Deriving the Full Conditionals

- The goal of any Bayesian analysis is to determine the joint probability of the parameters $\boldsymbol{\theta}$ and the observed data $\boldsymbol{y}$,

$$p(\boldsymbol{\theta}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

- We can obtain the [joint] **posterior distribution** using Baye’s theorem:

$$\begin{aligned} p(\boldsymbol{\theta}|\boldsymbol{y}) &= \frac{p(\boldsymbol{\theta}, \boldsymbol{y})}{p(\boldsymbol{y})} \\ &\propto p(\boldsymbol{\theta}, \boldsymbol{y}) \\ &= p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \\ &\propto L(\boldsymbol{\theta}|\boldsymbol{y})p(\boldsymbol{\theta}) \end{aligned}$$

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{p(\boldsymbol{y}|\boldsymbol{\theta})}{c}$$

where $L(\boldsymbol{\theta}|\boldsymbol{y})$ is the **likelihood** function and $p(\boldsymbol{\theta})$ is the **prior**.

The posterior distribution

- It is usually easier to summarise the posterior by considering the **marginal posterior distributions**:

$$p(\theta_j|\boldsymbol{y}) = \int \dots \int p(\theta_1, \dots, \theta_j, \dots, \theta_J|\boldsymbol{y}) d\boldsymbol{\theta}_{\setminus j}$$

$$= \int \dots \int p(\theta_j|\boldsymbol{\theta}_{\setminus j}, \boldsymbol{y}) p(\boldsymbol{\theta}_{\setminus j}|\boldsymbol{y}) d\boldsymbol{\theta}_{\setminus j}$$

(using the joint probability rule: $p(A, B) = p(A|B)p(B)$).

These terms $p(\theta_j|\boldsymbol{\theta}_{\setminus j}, \boldsymbol{y})$ for $j = 1, \dots, J$ are called the **full conditional posterior distributions**, or simply **full conditionals**.

Full conditional distributions

- Once the full conditionals have been determined, it is straightforward to sample from the posterior using MCMC – we need only sample from each of the full conditionals using the most recent estimate of the parameters.
- But how do we determine the full conditionals? Consider:

$$\boldsymbol{y}|\mu,\sigma^2 \sim \mathcal{N}(\mu,\sigma^2)$$
$$\mu|\eta \sim \mathcal{N}(\eta,5)$$
$$\sigma^2 \sim \mathcal{IG}(0.5, 0.05)$$
$$\eta \sim \mathrm{Uni}(80,120)$$

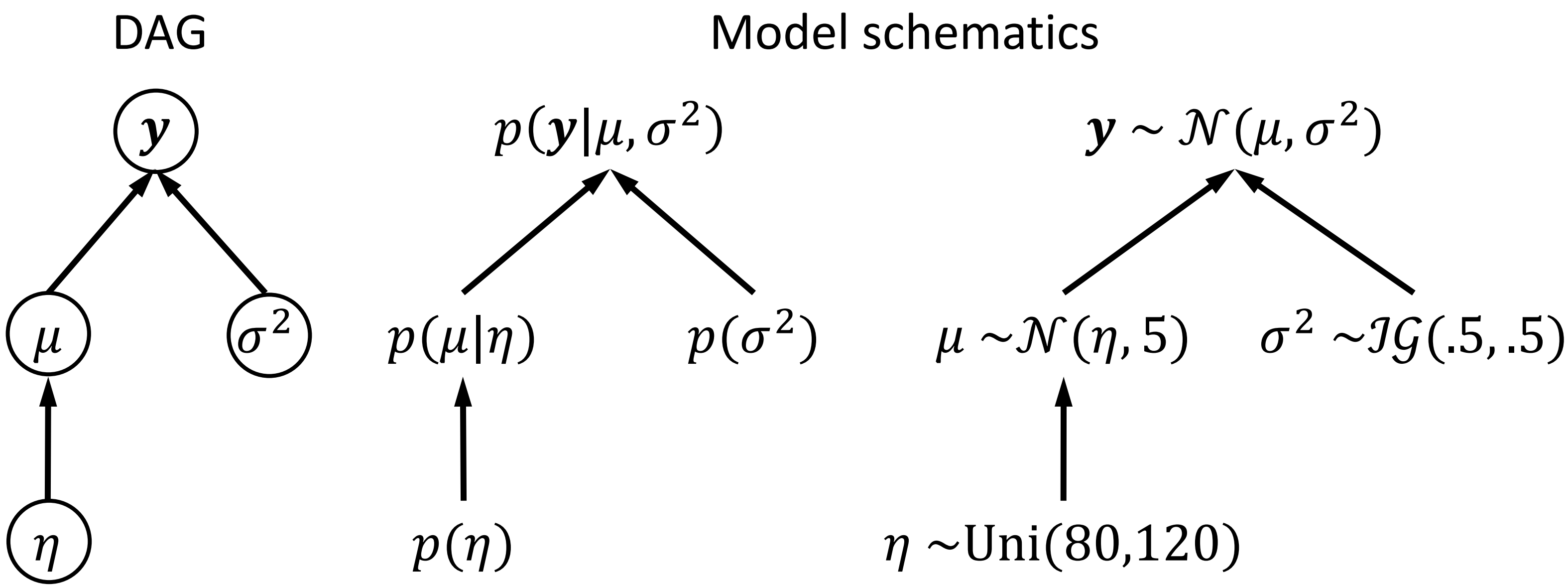
  - We could write our the joint posterior distribution, and then systematically remove all terms not involving the parameter of interest. E.g.

$$p(\boldsymbol{\theta},\boldsymbol{y}) = \frac{0.05^{0.5}(\sigma^2)^{-1.5}}{\Gamma(0.5)\sqrt{20\pi^2\sigma^2}} \exp\left\{\frac{(y-\mu)^2+0.1}{-2\sigma^2} - \frac{(\mu-\eta)^2}{10}\right\}$$

  - But this is tedious/untenable. Luckily, it's also unnecessary. ☺

Creating a model schematic

- **Tip 1**: It is usually very helpful to create a DAG or some other form of a model schematic (maybe showing more detail):



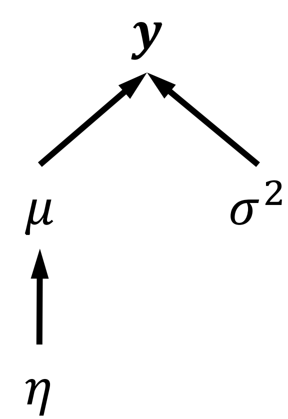
- The main objective of this diagram is to show the dependencies between nodes.

Removing unnecessary terms

- For this example, the full conditionals are:

$$
\begin{aligned}
&p(\mu|\sigma^2, \eta, \boldsymbol{y}) \\
&p(\sigma^2|\mu, \eta, \boldsymbol{y}) = p(\sigma^2|\mu, \boldsymbol{y}) \\
&p(\eta|\mu, \sigma^2, \boldsymbol{y}) = p(\eta|\mu\ )
\end{aligned}
$$

- **Tip 2**: the full conditionals can be simplified by removing dependence on terms ($\boldsymbol{y}$ or any $\boldsymbol{\theta}_{\setminus j}$) providing they are NOT:
  1. a child node;
  2. a parent node; or
  3. a 'sibling' node (another child node of a parent).



Deriving the FCs the long way…

- But how do we obtain the form of the full conditionals? We can obtain them ‘the long way’ using probability rules or via a shortcut using the DAG/model schematic.
- Recall probability rules:
  - $p(A|B) \propto p(B|A)p(A)$ conditional probability rule (CPR)
  - $p(A,B) = p(A|B)p(B)$ joint probability rule (JPR)
- Example using the long way:

$$
\begin{aligned}
p(\mu|\sigma^2,\eta,\boldsymbol{y}) &= p(\overbrace{(\mu|\eta)}^{A}|\overbrace{\sigma^2,\boldsymbol{y}}^{B}) \\
&\propto p(\overset{B}{\sigma^2},\overbrace{\boldsymbol{y}|\mu,\eta}^{A})p(\mu|\eta) && \text{by CPR} \\
&= p(\boldsymbol{y}|\mu,\eta,\sigma^2)p(\sigma^2)p(\mu|\eta) && \text{by JPR} \\
&= p(\boldsymbol{y}|\mu,\cancel{\eta},\sigma^2)\cancel{p(\sigma^2)}p(\mu|\eta) && \text{(tip 2)} \\
&= p(\boldsymbol{y}|\mu,\sigma^2)p(\mu|\eta)
\end{aligned}
$$

(A product of two distributions we know)

Deriving the FCs via the shortcut

- Doing this for the other two FCs, we have all three FCs:

$$
\begin{aligned}
p(\mu|\sigma^2, \eta, \boldsymbol{y}) &\propto p(\mu|\eta) p(\boldsymbol{y}|\mu, \sigma^2) \\
p(\sigma^2|\mu, \eta, \boldsymbol{y}) &\propto p(\sigma^2) p(\boldsymbol{y}|\mu, \sigma^2) \\
p(\eta|\mu, \sigma^2, \boldsymbol{y}) &\propto p(\eta) p(\mu|\eta)
\end{aligned}
$$

- What do you notice about the RHS terms?

$$
\begin{tikzcd}
 & p(\boldsymbol{y}|\mu, \sigma^2) & \\
p(\mu|\eta) \arrow[ur] & & p(\sigma^2) \arrow[ul] \\
p(\eta) \arrow[u] & &
\end{tikzcd}
$$

- **Tip 3** (shortcut method): the RHS is a product of the probability model for the node in question and its parent node(s).

FCs with non-standard form

- **Tip 4**: we need not concern ourselves with what the form of the full conditionals look like – this is only a matter of concern for the MCMC sampling method.
  - If the full conditional has a standard, recognisable form, we can use Gibbs sampling, e.g.

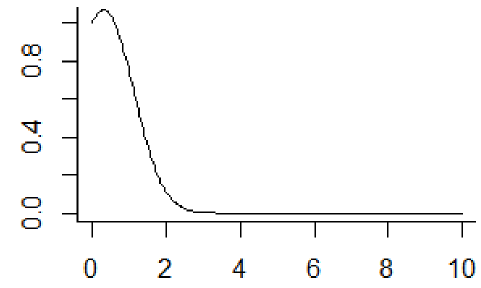
$$p(x) = \exp\left(\frac{x^2}{-2 \times 10^2}\right) \propto \mathcal{N}(0,100)$$

which is easy to sample from, e.g. in R:

```
x <- rnorm(1, mean = 0, sd = 10)
```

  - If the full conditional is 'obscure', e.g.

$$p(x) = e^{-x^2}\Gamma(x+2) \propto ???$$



we can use MH, slice sampling, etc.

Specific situations

- How do we deal with:
  - Non-stochastic nodes (e.g. regression equations)?
  - Truncated distributions?
  - Mixtures of distributions?

Non-stochastic nodes

- Example:

$$\begin{tikzcd}
 & Y_i \sim \mathrm{Po}(\mu_i) & \\
 & \log(\mu_i) = \alpha + \beta x_i \arrow[u] & \\
\alpha \sim \mathcal{N}(0, 10^3) \arrow[ur] & & \beta \sim \mathcal{N}(0, 10^3) \arrow[ul]
\end{tikzcd}$$

- We could remove non-stochastic term by replacing the term μ_i with $\alpha + \beta x_i$.
  - Not very convenient if regression equation is long.
- For the purpose of determining the parent node(s), ignore the non-stochastic nodes to identify the *real* parent(s).
  - Use μ_i for convenience, but keep in mind that μ_i is really shorthand for $\alpha + \beta x_i$, e.g. $\mathrm{Po}(\mu_i) = \mathrm{Po}(\alpha + \beta x_i)$.

Non-stochastic nodes cont…

$$
\begin{tikzcd}
 & Y_i \sim \mathrm{Po}(\mu_i) & \\
 & \log(\mu_i) = \alpha + \beta x_i \arrow[u] & \\
\alpha \sim \mathcal{N}(0, 10^3) \arrow[ur] & & \beta \sim \mathcal{N}(0, 10^3) \arrow[ul]
\end{tikzcd}
$$

- Using the shortcut method (tip 3), the FCs are:

$$
\begin{aligned}
p(\alpha|\beta, y_i) &\propto p(\alpha)p(y_i|\mu) \\
&= p(\alpha)p(y_i|\alpha, \beta) \\
p(\beta|\alpha, y_i) &\propto p(\beta)p(y_i|\mu) \\
&= p(\beta)p(y_i|\alpha, \beta)
\end{aligned}
$$

Truncated distributions

- Example:

$$Y_i \sim \mathrm{Po}(\eta_i)$$

$$\uparrow$$

$$\eta_i \sim \mathcal{N}(\mu_i, 10)\mathbb{I}_{(\mu_i, \infty^+)}$$

$$\uparrow$$

$$\mu_i \sim p(\mu_i)$$

- Using the shortcut method (tip 3), the FC for η_i is:

$$p(\eta_i|\mu_i, y_i) \propto p(\eta_i|\mu_i)p(y_i|\eta_i)$$

- Note the functional form of the truncated distribution is:

$$p(\eta_i|\mu_i) = \frac{1}{\sqrt{20\pi}} \exp\left(\frac{(\eta_i - \mu_i)^2}{-20}\right), \eta_i > \mu_i$$

Truncated distributions cont…

- **Tip 5**: If the truncated distribution is symmetric and the truncation occurs at the point of symmetry, the truncated distribution is proportional to the non-truncated form.

$$
\begin{aligned}
p(\eta_i|\mu_i) &= \mathcal{N}(\mu_i, 10)\mathbb{I}_{(\mu_i,\infty^+)} \\
&= \frac{1}{\sqrt{20\pi}} \exp\left(\frac{(\eta_i - \mu_i)^2}{-20}\right), \eta_i > \mu_i \\
&= 2 \times \frac{1}{\sqrt{20\pi}} \exp\left(\frac{(\eta_i - \mu_i)^2}{-20}\right) \\
&\propto \frac{1}{\sqrt{20\pi}} \exp\left(\frac{(\eta_i - \mu_i)^2}{-20}\right) \\
&= \mathcal{N}(\mu_i, 10)
\end{aligned}
$$

Truncated distributions cont…

- What if:
  - the truncated distribution is asymmetric; or
  - the truncated distribution is symmetric but the truncation does not occur at the point of symmetry?
- If the PDF of the truncated distribution is known, we can use that. E.g. we could have used the truncated Normal distribution:

$$\mathcal{N}(\eta_i; \mu_i, \sigma)\mathbb{I}_{(a,b)} = \frac{\frac{1}{\sigma}\phi\left(\frac{\eta_i - \mu_i}{\sigma}\right)}{\Phi\left(\frac{b-\mu_i}{\sigma}\right) - \Phi\left(\frac{a-\mu_i}{\sigma}\right)}$$

  where $\phi()$ is the standard Normal PDF, and $\Phi()$ is its CDF.
- If not, we can use rejection sampling or similar (see tip 4).
  - Note: this may be very inefficient, depending on the truncation. E.g. $\mathcal{N}(0,1)\mathbb{I}_{(3,\infty^+)} \Rightarrow$ 99.87% rejection.

Mixtures of distributions

- Example of mixture on the likelihood:

$$Y_i \sim \sum_{k=1}^{K} w_k \mathrm{Po}(\lambda_k)$$

$$\lambda_k \sim \mathrm{Gam}(2, 1) \qquad \boldsymbol{w} \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$$

- To sample from a mixture model, we typically introduce a latent allocation variable z_i which takes values in $\{1, \dots, K\}$ and indicates to which group y_i belongs.
- This variable z_i is actually missing (latent) data. If we knew the value of z_i, we would know which group y_i belongs to.
- Thus we can talk about the likelihood, or a full likelihood (the likelihood when the value of $\boldsymbol{z}$ is known).

Mixtures of distributions

- Example of mixture on the likelihood:

$$Y_i \sim \sum_{k=1}^{K} w_k \mathrm{Po}(\lambda_k)$$

$$\lambda_k \sim \mathrm{Gam}(2, 1) \qquad \boldsymbol{w} \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$$

- The likelihood is (prop. to):

$$p(y_i | \boldsymbol{\lambda}, \boldsymbol{w}) = \sum_{k=1}^{K} w_k \mathrm{Po}(y_i; \lambda_k)$$

- The full likelihood is (prop. to):

$$p\left(y_i, z_i \middle| \lambda_{z_i}, w_{z_i}\right) = w_{z_i} \mathrm{Po}(y_i; \lambda_{z_i})$$

- **Tip 6**: For the purpose of sampling, we use $p\left(y_i \middle| \lambda_{z_i}, z_i\right)$.

Mixtures of distributions

- Example of mixture on the likelihood:

$$Y_i \sim \sum_{k=1}^{K} w_k \mathrm{Po}(\lambda_k)$$

$$\lambda_k \sim \mathrm{Gam}(2,1) \qquad \boldsymbol{w} \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$$

$$Y_i \sim \mathrm{Po}(\lambda_{z_i})$$

$$\uparrow$$

$$\lambda_{z_i} \sim \mathrm{Gam}(2,1)$$

$$\uparrow$$

$$z_i \sim \mathrm{Cat}(w_1, \dots, w_K)$$

$$\uparrow$$

$$\boldsymbol{w} \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$$

- **Tip 7**: For mixtures, re-write the model schematic in terms of the latent allocation variable z_i.
- So the FC for $\boldsymbol{\lambda}$ is: $p\left(\lambda_{z_i} \middle| y_i, z_i\right) \propto p\left(\lambda_{z_i} \middle| z_i\right) p\left(y_i \middle| \lambda_{z_i}, z_i\right)$

Mixtures of distributions cont…

- Example of mixture on a random effect:

$$Y_i \sim \mathrm{Po}(\mu_i)$$

$$\log(\mu_i) = \gamma + \sum_{k=1}^{K} w_k \beta_k$$

$$\gamma \sim \mathcal{N}(0,5) \qquad \boldsymbol{w} \sim \mathcal{D}(\boldsymbol{\alpha}) \qquad \beta_k \sim \mathcal{N}(k,1)$$

$$Y_i \sim \mathrm{Po}(\mu_i)$$

$$\log(\mu_i) = \gamma + \beta_{z_i}$$

$$\gamma \sim \mathcal{N}(0,5) \qquad \beta_{z_i} \sim \mathcal{N}(z_i\,,1)$$

$$z_i \sim \mathrm{Cat}(w_1, \dots, w_K)$$

$$\boldsymbol{w} \sim \mathcal{D}(\alpha_1, \dots, \alpha_K)$$

- (Use tip 7 again).

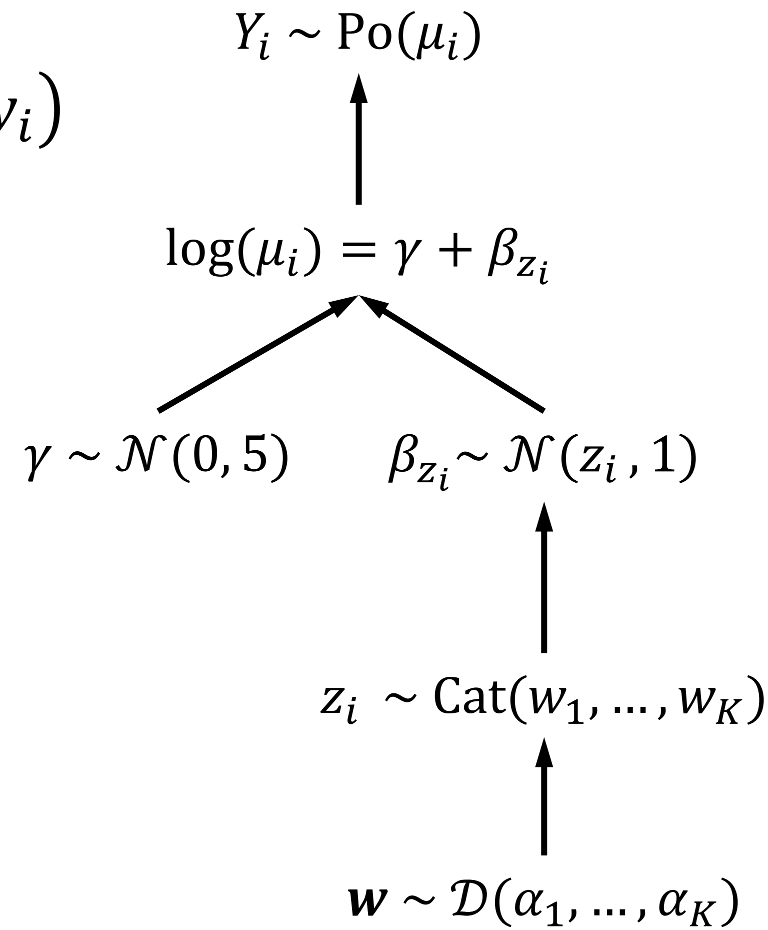
Mixtures of distributions cont…

- The FC for $\boldsymbol{\beta}$ is:

$$p(\boldsymbol{\beta}|\gamma, z_i, y_i) = \prod_{k=1}^{K} p\left(\beta_{z_i=k}|\gamma, z_i, y_i\right)$$

where

$$\begin{aligned} p\left(\beta_{z_i}|\gamma, z_i, y_i\right) & \\ &\propto p\left(\beta_{z_i}|z_i\right)p(y_i|\mu_i) \\ &= p\left(\beta_{z_i}|z_i\right)p\left(y_i|\gamma, \beta_{z_i}\right) \end{aligned}$$



Mixtures of distributions cont…

- The FC for $\boldsymbol{z}$ is:

$$p(\boldsymbol{z}|\boldsymbol{w},\boldsymbol{\beta}) = \prod_{i=1}^{N} p(z_i|\boldsymbol{w},\boldsymbol{\beta})$$

where

$$p(z_i|\boldsymbol{w},\boldsymbol{\beta}) \propto p(z_i|\boldsymbol{w})p\left(\beta_{z_i}|z_i\right)$$

and N is the data sample size.