Practice Writing Hierarchical Models

May 31, 2018

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 social science, 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:

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

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 90 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 we 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 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. 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.

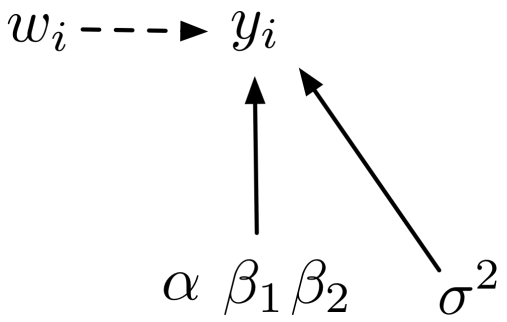
3 Problems

3.1 The Kuznets effect

You are interested in modeling the relationship between per capita income and an index of air pollution for $i = 1, ..., 80$ nations around the world. You hypothesize that air pollution increases then declines as per capita income increases (i.e., the Kuznets effect). Choose a deterministic model to represent this humped relationship, bearing in mind of course that air pollution cannot be negative.

3.1.1 A simple model

Start this problem by writing a model for the relationship between air quality for the i^{th} country measured as an air pollution index y_i, a continuous, non-negative response variable. The predictor variables is average per-capita income w_i. Assume both are measured perfectly.



$$g(\alpha, \boldsymbol{\beta}, w_i) = e^{\alpha+\beta_1 w_i+\beta_2 w_i^2}$$

$$\begin{aligned}
[\alpha, \boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}] &\propto \prod_{i=1}^{80} [y_i | g(\alpha, \boldsymbol{\beta}, w_i), \sigma^2][\alpha][\beta_1][\beta_2][\sigma^2] \\
y_i &\sim \text{gamma}\left(\frac{g(\alpha, \boldsymbol{\beta}, w_i)^2}{\sigma^2}, \frac{g(\alpha, \boldsymbol{\beta}, w_i)}{\sigma^2}\right) \\
\alpha &\sim \text{normal}(0, 10000) \\
\beta_1 &\sim \text{normal}(0, 10000) \\
\beta_2 &\sim \text{normal}(0,10000) \\
\sigma^2 &\sim \text{gamma}(.001, .001)
\end{aligned}$$

3.1.2 A hierarchical model with raw observations

Now imagine that you have a sample of n_i observations of the air pollution index in each country[1]. You also have a sample of annual incomes for 1000 individuals from each country. How would you model the effect of income on air pollution to include

[1]For now, we will ignore the possibility of spatial structure in the data within each country.

uncertainty in the response and the predictor?

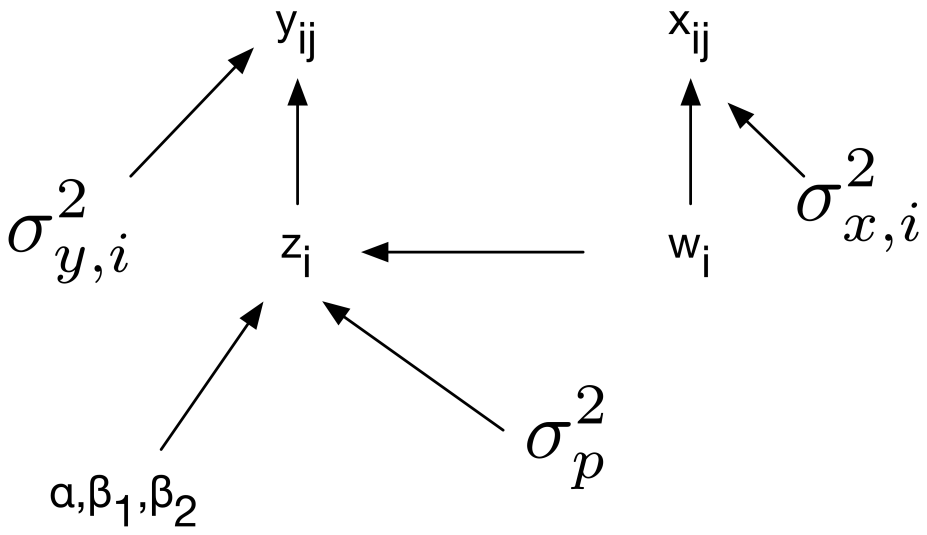


Figure 1: In this DAG, y_{ij} is the j^{th} observation of air quality in the i^{th} country ; x_{ik} is the k^{th}observation of per capita income in the i_{th} country; z_i is the unobserved mean air quality index, and w_i is the unobserved mean per-capita income.

$$\begin{aligned}\left[\boldsymbol{z}, \boldsymbol{w}, \alpha, \boldsymbol{\beta}, \sigma_p^2, \boldsymbol{\sigma}_x^2, \boldsymbol{\sigma}_y^2 \mid \boldsymbol{y}, \boldsymbol{x}\right] \propto & \prod_{i=1}^{80} \prod_{j=1}^{n_i} \prod_{k=1}^{1000} [y_{ij}|z_i, \sigma_{y,i}^2]\left[z_i \mid g\big(\alpha, \boldsymbol{\beta}, w_i\big), \sigma_p^2\right]\left[x_{ik} \mid w_i, \sigma_{x,i}^2\right] \\ & \times \left[w_i\right]\left[\alpha\right]\left[\beta_1\right]\left[\beta_2\right]\left[\sigma_p^2\right]\left[\sigma_{x,i}^2\right]\left[\sigma_{y,i}^2\right]\end{aligned}$$

$$
\begin{aligned}
g\big(\alpha, \boldsymbol{\beta}, w_i\big)\big) &= e^{\alpha+\beta_1 w_i+\beta_2 w_i^2} \\
z_i &\sim \text{gamma}\left(\frac{g\big(\alpha, \boldsymbol{\beta}, w_i\big)^2}{\sigma_p^2}, \frac{g\big(\alpha, \boldsymbol{\beta}, w_i\big)}{\sigma_p^2}\right) \\
y_{ij} &\sim \text{gamma}\left(\frac{z_i^2}{\sigma_{y,i}^2}, \frac{z_i}{\sigma_{y,i}^2}\right) \\
x_{ik} &\sim \text{gamma}\left(\frac{w_i^2}{\sigma_{x,i}^2}, \frac{w_i}{\sigma_{x,i}^2}\right) \\
w_i &\sim \text{gamma}\big(.001, .001\big)
\end{aligned}
$$

$$
\begin{aligned}
\alpha &\sim \text{normal}\big(0, 10000\big) \\
\beta_1 &\sim \text{normal}\big(0, 10000\big) \\
\beta_2 &\sim \text{normal}\big(0, 10000\big) \\
\sigma_p^2 &\sim \text{gamma}\big(.001, .001\big) \\
\sigma_{x,i}^2 &\sim \text{gamma}(.001, .001) \\
\sigma_{y,i}^2 &\sim \text{gamma}(.001, .001)
\end{aligned}
$$

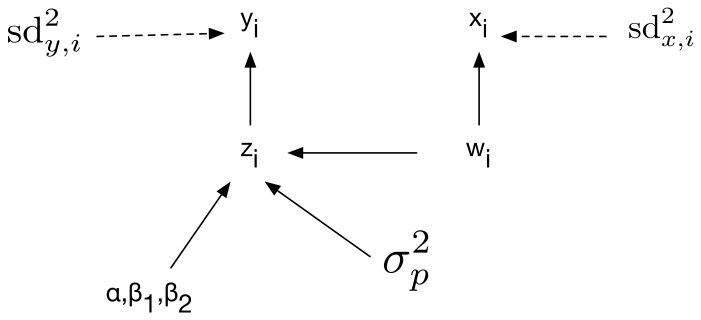
We used an exponentiated, quadratic model to represent our hypothesis to assert that the prediction of pollution is a humped function of income and is strictly non-negative. An un-exponentiated, quadratic model would have been a reasonable alternative, but you would need to be careful to constrain it to values greater than or equal to zero.

Why does the **x** now appear on the right hand side of the conditioning in the posterior distribution? Because it is a random variable before it is observed and fixed after, just like the **y**. The **w** appears on the left hand side because it is an unobserved, latent state – the mean per-captita income.

3.1.3 A hierarchical model with data summaries

Rewrite you model assuming that you didn't have the raw data, but rather the mean and the standard deviation of the mean for the air pollution index and annual income for each country.

The model would be modified as follows if we didn't have the raw data but did have data summaries. In this case, x_i is the observed income for country i and $\text{sd}_{x,i}$ is the standard deviation of the mean income; y_i is the observed air pollution index and $\text{sd}_{y,i}$ is the observed standard deviation of the mean air pollution.

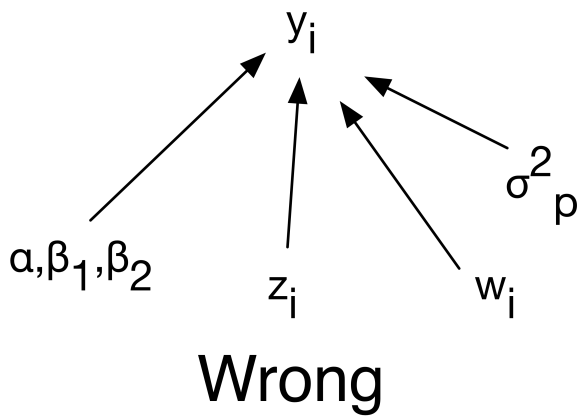


$$\left[\boldsymbol{z}, \boldsymbol{w}, \alpha, \boldsymbol{\beta}, \sigma_p^2, \mid \boldsymbol{y}, \mathbf{x}\right] \propto \prod_{i=1}^{80} [y_i | z_i, \text{sd}_{y,i}][z_i \mid g\big(\alpha, \boldsymbol{\beta}, w_i\big), \sigma_p^2]\left[x_i \mid w_i, \text{sd}_{x,i}\right] \times \left[w_i\right]\left[\alpha\right]\left[\beta_1\right]\left[\beta_2\right]\left[\sigma_p^2\right]$$

It is exceedingly important to think about how parameters are scaled when we use values taken from the literature, or values that are summarized as illustrated here. For example, we could have moment matched the lognormal distribution for z_i, w_i and y_i, but we must be careful to moment match for *both* parameters when using the standard deviations of the means. Matching for the mean alone will give the wrong answer (badly wrong). This is to say that moment matching for the first parameter using the log of the median would not work. Why? Because the second parameter is on the log scale and your standard deviations are on the exponential scale.

You might be tempted to use the data to put informative priors on w_i and z_i as in the incorrect Bayesian network below. This just doesn't work because now the y_i are arising from conflicting distributions, one with parameters $z_i, sd^2_{y,i}$ and the other with parameters $g(\alpha, \boldsymbol{\beta}, w_i), \sigma^2$, leading to a violation of the chain rule of probability because the y_i appear twice on the left hand side of conditioning. You could not fit

this model.



3.2 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?

    (a) What likelihood would you use for these 0 or 1 data?

    (b) 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.

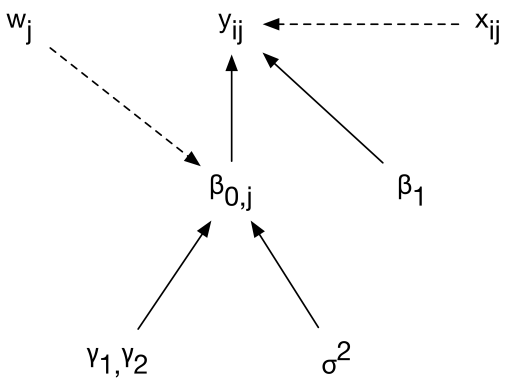


Figure 2: In this DAG, x_{ij} is the radon level and y_{ij} is an indicator that equals 1 if cancer is present and 0 if it is not in the i_{th} house in the j_{th} county, and w_{th} is the clay soil content in the j_{th} county.

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

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

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

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

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

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

We will learn later in the course that the prior of $\beta_1 \sim \text{normal}(0, 1000)$ has a substantially greater influence than we might like when we have a small number of observations. Stay tuned.

3.3 Controls on willow seedling establishment

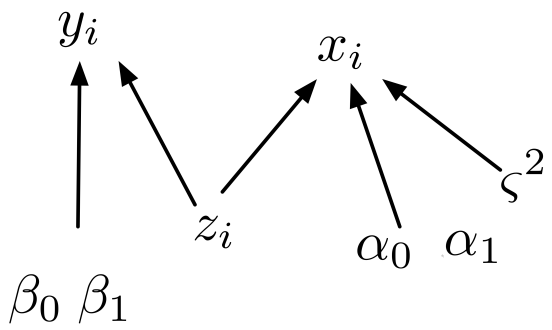
You are interested in the way that herbaceous plant cover influences establishment of willow seedlings in riparian communities. You have data on the number of willow seedlings that establish on 100 10 $\times$ 10 meter plots. Cover is estimated visually on each plot, creating a calibration problem – the visual estimate of cover is not the true cover. To deal with this problem, you obtained visual estimates of cover paired with the actual proportion of vegetated area (measured using many 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 $h(\boldsymbol{\alpha}, z_i)$:

$$\begin{aligned} h(\boldsymbol{\alpha}, z_i) &= \frac{e^{\alpha_o + \alpha_1 z_i}}{1 + e^{\alpha_o + \alpha_1 z_i}} \\ x_i &\sim \text{beta}\big(m(h(\boldsymbol{\alpha}, z_i), \varsigma^2)\big) \\ \alpha_o &\sim \text{normal}(.05,.006) \\ \alpha_1 &\sim \text{normal}(1.07, .13) \\ \varsigma^2 &\sim \text{inverse gamma}(10.2, 630) \end{aligned}$$

The function $m(\,)$ returns parameters of the beta distribution given moments as inputs. Note that this exercise provided informed priors on all of the unobserved quantities.

Write a model of willow establishment as a function of cover that models the observation process and the ecological process separately. Hints–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? Use informed prior distributions on α_0, α_1 and ς^2 to relate the true, unobserved cover to the observations of cover.



$$
\begin{aligned}
g(\boldsymbol{\beta}, z_i) &= e^{\beta_0 + \beta_1 z_i} \\
[\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}, \varsigma^2 \mid \mathbf{y}, \mathbf{x}] &\propto \prod_{i=1}^{100} [y_i \mid g(\boldsymbol{\beta}, z_i)][x_i | h(\boldsymbol{\alpha}, z_i), \varsigma^2][\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{z}, \varsigma^2] \\
y_i &\sim \text{Poisson}(g(\boldsymbol{\beta}, z_i)) \\
x_i &\sim \text{beta}\big(m(h(\boldsymbol{\alpha}, z_i), \varsigma^2)\big) \\
&\quad \times \text{appropriate priors}
\end{aligned}
$$

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., $\beta_{0,j}$), which you model as a linear or non-linear function of data of peak runoff.

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

1. You seek the marginal posterior distributions of $\boldsymbol{\phi}$, a vector consisting of elements ϕ_i, the proportions of the i^{th} species in the community, $i = 1, ..., R$. You will need to compose a likelihood for the vector of observed counts of species on each plot $\mathbf{y}_j$ conditional on $\boldsymbol{\phi}$ and n_j.

2. Take a look at the Dirichlet distribution as a way to form an prior on the vector $\boldsymbol{\phi}$. 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 ϕ_i and R, which will allow us to obtain a posterior distribution for H because any quantity that is a function of a random variable becomes a random variable in Bayesian analysis.

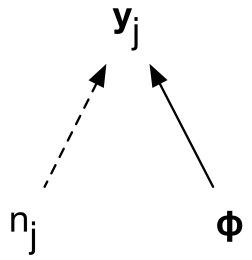


Figure 3: In this DAG, y_{ij} is the number of individuals in the i_{th} species observed in the j_{th} plot while n_j is the total number of individuals across all species observed in the j_{th} plot.

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

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

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

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

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