Model Building Exercises

May 24, 2018

Motivation

The ability of Bayesian methods to handle hierarchical models in an unusually tidy way is why they are becoming the first choice for complex problems in ecology and conservation biology, problems with multiple unknowns, sources of data and sources of uncertainty. Recall that the posterior distribution of all of the unobserved quantities is proportionate to the joint distributions of the unobserved quantities and the data:

$$\big[\boldsymbol{\theta} \mid \mathbf{y}\big] \propto \underbrace{\big[\boldsymbol{\theta}, \mathbf{y}\big]}_{\text{Factor into sensible parts}}$$

The starting point for developing hierarchical models is to write a properly factored expression for the proportionality between the posterior and joint distribution of the observed and unobserved quantities. Properly means that the expression for the factored joint distribution obeys the chain rule of probability after assumptions about independence have been made. Bayesian networks, also called directed acyclic graphs, offer a way to visually assure that your model does so. This will be true if there is one unknown and one data set or one hundred unknowns and ten data sets. This factored expression is all that is required to specify a "roll-your-own" MCMC algorithm or to write code in one of the current software packages that sample from the marginal posterior distributions, JAGS, STAN, OpenBUGS etc. The expression for posterior and joint is where you start discussions with statistical colleagues. It should be included in all papers and proposals using Bayesian methods because it communicates virtually everything about where your inferences come from.

It follows that learning to write proper mathematical and statistical expressions for Bayesian models is 90 percent of the battle. In this exercise, we practice that vital skill. The problems increase in difficulty as we proceed, so it will be important to understand what you did right and wrong before you proceed to the next problem. In addition to practice drawing Bayesian networks and writing posterior and joint distributions, the problems will challenge you to:

- Choose distributions appropriate for the support of the random variable.

- Deftly use moment matching to convert means and standard deviations to parameters of distributions.

- Make inferences on derived quantities.

Problems

1. You are interested in modeling the relationship between per capita income and an index of air pollution for 80 nations around the world (i.e., the Kuznets effect). You hypothesize that air pollution increases then declines as per capita income increases. You have data on the mean (and the standard deviation of the mean) for each country's air quality index (a continuous non-negative variable). The predictor variable (income) is not measured perfectly because it is based on a sample – you have a mean and a standard deviation of the mean for the per capita income of each county. How would you analyze these data?
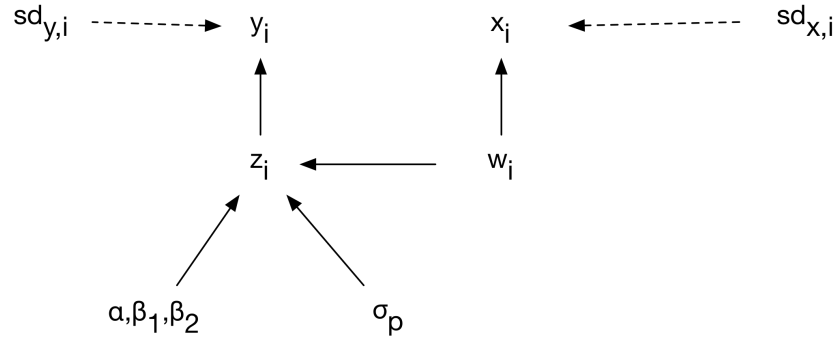
Figure 1: In this DAG, y_i and $sd_{y,i}$ and x_i and $sd_{x,i}$ and are the means (and standard deviations of the means) of air quality and per capita income in the i_{th} county. All these quantites are *known*.

$$\left[\boldsymbol{z}, \boldsymbol{w}, \alpha, \boldsymbol{\beta}, \sigma_p \mid \boldsymbol{y}, \boldsymbol{x}\right] \propto \prod_{i=1}^{n} \left[z_i \mid g\left(\alpha, \boldsymbol{\beta}, w_i\right), \sigma_p^2\right] \left[x_i \mid w_i, sd_{x,i}^2\right] \left[y_i \mid z_i, sd_{y,i}\right]$$

$$\times \left[w_i\right] \left[\alpha\right] \left[\beta_1\right] \left[\beta_2\right] \left[\sigma_p\right]$$

$$g\left(\alpha, \boldsymbol{\beta}, w_i\right)) = e^{\alpha + \beta_1 w_i + \beta_2 w_i^2} \qquad\qquad \alpha \sim \text{normal}\left(0, 1000\right)$$

$$z_i \sim \text{lognormal}\left(\log\left(g\left(\alpha, \boldsymbol{\beta}, w_i\right)\right), \sigma_p^2\right) \qquad \beta_1 \sim \text{normal}\left(0, 1000\right)$$

$$y_i \sim \text{lognormal}\left(\log\left(z_i\right), sd_{y,i}^2\right) \qquad\qquad \beta_2 \sim \text{normal}\left(0, 1000\right)$$

$$x_i \sim \text{lognormal}\left(\log\left(w_i\right), sd_{x,i}^2\right) \qquad\qquad \sigma_p \sim \text{uniform}\left(0, 10\right)$$

$$w_i \sim \text{gamma}\left(.001, .001\right)$$

2. You have data on the relationship between incidence of lung cancer in households (1 if cancer is present and 0 if no cancer) and radon levels in the house for 1000 houses in 40 counties within a state. You also have data on the clay soil content at the county level. You heroically assume both clay content and radon are know without error. How would you model the effect of radon and soil type on the probability of lung cancer?
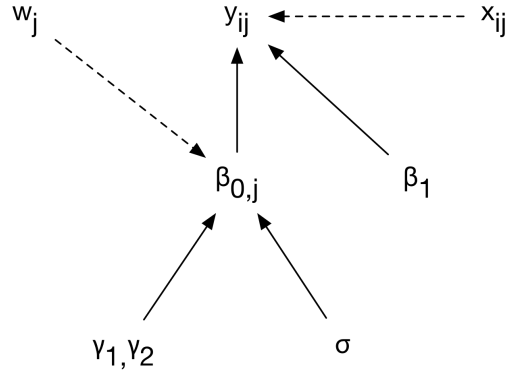
Figure 2: In this DAG, x_{ij} is the radon level and y_{ij} is an indicator that equals 1 if cancer is present and 0 if it is not in the i_{th} house in the j_{th} county, and w_{th} is the clay soil content in the j_{th} county.

$$\big[\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma \mid \boldsymbol{y}\big] \propto \prod_{i=1}^{M_j} \prod_{j=1}^{N} \big[y_{ij} \mid g(\boldsymbol{\beta}, x_{ij})\big] \big[\beta_{0,ij} \mid h(\boldsymbol{\gamma}, w_j), \sigma^2\big] \big[\boldsymbol{\gamma}\big] \big[\beta_1\big] \big[\sigma\big]$$

$$g(\boldsymbol{\beta}, x_{ij}) = \frac{e^{\beta_{0,ij} + \beta_1 x_{ij}}}{1 + e^{\beta_{0,ij} + \beta_1 x_{ij}}} \qquad\qquad \gamma_0 \sim \text{normal}(0, 1000)$$

$$h(\boldsymbol{\gamma}, w_j) = \gamma_0 + \gamma_1 w_j \qquad\qquad\qquad \gamma_1 \sim \text{normal}(0, 1000)$$

$$y_{ij} \sim \text{Bernoulli}\big(g(\boldsymbol{\beta}, x_{ij})\big) \qquad\qquad \sigma \sim \text{uniform}(0, 100)$$

$$\beta_0 \sim \text{normal}\big(h(\boldsymbol{\gamma}, w_j), \sigma^2\big)$$

$$\beta_1 \sim \text{normal}(0, 1000)$$

3. You have plot level data on diversity of plant communities. The data consist of counts y_{ij} of the number of individuals of species i on $j = 1, \ldots, J$ same-sized plots, and the total number of individuals on plot j is reported as n_j. How would you estimate an index (H) of species diversity across the community, where $H = -\sum_{i=1}^{R} p_i \log(p_i)$, p_i is the mean proportion of the i_{th} species in in the community, and R is the total number of species present? All estimates must include proper accounting for uncertainty.
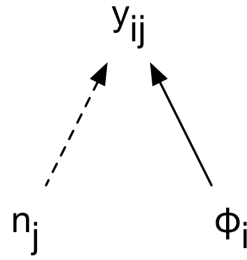
$$y_{ij}$$

Figure 3: In this DAG, y_{ij} is the number of individuals in the i_{th} species observed in the j_{th} plot while n_j is the total number of individuals across all species observed in the j_{th} plot.

$$\left[\boldsymbol{\phi} \mid \mathbf{Y}\right] \propto \prod_{j=1}^{J} \left[\boldsymbol{y}_j \mid n_j, \boldsymbol{\phi}\right]\left[\boldsymbol{\phi}\right]$$

$$H = -\sum_{i=1}^{R} \phi_i \log\!\left(\phi_i\right)$$

$$\boldsymbol{y}_j \sim \text{multinomial}\!\left(n_j, \boldsymbol{\phi}\right)$$

$$\boldsymbol{\phi} \sim \text{Dirichlet} \underbrace{\left(1, 1, \cdots, 1\right)'}_{\text{a vector of length R}}$$

where R is the the total number of species across all plots (this comes from the data).

4. You have data on the number of willow seedlings that establish on 100 different 10×10 meter plots. Assume these data are measured perfectly (i.e., you did not over or under count seedlings). You also have 5 measurements of soil water and 1 measurement of percent cover (estimated visually) on each plot. How would you model the effect of soil water and percent bare soil on the number of plants established? You also have paired data of visually estimated cover and actual cover on 15 calibration plots (separate from the 100 plots with seedling and soil water data), where the actual proportion of vegetated soil was laboriously measured using 1×1 meter gridded quadrats. How would you incorporate uncertainty in the percent cover and soil water covariates in your analysis?

5. Now presume that the 100 plots are arranged in 5 different stream reaches. You have data on peak runoff in each of the reaches, which you may assume is measured perfectly. You want to understand variation at the catchment scale created by peak runoff. How would you include these data in the model you developed in problem 4?
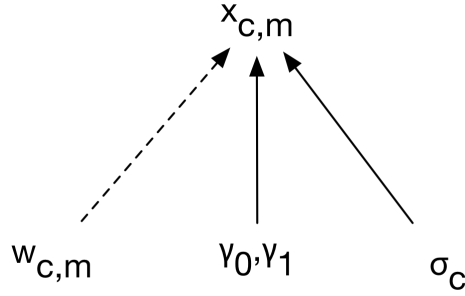
Figure 4: In this DAG, $x_{c,m}$ is the percent cover estimated visually and $w_{c,m}$ is the actual percent cover measured with a quadrat in the m_{th} calibration plot.

$$\big[\boldsymbol{\gamma}, \sigma_c \mid \boldsymbol{x_c}\big] \propto \prod_{m=1}^{15} \big[\boldsymbol{x}_{c,m} \mid \mu_m, \sigma_c^2\big] \big[\gamma_0\big] \big[\gamma_1\big] \big[\sigma_c\big]$$

$$\mu_m = \frac{e^{\gamma_0 + \gamma_1 w_{c,m}}}{1 + e^{\gamma_0 + \gamma_1 w_{c,m}}}$$

$$\alpha_m = \frac{\mu_m^2 - \mu_m^3 - \mu_m \sigma_c^2}{\sigma_c^2}$$

$$\beta_m = \frac{\mu_m^2 - 2\mu_m^2 + \mu_m^3 - \sigma_c^2 + \mu_m \sigma_c^2}{\sigma_c^2}$$

$$x_{c,m} \sim \text{beta}\big(\alpha_m, \beta_m\big)$$

$$\gamma_0 \sim \text{normal}\big(0, 1000\big)$$

$$\gamma_1 \sim \text{normal}\big(0, 1000\big)$$

$$\sigma_c \sim \text{uniform}\big(0, 100\big)$$

To use informed priors on γ_0, γ_1, and σ_c we take the mean and standard deviation of each of their MCMC chains. For σ_c, we must moment match to the parameters of a probability distribution whose support is all positive, such as gamma or lognormal. Please see the appendix for details.
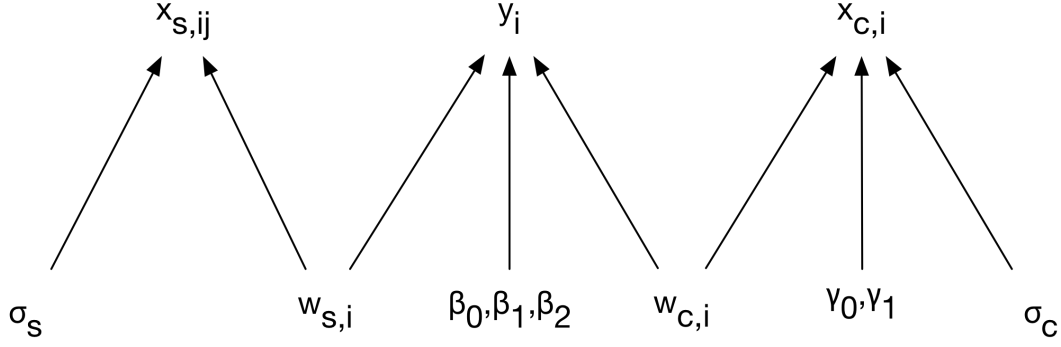
Figure 5: In this DAG, y_i is the number of willow seedlings, $x_{s,ij}$ is the j_{th} measurement of soil water content, and $x_{c,i}$ is a visual estimate of percent cover on the i_{th} plot.

$$\big[\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{w_s}, \boldsymbol{w_c}, \sigma_s, \sigma_c \mid \boldsymbol{y}, \boldsymbol{x_s}, \boldsymbol{x_c}\big] \propto \prod_{i=1}^{100} \big[y_i \mid g\big(\boldsymbol{\beta}, w_{s,i}, w_{c,i}\big)\big] \prod_{j=1}^{5} \big[x_{s,ij} \mid \log\big(w_{s,i}\big), \sigma_s\big]$$

$$\times \big[\boldsymbol{x_{c,i}} \mid \mu_i, \sigma_c^2\big] \big[\beta_0\big] \big[\beta_1\big] \big[\gamma_0\big] \big[\gamma_1\big] \big[w_{s,i}\big] \big[w_{c,i}\big] \big[\sigma_s\big] \big[\sigma_c\big]$$

$$g\big(\boldsymbol{\beta}, w_{s,i}, w_{c,i}\big) = e^{\beta_0 + \beta_1 w_{s,i} + \beta_2 w_{c,i}} \qquad \beta_0 \sim \text{normal}\big(0, 1000\big)$$

$$\mu_i = \frac{e^{\gamma_0 + \gamma_1 w_{c,i}}}{1 + e^{\gamma_0 + \gamma_1 w_{c,i}}} \qquad\qquad \beta_1 \sim \text{normal}\big(0, 1000\big)$$

$$\alpha_i = \frac{\mu_i^2 - \mu_i^3 - \mu_i \sigma_c^2}{\sigma_c^2} \qquad\qquad w_{s,i} \sim \text{gamma}\big(.001, .001\big)$$

$$\qquad\qquad\qquad\qquad\qquad\qquad w_{c,i} \sim \text{uniform}\big(0, 1\big)$$

$$\beta_i = \frac{\mu_i^2 - 2\mu_i^2 + \mu_i^3 - \sigma_c^2 + \mu_i \sigma_c^2}{\sigma_c^2} \qquad \gamma_0 \sim \text{normal}\big(\gamma_{0,mean}, \gamma_{0,var}\big)$$

$$y_i \sim \text{Poisson}\big(g\big(\boldsymbol{\beta}, w_{s,i}, w_{c,i}\big)\big) \qquad \gamma_1 \sim \text{normal}\big(\gamma_{1,mean}, \gamma_{1,var}\big)$$

$$x_{s,ij} \sim \text{lognormal}\big(\log\big(w_{s,i}\big), \sigma_s^2\big) \qquad \sigma_s \sim \text{uniform}\big(0, 100\big)$$

$$x_{c,i} \sim \text{beta}\big(\alpha_i, \beta_i\big) \qquad\qquad \sigma_c \sim \text{gamma}\big(a_c, b_c\big)$$

See appendix to understand how we determined the numerical priors on γ_0, γ_1, and σ_c.
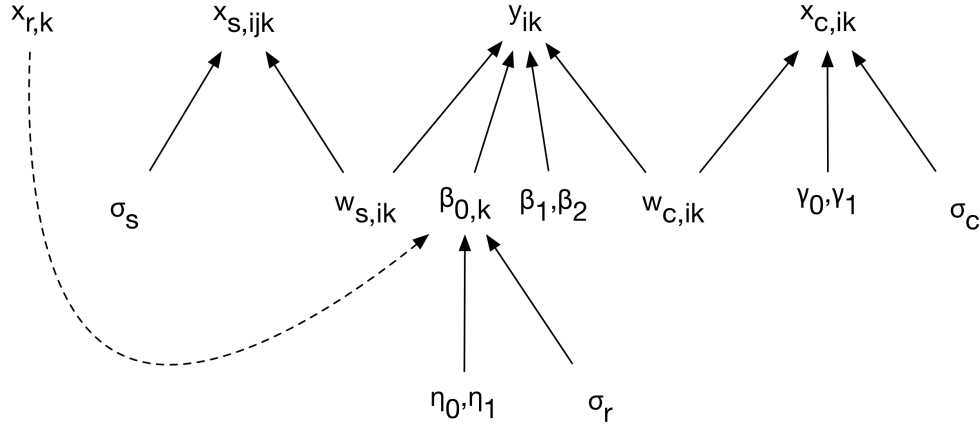
Figure 6: In this DAG, y_{ik} is the number of willow seedlings, $x_{s,ijk}$ is the j_{th} measurement of soil water content, $x_{c,ik}$ is a visual estimate of percent cover, and $x_{r,k}$ is peak runoff on the i_{th} plot in the k_{th} stream reach.

$$
\left[\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{w_s}, \boldsymbol{w_c}, \sigma_s, \sigma_c, \sigma_r \mid \boldsymbol{y}, \boldsymbol{x_s}, \boldsymbol{x_c}, \boldsymbol{x_r}\right] \propto
$$

$$
\prod_{i=1}^{20}\prod_{k=1}^{5} \left[y_{ik} \mid g\left(\boldsymbol{\beta}, w_{s,ik}, w_{c,ik}\right)\right]\left[\beta_{0,k} \mid h\left(\boldsymbol{\eta}, x_{r,k}\right)\right]
$$

$$
\times \prod_{j=1}^{5} \left[x_{s,ijk} \mid \log\left(w_{s,ik}\right), \sigma_s^2\right]\left[\boldsymbol{x}_{c,ik} \mid \mu_{ik}, \sigma_c^2\right]
$$

$$
\times \left[\beta_1\right]\left[\gamma_0\right]\left[\gamma_1\right]\left[\eta_0\right]\left[\eta_1\right]\left[w_{s,ik}\right]\left[w_{c,ik}\right]\left[\sigma_s\right]\left[\sigma_c\right]\left[\sigma_r\right]
$$

$$g\big(\boldsymbol{\beta}, w_{s,ik}, w_{c,i}\big) = e^{\beta_{0,k} + \beta_1 w_{s,ik} + \beta_2 w_{c,ik}}$$

$$h\big(\boldsymbol{\eta}, x_{r,k}\big) = \eta_0 + \eta_1 x_{r,k}$$

$$\mu_{ik} = \frac{e^{\gamma_0 + \gamma_1 w_{c,ik}}}{1 + e^{\gamma_0 + \gamma_1 w_{c,ik}}}$$

$$\alpha_{ik} = \frac{\mu_{ik}^2 - \mu_{ik}^3 - \mu_{ik}\sigma_c^2}{\sigma_c^2}$$

$$\beta_{ik} = \frac{\mu_{ik}^2 - 2\mu_{ik}^2 + \mu_{ik}^3 - \sigma_c^2 + \mu_i\sigma_c^2}{\sigma_c^2}$$

$$y_{ik} \sim \text{Poisson}\big(g\big(\boldsymbol{\beta}, w_{s,ik}, w_{c,ik}\big)\big)$$

$$\beta_{0,k} \sim \text{normal}\big(h\big(\boldsymbol{\eta}, x_{r,k}\big), \sigma_r^2\big)$$

$$x_{s,ijk} \sim \text{lognormal}\big(\log\big(w_{s,ik}\big), \sigma_s^2\big)$$

$$x_{c,ik} \sim \text{beta}\big(\alpha_{ik}, \beta_{ik}\big)$$

$$\beta_1 \sim \text{normal}\big(0, 1000\big)$$

$$w_{s,ik} \sim \text{gamma}\big(.001, .001\big)$$

$$w_{c,ik} \sim \text{uniform}\big(0, 1\big)$$

$$\eta_0 \sim \text{normal}\big(0, 1000\big)$$

$$\eta_1 \sim \text{normal}\big(0, 1000\big)$$

$$\gamma_0 \sim \text{normal}\big(\gamma_{0,mean}, \gamma_{0,var}\big)$$

$$\gamma_1 \sim \text{normal}\big(\gamma_{1,mean}, \gamma_{1,var}\big)$$

$$\sigma_s \sim \text{uniform}\big(0, 100\big)$$

$$\sigma_c \sim \text{gamma}\big(a_c, b_c\big)$$

$$\sigma_r \sim \text{uniform}\big(0, 100\big)$$

See appendix to understand how we determined the numerical priors on γ_0, γ_1, and σ_c.

Appendix

After convergence, select $t = 1, 2, ..., T$ samples from the chains for γ_0, γ_1, and σ_c. The math below might look obtuse, but all you are doing is approximating the means and variances of γ_0, γ_1, and σ_c from their MCMC chains.

$$\gamma_{0,mean} = \frac{1}{T} \sum_{t=1}^{T} \gamma_0^{(t)} \approx \mathrm{E}[\gamma_0] \qquad \gamma_{0,var} = \frac{1}{T} \sum_{t=1}^{T} \left(\gamma_0^{(t)} - \gamma_{0,mean}\right)^2 \approx \mathrm{E}\left[\left(\gamma_0 - \mathrm{E}[\gamma_0]\right)^2\right]$$

$$\gamma_{1,mean} = \frac{1}{T} \sum_{t=1}^{T} \gamma_1^{(t)} \approx \mathrm{E}[\gamma_0] \qquad \gamma_{1,var} = \frac{1}{T} \sum_{t=1}^{T} \left(\gamma_1^{(t)} - \gamma_{1,mean}\right)^2 \approx \mathrm{E}\left[\left(\gamma_1 - \mathrm{E}[\gamma_1]\right)^2\right]$$

$$\sigma_{c,mean} = \frac{1}{T} \sum_{t=1}^{T} \sigma_c^{(t)} \approx \mathrm{E}[\sigma_c] \qquad \sigma_{c,var} = \frac{1}{T} \sum_{t=1}^{T} \left(\sigma_c^{(t)} - \sigma_{c,mean}\right)^2 \approx \mathrm{E}\left[\left(\sigma_c - \mathrm{E}[\sigma_c]\right)^2\right]$$

Lastly, for σ_c we moment match to the a and b parameters in a gamma distribution.

$$a_c = \frac{\left(\sigma_{c,mean}\right)^2}{\left(\sigma_{c,var}\right)^2}$$

$$b_c = \frac{\sigma_{c,mean}}{\left(\sigma_{c,var}\right)^2}$$