

# Model Building Exercises

March 6, 2019

## 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 many unknowns, different type of data, and multiple 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:

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

It follows that 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 (or, unattractively in my view, DAGs), 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 must be included in all papers and proposals using Bayesian methods because it communicates virtually everything about where your inferences come from.

Learning to write proper mathematical and statistical expressions for Bayesian models is 70 percent of the battle of learning how to do Bayesian analysis. We will return to this battle time and time again during this course. In this exercise, we begin to learn the vital skill of model building. 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.

## 2 Preliminaries

- Review your notes on the Light Limitation of Trees lecture, where I illustrated several Bayesian models. The problems were chosen to align with the material covered in lecture.
- Read Chapters 1.1 (Preview), 6.1-6.21 (What is a Hierarchical model through Fecundity of spotted owls)<sup>1</sup>, 6.2.2 (Controls on nitrous oxide emissions of agricultural soils) and 10.1 and 10.2 (General approach, and An example of model building) in Hobbs and Hooten.
- Do problem 12.1 (Fisher's ticks) and consult the answers after struggling with each part. No write-up required on this one. It's a warmup.

---

<sup>1</sup>Note that in box 6.2.3, the  $x_i$  in panel B should be  $x_{ij}$  implying that there are reps of observation of  $x_{ij}$  arising from a distribution with mean  $\chi_i$ .

### 3 Instructions

- For each problem below, draw the Bayesian network, write the posterior and joint distributions using generic bracket notation with appropriate products. Next, choose specific distributions following the general flow that I illustrated in the Light Limitation of Trees lecture. At this point, don't worry too much about the specific forms for prior distributions. We will learn more about composing these as the course proceeds. You may use uniform distributions with bounds that are vague for non-negative parameters. Use normal distributions centered on zero with large variances for real-valued parameters. Again, don't sweat this too much.
- Work in groups to allow discussion and to teach each other. Prepare a write up, one per group. You may use pencil and paper for drawing DAGs and writing models. I don't want you to struggle with LaTeX at the same time you are struggling with model building. Scan your drawings and equations and turn them in as a pdf. The are due Friday, 3/22. We will go over the problems in class on Monday 3/25.
- I urge you to do a problem as completely as you possibly can, then consult the answer before going to the next problem. Don't correct your answer after consulting mine because I need to see how your are doing. The point is not to get the model right the first time, but rather to learn by *trying* to get it right.
- If you think you have found a mistake, good for you! There will be some lurking errors because some of these problems have not been vetted. There is no better way to show that you are learning than to find mistakes.
- Accumulate questions.

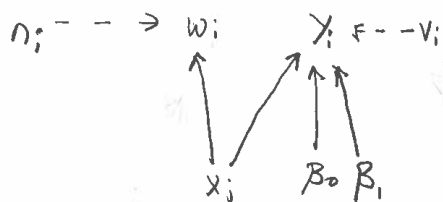
## 4 Problems

### 4.1 Cover of invasives on transects

This problem is taken with minor modification from work I am doing for the National Park Service Inventory and Monitoring Program. The data consist of point intercepts taken at .1 m intervals along 30 meter transects. Each transect has 300 points where observers classify a point defined by a plumb line dropped vertically from the transect as intersecting with exotic vegetation, native vegetation, or bare ground. There are 100 transects. You seek to estimate the proportion of total vegetative cover that is native using the proportion of a transect that is bare ground as a predictor variable. Develop a Bayesian model for this problem that properly models uncertainty in the predictor variable (proportion bare ground) and the response variable (proportion exotic cover). Hints - think about two data models, one for the response and one for the predictor variable – model the  $x$ 's using the same approach you use for modeling the  $y$ 's. The sample unit is a transect. Start by thinking about what is observed, what is unobserved, and what is known for each transect.

Define:

- $y_i$  = observed number of hits of exotics on transect  $i$
- $v_i$  = number of hits of exotic and native vegetation on transect  $i$ , presumed to be known because it observed without error.
- $x_i$  = the true, unobserved proportion of bare ground on transect  $i$
- $w_i$  = observed number of hits with bare soil
- $n_i$  = total number of intercepts on a transect, known.
- $N$  = total number of transects



Our deterministic  
model predicts the  
"probability of success"  
on  $v_i$  trials

$y_i$  = observed number of hits of  
exotics on transect  $i$

$v_i$  = observed number of  
hits of exotic and  
native vegetation  
on transect  $i$

$x_i$  = unobserved, true  
proportion of bare  
ground on transect  $i$

$\beta_0, \beta_1$  = regression coefficients

$w_i$  = observed number of  
hits with bare soil

$n_i$  = total number of  
intercepts on transect

$$[\underline{\beta}, \underline{x} | \underline{w}, \underline{y}] \propto \prod_{i=1}^N \underbrace{[y_i | v_i, g(\beta_0, \beta_1, x_i)]}_{\text{binomial}} \underbrace{[w_i | n_i, x_i]}_{\text{binomial}}$$

$$g(\beta_0, \beta_1, x_i) = \text{inverse logit}(\beta_0 + \beta_1 x_i)$$

$$y_i \sim \text{binomial}(v_i, g(\beta_0 + \beta_1 x_i))$$

$$w_i \sim \text{binomial}(n_i, x_i)$$

$$x_i \sim \text{uniform}(0, 1)$$

$$\beta_0, \beta_1 \sim \text{normal}(0, 2.7)$$

produce flat prior of  
probability of success  
will explain soon  
Also see H011 prior's  
chpt

## 4.2 Effect of radon on cancer risk

You seek to understand how radon levels influence risk of cancer. You have data on the incidence of lung cancer in households (1 if cancer is present and 0 if no cancer) and radon levels (a continuous, non-negative number) for 1000 houses in each of 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 levels are known without error. How would you model the effect of radon and soil type on the probability of lung cancer? Some hints–

1. What deterministic model would you use to predict the probability of cancer in a household as a function of radon level?
2. What likelihood would you use for these 0 or 1 data?
3. Assume that the intercept in your deterministic model of the effect of radon level on probability of cancer in a household is a linear function of county level clay soil content.

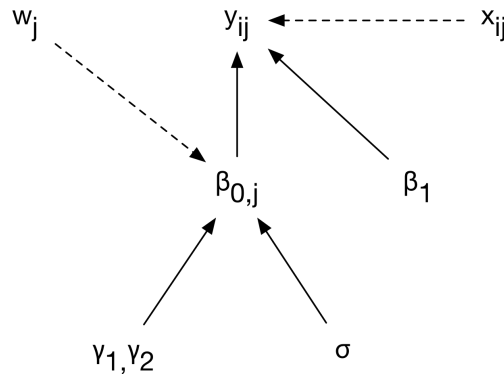


Figure 1: In this DAG,  $x_{ij}$  is the radon level and  $y_{ij}$  is an indicator that equals 1 if cancer is present and 0 if it is not in the  $i_{\text{th}}$  house in the  $j_{\text{th}}$  county, and  $w_{\text{th}}$  is the clay soil content in the  $j_{\text{th}}$  county.

$$[\gamma, \beta, \sigma \mid \mathbf{y}] \propto \prod_{i=1}^{1000} \prod_{j=1}^{40} [y_{ij} \mid g(\beta, x_{ij})] [\beta_{0,j} \mid h(\gamma, w_j), \sigma^2] [\gamma] [\beta_1] [\sigma]$$

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

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

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

$$\beta_{0,j} \sim \text{normal}(h(\boldsymbol{\gamma}, w_j), \sigma^2)$$

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

### 4.3 Diversity of a plant community

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  $j = 1, \dots, J$  same-sized plots, and the total number of individuals on plot  $j$  is reported as  $n_j$ . How would you model an index ( $H$ ) of species diversity across the community, where  $H = -\sum_{i=1}^R \phi_i \log(\phi_i)$ ,  $\phi_i$  is the unobserved proportion of the  $i_{\text{th}}$  species in the community, and  $R$  is the total number of species present? Hints–

1. Model the observed count data as a random variable (a vector) arising from the unobserved vector  $\phi$  of proportions.
2. Take a look at the Dirichlet distribution as a way to form a prior on the vector  $\phi$ . The Dirichlet is to the multinomial likelihood as the beta distribution is to the binomial likelihood. A vague Dirichlet has parameters = 1 for all categories.
3. Calculate  $H$  as a derived quantity of the  $\phi_i$  and  $R$ , which will allow us to obtain a marginal posterior distribution for  $H$  because the equivariance property of MCMC means that any quantity that is a function of a random variable becomes a random variable.



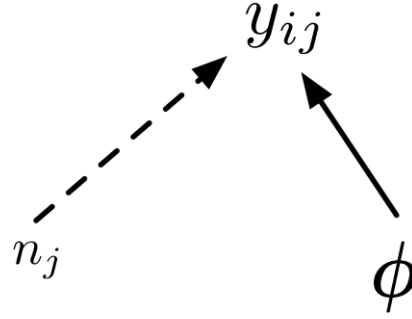


Figure 2: In this DAG,  $y_{ij}$  is the number of individuals in the  $i_{\text{th}}$  species observed in the  $j_{\text{th}}$  plot while  $n_j$  is the total number of individuals across all species observed in the  $j_{\text{th}}$  plot.

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

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

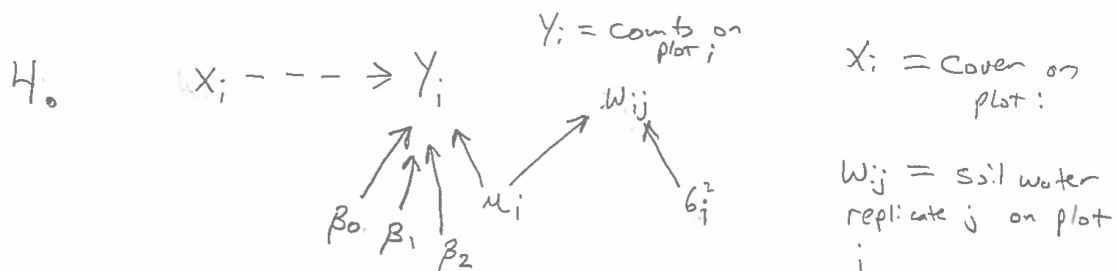
$$\mathbf{y}_j \sim \text{multinomial}(n_j, \phi)$$

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

where  $R$  is the the observed, total number of species across all plots.

## 4.4 Controls on willow seedling establishment

1. You are interested in the way that soil water and herbaceous plant cover influence establishment of willow seedlings in riparian communities. You have data on the number of willow seedlings that establish on 100  $10 \times 10$  meter plots, which of course are a random variables before they are observed. 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 herbaceous cover (estimated visually) on each plot. Assume for now that herbaceous cover is measured perfectly, but you want to include sampling variation in soil water for each plot in your model. How would you model the effect of soil water and herbaceous cover on the number of plants established?



$$\underline{\mu} \mid \underline{\beta}, \underline{\mu}, \underline{G^2} \mid \underline{W}, \underline{y} \propto \prod_{i=1}^{100} \prod_{j=1}^5 [Y_i \mid e^{\beta_0 + \beta_1 \mu_i + \beta_2 X_i}]$$

— underline indicates  
vectors or  
matrices

$$\times [W_{ij} \mid \mu_i, G_i^2]$$

$$\times [\mu_i] [G_i^2] [\beta_0] [\beta_1]$$

$$Y_i \sim \text{Poisson}(e^{\beta_0 + \beta_1 \mu_i + \beta_2 X_i})$$

$$W_{ij} \sim \text{gamma}\left(\frac{\mu_i^2}{G_i^2}, \frac{\mu_i}{G_i^2}\right)$$

$$G_i^2 \sim \text{unif}(0, 100)$$

$$\beta_0 \sim \text{normal}(0, 10000)$$

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

$$\beta_2 \sim \text{normal}(0, 1000)$$

$$\mu_i \sim \text{unif}(0, 100)$$

← could be  
inverse  
gamma on  
 $G_i^2$

$\beta_2 \sim \text{normal}(0, 1000)$

all vague priors scaled  
properly

2. Your major professor objected to your assumption of cover observed perfectly by eye, insisting, reasonably I think, that you develop a data model relating your ocular estimate to the true cover in a plot. So, you obtained visual estimates of cover paired with the actual proportion of vegetated area (measured using small sub-plots) on 15 10 x 10 m plots. After days of sweaty labor, you regressed visual estimates ( $x_i$ ) on the true cover ( $z_i$ ) and developed a calibration equation  $h(\boldsymbol{\alpha}, z_i)$ :

$$h(\boldsymbol{\alpha}, z_i) = \frac{e^{\alpha_o + \alpha_1 z_i}}{1 + e^{\alpha_o + \alpha_1 z_i}} \quad (1)$$

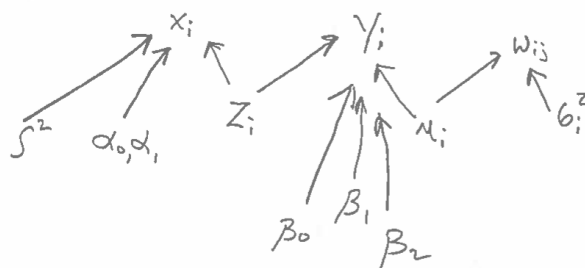
$$x_i \sim \text{beta}(m(h(\boldsymbol{\alpha}, z_i), \varsigma^2)) \quad (2)$$

$$\alpha_o \sim \text{normal}(.05, .006) \quad (3)$$

$$\alpha_1 \sim \text{normal}(1.07, .13) \quad (4)$$

$$\varsigma^2 \sim \text{inverse gamma}(10.2, 630) \quad (5)$$

5.

 $y_i$  = count on plot  $i$  $w_{ij}$  = soil water rep  $j$  on plot  $i$  $x_i$  = accurate estimate of cover on plot  $i$  $z_i$  = true, unobserved cover on plot  $i$  $\mu_i$  = mean soil water on plot  $i$ 

$$[\underline{\beta}, \underline{\mu}, \underline{b^2}, \underline{z}, \underline{\alpha}, S^2 | \underline{x}, \underline{y}, \underline{w}] \propto \prod_{i=1}^{100} \prod_{j=1}^5 [y_i | e^{\beta_0 + \beta_1 \mu_i + \beta_2 z_i}]$$

data model:

$$h(\alpha_0, \alpha_1, z_i) = \frac{e^{\alpha_0 + \alpha_1 z_i}}{1 + e^{\alpha_0 + \alpha_1 z_i}}$$

$$\times [w_{ij} | \mu_i, b_i^2]$$

$$\times [x_i | h(\alpha_0, \alpha_1, z_i), S^2]$$

$$\times [\mu_i] [b_i^2] [\beta_0] [\beta_1] [\beta_2] [\alpha_0] [\alpha_1] [S^2] [z_i]$$

process model

$$y_i \sim \text{Poisson} \left( e^{\beta_0 + \beta_1 \mu_i + \beta_2 z_i} \right) [\beta_2]$$

$$x_i \sim \text{beta} \left( m(h(\alpha_0, \alpha_1, z_i), S^2) \right) \rightarrow w_{ij} \sim \text{gamma} \left( \frac{\mu_i^2}{b_i^2}, \frac{\mu_i}{b_i^2} \right)$$

$$b_i^2 \sim \text{unif} (0, 100)$$

$$\beta_0 \sim \text{normal} (0, 10000)$$

$$\beta_1 \sim \text{normal} (0, 1000) \quad \beta_2 \sim \text{normal} (0, 1000)$$

$$\alpha_0 \sim \text{normal} (.05, .006)$$

$$\alpha_1 \sim \text{normal} (1.07, .13)$$

$$S^2 \sim \text{inverse gamma} (10.2, 630)$$

$$z_i \sim \text{uniform} (0, 1)$$

$$\mu_i \sim \text{uniform} (0, 2)$$

$m()$  is moment  
matching function  
returning parameters  
from moments

3. The function  $m()$  returns parameters of the beta distribution given moments as inputs.

Include the calibration equation in your model of effects of soil water and herbaceous cover on seedling establishment using informed priors on  $\alpha_0, \alpha_1$  and  $\varsigma^2$ . Hint—think about the predictor variable for herbaceous cover. Do you want to use the observed value of cover ( $x_i$ ) or the true value ( $z_i$ ) to model its effect on establishment?

4. Now presume that the 100 plots are arranged in 5 different stream reaches, 20 plots in each reach. You have data on peak runoff in each of the reaches, which you may assume is measured perfectly. Describe verbally how you might model variation at the catchment scale created by peak runoff.

You could allow each stream reach to have its own intercept (i.e.,  $\beta_{0,k}$ ), which you model as a linear or non-linear function of data of peak runoff.