# Meta-analysis and synthesis

Bayesian Modeling for Socio-Environmental Data

N. Thompson Hobbs

August 9, 2016

# Why this course?

SESYNC is dedicated to fostering synthetic, actionable science related to the structure, functioning, and sustainability of socio-environmental systems.

SESYNC

Synthesis: gaining new insight by combining information from multiple sources in novel ways.

- Multiple sources of data can enter through likelihoods.
- Results of previous studies can enter through priors.
- Hierarchical models allow "borrowing strength."

# Multiple sources of data

$$[\boldsymbol{\theta} \mid \mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3] \propto \prod_{i=1}^{n_1} [y_{1,i} \mid \boldsymbol{\theta}] \prod_{i=1}^{n_2} [y_{2,i} \mid \boldsymbol{\theta}] \prod_{i=1}^{n_3} [y_{3,i} \mid \boldsymbol{\theta}][\boldsymbol{\theta}] \qquad (1)$$

- Can have different $\theta' s$ in the different likelihoods.
- Likelihoods can have different distributions.

## Example

- Deterministic model of process: $g\left(\boldsymbol{\theta}, \mathbf{z}_{t-1}\right) = \mathbf{A}\mathbf{z}_{t-1}$
- Stochastic model of process:

$$\log\left(\mathbf{z}_t\right) \sim \text{multivariate normal}\left(\log\left(g\left(\boldsymbol{\theta}, \mathbf{z}_{t-1}\right)\right), \sigma_p^2 \mathbf{I}\right)$$

- Posterior and joint distribution:

$$\begin{aligned}\left[\boldsymbol{\theta}, \mathbf{z}, \sigma_p^2 | \mathbf{y}_1, \mathbf{Y}_2\right] \quad \propto \quad & \underbrace{\prod_{t=2}^{T} \text{Poisson}\left(y_{1,t} \,\middle|\, \sum_{i=1}^{m} z_{it}\right)}_{\text{likelihood for census data}} \times \\ & \times \underbrace{\text{multinomial}\left(\mathbf{y}_{2,t} \,\middle|\, \sum y_{2,it}, \frac{\mathbf{z}_t}{\sum_{i=1}^{m} z_{it}}\right)}_{\text{likelihood for classification data}} \\ & \times \underbrace{\text{multivariate normal}\left(\log(\mathbf{z}_t) | \log\left(g\left(\boldsymbol{\theta}, \mathbf{z}_{t-1}\right)\right), \sigma_p^2 \mathbf{I}\right)}_{\text{process model}} \\ & \times [\boldsymbol{\theta}] [\mathbf{z}_1] \left[\sigma_p^2\right].\end{aligned}$$

# Combining priors from multiple studies I

As derived quantity:

$$
\begin{aligned}
[\boldsymbol{\theta}_{meta}] &= \sum_{i=1}^{n} w_i[\boldsymbol{\theta}] \\
\sum_{i=1}^{n} w_i &= 1
\end{aligned}
$$

The $w_i$ are chosen by the investigator based on quality of study, sample size, etc.

# Combining priors from multiple studies II

As hyper-parameter:

$$[\mathbf{z}, \theta_{\text{meta}}, \varsigma^2 \mid \boldsymbol{\theta}] \quad \propto \quad \prod_{j=1}^{J} [\theta_i | z_i, \boldsymbol{\sigma}^2_{\boldsymbol{\theta}_i}][z_i | \theta_{\text{meta}}, \varsigma^2][\theta_{\text{meta}}][\varsigma^2] \quad (2)$$

## Example: multiple priors on an intercept $\beta_0$

Step 1: Do MCMC to approximate mean $\tilde{\beta}_0$ and variance of the mean $\sigma^2_{\tilde{\beta}_0}$.

$$
\begin{aligned}
\beta_{0,1} &\sim \text{normal}(3.2,\ .35) \\
\beta_{0,2} &\sim \text{normal}(5.6,\ 1.2) \\
\beta_{0,3} &\sim \text{normal}(1.2, .07) \\
w_i &= \frac{1}{3} \\
\tilde{\beta}_0 &= \sum_{i=1}^{3} w_i \beta_{0,i}
\end{aligned}
$$

You will have a numeric value for the mean of $\tilde{\beta}$ and a standard deviation of the distribution of $\tilde{\beta}_0$.

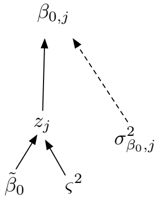
## Example: multiple priors on intercept $\beta_0$

Step 2: Using the numeric values for $\tilde{\beta}_0$ and $\sigma^2_{\tilde{\beta}_0}$, find the marginal posterior distribution of the intercept based on new data and prior information ($\beta_0^{\text{new}}$) in a separate MCMC algorithm:

$$
\begin{aligned}
[\beta_0^{\text{new}}, \boldsymbol{\beta}_0 \beta_1, \sigma^2_{\text{reg}} \mid \mathbf{y}] \quad \propto \prod_{i=1}^{n} \quad & \text{normal}(y_i \mid \beta_0^{\text{new}} + \beta_1 x_i, \sigma^2_{\text{reg}}) \\
& \times \text{normal}(\beta_0^{\text{new}} \mid \underbrace{\tilde{\beta}_0, \tilde{\sigma}^2_{\beta_0}}_{\text{numeric}}) \\
& \times \text{normal}(\beta_1 \mid 0, 10000) \\
& \times \text{uniform}(\varsigma^2 \mid 0, 2000)
\end{aligned}
$$

# Another approach: a hierarchical prior

<u>Step1</u>: Do MCMC to approximate mean $\tilde{\beta}_0$ and variance of the mean $\tilde{\sigma}^2_{\tilde{\beta}_0}$.



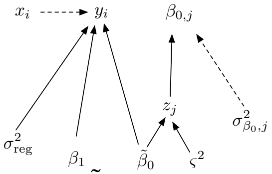

$$[\tilde{\beta}_0, \mathbf{z}, \varsigma^2 | \boldsymbol{\beta}_0] \propto \prod_{j=1}^{J} [\beta_{0,j} | z_i, \boldsymbol{\sigma}^2_{\beta_0,j}][z_i | \tilde{\beta}_0, \varsigma^2][\mathbf{z}][\tilde{\beta}_0][\varsigma^2]$$

## Another approach: a hierarchical prior

Step 2: As before, use the numeric values for $\tilde{\beta_0}$ and $\sigma^2_{\tilde{\beta}_0}$, find the marginal posterior distribution of the intercept based on new data and prior information ($\beta_0^{\text{new}}$) in a separate MCMC algorithm:

$$
\begin{aligned}
[\beta_0^{\text{new}}, \beta_1, \sigma^2_{\text{reg}} \mid \mathbf{y}] \quad \propto \prod_{i=1}^{n} \quad & \text{normal}(y_i \mid \beta_0^{\text{new}} + \beta_1 x_i, \sigma^2_{\text{reg}}) \\
& \times \text{normal}(\beta_0^{\text{new}} \mid \underbrace{\tilde{\beta_0}, \sigma^2_{\beta_0}}_{\text{numeric}}) \\
& \times \text{normal}(\beta_1 \mid 0, 10000) \\
& \times \text{uniform}(\varsigma^2 \mid 0, 2000)
\end{aligned}
$$

# Including a hierarchical prior in a single model




$$[\tilde{\beta}_0, \beta_1, \sigma_{\text{reg}}, \varsigma \mid \mathbf{y}] \propto \prod_{i=1}^{n} \text{normal}(y_i \mid \tilde{\beta}_0 + \beta_1 x_i, \sigma^2_{\text{reg}}) \quad (3)$$

$$\times \prod_{j=1}^{J} \text{normal}(\beta_{0,j} \mid z_j, \sigma^2_{\beta_{0,j}}) \quad (4)$$

$$\times \text{normal}(z_j \mid \tilde{\beta}_0, \varsigma^2) \quad (5)$$

$$\times \text{normal}(\tilde{\beta}_0 \mid 0, 1000) \quad (6)$$

$$\times \text{normal}(\beta_1 \mid 0, 10000) \quad (7)$$

$$\times \text{uniform}(\varsigma \mid 0, 200) \quad (8)$$

$$\times \text{uniform}(\sigma_{\text{reg}} \mid 0, 200) \quad (9)$$

These two approaches give results that are very similar, but not exactly the same. Why?

- Use the first approach when you want to weight studies with specific values for the $w_i$.
- Otherwise, the second approach which is more Bayesian and perhaps more easily explained.