi(\phi)$. The posterior distribution of θ is

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

where

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

Remark 9. It has important consequences in terms of the computation of Bayes estimators, though, since it shows that $\pi(\theta|y)$ can be simulated by generating, first, ϕ from $\pi(\phi|y)$ and then θ 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 θ 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 (3) such that $E_\pi(h(\theta)|y, \phi)$ can be computed analytically, and ϕ 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. 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. \square

Example 12. (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

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

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

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

115 Notice that δ_i are a postriori independent given all the resp unknown parameters
 116 $(\theta_{1:I}, \xi_{1:I}, \sigma_c^2, \sigma_a^2, \sigma_t^2, \sigma_\theta^2, \sigma_\delta^2, \sigma_P^2, \sigma_D^2, \mu_\theta, \mu_\delta, \mu_P, \mu_D)$. Notice that the prior $\delta_i \sim N(\mu_\delta, \sigma_\delta^2)$ in Example 7 is
 117 conditional conjugate prior of δ_i .

118 Try to compute the rest

119	$\pi(\theta_{1:I} \text{rest}, \dots) \sim ? ;$	$\pi(\sigma_t^2 \text{rest}, \dots) \sim ?$	<i>etc...</i>
120	$\pi(\xi_{1:I} \text{rest}, \dots) \sim ? ;$	$\pi(\sigma_\theta^2 \text{rest}, \dots) \sim ?$	
121	$\pi(\sigma_c^2 \text{rest}, \dots) \sim ? ;$	$\pi(\sigma_\delta^2 \text{rest}, \dots) \sim ?$	
122	$\pi(\sigma_a^2 \text{rest}, \dots) \sim ? ;$	$\pi(\sigma_P^2 \text{rest}, \dots) \sim ?$	

- 123 • See the solutions in: Robert, C. P., & Reber, A. (1998). Bayesian modelling of a pharmaceutical experiment
 124 with heterogeneous responses. Sankhy: The Indian Journal of Statistics, Series B, 145-160. from the link
 125 (<https://www.jstor.org/stable/pdf/25053027.pdf>).
- 126 • I have an R script with a demo in [https://github.com/georgios-stats/](https://github.com/georgios-stats/Bayesian_Statistics/blob/master/LectureHandouts/Rscripts/HierarchicalBayes/HierarchicalBayesPharmaceutical.R)
 127 `Bayesian_Statistics/blob/master/LectureHandouts/Rscripts/HierarchicalBayes/`
 128 `HierarchicalBayesPharmaceutical.R`

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 θ_1 ; i.e. $f(y|\theta_1, \theta_2) = f(y|\theta_2)$. The fact that the likelihood does not depend on θ_1 suggests that y does not provide information about θ_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 θ_2 may change his belief about θ_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

$$\begin{aligned} \pi(\theta_1, \theta_2|y) &\propto f(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2) = f(y|\theta_2)\pi(\theta_1|\theta_2)\pi(\theta_2) \implies \\ \pi(\theta_1, \theta_2|y) &= \pi(\theta_2|y)\pi(\theta_1|\theta_2) \implies \\ \pi(\theta_1|y, \theta_2) &= \pi(\theta_1|\theta_2) \end{aligned} \quad (8)$$

$$\begin{aligned} \pi(\theta_2|y) &= \frac{f(y|\theta_2)\pi(\theta_2)}{\int_{\Theta_2} f(y|\theta_2)\pi(\theta_2)d\theta_2} \cdot \\ \pi(\theta_1|y) &= \int_{\Theta_2} \pi(\theta_1|\theta_2)\pi(\theta_2)d\theta_2 \end{aligned} \quad (9)$$

Here, θ_1 is said to be non-identifiable parameter from the data y , because y provides no direct information about θ_1 . Inference about θ_1 based on marginal posterior $\pi(\theta_1|y)$ depends on y but the information provided about θ_1 comes indirectly through the marginal posterior of θ_2 , see (9). Equivalently, (9) implies that y provides no information about θ_1 given θ_2 .

If we a priori specify that learning the value of θ_2 does not change our belief about θ_1 $\pi(\theta_1|\theta_2) = \pi(\theta_1)$, then (9) becomes $\pi(\theta_1|y) = \pi(\theta_1)$ and hence data y provide no information about θ_1 at all.

Example 13. (A simple example) Consider a production process where manufactured items are classified as acceptable, with probability $1 - \theta_1 - \theta_2$, or defective, with probability $\theta_1 + \theta_2$. Assume that there are two exclusive assignable causes of failure that occur with probabilities θ_1 and θ_2 , respectively, $\theta_1, \theta_2 > 0$ with $\theta_1 + \theta_2 < 1$.

- For a random sample y , the statistical model for the total number of defective items may be considered as $r_n \sim \text{Bn}(n, \theta_1 + \theta_2)$.
- The data are fully informative for $\theta_1 + \theta_2$, however the individual parameters of interest, (θ_1, θ_2) , are non-identifiable.
- The problem may be mitigated if a suitable a priori on θ is assigned, e.g., $\pi(\theta_1, \theta_2) = \text{Di}_2(\theta|a)$.

Hint: Dirichlet distribution, $\theta \sim \text{Di}_k(a)$ has PDF,

$$\text{Di}_k(\theta|a) = \frac{\Gamma(\sum_{j=1}^{k+1} a_j)}{\prod_{j=1}^{k+1} \Gamma(a_j)} \prod_{j=1}^k \theta_j^{a_j-1} (1 - \sum_{j=1}^k \theta_j) 1(\{\sum_{j=1}^k \theta_j \in (0, 1)\} \cap \{\theta_j \in (0, 1)\})$$

and $a_j > 0$ for all $j = 1, \dots, k+1$. It is a generalization of Beta distribution in many dimensions.