# Variational Inference Masterclass: Variational Message Passing

https://github.com/tui-nolan/Variational-Inference-Masterclass

**Tui Nolan**

8 November 2022

UNIVERSITY OF CAMBRIDGE

UKRI MRC Biostatistics Unit

# Outline

Part I: Message Passing for Variational Inference



Part II: Efficient Model Extensions

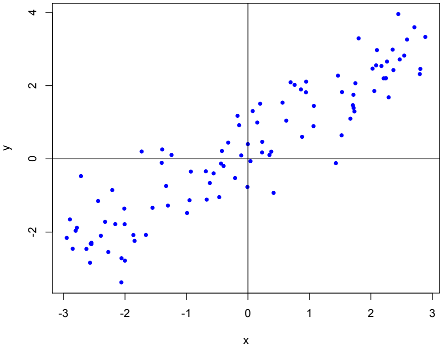


Part III: Nonconjugate Models

Part I

# Introduction and Motivation

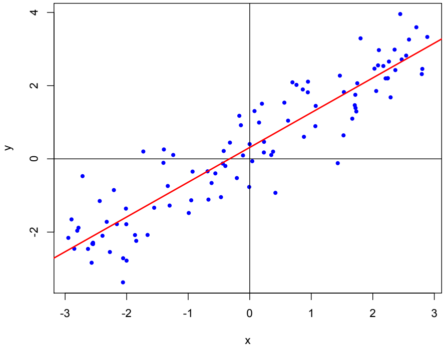
Consider the following data:




In a typical regression problem we solve using:

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^2}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\beta\|_2$$

Consider the following data:




In a typical regression problem we solve using:

$$\hat{\boldsymbol{\beta}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^2} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2$$

# Mean Field Variational Bayes

Tractable Bayesian inference is achieved by introducing a latent variable (Gelman, 2006; Huang and Wand, 2013):

$$
\begin{aligned}
y_i | x_i, \boldsymbol{\beta}, \sigma^2 &\overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n \\
\boldsymbol{\beta} &\sim N(\mathbf{0}, \Sigma_0) \\
\sigma^2 | a &\sim \text{Inverse} - \chi^2(1, 1/a) \\
a &\sim \text{Inverse} - \chi^2(1, 1/A^2)
\end{aligned}
$$

where

$$x \sim \text{Inverse} - \chi^2(\kappa, \lambda) \quad \text{if and only if} \quad x \sim \text{Inverse-Gamma}(\kappa/2, \lambda/2)$$

Note that (Gelman, 2006) if

$$\sigma^2 | a \sim \text{Inverse} - \chi^2(1, 1/a) \quad \text{and} \quad a \sim \text{Inverse} - \chi^2(1, 1/A^2)$$

then

$$\sigma \sim \text{Half-Cauchy}(A)$$

We use the mean field assumption (Menictas and Wand, 2013):

$$\{\boldsymbol{\beta}, a\} \perp\!\!\!\perp \sigma^2$$

That is,

$$p(\boldsymbol{\beta}, \sigma^2, a | \boldsymbol{y}) \approx p(\boldsymbol{\beta}, a | \boldsymbol{y}) p(\sigma^2 | \boldsymbol{y})$$

# Mean Field Variational Bayes

Tractable Bayesian inference is achieved by introducing a latent variable (Gelman, 2006; Huang and Wand, 2013):

$$
\begin{aligned}
y_i | x_i, \boldsymbol{\beta}, \sigma^2 &\overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n \\
\boldsymbol{\beta} &\sim N(\mathbf{0}, \Sigma_0) \\
\sigma^2 | a &\sim \text{Inverse} - \chi^2(1, 1/a) \\
a &\sim \text{Inverse} - \chi^2(1, 1/A^2)
\end{aligned}
$$

where

$$x \sim \text{Inverse} - \chi^2(\kappa, \lambda) \quad \text{if and only if} \quad x \sim \text{Inverse-Gamma}(\kappa/2, \lambda/2)$$

Note that (Gelman, 2006) if

$$\sigma^2 | a \sim \text{Inverse} - \chi^2(1, 1/a) \quad \text{and} \quad a \sim \text{Inverse} - \chi^2(1, 1/A^2)$$

then

$$\sigma \sim \text{Half-Cauchy}(A)$$

We use the mean field assumption (Menictas and Wand, 2013):

$$\{\boldsymbol{\beta}, a\} \perp\!\!\!\perp \sigma^2$$

That is,

$$p(\boldsymbol{\beta}, \sigma^2, a | \boldsymbol{y}) \approx p(\boldsymbol{\beta}, a | \boldsymbol{y}) p(\sigma^2 | \boldsymbol{y})$$

# Mean Field Variational Bayes

Tractable Bayesian inference is achieved by introducing a latent variable (Gelman, 2006; Huang and Wand, 2013):

$$\begin{aligned} y_i|x_i,\boldsymbol{\beta},\sigma^2 &\stackrel{\text{ind.}}{\sim} N(\beta_0+\beta_1 x_i,\sigma^2) \quad i=1,\dots,n \\ \boldsymbol{\beta} &\sim N(\mathbf{0},\Sigma_0) \\ \sigma^2|a &\sim \text{Inverse}-\chi^2(1,1/a) \\ a &\sim \text{Inverse}-\chi^2(1,1/A^2) \end{aligned}$$

where

$$x \sim \text{Inverse}-\chi^2(\kappa,\lambda) \quad \text{if and only if} \quad x \sim \text{Inverse-Gamma}(\kappa/2,\lambda/2)$$

Note that (Gelman, 2006) if

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad \text{and} \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

then

$$\sigma \sim \text{Half-Cauchy}(A)$$

We use the mean field assumption (Menictas and Wand, 2013):

$$\{\boldsymbol{\beta},a\} \perp\!\!\!\perp \sigma^2$$

That is,

$$p(\boldsymbol{\beta},\sigma^2,a|\boldsymbol{y}) \approx p(\boldsymbol{\beta},a|\boldsymbol{y})p(\sigma^2|\boldsymbol{y})$$

# Mean Field Variational Bayes

In variational inference, we approximate the posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})$ by another density function $q(\boldsymbol{\beta}, \sigma^2, a)$.

The "best" approximate density function $q^*(\boldsymbol{\beta}, \sigma^2, a)$ is the one that is "closest" to the true posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})$:

$$q^*(\boldsymbol{\beta}, \sigma^2, a) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\, D_{\text{KL}}\left\{q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})\right\}$$

Our mean field assumption combined with results from Graph Theory, see Chapter 8 of Bishop (2006), allow us to enforce the product form:

$$p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \sigma^2, a) = q(\boldsymbol{\beta}, a)q(\sigma^2) = q(\boldsymbol{\beta})q(\sigma^2)q(a)$$

For $\boldsymbol{\beta}$, the optimisation problem has the following solution:

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C_1 \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}\right] \\ &= C_2 \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, a)p(\boldsymbol{\beta})\right\}\right] \end{aligned}$$

where

$$\mathbb{E}_{-q(\boldsymbol{\beta})} f(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) \equiv \int\int f(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)q(\sigma^2)q(a)\,da\,d\sigma^2$$

In variational inference, we approximate the posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})$ by another density function $q(\boldsymbol{\beta}, \sigma^2, a)$.

The "best" approximate density function $q^*(\boldsymbol{\beta}, \sigma^2, a)$ is the one that is "closest" to the true posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})$:

$$q^*(\boldsymbol{\beta}, \sigma^2, a) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\, D_{\text{KL}} \left\{ q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y}) \right\}$$

Our mean field assumption combined with results from Graph Theory, see Chapter 8 of Bishop (2006), allow us to enforce the product form:

$$p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \sigma^2, a) = q(\boldsymbol{\beta}, a) q(\sigma^2) = q(\boldsymbol{\beta}) q(\sigma^2) q(a)$$

For $\boldsymbol{\beta}$, the optimisation problem has the following solution:

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C_1 \exp\left[ \mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{ p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) \right\} \right] \\ &= C_2 \exp\left[ \mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{ p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, a) p(\boldsymbol{\beta}) \right\} \right] \end{aligned}$$

where

$$\mathbb{E}_{-q(\boldsymbol{\beta})} f(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) \equiv \int \int f(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) q(\sigma^2) q(a) \, da \, d\sigma^2$$

In variational inference, we approximate the posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})$ by another density function $q(\boldsymbol{\beta}, \sigma^2, a)$.

The "best" approximate density function $q^*(\boldsymbol{\beta}, \sigma^2, a)$ is the one that is "closest" to the true posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})$:

$$q^*(\boldsymbol{\beta}, \sigma^2, a) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}} D_{\text{KL}} \left\{ q(\boldsymbol{\beta}, \sigma^2, a) || p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y}) \right\}$$

Our mean field assumption combined with results from Graph Theory, see Chapter 8 of Bishop (2006), allow us to enforce the product form:

$$p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y}) \approx q(\boldsymbol{\beta}, \sigma^2, a) = q(\boldsymbol{\beta}, a)q(\sigma^2) = q(\boldsymbol{\beta})q(\sigma^2)q(a)$$

For $\boldsymbol{\beta}$, the optimisation problem has the following solution:

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C_1 \exp\left[ \mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{ p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) \right\} \right] \\ &= C_2 \exp\left[ \mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{ p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, a) p(\boldsymbol{\beta}) \right\} \right] \end{aligned}$$

where

$$\mathbb{E}_{-q(\boldsymbol{\beta})} f(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) \equiv \int \int f(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a) q(\sigma^2) q(a) da d\sigma^2$$

# Mean Field Variational Bayes

In variational inference, we approximate the posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\mathbf{y})$ by another density function $q(\boldsymbol{\beta}, \sigma^2, a)$.

The "best" approximate density function $q^*(\boldsymbol{\beta}, \sigma^2, a)$ is the one that is "closest" to the true posterior density function $p(\boldsymbol{\beta}, \sigma^2, a|\mathbf{y})$:

$$q^*(\boldsymbol{\beta}, \sigma^2, a) = \underset{q \in \mathcal{Q}}{\operatorname{argmin}}\, D_{\mathrm{KL}} \left\{ q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a|\mathbf{y}) \right\}$$

Our mean field assumption combined with results from Graph Theory, see Chapter 8 of Bishop (2006), allow us to enforce the product form:

$$p(\boldsymbol{\beta}, \sigma^2, a|\mathbf{y}) \approx q(\boldsymbol{\beta}, \sigma^2, a) = q(\boldsymbol{\beta}, a) q(\sigma^2) = q(\boldsymbol{\beta}) q(\sigma^2) q(a)$$

For $\boldsymbol{\beta}$, the optimisation problem has the following solution:

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C_1 \exp\left[ \mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{ p(\mathbf{y}, \boldsymbol{\beta}, \sigma^2, a) \right\} \right] \\ &= C_2 \exp\left[ \mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{ p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2, a) p(\boldsymbol{\beta}) \right\} \right] \end{aligned}$$

where

$$\mathbb{E}_{-q(\boldsymbol{\beta})} f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2, a) \equiv \int \int f(\mathbf{y}, \boldsymbol{\beta}, \sigma^2, a) q(\sigma^2) q(a)\, da\, d\sigma^2$$

# Mean Field Variational Bayes

$$
\begin{aligned}
&\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\} \\
&\qquad = \mathbb{E}_{-q(\boldsymbol{\beta})} \log p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) + \log p(\boldsymbol{\beta}) \\
&\qquad = \mathbb{E}_{-q(\boldsymbol{\beta})} \left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\qquad = -\frac{1}{2}\,\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\qquad = -\frac{1}{2}\boldsymbol{\beta}^T\left\{\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \text{const.}
\end{aligned}
$$

Remember that $q^*(\boldsymbol{\beta}) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\}\right]$

That is, $q^*(\boldsymbol{\beta})$ is a normal density function.

Similar arguments can be used to show that $q^*(\sigma^2)$ and $q^*(a)$ are inverse chi-squared density functions.

$$
\begin{aligned}
&\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\} \\
&\qquad = \mathbb{E}_{-q(\boldsymbol{\beta})} \log p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) + \log p(\boldsymbol{\beta}) \\
&\qquad = \mathbb{E}_{-q(\boldsymbol{\beta})} \left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\qquad = -\frac{1}{2}\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\qquad = -\frac{1}{2}\boldsymbol{\beta}^T\left\{\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \text{const.}
\end{aligned}
$$

Remember that $q^*(\boldsymbol{\beta}) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\}\right]$

That is, $q^*(\beta)$ is a normal density function.

Similar arguments can be used to show that $q^*(\sigma^2)$ and $q^*(a)$ are inverse chi-squared density functions.

# Mean Field Variational Bayes

$$
\begin{aligned}
&\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\} \\
&\quad = \mathbb{E}_{-q(\boldsymbol{\beta})} \log p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) + \log p(\boldsymbol{\beta}) \\
&\quad = \mathbb{E}_{-q(\boldsymbol{\beta})} \left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\right\} - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\quad = -\frac{1}{2}\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\quad = -\frac{1}{2}\boldsymbol{\beta}^T\left\{\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \text{const.}
\end{aligned}
$$

Remember that $q^*(\boldsymbol{\beta}) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\}\right]$

That is, $q^*(\boldsymbol{\beta})$ is a normal density function.

Similar arguments can be used to show that $q^*(\sigma^2)$ and $q^*(a)$ are inverse chi-squared density functions.

$$
\begin{aligned}
&\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\} \\
&\qquad = \mathbb{E}_{-q(\boldsymbol{\beta})} \log p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) + \log p(\boldsymbol{\beta}) \\
&\qquad = \mathbb{E}_{-q(\boldsymbol{\beta})} \left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})\right\} - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\qquad = -\frac{1}{2}\,\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)(\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta}) - \frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta} + \text{const.} \\
&\qquad = -\frac{1}{2}\boldsymbol{\beta}^T\left\{\mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q\left(\frac{1}{\sigma^2}\right)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \text{const.}
\end{aligned}
$$

Remember that $q^*(\boldsymbol{\beta}) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})} \log\left\{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)p(\boldsymbol{\beta})\right\}\right]$

That is, $q^*(\boldsymbol{\beta})$ is a normal density function.

Similar arguments can be used to show that $q^*(\sigma^2)$ and $q^*(a)$ are inverse chi-squared density functions.

# Mean Field Variational Bayes

$q^*(\boldsymbol{\beta})$ is the $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \Sigma_{q(\boldsymbol{\beta})})$ density function, where

$$\Sigma_{q(\boldsymbol{\beta})} \longleftarrow \left\{ \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) \boldsymbol{X}^T \boldsymbol{X} + \Sigma_0^{-1} \right\}^{-1}$$

$$\text{and} \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) \Sigma_{q(\boldsymbol{\beta})} \boldsymbol{X}^T \boldsymbol{y}$$

$q^*(\sigma^2)$ is the Inverse $-\chi^2(N+1, \lambda_{q(\sigma^2)})$ density function, where

$$\lambda_{q(\sigma^2)} \longleftarrow \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \boldsymbol{X}^T \boldsymbol{y} + \text{tr}\left\{ \boldsymbol{X}^T \boldsymbol{X} (\Sigma_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T) \right\} + \frac{2}{\lambda_{q(a)}}$$

$q^*(a)$ is the Inverse $-\chi^2(2, \lambda_{q(a)})$ density function, where

$$\lambda_{q(a)} \longleftarrow 2\left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) + \frac{1}{A^2}$$

$q^*(\boldsymbol{\beta})$ is the $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \Sigma_{q(\boldsymbol{\beta})})$ density function, where

$$\Sigma_{q(\boldsymbol{\beta})} \longleftarrow \left\{ \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) \boldsymbol{X}^T \boldsymbol{X} + \Sigma_0^{-1} \right\}^{-1}$$

$$\text{and} \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) \Sigma_{q(\boldsymbol{\beta})} \boldsymbol{X}^T \boldsymbol{y}$$

$q^*(\sigma^2)$ is the Inverse $-\,\chi^2(N+1, \lambda_{q(\sigma^2)})$ density function, where

$$\lambda_{q(\sigma^2)} \longleftarrow \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \boldsymbol{X}^T \boldsymbol{y} + \operatorname{tr}\left\{ \boldsymbol{X}^T \boldsymbol{X} (\Sigma_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T) \right\} + \frac{2}{\lambda_{q(a)}}$$

$q^*(a)$ is the Inverse $-\,\chi^2(2, \lambda_{q(a)})$ density function, where

$$\lambda_{q(a)} \longleftarrow 2 \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) + \frac{1}{A^2}$$

# Mean Field Variational Bayes

$q^*(\boldsymbol{\beta})$ is the $N(\boldsymbol{\mu}_{q(\boldsymbol{\beta})}, \Sigma_{q(\boldsymbol{\beta})})$ density function, where

$$\Sigma_{q(\boldsymbol{\beta})} \longleftarrow \left\{ \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) \boldsymbol{X}^T \boldsymbol{X} + \Sigma_0^{-1} \right\}^{-1}$$

$$\text{and} \quad \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \longleftarrow \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) \Sigma_{q(\boldsymbol{\beta})} \boldsymbol{X}^T \boldsymbol{y}$$

$q^*(\sigma^2)$ is the Inverse $-\chi^2(N+1, \lambda_{q(\sigma^2)})$ density function, where

$$\lambda_{q(\sigma^2)} \longleftarrow \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\mu}_{q(\boldsymbol{\beta})} \boldsymbol{X}^T \boldsymbol{y} + \operatorname{tr}\left\{ \boldsymbol{X}^T \boldsymbol{X} (\Sigma_{q(\boldsymbol{\beta})} + \boldsymbol{\mu}_{q(\boldsymbol{\beta})} \boldsymbol{\mu}_{q(\boldsymbol{\beta})}^T) \right\} + \frac{2}{\lambda_{q(a)}}$$

$q^*(a)$ is the Inverse $-\chi^2(2, \lambda_{q(a)})$ density function, where

$$\lambda_{q(a)} \longleftarrow 2 \left( \frac{N+1}{\lambda_{q(\sigma^2)}} \right) + \frac{1}{A^2}$$

## MFVB for the Bayesian linear regression model

1. Initialise all optimal posterior density functions
2. Cycle:

$$q^*(a) \propto \exp\left\{\mathbb{E}_{-q^*(a)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\sigma^2) \propto \exp\left\{\mathbb{E}_{-q^*(\sigma^2)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\boldsymbol{\beta}) \propto \exp\left\{\mathbb{E}_{-q^*(\boldsymbol{\beta})} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$

3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a | \boldsymbol{y})\right\}$ converges.

## MFVB for the Bayesian linear regression model

1. Initialise all optimal posterior density functions
2. Cycle:

$$q^*(a) \propto \exp\left\{\mathbb{E}_{-q^*(a)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\sigma^2) \propto \exp\left\{\mathbb{E}_{-q^*(\sigma^2)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\boldsymbol{\beta}) \propto \exp\left\{\mathbb{E}_{-q^*(\boldsymbol{\beta})} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$

3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a | \boldsymbol{y})\right\}$ converges.

# Mean Field Variational Bayes

## MFVB for the Bayesian linear regression model

1. Initialise all optimal posterior density functions
2. Cycle:

$$q^*(a) \propto \exp\left\{\mathbb{E}_{-q^*(a)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\sigma^2) \propto \exp\left\{\mathbb{E}_{-q^*(\sigma^2)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\boldsymbol{\beta}) \propto \exp\left\{\mathbb{E}_{-q^*(\boldsymbol{\beta})} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$

3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a | \boldsymbol{y})\right\}$ converges.

## MFVB for the Bayesian linear regression model

1. Initialise all optimal posterior density functions
2. Cycle:

$$q^*(a) \propto \exp\left\{\mathbb{E}_{-q^*(a)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$

$$q^*(\sigma^2) \propto \exp\left\{\mathbb{E}_{-q^*(\sigma^2)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$

$$q^*(\boldsymbol{\beta}) \propto \exp\left\{\mathbb{E}_{-q^*(\boldsymbol{\beta})} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$

3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta}, \sigma^2, a) || p(\boldsymbol{\beta}, \sigma^2, a|\boldsymbol{y})\right\}$ converges.

## MFVB for the Bayesian linear regression model

1. Initialise all optimal posterior density functions
2. Cycle:

$$
\begin{aligned}
q^*(a) &\propto \exp\left\{\mathbb{E}_{-q^*(a)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\} \\
q^*(\sigma^2) &\propto \exp\left\{\mathbb{E}_{-q^*(\sigma^2)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\} \\
q^*(\boldsymbol{\beta}) &\propto \exp\left\{\mathbb{E}_{-q^*(\boldsymbol{\beta})} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}
\end{aligned}
$$

3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a | \boldsymbol{y})\right\}$ converges.

## MFVB for the Bayesian linear regression model

1. Initialise all optimal posterior density functions
2. Cycle:

$$q^*(a) \propto \exp\left\{\mathbb{E}_{-q^*(a)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\sigma^2) \propto \exp\left\{\mathbb{E}_{-q^*(\sigma^2)} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$
$$q^*(\boldsymbol{\beta}) \propto \exp\left\{\mathbb{E}_{-q^*(\boldsymbol{\beta})} \log p(\boldsymbol{y}, \boldsymbol{\beta}, \sigma^2, a)\right\}$$

3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta}, \sigma^2, a) \| p(\boldsymbol{\beta}, \sigma^2, a | \boldsymbol{y})\right\}$ converges.

# Limitations of MFVB

The advantage of MFVB over Monte Carlo alternatives is that it is a much faster algorithm

However, for Bayesian inference on more complex models, we have to re-do all the derivations.

Take for example, the Bayesian Gaussian-response linear mixed model:

$$\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{u}_i, \sigma^2 \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{X}_i\boldsymbol{u}_i, \sigma^2\boldsymbol{I}), \qquad \boldsymbol{u}_i \mid \Sigma \sim N(\boldsymbol{0}, \Sigma), \quad \text{for } 1 \le i \le m$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2\boldsymbol{I}); \qquad \sigma^2 \mid a \sim \text{Inverse} - \chi^2(1, 1/a)$$

$$a \sim \text{Inverse} - \chi^2(1, 1/2), \quad \Sigma|\boldsymbol{A} \sim \text{Inverse G-Wishart}(G_{\text{full}}, 2q, \boldsymbol{A}^{-1})$$

$$\boldsymbol{A} \sim \text{Inverse G-Wishart}(G_{\text{diag}}, 1, \tfrac{1}{2}\boldsymbol{I})$$

# Limitations of MFVB

The advantage of MFVB over Monte Carlo alternatives is that it is a much faster algorithm

However, for Bayesian inference on more complex models, we have to re-do all the derivations.

Take for example, the Bayesian Gaussian-response linear mixed model:

$$\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{u}_i, \sigma^2 \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{X}_i\boldsymbol{u}_i, \sigma^2\boldsymbol{I}), \qquad \boldsymbol{u}_i \mid \Sigma \sim N(\boldsymbol{0}, \Sigma), \quad \text{for } 1 \le i \le m$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_{\beta}^2\boldsymbol{I}); \qquad \sigma^2 \mid a \sim \text{Inverse} - \chi^2(1, 1/a)$$

$$a \sim \text{Inverse} - \chi^2(1, 1/2), \quad \Sigma|\boldsymbol{A} \sim \text{Inverse G-Wishart}(G_{\text{full}}, 2q, \boldsymbol{A}^{-1})$$

$$\boldsymbol{A} \sim \text{Inverse G-Wishart}(G_{\text{diag}}, 1, \tfrac{1}{2}\boldsymbol{I})$$

# Limitations of MFVB

The advantage of MFVB over Monte Carlo alternatives is that it is a much faster algorithm

However, for Bayesian inference on more complex models, we have to re-do all the derivations.

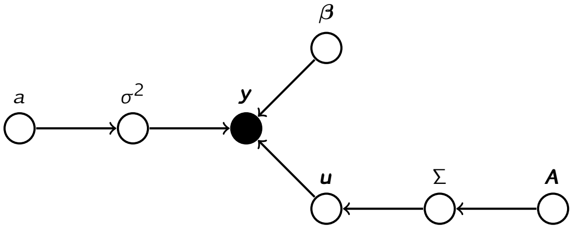
Take for example, the Bayesian Gaussian-response linear mixed model:

$$\boldsymbol{y}_i \mid \boldsymbol{\beta}, \boldsymbol{u}_i, \sigma^2 \sim N(\boldsymbol{X}_i\boldsymbol{\beta} + \boldsymbol{X}_i\boldsymbol{u}_i, \sigma^2\boldsymbol{I}), \qquad \boldsymbol{u}_i \mid \Sigma \sim N(\boldsymbol{0}, \Sigma), \quad \text{for } 1 \le i \le m$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2\boldsymbol{I}); \qquad \sigma^2 \mid a \sim \text{Inverse} - \chi^2(1, 1/a)$$

$$a \sim \text{Inverse} - \chi^2(1, 1/2), \quad \Sigma|\boldsymbol{A} \sim \text{Inverse G-Wishart}(G_{\text{full}}, 2q, \boldsymbol{A}^{-1})$$

$$\boldsymbol{A} \sim \text{Inverse G-Wishart}(G_{\text{diag}}, 1, \tfrac{1}{2}\boldsymbol{I})$$



Prof. Matt Wand
Distinguished professor of statistics, University of Technology Sydney
*Fast Approximate Inference for Arbitrarily Large Semiparametric Regression Models via Message Passing, Journal of the American Statistical Association, 2017, VOL. 112, NO. 517, 137–168, Theory and Methods*

# Steps Towards Message Passing Algorithms

Let's reconsider the Bayesian linear regression model:

$$y_i|x_i,\boldsymbol{\beta},\sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0+\beta_1 x_i,\sigma^2) \quad i=1,\ldots,n, \quad \boldsymbol{\beta}\sim N(\mathbf{0},\Sigma_0)$$
$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C\exp\left[-\frac{1}{2}\boldsymbol{\beta}^T\left\{\mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{X}+\Sigma_0^{-1}\right\}\boldsymbol{\beta}+\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right] \\ &= C\exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}+\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\} \end{aligned}$$

That is, the optimal posterior density function can be factorised into a product of a Gaussian contribution from the likelihood and a Gaussian contribution from the prior distribution.

Let's reconsider the Bayesian linear regression model:

$$y_i | x_i, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n, \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_0)$$
$$\sigma^2 | a \sim \text{Inverse} - \chi^2(1, 1/a) \quad a \sim \text{Inverse} - \chi^2(1, 1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$q^*(\boldsymbol{\beta}) = C \exp\left[-\frac{1}{2}\boldsymbol{\beta}^T \left\{\mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right]$$
$$= C \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

That is, the optimal posterior density function can be factorised into a product of a Gaussian contribution from the likelihood and a Gaussian contribution from the prior distribution.

Let's reconsider the Bayesian linear regression model:

$$y_i | x_i, \boldsymbol{\beta}, \sigma^2 \stackrel{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n, \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_0)$$
$$\sigma^2 | a \sim \text{Inverse} - \chi^2(1, 1/a) \quad a \sim \text{Inverse} - \chi^2(1, 1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C \exp\left[-\frac{1}{2}\boldsymbol{\beta}^T \left\{\mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right] \\ &= C \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\} \end{aligned}$$

That is, the optimal posterior density function can be factorised into a product of a Gaussian contribution from the likelihood and a Gaussian contribution from the prior distribution.

Let's reconsider the Bayesian linear regression model:

$$y_i | x_i, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n, \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_0)$$
$$\sigma^2 | a \sim \text{Inverse} - \chi^2(1, 1/a) \quad a \sim \text{Inverse} - \chi^2(1, 1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$q^*(\boldsymbol{\beta}) = C \exp\left[ -\frac{1}{2} \boldsymbol{\beta}^T \left\{ \mathbb{E}_q(1/\sigma^2) \boldsymbol{X}^T \boldsymbol{X} + \Sigma_0^{-1} \right\} \boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2) \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} \right]$$
$$= C \exp\left\{ -\frac{1}{2} \mathbb{E}_q(1/\sigma^2) \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2) \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{y} \right\} \exp\left\{ -\frac{1}{2} \boldsymbol{\beta}^T \Sigma_0^{-1} \boldsymbol{\beta} \right\}$$

That is, the optimal posterior density function can be factorised into a product of a Gaussian contribution from the likelihood and a Gaussian contribution from the prior distribution.

# Steps Towards Message Passing Algorithms

Let's reconsider the Bayesian linear regression model:

$$y_i | x_i, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n, \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_0)$$
$$\sigma^2 | a \sim \text{Inverse} - \chi^2(1, 1/a) \quad a \sim \text{Inverse} - \chi^2(1, 1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C \exp\left[-\frac{1}{2}\boldsymbol{\beta}^T \left\{\mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right] \\ &= C \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\} \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\} \end{aligned}$$

That is, the optimal posterior density function can be factorised into a product of a Gaussian contribution from the likelihood and a Gaussian contribution from the prior distribution.

# Steps Towards Message Passing Algorithms

Let's reconsider the Bayesian linear regression model:

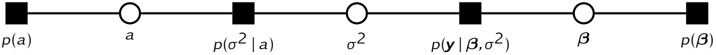
$$y_i | x_i, \boldsymbol{\beta}, \sigma^2 \overset{\text{ind.}}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2) \quad i = 1, \ldots, n, \quad \boldsymbol{\beta} \sim N(\mathbf{0}, \Sigma_0)$$
$$\sigma^2 | a \sim \text{Inverse} - \chi^2(1, 1/a) \quad a \sim \text{Inverse} - \chi^2(1, 1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$\begin{aligned} q^*(\boldsymbol{\beta}) &= C \exp\left[-\frac{1}{2}\boldsymbol{\beta}^T \left\{\mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{X} + \Sigma_0^{-1}\right\}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right] \\ &= C \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\} \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\} \end{aligned}$$

That is, the optimal posterior density function can be factorised into a product of a Gaussian contribution from the likelihood and a Gaussian contribution from the prior distribution.

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\beta$ is

$$q^*(\beta) = C\exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta + \mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

We define the messages as

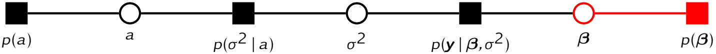
$$m_{p(y|\beta,\sigma^2)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T X^T X\beta + \mathbb{E}_q(1/\sigma^2)\beta^T X^T y\right\}$$

$$m_{p(\beta)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

# Steps Towards Message Passing Algorithms

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse} - \chi^2(1,1/a) \quad a \sim \text{Inverse} - \chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\beta$ is

$$q^*(\beta) = C\exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta + \mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

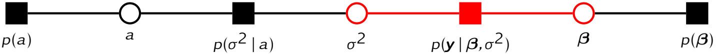
We define the messages as

$$m_{p(y|\beta,\sigma^2)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T X^T X\beta + \mathbb{E}_q(1/\sigma^2)\beta^T X^T y\right\}$$

$$m_{p(\beta)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\beta$ is

$$q^*(\beta) = C\exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta + \mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

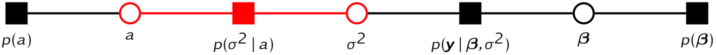
We define the messages as

$$m_{p(y|\beta,\sigma^2)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T X^T X\beta + \mathbb{E}_q(1/\sigma^2)\beta^T X^T y\right\}$$

$$m_{p(\beta)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\beta$ is

$$q^*(\beta) = C\exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta + \mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

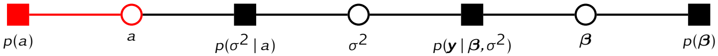
We define the messages as

$$m_{p(y|\beta,\sigma^2)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T X^T X\beta + \mathbb{E}_q(1/\sigma^2)\beta^T X^T y\right\}$$

$$m_{p(\beta)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\beta$ is

$$q^*(\beta) = C\exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{X}\beta + \mathbb{E}_q(1/\sigma^2)\beta^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

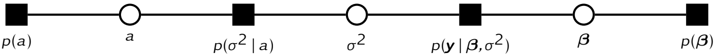
We define the messages as

$$m_{p(y|\beta,\sigma^2)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\mathbb{E}_q(1/\sigma^2)\beta^T X^T X\beta + \mathbb{E}_q(1/\sigma^2)\beta^T X^T y\right\}$$

$$m_{p(\beta)\to\beta}(\beta) \equiv \exp\left\{-\frac{1}{2}\beta^T\Sigma_0^{-1}\beta\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$q^*(\boldsymbol{\beta}) = C\exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

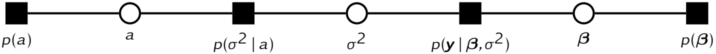
We define the messages as

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}$$

$$m_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$q^*(\boldsymbol{\beta}) = C\exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

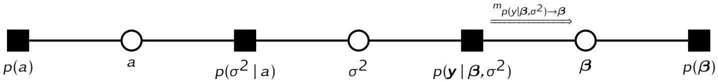
We define the messages as

$$m_{p(y|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}$$

$$m_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

This factorised approach is key to the message passing infrastructure:




$$\mathbf{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2 \boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\mathbf{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$q^*(\boldsymbol{\beta}) = C\exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\mathbf{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

We define the messages as

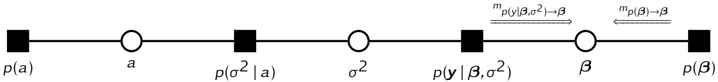
$$m_{p(y|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\mathbf{y}\right\}$$

$$m_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

# Steps Towards Message Passing Algorithms

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n),\quad \boldsymbol{\beta}\sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a)\quad a\sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

$$q^*(\boldsymbol{\beta}) = C\exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}+\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

We define the messages as

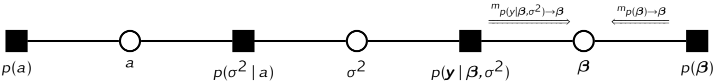
$$m_{p(y|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}+\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}$$

$$m_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:

# Steps Towards Message Passing Algorithms

This factorised approach is key to the message passing infrastructure:




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I}_n), \quad \boldsymbol{\beta} \sim N(\boldsymbol{0},\Sigma_0)$$

$$\sigma^2|a \sim \text{Inverse}-\chi^2(1,1/a) \quad a \sim \text{Inverse}-\chi^2(1,1/A^2)$$

We showed that the optimal posterior density function for $\boldsymbol{\beta}$ is

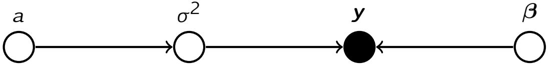
$$q^*(\boldsymbol{\beta}) = C\exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}\exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

We define the messages as

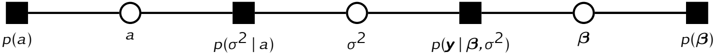
$$m_{p(y|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} + \mathbb{E}_q(1/\sigma^2)\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}\right\}$$

$$m_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \equiv \exp\left\{-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_0^{-1}\boldsymbol{\beta}\right\}$$

The representation via a directed acyclic graph does not facilitate this factorised form:



# General Message Passing Rules




The message passed from a factor $f$ to its neighbouring parameter $\boldsymbol{\theta}$ is

$$m_{f \to \boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \exp\left\{ \mathbb{E}_{-q(\boldsymbol{\theta})}(\log f) \right\}$$

For example, the message from the Gaussian likelihood factor $p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)$ to the parameter $\boldsymbol{\beta}$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) \to \boldsymbol{\beta}}(\boldsymbol{\beta}) \propto \exp\left[ \mathbb{E}_{-q(\boldsymbol{\beta})}\{\log p(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2)\} \right]$$

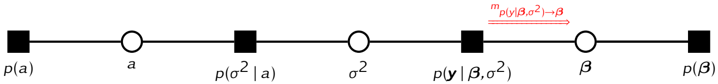
The message that $\boldsymbol{\theta}$ will pass on to another neighbouring factor is simply the message that it received from the factor $f$. For example, the message that $\boldsymbol{\beta}$ passes on to $p(\boldsymbol{\beta})$ is

$$m_{\boldsymbol{\beta} \to p(\boldsymbol{\beta})}(\boldsymbol{\beta}) = m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) \to \boldsymbol{\beta}}(\boldsymbol{\beta})$$

The approximate $q$-density function for a parameter $\boldsymbol{\theta}$ is the product of all messages that it received. For example, $q(\boldsymbol{\beta})$ is

$$q(\boldsymbol{\beta}) = m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2) \to \boldsymbol{\beta}}(\boldsymbol{\beta}) \times m_{\boldsymbol{\beta} \to p(\boldsymbol{\beta})}(\boldsymbol{\beta})$$

# General Message Passing Rules




The message passed from a factor $f$ to its neighbouring parameter $\boldsymbol{\theta}$ is

$$m_{f\to\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \exp\left\{\mathbb{E}_{-q(\boldsymbol{\theta})}(\log f)\right\}$$

For example, the message from the Gaussian likelihood factor $p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)$ to the parameter $\boldsymbol{\beta}$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\{\log p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\}\right]$$

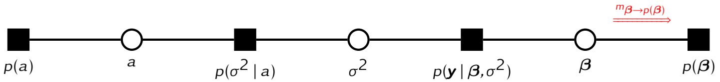
The message that $\boldsymbol{\theta}$ will pass on to another neighbouring factor is simply the message that it received from the factor $f$. For example, the message that $\boldsymbol{\beta}$ passes on to $p(\boldsymbol{\beta})$ is

$$m_{\boldsymbol{\beta}\to p(\boldsymbol{\beta})}(\boldsymbol{\beta}) = m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta})$$

The approximate $q$-density function for a parameter $\boldsymbol{\theta}$ is the product of all messages that it received. For example, $q(\boldsymbol{\beta})$ is

$$q(\boldsymbol{\beta}) = m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \times m_{\boldsymbol{\beta}\to p(\boldsymbol{\beta})}(\boldsymbol{\beta})$$

# General Message Passing Rules




The message passed from a factor $f$ to its neighbouring parameter $\boldsymbol{\theta}$ is

$$m_{f\rightarrow\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \exp\left\{\mathbb{E}_{-q(\boldsymbol{\theta})}(\log f)\right\}$$

For example, the message from the Gaussian likelihood factor $p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)$ to the parameter $\boldsymbol{\beta}$ is

$$m_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\{\log p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\}\right]$$

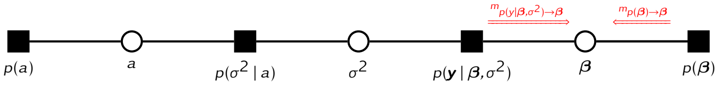
The message that $\boldsymbol{\theta}$ will pass on to another neighbouring factor is simply the message that it received from the factor $f$. For example, the message that $\boldsymbol{\beta}$ passes on to $p(\boldsymbol{\beta})$ is

$$m_{\boldsymbol{\beta}\rightarrow p(\boldsymbol{\beta})}(\boldsymbol{\beta}) = m_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}}(\boldsymbol{\beta})$$

The approximate $q$-density function for a parameter $\boldsymbol{\theta}$ is the product of all messages that it received. For example, $q(\boldsymbol{\beta})$ is

$$q(\boldsymbol{\beta}) = m_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}}(\boldsymbol{\beta}) \times m_{\boldsymbol{\beta}\rightarrow p(\boldsymbol{\beta})}(\boldsymbol{\beta})$$

# General Message Passing Rules




The message passed from a factor $f$ to its neighbouring parameter $\boldsymbol{\theta}$ is

$$m_{f\to\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \exp\left\{\mathbb{E}_{-q(\boldsymbol{\theta})}(\log f)\right\}$$

For example, the message from the Gaussian likelihood factor $p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)$ to the parameter $\boldsymbol{\beta}$ is

$$m_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto \exp\left[\mathbb{E}_{-q(\boldsymbol{\beta})}\{\log p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\}\right]$$

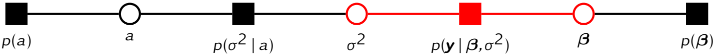
The message that $\boldsymbol{\theta}$ will pass on to another neighbouring factor is simply the message that it received from the factor $f$. For example, the message that $\boldsymbol{\beta}$ passes on to $p(\boldsymbol{\beta})$ is

$$m_{\boldsymbol{\beta}\to p(\boldsymbol{\beta})}(\boldsymbol{\beta}) = m_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta})$$

The approximate $q$-density function for a parameter $\boldsymbol{\theta}$ is the product of all messages that it received. For example, $q(\boldsymbol{\beta})$ is

$$q(\boldsymbol{\beta}) = m_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \times m_{\boldsymbol{\beta}\to p(\boldsymbol{\beta})}(\boldsymbol{\beta})$$

# The Gaussian Likelihood Fragment




The message from the likelihood specification to the parameter $\boldsymbol{\beta}$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\end{bmatrix}^T \begin{bmatrix}\mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{y}\\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\,\mathrm{vec}(\boldsymbol{X}^T\boldsymbol{X})\end{bmatrix}\right\}$$

That is, $m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta})$ is a multivariate normal density function in exponential family form:

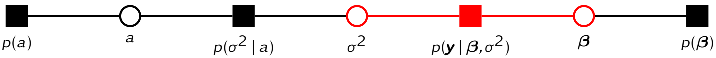
$$p(\boldsymbol{x}) = C\exp\{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta}\}.$$

The message to $\sigma^2$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \begin{bmatrix}-n/2\\ -\frac{1}{2}\,\mathbb{E}_q(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2)\end{bmatrix}\right\},$$

which is an inverse chi-squared density function in exponential family form.

# The Gaussian Likelihood Fragment




The message from the likelihood specification to the parameter $\boldsymbol{\beta}$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto \exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \text{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T)\end{bmatrix}^T \begin{bmatrix}\mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{y}\\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\,\text{vec}(\boldsymbol{X}^T\boldsymbol{X})\end{bmatrix}\right\}$$

That is, $m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}}(\boldsymbol{\beta})$ is a multivariate normal density function in exponential family form:

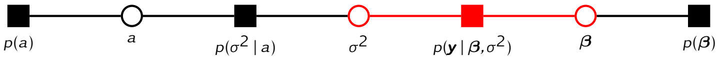
$$p(\boldsymbol{x}) = C\exp\{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta}\}.$$

The message to $\sigma^2$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix}\log(\sigma^2)\\ 1/\sigma^2\end{bmatrix}^T \begin{bmatrix}-n/2\\ -\frac{1}{2}\,\mathbb{E}_q(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2)\end{bmatrix}\right\},$$

which is an inverse chi-squared density function in exponential family form.

# The Gaussian Likelihood Fragment




The message from the likelihood specification to the parameter $\boldsymbol{\beta}$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \propto \exp\left\{\begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T \begin{bmatrix} \mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{y} \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\,\mathrm{vec}(\boldsymbol{X}^T\boldsymbol{X}) \end{bmatrix}\right\}$$

That is, $m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta})$ is a multivariate normal density function in exponential family form:

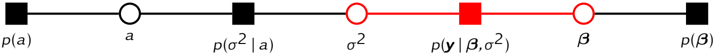
$$p(\boldsymbol{x}) = C\exp\{\boldsymbol{T}(\boldsymbol{x})^T\boldsymbol{\eta}\}.$$

The message to $\sigma^2$ is

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\sigma^2}(\sigma^2) = \exp\left\{\begin{bmatrix} \log(\sigma^2) \\ 1/\sigma^2 \end{bmatrix}^T \begin{bmatrix} -n/2 \\ -\frac{1}{2}\,\mathbb{E}_q(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2) \end{bmatrix}\right\},$$

which is an inverse chi-squared density function in exponential family form.

# The Gaussian Likelihood Fragment




By restricting the form of the messages passing updates to distributions in the exponential family form, we can completely characterise these message updates by their natural parameter vectors:

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}} = \begin{bmatrix} \mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{y} \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\,\text{vec}(\boldsymbol{X}^T\boldsymbol{X}) \end{bmatrix}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\sigma^2} = \begin{bmatrix} -n/2 \\ -\frac{1}{2}\,\mathbb{E}_q(\|\boldsymbol{y} - \boldsymbol{X}\beta\|^2) \end{bmatrix}$$

# Gaussian Density Function in Exponential Family Form

$$\boldsymbol{x}|\boldsymbol{\mu},\Sigma \sim \mathrm{N}(\boldsymbol{\mu},\Sigma)$$

The density function in exponential family form is:

$$p(\boldsymbol{x}|\boldsymbol{\mu},\Sigma) = C\exp\left\{\begin{bmatrix}\boldsymbol{\beta}\\ \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^{\mathsf{T}})\end{bmatrix}^{\mathsf{T}}\boldsymbol{\eta}\right\}, \quad \boldsymbol{\eta} = \begin{bmatrix}\boldsymbol{\eta}_1\\ \boldsymbol{\eta}_2\end{bmatrix} = \begin{bmatrix}\Sigma^{-1}\boldsymbol{\mu}\\ -\frac{1}{2}\,\mathrm{vec}(\Sigma^{-1})\end{bmatrix}$$

The transformation back to the common parameters is

$$\Sigma = -\frac{1}{2}\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_2)\}^{-1}, \quad \boldsymbol{\mu} = \Sigma\boldsymbol{\eta}_1$$

# Inverse-$\chi^2$ Density Function in Exponential Family Form

$$x|\kappa, \lambda \sim \text{Inverse} - \chi^2(\kappa, \lambda)$$

The density function in exponential family form is:

$$p(x|\kappa,\lambda) = C \exp\left\{ \begin{bmatrix} \log|x| \\ 1/x \end{bmatrix}^{\mathsf{T}} \boldsymbol{\eta} \right\}, \quad \boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \end{bmatrix} = \begin{bmatrix} -\frac{1}{2}(\kappa+2) \\ -\frac{\lambda}{2} \end{bmatrix}$$

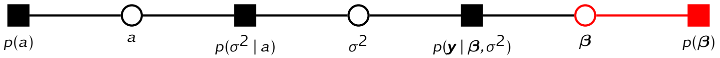
The transformation back to the common parameters is

$$\kappa = -2\eta_1 - 2, \quad \lambda = -2\eta_2$$

For variational inference, we require:

$$\mathbb{E}(1/\sigma^2) = \frac{\eta_1 + 1}{\eta_2}$$

# Fragments



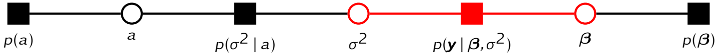

Gaussian prior fragment (Wand, 2017)

Gaussian likelihood fragment (Wand, 2017)

Iterated inverse chi-squared fragment (Maestrini and Wand, 2020)

Inverse chi-squared prior fragment (Maestrini and Wand, 2020)

# Fragments



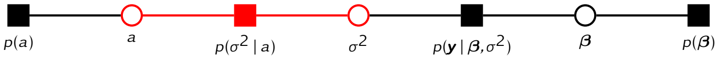

Gaussian prior fragment (Wand, 2017)

Gaussian likelihood fragment (Wand, 2017)

Iterated inverse chi-squared fragment (Maestrini and Wand, 2020)

Inverse chi-squared prior fragment (Maestrini and Wand, 2020)

# Fragments



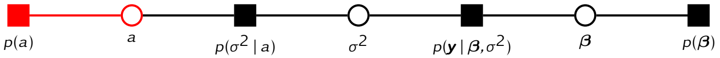

Gaussian prior fragment (Wand, 2017)

Gaussian likelihood fragment (Wand, 2017)

Iterated inverse chi-squared fragment (Maestrini and Wand, 2020)

Inverse chi-squared prior fragment (Maestrini and Wand, 2020)

# Fragments



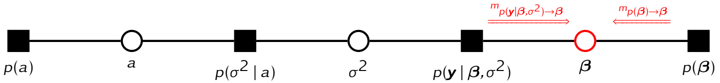

Gaussian prior fragment (Wand, 2017)

Gaussian likelihood fragment (Wand, 2017)

Iterated inverse chi-squared fragment (Maestrini and Wand, 2020)

Inverse chi-squared prior fragment (Maestrini and Wand, 2020)

# $q$-Density Functions




Recall that the approximate $q$-density function for each parameter is the product of all the messages that it received.

The computation for $q(\boldsymbol{\beta})$ is

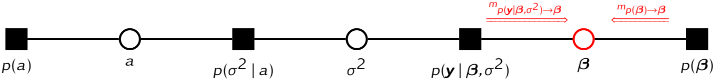
$$q(\boldsymbol{\beta}) = m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \times m_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta})$$

$$= \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T (\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}) \right\}$$

So the $q$-density function is also in the exponential family of density functions with natural parameter vector:

$$\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})} = \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}$$

# $q$-Density Functions




Recall that the approximate $q$-density function for each parameter is the product of all the messages that it received.

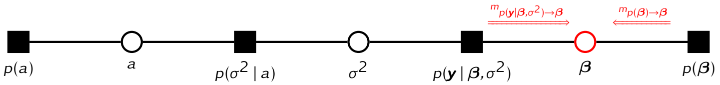
The computation for $q(\boldsymbol{\beta})$ is

$$
\begin{aligned}
q(\boldsymbol{\beta}) &= m_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}}(\boldsymbol{\beta}) \times m_{p(\boldsymbol{\beta})\rightarrow\boldsymbol{\beta}}(\boldsymbol{\beta}) \\
&= \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \operatorname{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T (\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\rightarrow\boldsymbol{\beta}}) \right\}
\end{aligned}
$$

So the $q$-density function is also in the exponential family of density functions with natural parameter vector:

$$
\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})} = \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\rightarrow\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\rightarrow\boldsymbol{\beta}}
$$

# $q$-Density Functions




Recall that the approximate $q$-density function for each parameter is the product of all the messages that it received.

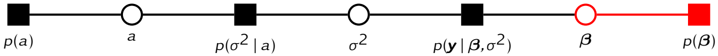
The computation for $q(\boldsymbol{\beta})$ is

$$
\begin{aligned}
q(\boldsymbol{\beta}) &= m_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \times m_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \\
&= \exp\left\{ \begin{bmatrix} \boldsymbol{\beta} \\ \mathrm{vec}(\boldsymbol{\beta}\boldsymbol{\beta}^T) \end{bmatrix}^T \left( \boldsymbol{\eta}_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}} \right) \right\}
\end{aligned}
$$

So the $q$-density function is also in the exponential family of density functions with natural parameter vector:

$$
\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})} = \boldsymbol{\eta}_{p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}} + \boldsymbol{\eta}_{p(\boldsymbol{\beta})\to\boldsymbol{\beta}}
$$

# The Gaussian Prior Fragment




$$\boldsymbol{\beta} \sim N(\mathbf{0}, \sigma_\beta^2 \boldsymbol{I})$$

Inputs:

$$\sigma_\beta^2$$

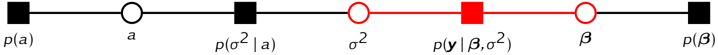
Updates:

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta}) \to \boldsymbol{\beta}} = \begin{bmatrix} \mathbf{0} \\ -\frac{1}{2\sigma_\beta^2} \operatorname{vec}(\boldsymbol{I}) \end{bmatrix}$$

Outputs:

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta}) \to \boldsymbol{\beta}}$$

# The Gaussian Likelihood Fragment




$$\boldsymbol{y}|\boldsymbol{\beta},\sigma^2 \sim N(\boldsymbol{X}\boldsymbol{\beta},\sigma^2\boldsymbol{I})$$

Inputs:

$$\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})}, \quad \boldsymbol{\eta}_{q^*(\sigma^2)}$$

Updates:

$$\mathrm{Cov}_q(\boldsymbol{\beta}) = -\frac{1}{2}\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})})_2\}^{-1}, \quad \mathbb{E}_q(\boldsymbol{\beta}) = \mathrm{Cov}_q(\boldsymbol{\beta})(\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})})_1,$$

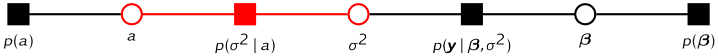
$$\mathbb{E}_q(1/\sigma^2) = \frac{(\boldsymbol{\eta}_{q^*(\sigma^2)})_1 + 1}{(\boldsymbol{\eta}_{q^*(\sigma^2)})_2}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}} = \begin{bmatrix} \mathbb{E}_q(1/\sigma^2)\boldsymbol{X}^T\boldsymbol{y} \\ -\frac{1}{2}\,\mathbb{E}_q(1/\sigma^2)\mathrm{vec}(\boldsymbol{X}^T\boldsymbol{X}) \end{bmatrix}, \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\sigma^2} = \begin{bmatrix} -n/2 \\ -\frac{1}{2}\,\mathbb{E}_q(\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|^2) \end{bmatrix}$$

Outputs:

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\boldsymbol{\beta}}, \quad \boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2)\to\sigma^2}$$

# Iterated Inverse-$\chi^2$ Fragment




$$\sigma^2|a \sim \text{Inverse-}\chi^2(1, 1/a)$$

Inputs:

$$\boldsymbol{\eta}_{q^*(\Sigma)}, \quad \boldsymbol{\eta}_{q^*(\boldsymbol{A})}$$

Updates:

$$\mathbb{E}_q(1/\sigma^2) = \frac{(\boldsymbol{\eta}_{q^*(\sigma^2)})_1 + 1}{(\boldsymbol{\eta}_{q^*(\sigma^2)})_2},$$

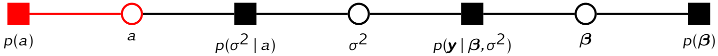
$$\mathbb{E}_q(1/a) = \frac{(\boldsymbol{\eta}_{q^*(a)})_1 + 1}{(\boldsymbol{\eta}_{q^*(a)})_2},$$

$$\boldsymbol{\eta}_{p(\sigma^2|a)\to\sigma^2} = \begin{bmatrix} -\frac{3}{2} \\ -\frac{1}{2}\,\mathbb{E}_q(1/a) \end{bmatrix}, \quad \boldsymbol{\eta}_{p(\sigma^2|a)\to a} = \begin{bmatrix} -\frac{1}{2} \\ -\frac{1}{2}\{\text{vec}\,\mathbb{E}_q(1/\sigma^2)\} \end{bmatrix}$$

Outputs:

$$\boldsymbol{\eta}_{p(\sigma^2|a)\to\sigma^2}, \quad \boldsymbol{\eta}_{p(\sigma^2|a)\to a}$$

# Inverse-$\chi^2$ Prior Fragment




$$a \sim \text{Inverse-}\chi^2(\kappa, \lambda)$$

Inputs:

$$\kappa, \lambda$$

Updates:

$$\boldsymbol{\eta}_{p(a)\to a} = \begin{bmatrix} -\frac{1}{2}(\kappa+2) \\ -\frac{\lambda}{2} \end{bmatrix}$$

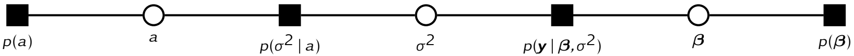
Outputs:

$$\boldsymbol{\eta}_{p(a)\to a}$$

# Variational Message Passing

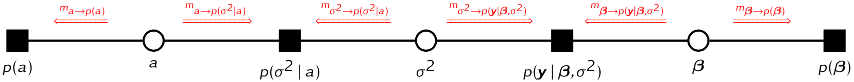
## VMP for the Bayesian linear regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\sigma^2,a)\|p(\boldsymbol{\beta},\sigma^2,a|\boldsymbol{y})\right\}$ converges.



## VMP for the Bayesian linear regression model

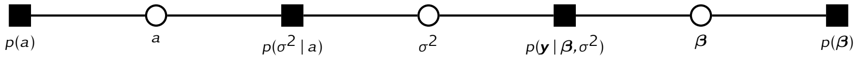
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\sigma^2,a)||p(\boldsymbol{\beta},\sigma^2,a|\mathbf{y})\right\}$ converges.



# Variational Message Passing

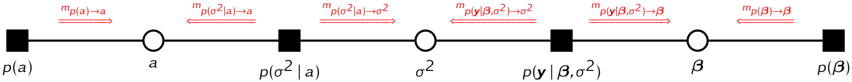
## VMP for the Bayesian linear regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\sigma^2,a)\|p(\boldsymbol{\beta},\sigma^2,a|\boldsymbol{y})\right\}$ converges.



## VMP for the Bayesian linear regression model

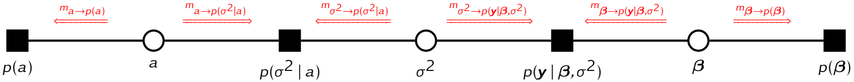
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\sigma^2,a)\|p(\boldsymbol{\beta},\sigma^2,a|\mathbf{y})\right\}$ converges.



# Variational Message Passing

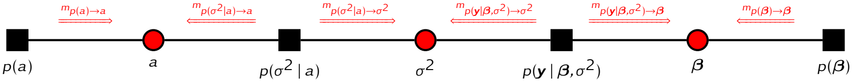
## VMP for the Bayesian linear regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\sigma^2,a)||p(\boldsymbol{\beta},\sigma^2,a|\mathbf{y})\right\}$ converges.



## VMP for the Bayesian linear regression model

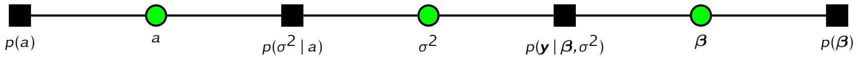
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\sigma^2,a)\|p(\boldsymbol{\beta},\sigma^2,a|\mathbf{y})\right\}$ converges.



# Variational Message Passing

## VMP for the Bayesian linear regression model

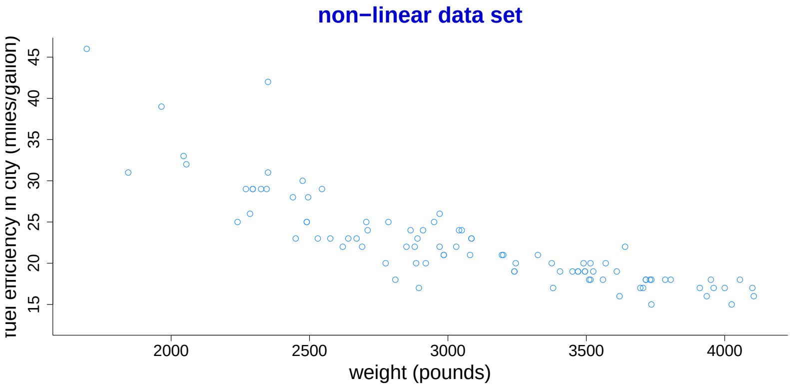
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\sigma^2,a)\|p(\boldsymbol{\beta},\sigma^2,a|\mathbf{y})\right\}$ converges.



Part II

# Building Model Complexity

# Nonlinear Data



We can address nonlinear data sets using semiparametric regression techniques

Let $x_i$ be the $i$th vehicle weight and $y_i$ be its corresponding fuel efficiency score. The corresponding vectors are $\boldsymbol{x}$ and $\boldsymbol{y}$.

We construct a fixed effects matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Nonlinear effects are incorporated through a random effects matrix

$$\boldsymbol{Z} = \begin{bmatrix} z_1(x_1) & \dots & z_K(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_N) & \dots & z_K(x_N) \end{bmatrix},$$

where $\{z_k\}_{k=1}^K$ is a suitable spline basis – see Wand and Ormerod (2008) for a description of canonical cubic O'Sullivan penalised splines.

# Bayesian Semiparametric Regression

We can address nonlinear data sets using semiparametric regression techniques

Let $x_i$ be the $i$th vehicle weight and $y_i$ be its corresponding fuel efficiency score. The corresponding vectors are $\boldsymbol{x}$ and $\boldsymbol{y}$.

We construct a fixed effects matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Nonlinear effects are incorporated through a random effects matrix

$$\boldsymbol{Z} = \begin{bmatrix} z_1(x_1) & \dots & z_K(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_N) & \dots & z_K(x_N) \end{bmatrix},$$

where $\{z_k\}_{k=1}^{K}$ is a suitable spline basis – see Wand and Ormerod (2008) for a description of canonical cubic O'Sullivan penalised splines.

# Bayesian Semiparametric Regression

We can address nonlinear data sets using semiparametric regression techniques

Let $x_i$ be the $i$th vehicle weight and $y_i$ be its corresponding fuel efficiency score. The corresponding vectors are $\boldsymbol{x}$ and $\boldsymbol{y}$.

We construct a fixed effects matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Nonlinear effects are incorporated through a random effects matrix

$$Z = \begin{bmatrix} z_1(x_1) & \dots & z_K(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_N) & \dots & z_K(x_N) \end{bmatrix},$$

where $\{z_k\}_{k=1}^K$ is a suitable spline basis – see Wand and Ormerod (2008) for a description of canonical cubic O'Sullivan penalised splines.

We can address nonlinear data sets using semiparametric regression techniques

Let $x_i$ be the $i$th vehicle weight and $y_i$ be its corresponding fuel efficiency score. The corresponding vectors are $\boldsymbol{x}$ and $\boldsymbol{y}$.

We construct a fixed effects matrix

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$$

Nonlinear effects are incorporated through a random effects matrix

$$\boldsymbol{Z} = \begin{bmatrix} z_1(x_1) & \dots & z_K(x_1) \\ \vdots & \ddots & \vdots \\ z_1(x_N) & \dots & z_K(x_N) \end{bmatrix},$$

where $\{z_k\}_{k=1}^{K}$ is a suitable spline basis – see Wand and Ormerod (2008) for a description of canonical cubic O′Sullivan penalised splines.

A fundamental ingredient, which facilitates the incorporation of nonlinear predictor effects, is that of mixed model-based penalized splines:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x), \quad u_k|\sigma_u \overset{\text{ind.}}{\sim} \mathrm{N}(0, \sigma_u^2), \quad k = 1, \ldots, K.$$

The Bayesian semiparametric regression model is

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{X\beta} + \boldsymbol{Zu}, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_2)$$

$$\boldsymbol{u}|\sigma_u^2 \sim \mathrm{N}(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}_K)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1, 1/a_u), \quad a_u \sim \text{inverse-}\chi^2(1, 1/A^2)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1, 1/a), \quad a \sim \text{inverse-}\chi^2(1, 1/A^2)$$

# Bayesian Semiparametric Regression

A fundamental ingredient, which facilitates the incorporation of nonlinear predictor effects, is that of mixed model-based penalized splines:

$$f(x) = \beta_0 + \beta_1 x + \sum_{k=1}^{K} u_k z_k(x), \quad u_k|\sigma_u \overset{\text{ind.}}{\sim} \mathrm{N}(0, \sigma_u^2), \quad k = 1, \ldots, K.$$

The Bayesian semiparametric regression model is

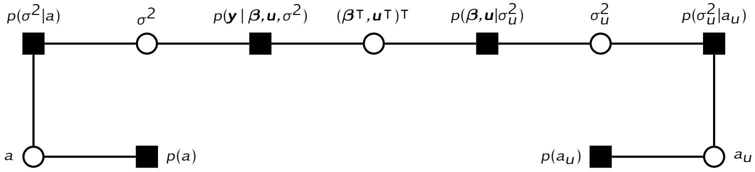
$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2 \sim \mathrm{N}(\boldsymbol{X\beta} + \boldsymbol{Zu}, \sigma^2 \boldsymbol{I}), \quad \boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I}_2)$$
$$\boldsymbol{u}|\sigma_u^2 \sim \mathrm{N}(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}_K)$$
$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1, 1/a_u), \quad a_u \sim \text{inverse-}\chi^2(1, 1/A^2)$$
$$\sigma^2|a \sim \text{Inverse-}\chi^2(1, 1/a), \quad a \sim \text{inverse-}\chi^2(1, 1/A^2)$$

# Bayesian Semiparametric Regression

$$y|\boldsymbol{\beta},\boldsymbol{u},\sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma^2\boldsymbol{I}), \quad \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\Bigg|\, \sigma_u^2 \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1,1/A^2)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

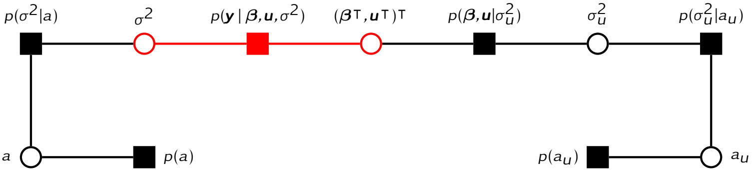
Inverse Wishart prior fragment

# Bayesian Semiparametric Regression

$$y|\boldsymbol{\beta},\boldsymbol{u},\sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma^2\boldsymbol{I}), \quad \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\Bigg|\ \sigma_u^2 \sim \mathrm{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1,1/A^2)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

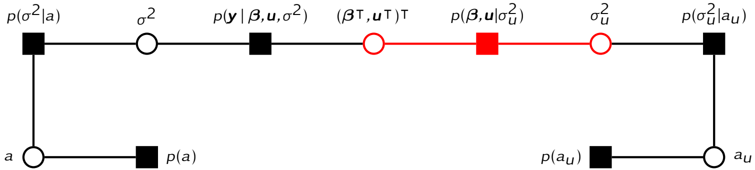
Inverse Wishart prior fragment

# Bayesian Semiparametric Regression

$$y|\boldsymbol{\beta},\boldsymbol{u},\sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma^2\boldsymbol{I}), \quad \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\Bigg|\ \sigma_u^2 \sim \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1,1/A^2)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

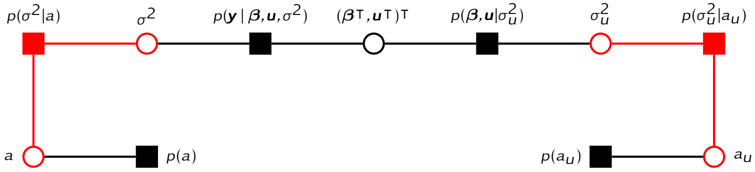
Inverse Wishart prior fragment

# Bayesian Semiparametric Regression

$$\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{u},\sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma^2\boldsymbol{I}), \quad \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\Bigg|\ \sigma_u^2 \sim \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1,1/A^2)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

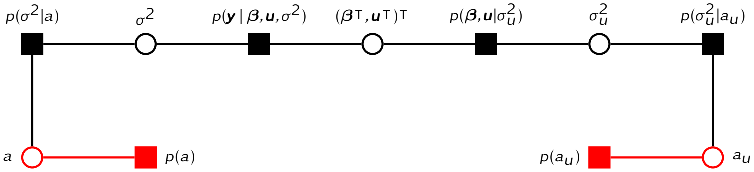
Inverse Wishart prior fragment

# Bayesian Semiparametric Regression

$$y|\boldsymbol{\beta},\boldsymbol{u},\sigma^2 \sim \mathrm{N}(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{Z}\boldsymbol{u},\sigma^2\boldsymbol{I}), \quad \left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right|\sigma_u^2 \sim \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad {\color{red}a_u \sim \text{Inverse-}\chi^2(1,1/A^2)}$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad {\color{red}a \sim \text{Inverse-}\chi^2(1,1/A^2)}$$



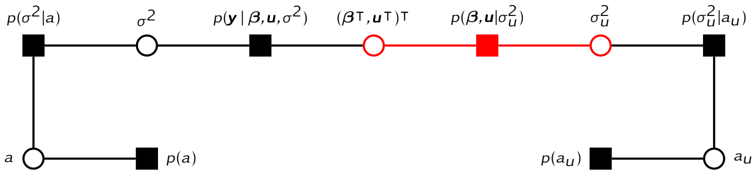

Gaussian likelihood fragment

Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

Inverse Wishart prior fragment

# Gaussian Penalization Fragment




$$\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\Bigg|\ \sigma_u^2\boldsymbol{I}_K \sim \mathrm{N}\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\boldsymbol{I}_K\end{bmatrix}\right)$$

Inputs:

$$\boldsymbol{\eta}_{q^*(\boldsymbol{\beta},\boldsymbol{u})},\quad \boldsymbol{\eta}_{q^*(\sigma_u^2)}$$

Updates:

$$\mathrm{Cov}_q\{(\boldsymbol{\beta}^\top,\boldsymbol{u}^\top)^\top\} = -\tfrac{1}{2}\{\mathrm{vec}^{-1}(\boldsymbol{\eta}_{q^*(\boldsymbol{\beta},\boldsymbol{u})})_2\}^{-1},\quad \mathbb{E}_q\{(\boldsymbol{\beta}^\top,\boldsymbol{u}^\top)^\top\} = \mathrm{Cov}_q\{(\boldsymbol{\beta}^\top,\boldsymbol{u}^\top)^\top\}(\boldsymbol{\eta}_{q^*(\boldsymbol{\beta},\boldsymbol{u})})_1,$$

$$\mathbb{E}_q(1/\sigma_u^2) = \frac{(\boldsymbol{\eta}_{q^*(\sigma_u^2)})_1+1}{(\boldsymbol{\eta}_{q^*(\sigma_u^2)})_2},$$

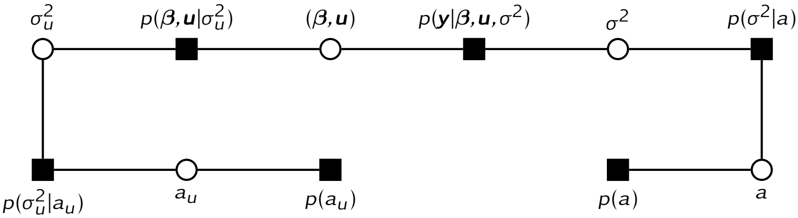
$$\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\sigma_u^2)\to(\boldsymbol{\beta},\boldsymbol{u})} = \begin{bmatrix}\boldsymbol{0}_{K+2}\\ -\frac{1}{2}\,\mathrm{vec}\left(\begin{bmatrix}(1/\sigma_\beta^2)\boldsymbol{I}_2 & \boldsymbol{O}\\ \boldsymbol{O} & \mathbb{E}_q(1/\sigma_u^2)\boldsymbol{I}_K\end{bmatrix}\right)\end{bmatrix},\quad \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\sigma_u^2)\to(\sigma_u^2)} = \begin{bmatrix}-K/2\\ -\frac{1}{2}\,\mathrm{vec}\{\mathbb{E}_q(\boldsymbol{u}^\top\boldsymbol{u})\}\end{bmatrix}$$

Outputs:

$$\boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\sigma_u^2)\to(\boldsymbol{\beta},\boldsymbol{u})},\quad \boldsymbol{\eta}_{p(\boldsymbol{\beta},\boldsymbol{u}|\sigma_u^2)\to(\sigma_u^2)}$$

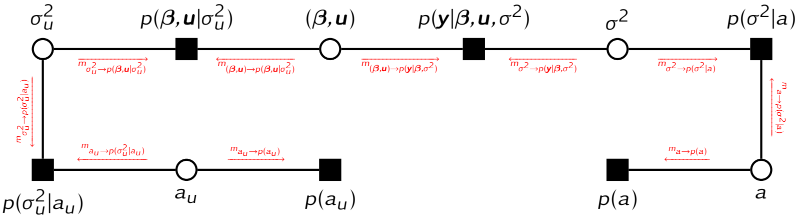
## VMP for the Bayesian semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u|\boldsymbol{y})\right\}$ converges.



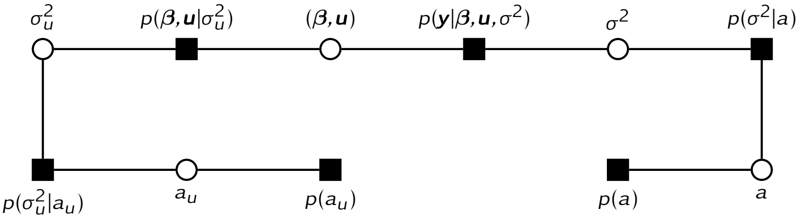
## VMP for the Bayesian semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u|\boldsymbol{y})\right\}$ converges.



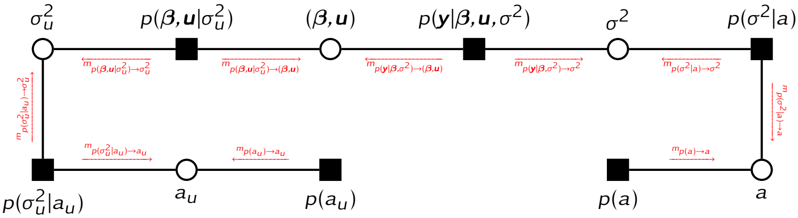
## VMP for the Bayesian semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u|\boldsymbol{y})\right\}$ converges.



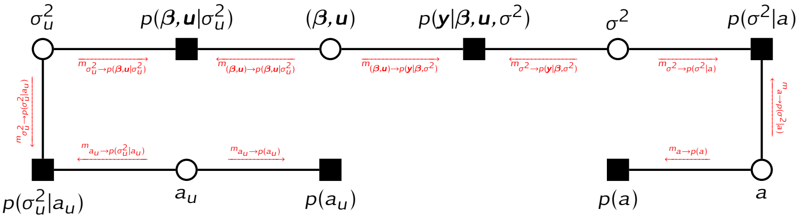
## VMP for the Bayesian semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\mathbf{u},\sigma^2,\sigma_u^2,a,a_u)\|p(\boldsymbol{\beta},\mathbf{u},\sigma^2,\sigma_u^2,a,a_u|\mathbf{y})\right\}$ converges.



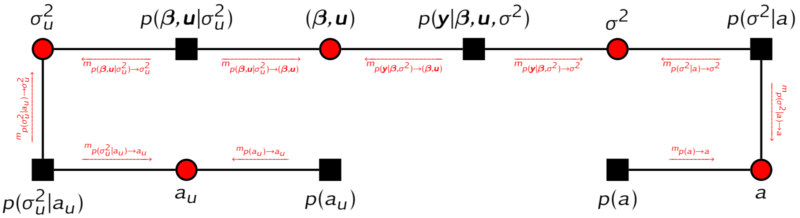
## VMP for the Bayesian semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u|\boldsymbol{y})\right\}$ converges.



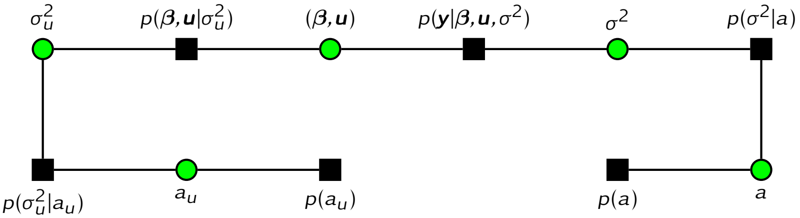
## VMP for the Bayesian semiparametric regression model

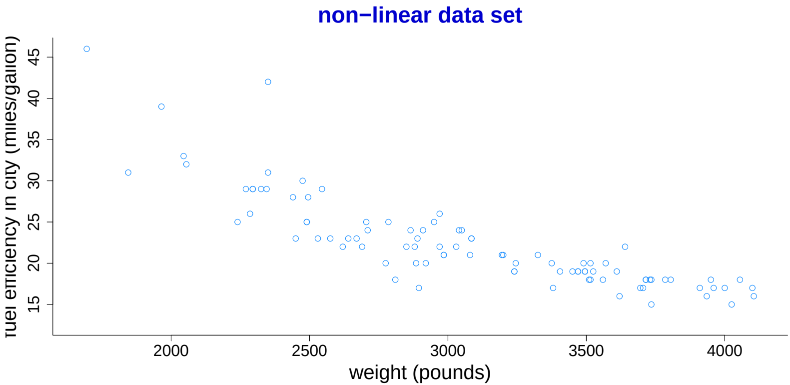
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u|\boldsymbol{y})\right\}$ converges.

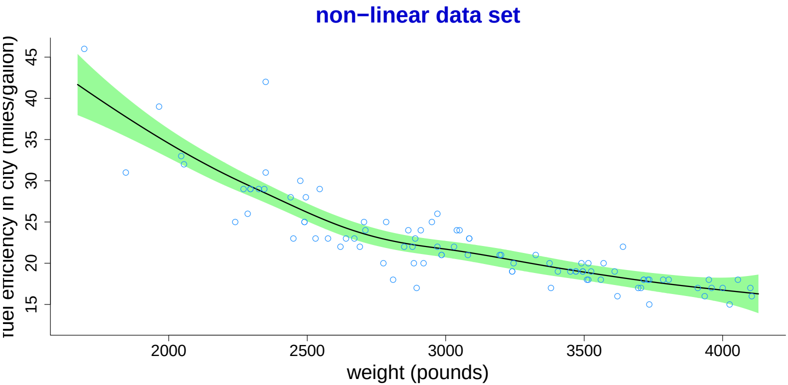


## VMP for the Bayesian semiparametric regression model

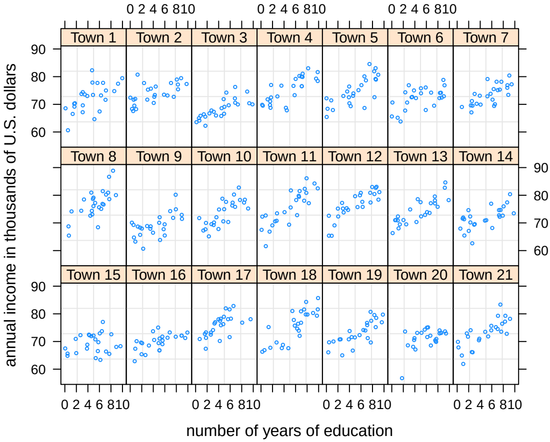
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\sigma_u^2,a,a_u|\boldsymbol{y})\right\}$ converges.




non−linear data set
fuel efficiency in city (miles/gallon)
15
20
25
30
35
40
45
2000
2500
3000
3500
4000
weight (pounds)


non−linear data set
fuel efficiency in city (miles/gallon)
15
20
25
30
35
40
45
2000
2500
3000
3500
4000
weight (pounds)

# Multilevel Data



# Multilevel Data

Suppose we have $m$ groups and $n_i$ subjects in the $i$th group.

Let $x_{ij}$ be the predictor for the $j$th subject in the $i$th, and let $y_{ij}$ be the corresponding observation.

A Gaussian response linear mixed model for such data consists of

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{i1} + u_{i2} x_{ij} + \epsilon_{ij}, \quad \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \overset{\text{ind.}}{\sim} \mathrm{N}(\mathbf{0}_2, \Sigma_u),$$

$$\epsilon_{ij} \overset{\text{ind.}}{\sim} \mathrm{N}(0, \sigma^2), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m.$$

Next, set

$$\boldsymbol{y}_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \boldsymbol{x}_i \equiv \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{bmatrix}, \quad \boldsymbol{X}_i \equiv \begin{bmatrix} \mathbf{1}_{n_i} & \boldsymbol{x}_i \end{bmatrix} \quad \text{and} \quad \boldsymbol{u}_i \equiv \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix}.$$

# Multilevel Data

Suppose we have $m$ groups and $n_i$ subjects in the $i$th group.

Let $x_{ij}$ be the predictor for the $j$th subject in the $i$th, and let $y_{ij}$ be the corresponding observation.

A Gaussian response linear mixed model for such data consists of

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{i1} + u_{i2} x_{ij} + \epsilon_{ij}, \quad \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \overset{\text{ind.}}{\sim} \mathrm{N}(\mathbf{0}_2, \Sigma_u),$$

$$\epsilon_{ij} \overset{\text{ind.}}{\sim} \mathrm{N}(0, \sigma^2), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m.$$

Next, set

$$\boldsymbol{y}_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \boldsymbol{x}_i \equiv \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{bmatrix}, \quad \boldsymbol{X}_i \equiv \begin{bmatrix} \mathbf{1}_{n_i} & \boldsymbol{x}_i \end{bmatrix} \quad \text{and} \quad \boldsymbol{u}_i \equiv \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix}.$$

Suppose we have $m$ groups and $n_i$ subjects in the $i$th group.

Let $x_{ij}$ be the predictor for the $j$th subject in the $i$th, and let $y_{ij}$ be the corresponding observation.

A Gaussian response linear mixed model for such data consists of

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{i1} + u_{i2} x_{ij} + \epsilon_{ij}, \quad \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \overset{\text{ind.}}{\sim} N(\mathbf{0}_2, \Sigma_u),$$

$$\epsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma^2), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m.$$

Next, set

$$\boldsymbol{y}_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \boldsymbol{x}_i \equiv \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{bmatrix}, \quad \boldsymbol{X}_i \equiv \begin{bmatrix} \mathbf{1}_{n_i} & \boldsymbol{x}_i \end{bmatrix} \quad \text{and} \quad \boldsymbol{u}_i \equiv \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix}.$$

Suppose we have $m$ groups and $n_i$ subjects in the $i$th group.

Let $x_{ij}$ be the predictor for the $j$th subject in the $i$th, and let $y_{ij}$ be the corresponding observation.

A Gaussian response linear mixed model for such data consists of

$$y_{ij} = \beta_0 + \beta_1 x_{ij} + u_{i1} + u_{i2} x_{ij} + \epsilon_{ij}, \quad \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix} \overset{\text{ind.}}{\sim} N(\mathbf{0}_2, \Sigma_u),$$

$$\epsilon_{ij} \overset{\text{ind.}}{\sim} N(0, \sigma^2), \quad j = 1, \ldots, n_i, \quad i = 1, \ldots, m.$$

Next, set

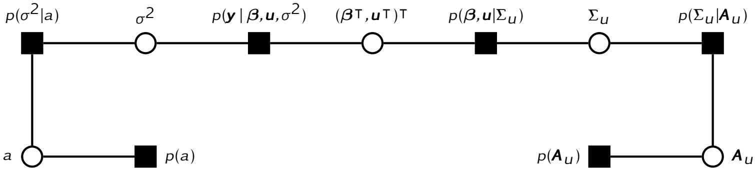
$$\boldsymbol{y}_i \equiv \begin{bmatrix} y_{i1} \\ \vdots \\ y_{in_i} \end{bmatrix}, \quad \boldsymbol{x}_i \equiv \begin{bmatrix} x_{i1} \\ \vdots \\ x_{in_i} \end{bmatrix}, \quad \boldsymbol{X}_i \equiv \begin{bmatrix} \mathbf{1}_{n_i} & \boldsymbol{x}_i \end{bmatrix} \quad \text{and} \quad \boldsymbol{u}_i \equiv \begin{bmatrix} u_{i1} \\ u_{i2} \end{bmatrix}.$$

# Bayesian Multilevel Data Model

$$\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{u}_i,\sigma^2 \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{X}_i\boldsymbol{u}_i,\sigma^2\boldsymbol{I}), \quad \left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}_i\end{bmatrix}\right| \Sigma_u \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \Sigma_u\end{bmatrix}\right)$$

$$\Sigma_u|\boldsymbol{A}_u \sim \text{Inverse Wishart}(1,\boldsymbol{A}_u^{-1}), \quad \boldsymbol{A}_u \sim \text{Inverse Wishart}\left(1,\frac{1}{A^2}\boldsymbol{I}\right)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment

Iterated inverse Wishart fragment

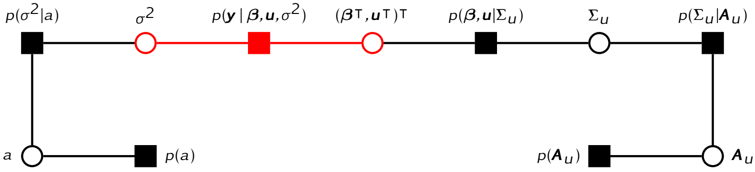
Inverse Wishart prior fragment

# Bayesian Multilevel Data Model

$$\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{u}_i,\sigma^2 \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{X}_i\boldsymbol{u}_i,\sigma^2\boldsymbol{I}), \quad \left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}_i\end{bmatrix}\right| \Sigma_u \stackrel{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \Sigma_u\end{bmatrix}\right)$$

$$\Sigma_u|\boldsymbol{A}_u \sim \text{Inverse Wishart}(1,\boldsymbol{A}_u^{-1}), \quad \boldsymbol{A}_u \sim \text{Inverse Wishart}\left(1,\frac{1}{A^2}\boldsymbol{I}\right)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment

Iterated inverse Wishart fragment

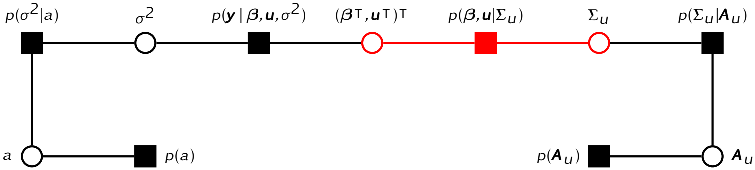
Inverse Wishart prior fragment

# Bayesian Multilevel Data Model

$$\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{u}_i,\sigma^2 \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{X}_i\boldsymbol{u}_i,\sigma^2\boldsymbol{I}), \quad \left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}_i\end{bmatrix}\right| \Sigma_u \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \Sigma_u\end{bmatrix}\right)$$

$$\Sigma_u|\boldsymbol{A}_u \sim \text{Inverse Wishart}(1,\boldsymbol{A}_u^{-1}), \quad \boldsymbol{A}_u \sim \text{Inverse Wishart}\left(1,\frac{1}{A^2}\boldsymbol{I}\right)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment

Iterated inverse Wishart fragment

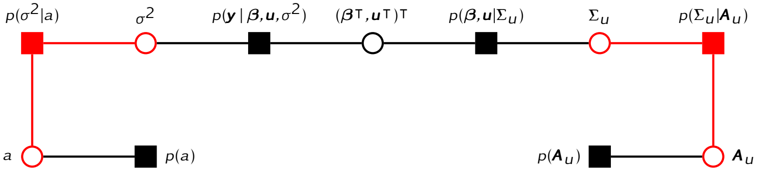
Inverse Wishart prior fragment

# Bayesian Multilevel Data Model

$$\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{u}_i,\sigma^2 \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{X}_i\boldsymbol{u}_i,\sigma^2\boldsymbol{I}), \quad \left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}_i\end{bmatrix}\right|\Sigma_u \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \Sigma_u\end{bmatrix}\right)$$

$$\Sigma_u|\boldsymbol{A}_u \sim \text{Inverse Wishart}(1,\boldsymbol{A}_u^{-1}), \quad \boldsymbol{A}_u \sim \text{Inverse Wishart}\left(1,\frac{1}{A^2}\boldsymbol{I}\right)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment

Iterated inverse Wishart fragment

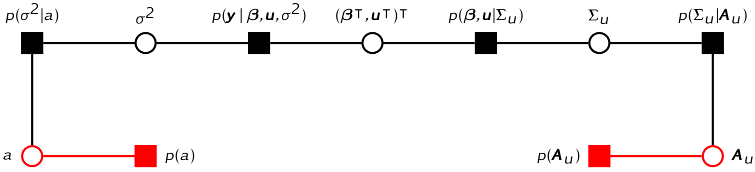
Inverse Wishart prior fragment

# Bayesian Multilevel Data Model

$$\boldsymbol{y}_i|\boldsymbol{\beta},\boldsymbol{u}_i,\sigma^2 \sim \mathrm{N}(\boldsymbol{X}_i\boldsymbol{\beta}+\boldsymbol{X}_i\boldsymbol{u}_i,\sigma^2\boldsymbol{I}), \quad \left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}_i\end{bmatrix}\right| \Sigma_u \overset{\text{ind.}}{\sim} \mathrm{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \Sigma_u\end{bmatrix}\right)$$

$$\Sigma_u|\boldsymbol{A}_u \sim \text{Inverse Wishart}(1,\boldsymbol{A}_u^{-1}), \quad \boldsymbol{A}_u \sim \text{Inverse Wishart}\left(1,\frac{1}{A^2}\boldsymbol{I}\right)$$

$$\sigma^2|a \sim \text{Inverse-}\chi^2(1,1/a), \quad a \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Gaussian likelihood fragment

Gaussian penalization fragment

Iterated inverse Wishart fragment

Inverse Wishart prior fragment

The parameters for the posterior of $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}^{\top}, \boldsymbol{u}^{\top})^{\top}$ take the form - Lee and Wand (2016):

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}\boldsymbol{a}, \quad \text{Cov}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known

The posterior of $\sigma^2$ also depends on norms and determinants involving $\boldsymbol{A}$.

What is the issue with $\boldsymbol{A}$?

Consider our fictitious model of income against number of years of education for 21 towns.

# Multilevel Model

The parameters for the posterior of $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}^\intercal, \boldsymbol{u}^\intercal)^\intercal$ take the form - Lee and Wand (2016):

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}\boldsymbol{a}, \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known

The posterior of $\sigma^2$ also depends on norms and determinants involving $\boldsymbol{A}$.

What is the issue with $\boldsymbol{A}$?

Consider our fictitious model of income against number of years of education for 21 towns.

The parameters for the posterior of $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}^\top, \boldsymbol{u}^\top)^\top$ take the form - Lee and Wand (2016):

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}\boldsymbol{a}, \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known

The posterior of $\sigma^2$ also depends on norms and determinants involving $\boldsymbol{A}$.

What is the issue with $\boldsymbol{A}$?

Consider our fictitious model of income against number of years of education for 21 towns.

# Multilevel Model

The parameters for the posterior of $\boldsymbol{\nu} \equiv (\boldsymbol{\beta}^\top, \boldsymbol{u}^\top)^\top$ take the form - Lee and Wand (2016):

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}\boldsymbol{a}, \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \boldsymbol{A}^{-1}$$

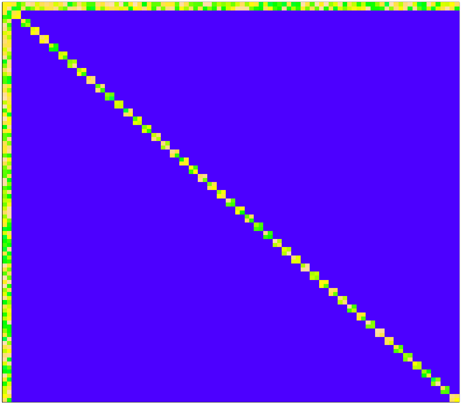
where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known

The posterior of $\sigma^2$ also depends on norms and determinants involving $\boldsymbol{A}$.

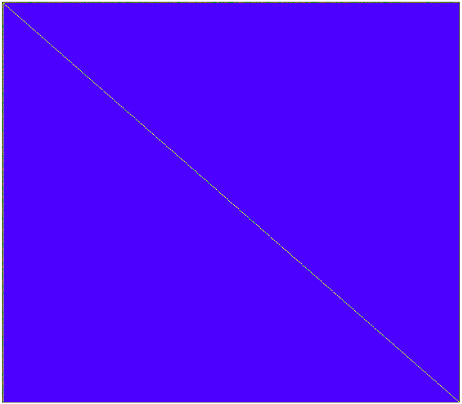
What is the issue with $\boldsymbol{A}$?

Consider our fictitious model of income against number of years of education for 21 towns.

# Multilevel Model (21 Towns)

$\boldsymbol{A} =$ 

# Multilevel Model (1000 Towns)

$A =$ 

# Multilevel Model

The solution comes from a simple observation:

$$\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1} \implies \boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1}\boldsymbol{a} \implies \boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known.

Matrices and vectors are partitioned as:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

# Multilevel Model

The solution comes from a simple observation:

$$\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1} \implies \boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1}\boldsymbol{a} \implies \boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known.

Matrices and vectors are partitioned as:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

The solution comes from a simple observation:

$$\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1} \implies \boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1}\boldsymbol{a} \implies \boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known.

Matrices and vectors are partitioned as:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

The solution comes from a simple observation:

$$\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1} \implies \boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1}\boldsymbol{a} \implies \boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known.

Matrices and vectors are partitioned as:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

# Multilevel Model

The solution comes from a simple observation:

$$\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1} \implies \boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1}\boldsymbol{a} \implies \boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known.

Matrices and vectors are partitioned as:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

The solution comes from a simple observation:

$$\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1} \implies \boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$$
$$\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1}\boldsymbol{a} \implies \boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known.

Matrices and vectors are partitioned as:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

# Multilevel Model

The solution comes from a simple observation:

$$\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1} \implies \boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{A}^{-1}\boldsymbol{a} \implies \boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$$

where $\boldsymbol{A}$ and $\boldsymbol{a}$ are known.

Matrices and vectors are partitioned as:

$$\boldsymbol{A} \equiv \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \quad \mathrm{Cov}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$\mathbb{E}_q(\boldsymbol{\nu}) \equiv \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} \quad \boldsymbol{a} \equiv \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^3 \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1,2,3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1,2,3$$

# Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$
\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}
$$

$$
= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}
$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1,2,3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1,2,3$$

# Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1,2,3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1,2,3$$

## Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1,2,3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1,2,3$$

# Multilevel Model ($\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \boldsymbol{I}$)

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1,2,3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1,2,3$$

$$\boldsymbol{A}\,\mathrm{Cov}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathrm{Cov}_q(\boldsymbol{\beta}) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2) & \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3) \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_1)^T & \mathrm{Cov}_q(\boldsymbol{u}_1) & \times & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_2)^T & \times & \mathrm{Cov}_q(\boldsymbol{u}_2) & \times \\ \mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_3)^T & \times & \times & \mathrm{Cov}_q(\boldsymbol{u}_3) \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{I} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\mathrm{Cov}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{I}$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1,2,3$$

$$\boldsymbol{A}_{12,i}^T\mathrm{Cov}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\mathrm{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = \boldsymbol{O}, \quad i = 1,2,3$$

# Multilevel Model ($\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$)

$$\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\,\mathbb{E}_q(\boldsymbol{\beta}) + \sum_{i=1}^{3} \boldsymbol{A}_{12,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{a}_1$$

$$\boldsymbol{A}_{12,i}^T\,\mathbb{E}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1,2,3$$

# Multilevel Model ($\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$)

$$\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\,\mathbb{E}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{a}_1$$

$$\boldsymbol{A}_{12,i}^T\,\mathbb{E}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$)

$$\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\,\mathbb{E}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{a}_1$$

$$\boldsymbol{A}_{12,i}^T\,\mathbb{E}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$)

$$\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\,\mathbb{E}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{a}_1$$

$$\boldsymbol{A}_{12,i}^T\,\mathbb{E}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$)

$$\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\,\mathbb{E}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{a}_1$$

$$\boldsymbol{A}_{12,i}^T\,\mathbb{E}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

# Multilevel Model ($\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \boldsymbol{a}$)

$$\boldsymbol{A}\,\mathbb{E}_q(\boldsymbol{\nu}) = \begin{bmatrix} \boldsymbol{A}_{11} & \boldsymbol{A}_{12,1} & \boldsymbol{A}_{12,2} & \boldsymbol{A}_{12,3} \\ \boldsymbol{A}_{12,1}^T & \boldsymbol{A}_{22,1} & \boldsymbol{O} & \boldsymbol{O} \\ \boldsymbol{A}_{12,2}^T & \boldsymbol{O} & \boldsymbol{A}_{22,2} & \boldsymbol{O} \\ \boldsymbol{A}_{12,3}^T & \boldsymbol{O} & \boldsymbol{O} & \boldsymbol{A}_{22,3} \end{bmatrix} \begin{bmatrix} \mathbb{E}_q(\boldsymbol{\beta}) \\ \mathbb{E}_q(\boldsymbol{u}_1) \\ \mathbb{E}_q(\boldsymbol{u}_2) \\ \mathbb{E}_q(\boldsymbol{u}_3) \end{bmatrix} = \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_{2,1} \\ \boldsymbol{a}_{2,2} \\ \boldsymbol{a}_{2,3} \end{bmatrix}$$

$$\boldsymbol{A}_{11}\,\mathbb{E}_q(\boldsymbol{\beta}) + \textstyle\sum_{i=1}^{3} \boldsymbol{A}_{12,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{a}_1$$

$$\boldsymbol{A}_{12,i}^T\,\mathbb{E}_q(\boldsymbol{\beta}) + \boldsymbol{A}_{22,i}\,\mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{I}, \quad i = 1, 2, 3$$

## Theorem: Nolan and Wand (2020)

For the two-level mean field variational Bayes model, the solutions for the required sub-blocks of $\text{Cov}_q(\boldsymbol{\nu})$ are

$$\text{Cov}_q(\boldsymbol{\beta}) = \left(\boldsymbol{A}_{11} - \sum_{i=1}^{m} \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^{T}\right)^{-1}$$

$$\text{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = -(\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^{T}\text{Cov}_q(\boldsymbol{\beta}))^{T}, \quad \text{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{I} - \boldsymbol{A}_{12.i}^{T}\text{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i)), \quad 1 \le i \le m.$$

The determinant of $\text{Cov}_q(\boldsymbol{\nu})$ is

$$|\text{Cov}_q(\boldsymbol{\nu})| = \left|\text{Cov}_q(\boldsymbol{\beta})\right| \prod_{i=1}^{m} \left|\text{Cov}_q(\boldsymbol{u}_i)\right|.$$

The solutions for the sub-vectors of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ are

$$E_q(\boldsymbol{\beta}) = \text{Cov}_q(\boldsymbol{\beta})\left(\boldsymbol{a}_1 - \sum_{i=1}^{m} \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{a}_{2,i}\right) \quad \text{and} \quad E_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{a}_{2,i} - \boldsymbol{A}_{12,i}^{T}\, E_q(\boldsymbol{\beta})).$$

## Theorem: Nolan and Wand (2020)

For the two-level mean field variational Bayes model, the solutions for the required sub-blocks of $\text{Cov}_q(\boldsymbol{\nu})$ are

$$\text{Cov}_q(\boldsymbol{\beta}) = \left(\boldsymbol{A}_{11} - \sum_{i=1}^{m} \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^{T}\right)^{-1}$$

$$\text{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = -(\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^{T}\text{Cov}_q(\boldsymbol{\beta}))^{T}, \quad \text{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{I} - \boldsymbol{A}_{12.i}^{T}\text{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i)), \quad 1 \le i \le m.$$

The determinant of $\text{Cov}_q(\boldsymbol{\nu})$ is

$$|\text{Cov}_q(\boldsymbol{\nu})| = \left|\text{Cov}_q(\boldsymbol{\beta})\right| \prod_{i=1}^{m} \left|\text{Cov}_q(\boldsymbol{u}_i)\right|.$$

The solutions for the sub-vectors of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ are

$$\mathbb{E}_q(\boldsymbol{\beta}) = \text{Cov}_q(\boldsymbol{\beta})\left(\boldsymbol{a}_1 - \sum_{i=1}^{m} \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{a}_{2,i}\right) \quad \text{and} \quad \mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{a}_{2,i} - \boldsymbol{A}_{12,i}^{T}\,\mathbb{E}_q(\boldsymbol{\beta})).$$

## Theorem: Nolan and Wand (2020)

For the two-level mean field variational Bayes model, the solutions for the required sub-blocks of $\text{Cov}_q(\boldsymbol{\nu})$ are

$$\text{Cov}_q(\boldsymbol{\beta}) = \left(\boldsymbol{A}_{11} - \sum_{i=1}^{m} \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^{T}\right)^{-1}$$

$$\text{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i) = -(\boldsymbol{A}_{22,i}^{-1}\boldsymbol{A}_{12,i}^{T}\text{Cov}_q(\boldsymbol{\beta}))^{T}, \quad \text{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{I} - \boldsymbol{A}_{12.i}^{T}\text{Cov}_q(\boldsymbol{\beta},\boldsymbol{u}_i)), \quad 1 \le i \le m.$$

The determinant of $\text{Cov}_q(\boldsymbol{\nu})$ is

$$|\text{Cov}_q(\boldsymbol{\nu})| = \left|\text{Cov}_q(\boldsymbol{\beta})\right| \prod_{i=1}^{m} \left|\text{Cov}_q(\boldsymbol{u}_i)\right|.$$

The solutions for the sub-vectors of $\boldsymbol{\mu}_{q(\boldsymbol{\beta},\boldsymbol{u})}$ are

$$\mathbb{E}_q(\boldsymbol{\beta}) = \text{Cov}_q(\boldsymbol{\beta})\left(\boldsymbol{a}_1 - \sum_{i=1}^{m} \boldsymbol{A}_{12,i}\boldsymbol{A}_{22,i}^{-1}\boldsymbol{a}_{2,i}\right) \quad \text{and} \quad \mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{a}_{2,i} - \boldsymbol{A}_{12,i}^{T}\,\mathbb{E}_q(\boldsymbol{\beta})).$$

## Theorem: Nolan and Wand (2020)

For the two-level mean field variational Bayes model, the solutions for the required sub-blocks of $\text{Cov}_q(\boldsymbol{\nu})$ are

$$\text{Cov}_q(\boldsymbol{\beta}) = \left( \boldsymbol{A}_{11} - \sum_{i=1}^{m} \boldsymbol{A}_{12,i} \boldsymbol{A}_{22,i}^{-1} \boldsymbol{A}_{12,i}^{T} \right)^{-1}$$

$$\text{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_i) = -(\boldsymbol{A}_{22,i}^{-1} \boldsymbol{A}_{12,i}^{T} \text{Cov}_q(\boldsymbol{\beta}))^T, \quad \text{Cov}_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{I} - \boldsymbol{A}_{12.i}^{T} \text{Cov}_q(\boldsymbol{\beta}, \boldsymbol{u}_i)), \quad 1 \le i \le m.$$

The determinant of $\text{Cov}_q(\boldsymbol{\nu})$ is

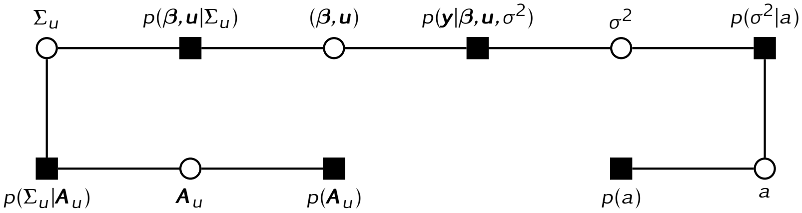
$$|\text{Cov}_q(\boldsymbol{\nu})| = \left|\text{Cov}_q(\boldsymbol{\beta})\right| \prod_{i=1}^{m} \left|\text{Cov}_q(\boldsymbol{u}_i)\right|.$$

The solutions for the sub-vectors of $\boldsymbol{\mu}_{q(\boldsymbol{\beta}, \boldsymbol{u})}$ are

$$\mathbb{E}_q(\boldsymbol{\beta}) = \text{Cov}_q(\boldsymbol{\beta}) \left( \boldsymbol{a}_1 - \sum_{i=1}^{m} \boldsymbol{A}_{12,i} \boldsymbol{A}_{22,i}^{-1} \boldsymbol{a}_{2,i} \right) \quad \text{and} \quad \mathbb{E}_q(\boldsymbol{u}_i) = \boldsymbol{A}_{22,i}^{-1}(\boldsymbol{a}_{2,i} - \boldsymbol{A}_{12,i}^{T} \, \mathbb{E}_q(\boldsymbol{\beta})).$$

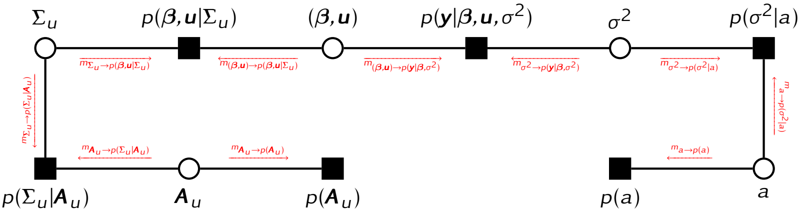
## VMP for the Bayesian multilevel data model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u|\boldsymbol{y})\right\}$ converges.



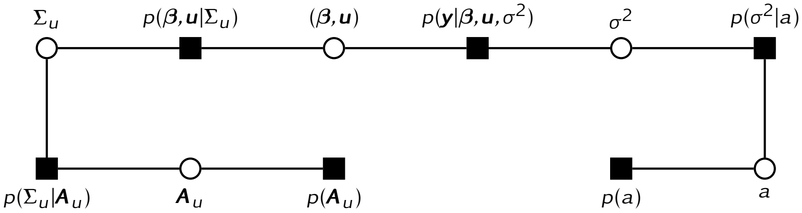
## VMP for the Bayesian multilevel data model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u|\boldsymbol{y})\right\}$ converges.



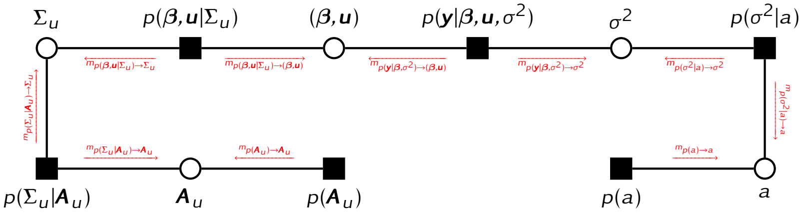
## VMP for the Bayesian multilevel data model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u|\boldsymbol{y})\right\}$ converges.



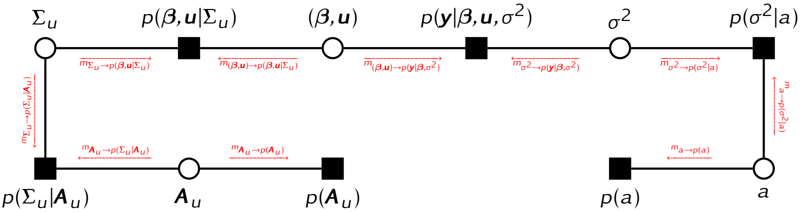
## VMP for the Bayesian multilevel data model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u|\boldsymbol{y})\right\}$ converges.



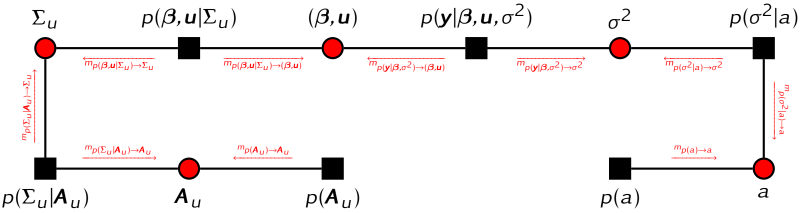
## VMP for the Bayesian multilevel data model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u|\boldsymbol{y})\right\}$ converges.



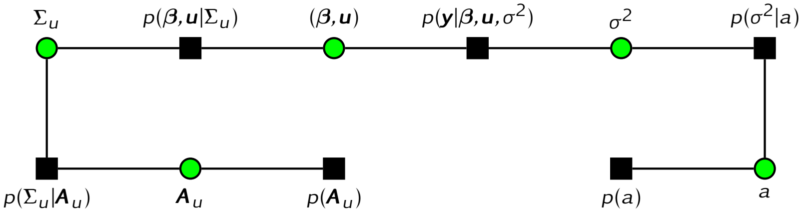
## VMP for the Bayesian multilevel data model

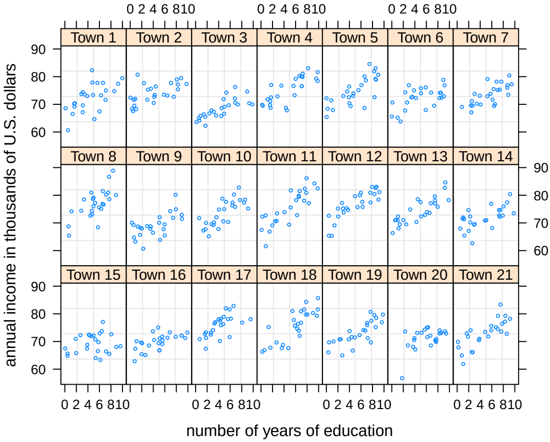
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u|\boldsymbol{y})\right\}$ converges.



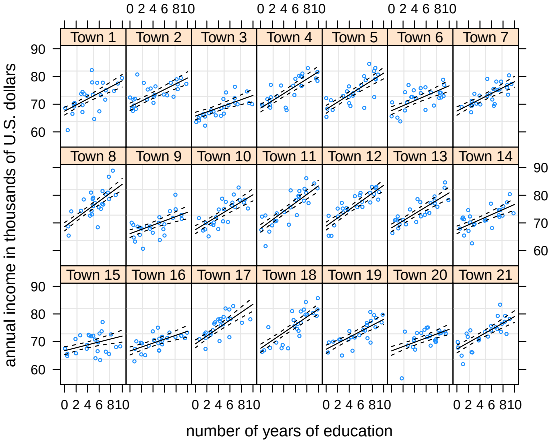
## VMP for the Bayesian multilevel data model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma^2,\Sigma_u,a,\boldsymbol{A}_u|\boldsymbol{y})\right\}$ converges.



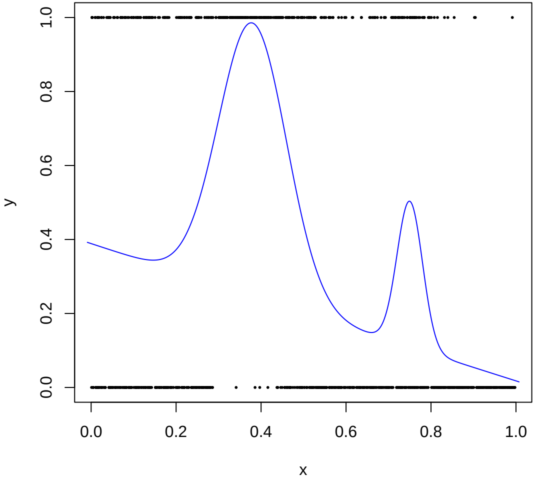

Town 1
Town 2
Town 3
Town 4
Town 5
Town 6
Town 7
Town 8
Town 9
Town 10
Town 11
Town 12
Town 13
Town 14
Town 15
Town 16
Town 17
Town 18
Town 19
Town 20
Town 21
0 2 4 6 810
60
70
80
90
annual income in thousands of U.S. dollars
number of years of education

# Multilevel Data



Part III

# Non-conjugate Models


y
x
0.0
0.2
0.4
0.6
0.8
1.0

Consider the Bayesian logistic regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left\{[1 + \exp\{-\boldsymbol{X}\boldsymbol{\beta}\}]^{-1}\right\}$$
$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$$

The factor graph is




Gaussian Prior Fragment

Logistic Likelihood Fragment (Nolan and Wand, 2017)

The message takes the form:

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{1}^T\mathbb{E}_q \log\left[\boldsymbol{1} + \exp(\boldsymbol{X}\boldsymbol{\beta})\right]\right\}$$

The messages that are typically passed to $\boldsymbol{\beta}$ from other fragments are Gaussian density messages.

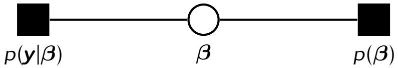
# Bayesian Logistic Regression Model

Consider the Bayesian logistic regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left\{[1 + \exp\{-\boldsymbol{X\beta}\}]^{-1}\right\}$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$$

The factor graph is




Gaussian Prior Fragment

Logistic Likelihood Fragment (Nolan and Wand, 2017)

The message takes the form:

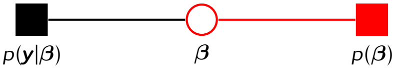
$$m_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{\boldsymbol{y}^T\boldsymbol{X\beta} - \boldsymbol{1}^T\mathbb{E}_q \log\left[\boldsymbol{1} + \exp(\boldsymbol{X\beta})\right]\right\}$$

The messages that are typically passed to $\boldsymbol{\beta}$ from other fragments are Gaussian density messages.

Consider the Bayesian logistic regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left\{[1 + \exp\{-\boldsymbol{X}\boldsymbol{\beta}\}]^{-1}\right\}$$
$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$$

The factor graph is




Gaussian Prior Fragment

Logistic Likelihood Fragment (Nolan and Wand, 2017)

The message takes the form:

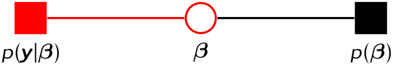
$$m_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{1}^T \mathbb{E}_q \log[\boldsymbol{1} + \exp(\boldsymbol{X}\boldsymbol{\beta})]\right\}$$

The messages that are typically passed to $\boldsymbol{\beta}$ from other fragments are Gaussian density messages.

Consider the Bayesian logistic regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left\{[1 + \exp\{-\boldsymbol{X}\boldsymbol{\beta}\}]^{-1}\right\}$$
$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$$

The factor graph is




Gaussian Prior Fragment

Logistic Likelihood Fragment (Nolan and Wand, 2017)

The message takes the form:

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{1}^T\mathbb{E}_q \log\left[\boldsymbol{1} + \exp(\boldsymbol{X}\boldsymbol{\beta})\right]\right\}$$

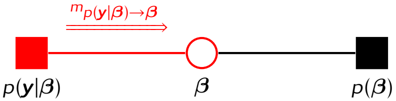
The messages that are typically passed to $\boldsymbol{\beta}$ from other fragments are Gaussian density messages.

# Bayesian Logistic Regression Model

Consider the Bayesian logistic regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left\{[1 + \exp\{-\boldsymbol{X}\boldsymbol{\beta}\}]^{-1}\right\}$$
$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$$

The factor graph is




Gaussian Prior Fragment

Logistic Likelihood Fragment (Nolan and Wand, 2017)

The message takes the form:

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{1}^T\mathbb{E}_q \log\left[\boldsymbol{1} + \exp(\boldsymbol{X}\boldsymbol{\beta})\right]\right\}$$

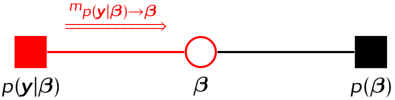
The messages that are typically passed to $\beta$ from other fragments are Gaussian density messages.

# Bayesian Logistic Regression Model

Consider the Bayesian logistic regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left\{[1 + \exp\{-\boldsymbol{X}\boldsymbol{\beta}\}]^{-1}\right\}$$
$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_{\beta}^2 \boldsymbol{I})$$

The factor graph is




Gaussian Prior Fragment

Logistic Likelihood Fragment (Nolan and Wand, 2017)

The message takes the form:

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{1}^T \mathbb{E}_q \log[\boldsymbol{1} + \exp(\boldsymbol{X}\boldsymbol{\beta})]\right\}$$

The messages that are typically passed to $\boldsymbol{\beta}$ from other fragments are Gaussian density messages.

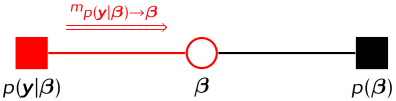
## Bayesian Logistic Regression Model

Consider the Bayesian logistic regression model:

$$\boldsymbol{y} \mid \boldsymbol{\beta} \sim \text{Bernoulli}\left\{\left[1+\exp\{-\boldsymbol{X}\boldsymbol{\beta}\}\right]^{-1}\right\}$$

$$\boldsymbol{\beta} \sim N(\boldsymbol{0}, \sigma_\beta^2 \boldsymbol{I})$$

The factor graph is



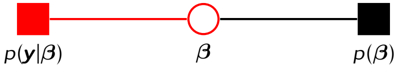

Gaussian Prior Fragment

Logistic Likelihood Fragment (Nolan and Wand, 2017)

The message takes the form:

$$m_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}(\boldsymbol{\beta}) \longleftarrow \exp\left\{\boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{1}^T\mathbb{E}_q \log\left[\boldsymbol{1}+\exp(\boldsymbol{X}\boldsymbol{\beta})\right]\right\}$$

The messages that are typically passed to $\boldsymbol{\beta}$ from other fragments are Gaussian density messages.



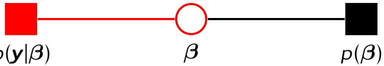

A fixed point iterative scheme for $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}$ (Wand, 2014):

$$\boldsymbol{\mu} = \boldsymbol{X}\,\mathbb{E}_q(\boldsymbol{\beta}), \quad \boldsymbol{\sigma}^2 = \text{diagonal}\left\{\boldsymbol{X}\text{Cov}_q(\boldsymbol{\beta})\boldsymbol{X}^T\right\}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{X}^T\left\{\boldsymbol{y} - \mathcal{B}_0(\boldsymbol{\mu},\boldsymbol{\sigma}^2) + \mathcal{B}_1(\boldsymbol{\mu},\boldsymbol{\sigma}^2)\odot\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right\} \\ -\frac{1}{2}\,\text{vec}\left\{\boldsymbol{X}^T\,\text{diag}(\omega_2)\boldsymbol{X}\right\} \end{bmatrix}$$

where

$$\mathcal{B}_r(\mu,\sigma^2) \equiv \int_{-\infty}^{\infty} x^r \frac{d}{d\mu}\log\{1+\exp(\mu+\sigma x)\}\,\phi(x)dx, \quad \text{for } r = 0,1$$




A fixed point iterative scheme for $\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}$ (Wand, 2014):

$$\boldsymbol{\mu} = \boldsymbol{X}\,\mathbb{E}_q(\boldsymbol{\beta}), \quad \boldsymbol{\sigma}^2 = \text{diagonal}\left\{\boldsymbol{X}\,\text{Cov}_q(\boldsymbol{\beta})\boldsymbol{X}^T\right\}$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{X}^T\left\{\boldsymbol{y} - \mathcal{B}_0(\boldsymbol{\mu},\boldsymbol{\sigma}^2) + \mathcal{B}_1(\boldsymbol{\mu},\boldsymbol{\sigma}^2)\odot\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right\} \\ -\frac{1}{2}\,\text{vec}\left\{\boldsymbol{X}^T\,\text{diag}(\omega_2)\boldsymbol{X}\right\} \end{bmatrix}$$

where

$$\mathcal{B}_r(\mu,\sigma^2) \equiv \int_{-\infty}^{\infty} x^r \frac{d}{d\mu}\log\{1+\exp(\mu+\sigma x)\}\,\phi(x)dx, \quad \text{for } r = 0, 1$$

# Normal Mixture Approximation

A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x)dx, \quad \text{for } r = 0,1$$

$$\Downarrow \text{ Monahan and Stefanski (1992)}$$

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right)\phi(x)dx, \quad \text{for } r = 0,1$$

# Normal Mixture Approximation

A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x)dx, \quad \text{for } r=0,1$$

$$\Downarrow \text{ Monahan and Stefanski (1992)}$$

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right)\phi(x)dx, \quad \text{for } r=0,1$$

# Normal Mixture Approximation

A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x)dx, \quad \text{for } r = 0, 1$$

$$\Downarrow \text{ Monahan and Stefanski (1992)}$$

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right) \phi(x)dx, \quad \text{for } r = 0, 1$$

# Normal Mixture Approximation

A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x)dx, \quad \text{for } r = 0, 1$$

$$\Downarrow \text{ Monahan and Stefanski (1992)}$$

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right) \phi(x)dx, \quad \text{for } r = 0, 1$$

# Normal Mixture Approximation

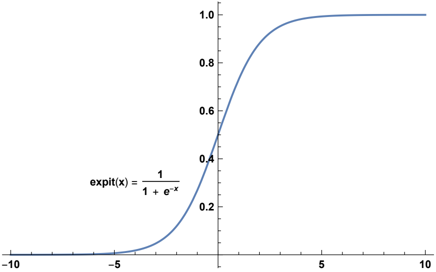
A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x) dx, \quad \text{for } r = 0,1$$

$$\Downarrow \text{ Monahan and Stefanski (1992)}$$

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right) \phi(x) dx, \quad \text{for } r = 0,1$$



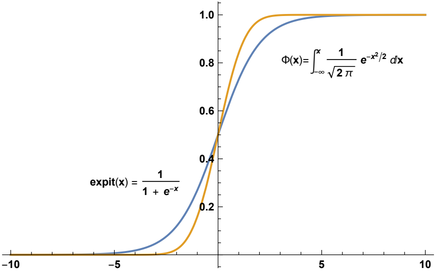
A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1}\phi(x)dx, \quad \text{for } r=0,1$$

$\Downarrow$ Monahan and Stefanski (1992)

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right)\phi(x)dx, \quad \text{for } r=0,1$$



# Normal Mixture Approximation

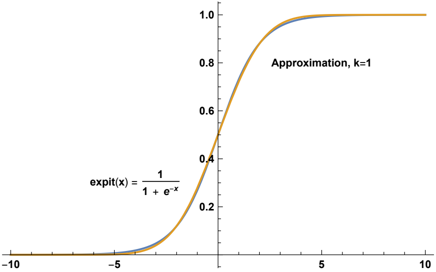
A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x)dx, \quad \text{for } r=0,1$$

$\Downarrow$ Monahan and Stefanski (1992)

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right)\phi(x)dx, \quad \text{for } r=0,1$$




| $p$ | $s$ |
|---|---|
| 1.00000 | 0.58763 |

# Normal Mixture Approximation

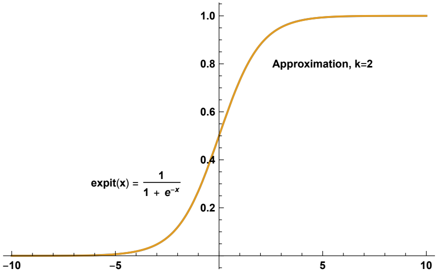
A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x)dx, \quad \text{for } r=0,1$$

$\Downarrow$ Monahan and Stefanski (1992)

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right)\phi(x)dx, \quad \text{for } r=0,1$$




| $p$ | $s$ |
|---|---|
| 0.56442 | 0.76862 |
| 0.43557 | 0.43525 |

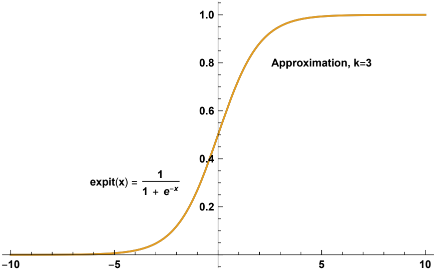
A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1} \phi(x)dx, \quad \text{for } r=0,1$$

$$\Downarrow \text{ Monahan and Stefanski (1992)}$$

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right)\phi(x)dx, \quad \text{for } r=0,1$$




| $p$ | $s$ |
|---|---|
| 0.25220 | 0.90793 |
| 0.58522 | 0.57778 |
| 0.16257 | 0.36403 |

# Normal Mixture Approximation

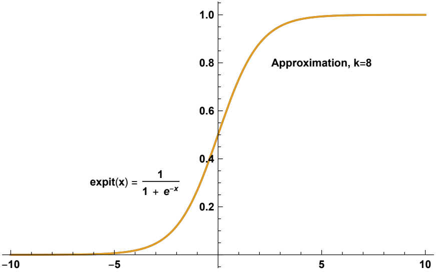
A fixed point iterative scheme for $q(\boldsymbol{\beta})$ - Wand (2014):

However, we come across the following integrals:

$$\mathcal{B}_r(\mu,\sigma) \equiv \int_{-\infty}^{\infty} x^r \{1+\exp(\mu+\sigma x)\}^{-1}\phi(x)dx, \quad \text{for } r=0,1$$

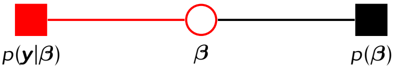
$\Downarrow$ Monahan and Stefanski (1992)

$$\mathcal{B}_r(\mu,\sigma) \approx \int_{-\infty}^{\infty} x^r \sum_{i=1}^{k} p_i \Phi\left(s_i \frac{x-\mu}{\sigma}\right)\phi(x)dx, \quad \text{for } r=0,1$$




| p | s |
|---|---|
| 0.00324 | 1.36534 |
| 0.05151 | 1.05952 |
| 0.19507 | 0.83079 |
| 0.31556 | 0.65073 |
| 0.27414 | 0.50813 |
| 0.13107 | 0.39631 |
| 0.02791 | 0.30890 |
| 0.00144 | 0.23821 |

# Logistic Likelihood Fragment




$$\boldsymbol{y}|\boldsymbol{\beta} \sim \text{Bernoulli}\left[\{\mathbf{1} + \exp(-\boldsymbol{X}\boldsymbol{\beta})\}^{-1}\right]$$

Inputs:

$$\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})}$$

Updates:

$$\text{Cov}_q\{(\boldsymbol{\beta})\} = -\tfrac{1}{2}\{\text{vec}^{-1}(\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})})_2\}^{-1}, \quad \mathbb{E}_q\{(\boldsymbol{\beta})\} = \text{Cov}_q\{(\boldsymbol{\beta})\}(\boldsymbol{\eta}_{q^*(\boldsymbol{\beta})})_1,$$

$$\boldsymbol{\mu} = \boldsymbol{X}\,\mathbb{E}_q(\boldsymbol{\beta}), \quad \boldsymbol{\sigma}^2 = \text{diagonal}\left\{\boldsymbol{X}\text{Cov}_q(\boldsymbol{\beta})\boldsymbol{X}^T\right\},$$

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}} = \begin{bmatrix} \boldsymbol{X}^T\left\{\boldsymbol{y} - \mathcal{B}_0(\boldsymbol{\mu},\boldsymbol{\sigma}^2) + \mathcal{B}_1(\boldsymbol{\mu},\boldsymbol{\sigma}^2)\odot\frac{\boldsymbol{\mu}}{\boldsymbol{\sigma}}\right\} \\ -\frac{1}{2}\,\text{vec}\left\{\boldsymbol{X}^T\,\text{diag}(\omega_2)\boldsymbol{X}\right\} \end{bmatrix}$$

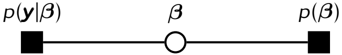
Outputs:

$$\boldsymbol{\eta}_{p(\boldsymbol{y}|\boldsymbol{\beta})\to\boldsymbol{\beta}}$$

# Bayesian Logistic Regression

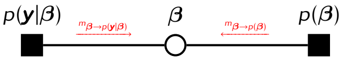
## VMP for the Bayesian logistic regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta})||p(\boldsymbol{\beta}|\boldsymbol{y})\right\}$ converges.



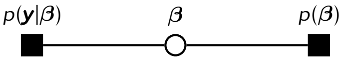
## VMP for the Bayesian logistic regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta}) \| p(\boldsymbol{\beta}|\boldsymbol{y})\right\}$ converges.



## VMP for the Bayesian logistic regression model

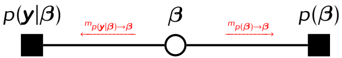
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\{q(\boldsymbol{\beta})||p(\boldsymbol{\beta}|\boldsymbol{y})\}$ converges.



# Bayesian Logistic Regression

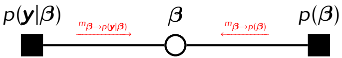
## VMP for the Bayesian logistic regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta})||p(\boldsymbol{\beta}|\boldsymbol{y})\right\}$ converges.



## VMP for the Bayesian logistic regression model

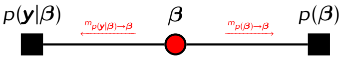
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\{q(\boldsymbol{\beta})\|p(\boldsymbol{\beta}|\boldsymbol{y})\}$ converges.



# Bayesian Logistic Regression

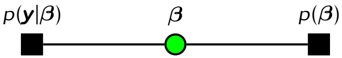
## VMP for the Bayesian logistic regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\{q(\boldsymbol{\beta})||p(\boldsymbol{\beta}|\boldsymbol{y})\}$ converges.



## VMP for the Bayesian logistic regression model

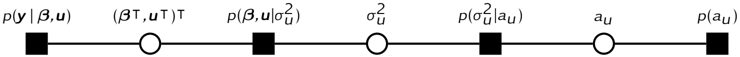
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\{q(\boldsymbol{\beta})\|p(\boldsymbol{\beta}|\boldsymbol{y})\}$ converges.



# Bayesian Logistic Semiparametric Regression

$$\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u} \sim \text{Bernoulli}\Big(\big[1+\exp\{-(\boldsymbol{X\beta}+\boldsymbol{Xu})\}\big]^{-1}\Big), \quad \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\Bigg|\ \sigma_u^2 \sim \text{N}\left(\begin{bmatrix}\boldsymbol{0}\\ \boldsymbol{0}\end{bmatrix}, \begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\,\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Logistic likelihood fragment

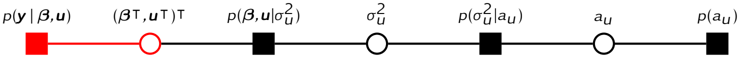
Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

Inverse Wishart prior fragment

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}\left\{\left[1 + \exp\{-(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{u})\}\right]^{-1}\right\}, \quad \begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \Bigg| \; \sigma_u^2 \sim \mathrm{N}\left( \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \sigma_\beta^2 \boldsymbol{I}_2 & \boldsymbol{O}^T \\ \boldsymbol{O} & \sigma_u^2 \boldsymbol{I}_m \end{bmatrix} \right)$$

$$\sigma_u^2 | a_u \sim \text{Inverse-}\chi^2(1, 1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1, 1/A^2)$$




Logistic likelihood fragment

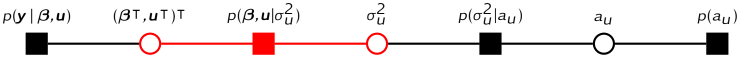
Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

Inverse Wishart prior fragment

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}\left\{\left[1+\exp\{-(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{u})\}\right]^{-1}\right\}, \quad \begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix} \Bigg| \; \sigma_u^2 \sim \text{N}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\sigma_\beta^2 \boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\,\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2 | a_u \sim \text{Inverse-}\chi^2(1, 1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1, 1/A^2)$$




Logistic likelihood fragment

Gaussian penalization fragment (Wand, 2017)

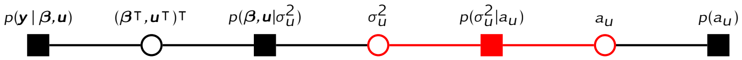
Iterated inverse Wishart fragment

Inverse Wishart prior fragment

# Bayesian Logistic Semiparametric Regression

$$\boldsymbol{y}\,|\,\boldsymbol{\beta},\boldsymbol{u} \sim \text{Bernoulli}\left(\left[1+\exp\{-(\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{X}\boldsymbol{u})\}\right]^{-1}\right),\quad \left.\begin{bmatrix}\boldsymbol{\beta}\\ \boldsymbol{u}\end{bmatrix}\right|\,\sigma_u^2 \sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix},\begin{bmatrix}\sigma_\beta^2\boldsymbol{I}_2 & \boldsymbol{O}^T\\ \boldsymbol{O} & \sigma_u^2\,\boldsymbol{I}_m\end{bmatrix}\right)$$

$$\sigma_u^2|a_u \sim \text{Inverse-}\chi^2(1,1/a_u),\quad a_u \sim \text{Inverse-}\chi^2(1,1/A^2)$$




Logistic likelihood fragment

Gaussian penalization fragment (Wand, 2017)

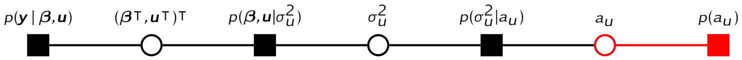
Iterated inverse Wishart fragment

Inverse Wishart prior fragment

# Bayesian Logistic Semiparametric Regression

$$\boldsymbol{y} \mid \boldsymbol{\beta}, \boldsymbol{u} \sim \text{Bernoulli}\left(\left[1 + \exp\{-(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}\boldsymbol{u})\}\right]^{-1}\right), \quad \left.\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{u} \end{bmatrix} \right| \sigma_u^2 \sim \text{N}\left(\begin{bmatrix} \boldsymbol{0} \\ \boldsymbol{0} \end{bmatrix}, \begin{bmatrix} \sigma_\beta^2 \boldsymbol{I}_2 & \boldsymbol{O}^T \\ \boldsymbol{O} & \sigma_u^2 \boldsymbol{I}_m \end{bmatrix}\right)$$

$$\sigma_u^2 | a_u \sim \text{Inverse-}\chi^2(1, 1/a_u), \quad a_u \sim \text{Inverse-}\chi^2(1, 1/A^2)$$




Logistic likelihood fragment

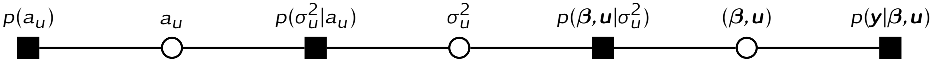
Gaussian penalization fragment (Wand, 2017)

Iterated inverse Wishart fragment

Inverse Wishart prior fragment

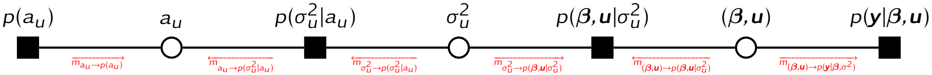
## VMP for the Bayesian logistic semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u|\boldsymbol{y})\right\}$ converges.



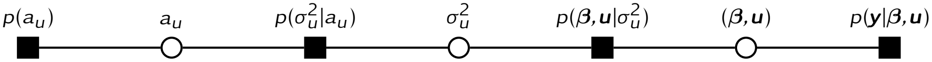
## VMP for the Bayesian logistic semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u|\boldsymbol{y})\right\}$ converges.



## VMP for the Bayesian logistic semiparametric regression model

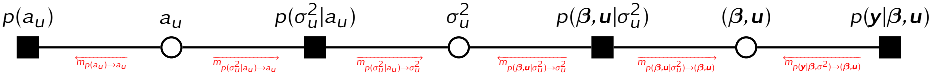
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u|\boldsymbol{y})\right\}$ converges.



# Bayesian Logistic Semiparametric Regression

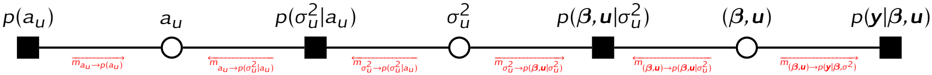
## VMP for the Bayesian logistic semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\text{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u)||p(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u|\boldsymbol{y})\right\}$ converges.



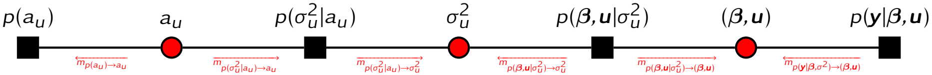
## VMP for the Bayesian logistic semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   - (i) Update all messages from factors to stochastic nodes
   - (ii) Update all messages from stochastic nodes to factors
   - (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u|\boldsymbol{y})\right\}$ converges.



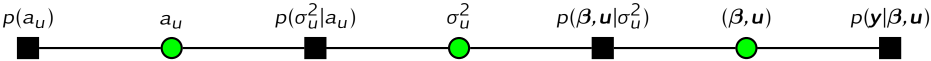
## VMP for the Bayesian logistic semiparametric regression model

1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u|\boldsymbol{y})\right\}$ converges.

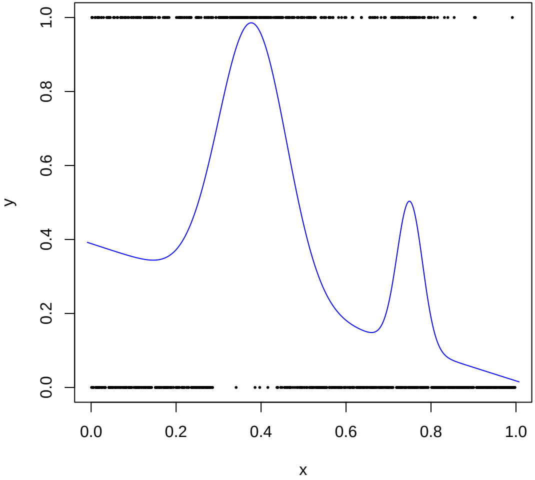


## VMP for the Bayesian logistic semiparametric regression model

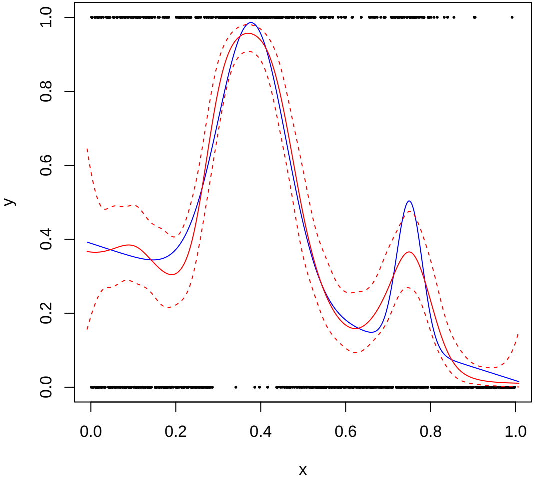
1. Initialise all messages from stochastic nodes to factors
2. Cycle:
   (i) Update all messages from factors to stochastic nodes
   (ii) Update all messages from stochastic nodes to factors
   (iii) Update all optimal posterior density functions
3. Stop: $D_{\mathrm{KL}}\left\{q(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u)\|p(\boldsymbol{\beta},\boldsymbol{u},\sigma_u^2,a_u|\boldsymbol{y})\right\}$ converges.



# Binary Response Data



# Binary Response Data



# Bibliography


Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer, New York.

Chamberlain, S., Anderson, B., Salmon, M., Erickson, A., Potter, N., Stachelek, J., Simmons, A., Ram, K., Edmund, H., and rOpenSci (2021). 'NOAA' weather data from r. R package version 1.3.0.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1:515–534.

Goldsmith, J., Zippunnikov, V., and Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics*, 71:344–353.

Happ, C. and Greven, S. (2018). Multivariate functional principal component analysis for data observed on different (dimensional) domains. *Journal of the American Statistical Association*, 113:649–659.

Huang, A. and Wand, M. P. (2013). Simple marginally noninformative prior distributions for covariance matrices. *Bayesian Analysis*, 8:439–452.

Lee, C. Y. Y. and Wand, M. P. (2016). Streamlined mean field variational Bayes for longitudinal and multilevel data analysis. *Biometrical Journal*, 58:868–895.

Maestrini, L. and Wand, M. P. (2020). The Inverse G-Wishart distribution and variational message passing. *arXiv e-prints*, page arXiv:2005.09876.

Menictas, M. and Wand, M. P. (2013). Variational inference for marginal longitudinal semiparametric regression. *Stat*, 2:61–71.

Minka, T. (2005). Divergence measures and message passing. Technical report, Microsoft Research Ltd., Cambridge, UK.

Monahan, J. F. and Stefanski, L. A. (1992). Normal scale mixture approximations to $F^*(z)$ and computation of the logistic-normal integral. In *Handbook of the Logistic Distribution*, pages 529–540. CRC Press.

Nolan, T. H., Goldsmith, J., and Ruppert, D. (2021). Bayesian functional principal components analysis via variational message passing. *arXiv e-prints*, page arXiv:2104.00645.

Nolan, T. H. and Wand, M. P. (2017). Accurate logistic variational message passing: Algebraic and numerical details. *Stat*, 6:102–112.

Nolan, T. H. and Wand, M. P. (2020). Streamlined solutions to multilevel sparse matrix problems. *ANZIAM Journal*, 62:18–41.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*. Springer, New York.

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*, 15:1351–1369.

Wand, M. P. (2017). Fast approximate inference for arbitrarily large semiparametric regression models via message passing (with discussion). *Journal of the American Statistical Association*, 112:137–168.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Australian & New Zealand Journal of Statistics*, 50:179–198.

Wang, J. L., Chiou, J. M., and Müller, H. G. (2016). Functional data analysis. *Annual Review of Statistics and Its Applications*, 3:257–295.