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 introduces 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/HandoutsSupplementary/HierarchicalBayes/HierarchicalBayesPharmaceutical.R`

[a] Author: 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

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

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

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

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

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

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

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

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

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

for $\phi = (\phi_1, ..., \phi_{m-1})$, 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

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

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 ith flight (0 for absence, and 1 for presence).

Assume that y_i can be modeled as observations generated independently from a Bernoulli distribution with 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 ith flight. Here are some possible models of interest:

$$\mathscr{M}^{I}: \ p(t; \beta_{\mathscr{M}^{I}}, \mathscr{M}^{I}) = \frac{\exp(\beta_0)}{1 + \exp(\beta_0)} \quad ; \mathscr{M}^{IV}: \ p(t; \beta_{\mathscr{M}^{IV}}, \mathscr{M}^{IV}) = \frac{\exp(\beta_0 + \beta_1 t + \beta_2 s)}{1 + \exp(\beta_0 + \beta_1 t + \beta_2 s)}$$

$$\mathscr{M}^{II}: \ p(t; \beta_{\mathscr{M}^{II}}, \mathscr{M}^{II}) = \frac{\exp(\beta_0 + \beta_1 t)}{1 + \exp(\beta_0 + \beta_1 t)} \quad ; \mathscr{M}^{V}: \ p(t; \beta_{\mathscr{M}^{V}}, \mathscr{M}^{V}) = \frac{\exp(\beta_0 + \beta_1 t + \beta_2 s + \beta_3 ts)}{1 + \exp(\beta_0 + \beta_1 t + \beta_2 s + \beta_3 ts)}$$

$$\mathscr{M}^{III}: \ p(t; \beta_{\mathscr{M}^{III}}, \mathscr{M}^{III}) = \frac{\exp(\beta_0 + \beta_2 s)}{1 + \exp(\beta_0 + \beta_2 s)} \qquad \text{etc...}$$

The Bayesian hierarchical model under consideration is:

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

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

- 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, compared to the computer practical handout example where we had set $\varpi = 0.5$.

Now the joint probability distribution has pdf

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

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

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

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

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

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

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

$$\theta_i \sim \text{N}(\mu_\theta, \sigma_\theta^2), \qquad \delta_i \sim \text{N}(\mu_\delta, \sigma_\delta^2), \qquad \xi_i|w_i \sim \begin{cases} \text{N}(\mu_P, \sigma_P^2) & , w_i = 0 \\ \text{N}(\mu_D, \sigma_D^2) & , w_i = 1 \end{cases}$$

$$\sigma_c \sim \pi(\sigma_c) \propto \frac{1}{\sigma_c}, \qquad \sigma_a \sim \pi(\sigma_a) \propto \frac{1}{\sigma_a}, \qquad \sigma_t \sim \pi(\sigma_t) \propto \frac{1}{\sigma_t}, \tag{4}$$

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.

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

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

The priors in lines (4), (5) and (6) are improper non-informative priors. One could have 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

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

where

$$\pi(\theta|y, \phi) = \frac{f(y|\theta)\pi_1(\theta|\phi)}{f_1(y|\phi)}$$

$$f_1(y|\phi) = \int_\Theta f(y|\theta)\pi_1(\theta|\phi)\mathrm{d}\theta$$

$$\pi(\phi|y) = \frac{f_1(y|\phi)\pi_2(\phi)}{f(y)}$$

$$f(y) = \int_\Theta f_1(y|\phi)\pi_2(\phi)\mathrm{d}\phi$$

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 $h : \Theta \to \mathbb{R}$, then

$$\mathrm{E}_\pi(h(\theta)|y) = \mathrm{E}_\pi\left(\mathrm{E}_\pi\left(h(\theta)|y, \phi\right)\,|y\right).$$

If $\mathrm{E}_\pi(h(\theta)|y) = \int h(\theta)\pi(\theta|y)\mathrm{d}\theta$ is intractable and θ has high dimensionality, one could possibly try to specify the prior decomposition $\pi(\theta) = \int_\Phi \pi_1(\theta|\phi)\pi_2(\phi|\phi_m)\mathrm{d}\phi$ in (3) such that $\mathrm{E}_\pi\left(h(\theta)|y, \phi\right)$ can be computed analytically, and ϕ has low conventionality. In that case one would have to compute the equivalent but lower dimensional (and hence easier) integral $\mathrm{E}_\pi\left(\mathrm{E}_\pi\left(h(\theta)|y, \phi\right)\,|y\right) = \int \mathrm{E}_\pi\left(h(\theta)|y, \phi\right)\pi(\phi|y)\mathrm{d}\phi$.

Example 11. Regarding the fully hierarchical model (1), the full conditionals distributions of each element of $\vartheta = (\theta, \phi_1, ..., \phi_{m-1}) \in \Theta \times \Phi$ are given as:

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

with the convention

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

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

Proof. Straightforward by using the Bayesian theorem. □

Example 12. (Cont...) You may use

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

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

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

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

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

with

$$\delta_i|\text{rest}, ... \overset{\text{ind}}{\sim} \text{N}\left(\mu_{\delta,i}^*, \left(\sigma_{\delta,i}^*\right)^2\right), \forall i = 1, ..., n$$

where

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

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

Try to compute the rest

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

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

$$
\begin{aligned}
\pi(\theta_1, \theta_2|y) &\propto f(y|\theta_1, \theta_2)\pi(\theta_1, \theta_2) = f(y|\theta_2)\pi(\theta_1|\theta_2)\pi(\theta_2) \implies \\
\pi(\theta_1, \theta_2|y) &= \pi(\theta_2|y)\pi(\theta_1|\theta_2) \implies \\
\pi(\theta_1|y, \theta_2) &= \pi(\theta_1|\theta_2) \qquad (8)\\
\pi(\theta_2|y) &= \frac{f(y|\theta_2)\pi(\theta_2)}{\int_{\Theta_2} f(y|\theta_2)\pi(\theta_2)\mathrm{d}\theta_2} \ . \\
\pi(\theta_1|y) &= \int_{\Theta_2} \pi(\theta_1|\theta_2)\pi(\theta_2)\mathrm{d}\theta_2 \qquad (9)
\end{aligned}
$$

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 andθ_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 $j = 1, ..., k+1$. It is a generalization of Beta distribution in many dimensions.