

# Model Building Exercises

July 31, 2016

## 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:

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

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?
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, which you heroically assume is known without error. How would you model the effect of radon and soil type on the probability of lung cancer?
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  $j = 1, \dots, J$  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  $i_{\text{th}}$  species in the community, and  $R$  is the total number of species present? All estimates must include proper accounting for uncertainty.

4. You have data on the number of willow seedlings that establish on 100 different  $10 \times 10$  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  $1 \times 1$  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?