Practice in writing hierarchical models

June 19, 2016

1 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:

$$[\boldsymbol{\theta}|\mathbf{y}] \propto \underbrace{[\boldsymbol{\theta}, \mathbf{y}]}_{\text{Factor into sensible parts.}} \tag{1}$$

The starting point for developing hierarchical models is to write a properly factored[1] expression for the proportionality between the posterior and joint distribution of the observed and unobserved quantities. 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 for you, JAGS, STAN, OpenBUGS etc. The expression for posterior and joint is where you start discussions with statistical colleague. It should be included in all papers and proposals using Bayesian methods because it communicates virtually everything about where your inferences come from.

[1] 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.

It follows that learning to write proper mathematical and statistical expressions for Bayesian models is 90% of the battle. In this exercise, we will 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

1) Choose distributions appropriate for the support of the random variable.

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

3) Make inferences on derived quantities.

2 Problems

1. Clark (2003) modeled fecundity of spotted owls using data on number of offspring produced per female for 119 individuals.

   (a) Write a model for the mean fecundity assuming that all individuals have the same mean fecundity and that differences among the observations arise from sampling error.

   (b) Now write a model that allows each individual to have its own fecundity.

   (c) Answers will be given on the board in class.

2. You have data on patch occupancy by a breeding bird based on three replicated searches of 200 transects at random locations on the landscape. The data consist of the number of the number of times the bird was detected visually or by hearing its call. For now, you may assume no spatial autocorrelation in the data. You have a set of landscape covariates associated with each patch and you seek to understand how those covariates influence the probability that a patch is occupied (ϕ_i).

   (a) Start by assuming that detection is perfect. Write a Bayesian model for the effect of covariates ($\mathbf{x}_i$) on the probability of patch occupancy (ϕ_i).

(b) Now relax the heroic assumption of perfect detection. You must first model the data (the number of successful detections on a given number of visits) and link your model of the data to the occupancy process, the true state of occupancy of the patch. Assume that detection probability (p) is constant across all patches.

(c) Discuss how you would need to change your model if you had visit-specific covariates allowing you to model detection probability for each patch.

(d) Answers will be given on the board in class.

3. 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 in a way that is honest about all sources of uncertainty?

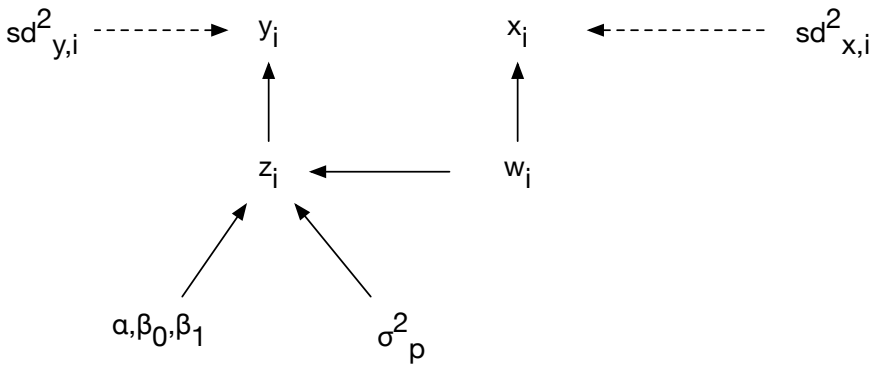

Figure 1: In this DAG, y_i and $sd_{y,i}$ and x_i and $sd_{x,i}$ and are the means (and standard deviations of the means) of air quality and per capita income in the i_{th} county.

$$\begin{aligned}\big[\boldsymbol{z}, \boldsymbol{w}, \alpha, \boldsymbol{\beta}, \sigma_p^2 \mid \boldsymbol{y}, \boldsymbol{x}\big] \propto \prod_{i=1}^{n} \big[y_i \mid z_i, sd_{y,i}^2\big] \big[z_i \mid g\big(\alpha, \boldsymbol{\beta}, w_i\big), \sigma_p^2\big] \big[x_i \mid w_i, sd_{x,i}^2\big] \\ \times \big[z_i\big] \big[w_i\big] \big[\alpha\big] \big[\beta_0\big] \big[\beta_1\big] \big[\sigma_p^2\big]\end{aligned}$$

$$\begin{aligned}
&g\big(\alpha, \boldsymbol{\beta}, w_i\big)\big) = e^{\alpha+\beta_1 w_i+\beta_2 w_i^2} && z_i \sim \text{gamma}\big(.001, .001\big) \\
&z_i \sim \text{lognormal}\big(\log\big(g\big(\alpha, \boldsymbol{\beta}, w_i\big)\big), \sigma_p^2\big) && \alpha \sim \text{normal}\big(0, 1000\big) \\
&y_i \sim \text{lognormal}\big(\log\big(z_i\big), sd_{y,i}^2\big) && \beta_1 \sim \text{normal}\big(0, 1000\big) \\
&x_i \sim \text{lognormal}\big(\log\big(w_i\big), sd_{x,i}^2\big) && \beta_2 \sim \text{normal}\big(0, 1000\big) \\
&w_i \sim \text{gamma}\big(.001, .001\big) && \sigma_p^2 \sim \text{uniform}\big(0, 100\big)
\end{aligned}$$

4. 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, which you heroically assume is know without error. How would you model the effect of radon and soil type on the probability of lung cancer?

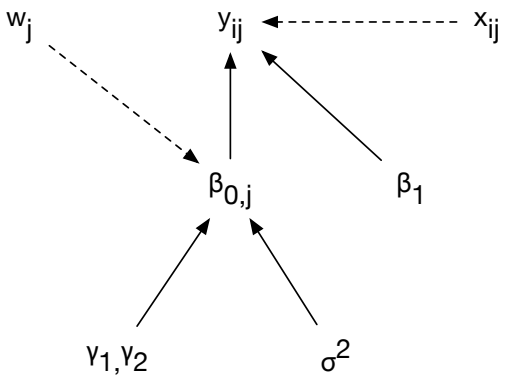

Figure 2: In this DAG, x_{ij} is the radon level and y_{ij} is an indicator that equals 1 if cancer is present and 0 if it is not in the i^{th} house in the j^{th} county, and w_j is the clay soil content in the j^{th} county.

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

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

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

$$y_{ij} \sim \text{Bernoulli}\big(g\big(\boldsymbol{\beta}, x_{ij}\big)\big) \qquad \sigma^2 \sim \text{uniform}\big(0, 1000\big)$$

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

$$\beta_1 \sim \text{normal}\big(0, .001\big)$$

5. 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 $j = 1...J$ 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 $H = -\sum_{i=1}^{R} p_i \log\big(p_i\big)$, (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.

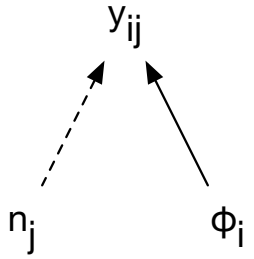

Figure 3: In this DAG, y_{ij} is the number of individuals in the i_{th} species observed in the j_{th} plot while n_j is the total number of individuals across all species observed in the j_{th} plot.

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

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

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

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

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

6. You have data on the number of willow seedings that establish on 100 10 $\times$ 10 meter plots. Assume these data are measured perfectly, i.e. you did not over- or under-count seedlings. You also have five measurements of soil water and one measurement of percent cover (estimated visually) on each plot. You also know the mean and standard deviation of the intercept and slope in an inverse logit calibration equation $\left(y_{\text{visual,i}} = \frac{e^{\gamma_0+\gamma_1 y_{\text{true},i}}}{1+e^{\gamma_0+\gamma_1 y_{\text{true},i}}}\right)$ regressing visually estimated cover ($y_{\text{visual},i}$) to measured cover ($y_{\text{true},i}$) as well as the standard error of the regression. Develop a Bayesian hierarchical model the the effect of soil water on counts of willow seedlings. For the speedy and ambitious, write a model for the calibration equation as well.

7. Now presume that the 100 plots are arranged in five 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 the data on runoff in the model you developed in problem 4?

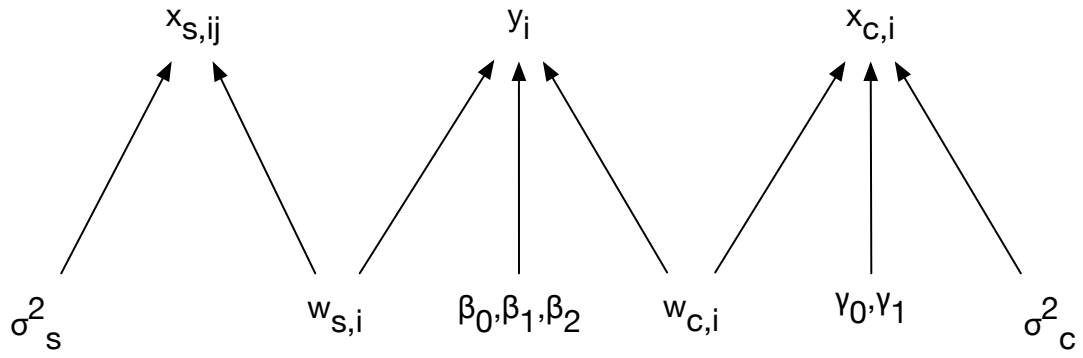

Figure 4: In this DAG, y_i is the number of willow seedlings, $x_{s,ij}$ is the j_{th} measurement of soil water content, and $x_{c,i}$ is a visual estimate of percent cover on the i_{th} plot.

$$\begin{aligned}\left[\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{w_s}, \boldsymbol{w_c}, \sigma_s^2, \sigma_c^2 \mid \boldsymbol{y}, \boldsymbol{X_s}, \boldsymbol{x_c}\right] \propto & \prod_{i=1}^{100} \left[y_i \mid g\big(\boldsymbol{\beta}, w_{s,i}, w_{c,i}\big)\right] \prod_{j=1}^{5} \left[x_{s,ij} \mid \log\big(w_{s,i}\big), \sigma_s^2\right] \\ & \times \left[x_{c,i} \mid \mu_i, \sigma_c^2\right] \left[\beta_0\right] \left[\beta_1\right] \left[\gamma_0\right] \left[\gamma_1\right] \left[w_{s,i}\right] \left[w_{c,i}\right] \left[\sigma_s^2\right] \left[\sigma_c^2\right]\end{aligned}$$

$$g\big(\boldsymbol{\beta}, w_{s,i}, w_{c,i}\big) = e^{\beta_0 + \beta_1 w_{s,i} + \beta_2 w_{c,i}}$$

$$\mu_i = \frac{e^{\gamma_0 + \gamma_1 w_{c,i}}}{1 + e^{\gamma_0 + \gamma_1 w_{c,i}}}$$

$$\alpha_i = \frac{\mu_i^2 - \mu_i^3 - \mu_i \sigma_c^2}{\sigma_c^2}$$

$$\beta_i = \frac{\mu_i^2 - 2\mu_i^2 + \mu_i^3 - \sigma_c^2 + \mu_i \sigma_c^2}{\sigma_c^2}$$

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

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

$$x_{c,i} \sim \text{beta}\big(\alpha_i, \beta_i\big)$$

$$\beta_0 \sim \text{normal}\big(0, 1000\big)$$

$$\beta_1 \sim \text{normal}\big(0, 100\big)$$

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

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

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

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

$$\sigma_s^2 \sim \text{uniform}\big(0, 100\big)$$

$$\sigma_c^2 \sim \text{gamma}\big(\sigma_{c,\alpha}^2, \sigma_{c,\beta}^2\big)$$

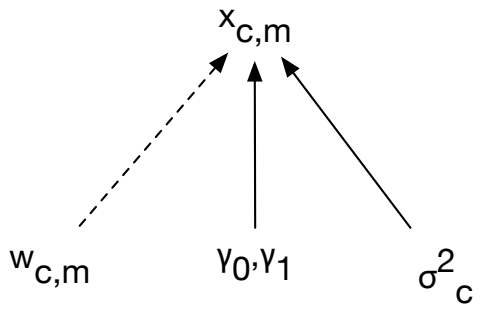

Figure 5: In this DAG, $x_{c,m}$ is the percent cover estimated visually and $w_{c,m}$ is the actual percent cover measured with a quadrat in the m_{th} calibration plot.

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

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

$$\alpha_m = \frac{\mu_m^2 - \mu_m^3 - \mu_m\sigma_c^2}{\sigma_c^2} \qquad \gamma_1 \sim \text{normal}\big(0, 1000\big)$$

$$\beta_m = \frac{\mu_m^2 - 2\mu_m^2 + \mu_m^3 - \sigma_c^2 + \mu_m\sigma_c^2}{\sigma_c^2} \qquad \sigma_c^2 \sim \text{uniform}\big(0, 100\big)$$

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

To use informed priors on γ_0, γ_1, and σ_c^2 we take the mean and standard deviation of each of their MCMC chains and moment match into the appropriate parameters for priors.

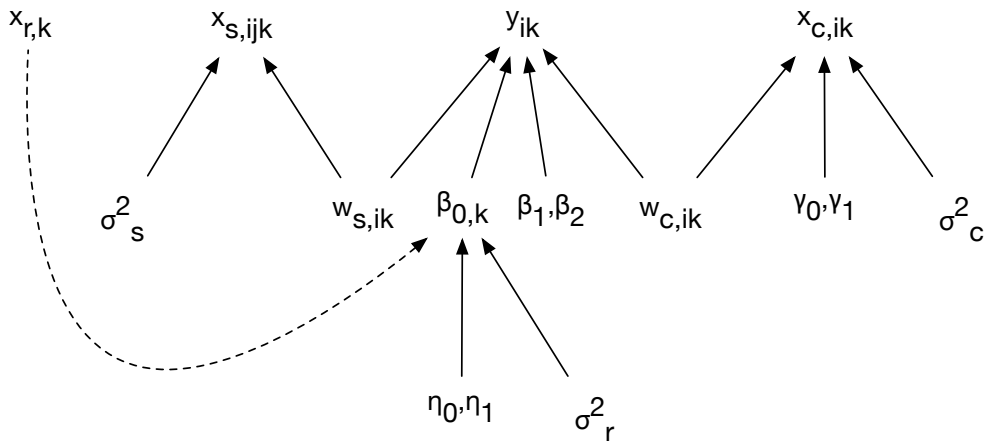

Figure 6: In this DAG, y_{ik} is the number of willow seedlings, $x_{s,ijk}$ is the j_{th} measurement of soil water content, $x_{c,ik}$ is a visual estimate of percent cover, and $x_{r,k}$ is peak runoff on the i_{th} plot in the k_{th} stream reach.

$$
\begin{aligned}
\left[\boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\eta}, \boldsymbol{w_s}, \boldsymbol{w_c}, \sigma_s^2, \sigma_c^2, \sigma_r^2 \mid \boldsymbol{y}, \boldsymbol{x_s}, \boldsymbol{x_c}, \boldsymbol{x_r}\right] \propto & \\
& \prod_{i=1}^{20}\prod_{k=1}^{5}\left[y_{ik} \mid g\big(\boldsymbol{\beta}, w_{s,ik}, w_{c,ik}\big)\right]\left[\beta_{0,k} \mid h\big(\boldsymbol{\eta}, x_{r,k}\big)\right] \\
& \times \prod_{j=1}^{5}\left[x_{s,ijk} \mid \log\big(w_{s,ik}\big), \sigma_s^2\right]\left[x_{c,ik} \mid \mu_{ik}, \sigma_c^2\right] \\
& \times \left[\beta_1\right]\left[\gamma_0\right]\left[\gamma_1\right]\left[\eta_0\right]\left[\eta_1\right]\left[w_{s,ik}\right]\left[w_{c,ik}\right]\left[\sigma_s^2\right]\left[\sigma_c^2\right]\left[\sigma_r^2\right]
\end{aligned}
$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

$$\sigma_s^2 \sim \text{uniform}\big(0, 100\big)$$

$$\sigma_c^2 \sim \text{gamma}\big(\alpha_c, \eta_c\big)$$

$$\sigma_r^2 \sim \text{uniform}\big(0, 100\big)$$

References

Clark, J. S., 2003. Uncertainty and variability in demography and population growth: A hierarchical approach. *Ecology* **84**:1370–1381.