Model Building Exercises

February 17, 2017

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:

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

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 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 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)[1], 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 require on this one. It's a warmup.

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, 3/1. We will go over the problems in

[1]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 χ_i.

class on Tuesday 3/7.

- 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 The Kuznets effect

You are interested in modeling the relationship between per capita income and an index of air pollution for 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. The response, an air pollution index, and the predictor variable, income, are based on a sample of observations from each country. You have data on the mean (and the standard deviation of the mean) for each country's air pollution index, a continuous, non-negative response variable. You also have a mean and a standard deviation of the mean for each county's income, which is also continuous and non-negative. How would you model the effect of income on air pollution to include uncertainty in the response and the predictor?

Hint – Think of the response and predictor data as arising from distributions of means with known standard deviations. You want to use the unobserved means of those distributions in your model of the Kuznets effect.

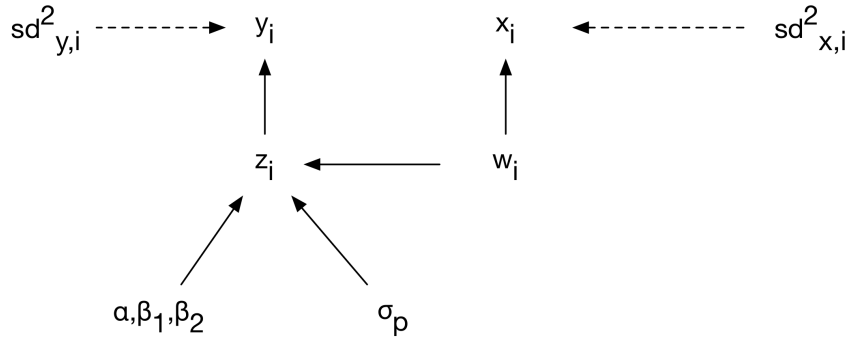


Figure 1: noneIn this DAG, y_i and $sd_{y,i}$ and x_i and $sd_{x,i}$ and are the observed means (and standard deviations of the means) of air quality and per capita income in the i_{th} country. The observed y_i and x_i are random variables drawn from distributions with unobserved mean z_i for pollution and unobserved mean w_i for income. We assume the standard deviations of those distributions are known as $sd_{y,i}$ and $sd_{x,i}$.

$$\begin{aligned}\big[\boldsymbol{z}, \boldsymbol{w}, \alpha, \boldsymbol{\beta}, \sigma_p \mid \boldsymbol{y}, \boldsymbol{x}\big] \propto &\prod_{i=1}^{n} \big[z_i \mid g\big(\alpha, \boldsymbol{\beta}, w_i\big), \sigma_p^2\big] \big[x_i \mid w_i, sd_{x,i}\big] \big[y_i \mid z_i, sd_{y,i}\big] \\ &\times \big[w_i\big] \big[\alpha\big] \big[\beta_1\big] \big[\beta_2\big] \big[\sigma_p\big]\end{aligned}$$

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

$$z_i \sim \text{gamma}\left(\frac{g\big(\alpha, \boldsymbol{\beta}, w_i\big)^2}{\sigma_p^2}, \frac{g\big(\alpha, \boldsymbol{\beta}, w_i\big)}{\sigma_p^2}\right)$$

$$y_i \sim \text{gamma}\left(\frac{z_i^2}{sd_{y,i}^2}, \frac{z_i}{sd_{y,i}^2}\right)$$

$$x_i \sim \text{gamma}\left(\frac{w_i^2}{sd_{x,i}^2}, \frac{w_i}{sd_{x,i}^2}\right)$$

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

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

$$\beta_1 \sim \text{normal}\big(0, 10000\big)$$

$$\beta_2 \sim \text{normal}\big(0, 10000\big)$$

$$\sigma_p \sim \text{gamma}\big(.001, .001\big)$$

Notes: We use 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. A linear model (not exponentiated) would have been a reasonable alternative. 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 in this case. 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. We also could have used models like

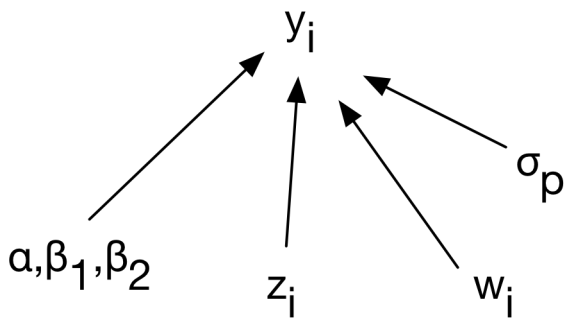
$$\log(z_i) \sim \text{normal}\left(\log(g\big(\alpha, \boldsymbol{\beta}, w_i\big)), \sigma_p^2\right) \quad (1)$$

$$\log(y_i) \sim \text{normal}(\log(z_i), sd_{y,i}^2) \quad (2)$$

$$\log(x_i) \sim \text{normal}(\log(w_i), sd_{x,i}^2) \quad (3)$$

$$(4)$$

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 $z_i, sd_{y,i}^2$ 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.



Wrong

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 know 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.

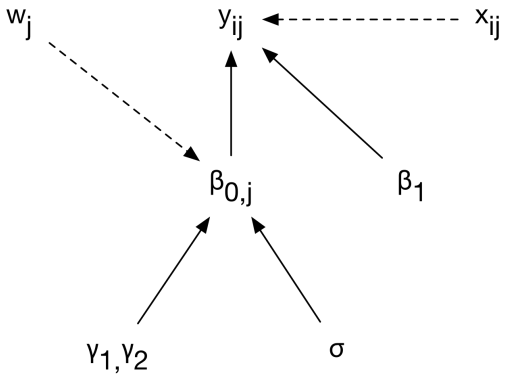


Figure 2: In this DAG, x_{ij} is the radon level and y_{ij} is an indicator that equals 1 if cancer is present and 0 if it is not in the i_{th} house in the j_{th} county, and w_{th} is the clay soil content in the j_{th} county.

Delete i

Add j subscript

$$[\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma \mid \boldsymbol{y}] \propto \prod_{i=1}^{1000} \prod_{j=1}^{40} [y_{ij} \mid g(\boldsymbol{\beta}, x_{ij})][\beta_0 \mid h(\boldsymbol{\gamma}, w_j), \sigma^2][\boldsymbol{\gamma}][\beta_1][\sigma]$$

Where is the process variance for effect of radon? Why not Binomial?

$$g(\boldsymbol{\beta}, x_{ij}) = \frac{e^{\beta_{0,ij}+\beta_1 x_{ij}}}{1 + e^{\beta_{0,ij}+\beta_1 x_{ij}}}$$

$$h(\boldsymbol{\gamma}, w_j) = \gamma_0 + \gamma_1 w_j$$

$$y_{ij} \sim \text{Bernoulli}\big(g(\boldsymbol{\beta}, x_{ij})\big)$$

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

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

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

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

$$\sigma \sim \text{uniform}\big(0, 1000\big)$$

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 $j = 1, \ldots, J$ 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 $H = -\sum_{i=1}^{R} \phi_i \log(\phi_i)$, 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. Model the observed count data as a random variable (a vector) arising from the unobserved vector $\boldsymbol{\phi}$ of proportions.

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.

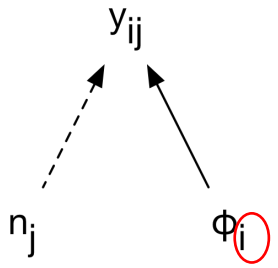


Figure 3: In this DAG, y_{ij} is the number of individuals in the i_{th} species observed in the j_{th} plot while n_j is the total number of individuals across all species observed in the j_{th} plot.

subscript is needed because it refers to species i. Could also be written by dropping the i subscripts from y and making \phi bold, indicating that both are vectors, , which I actually like better.

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

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

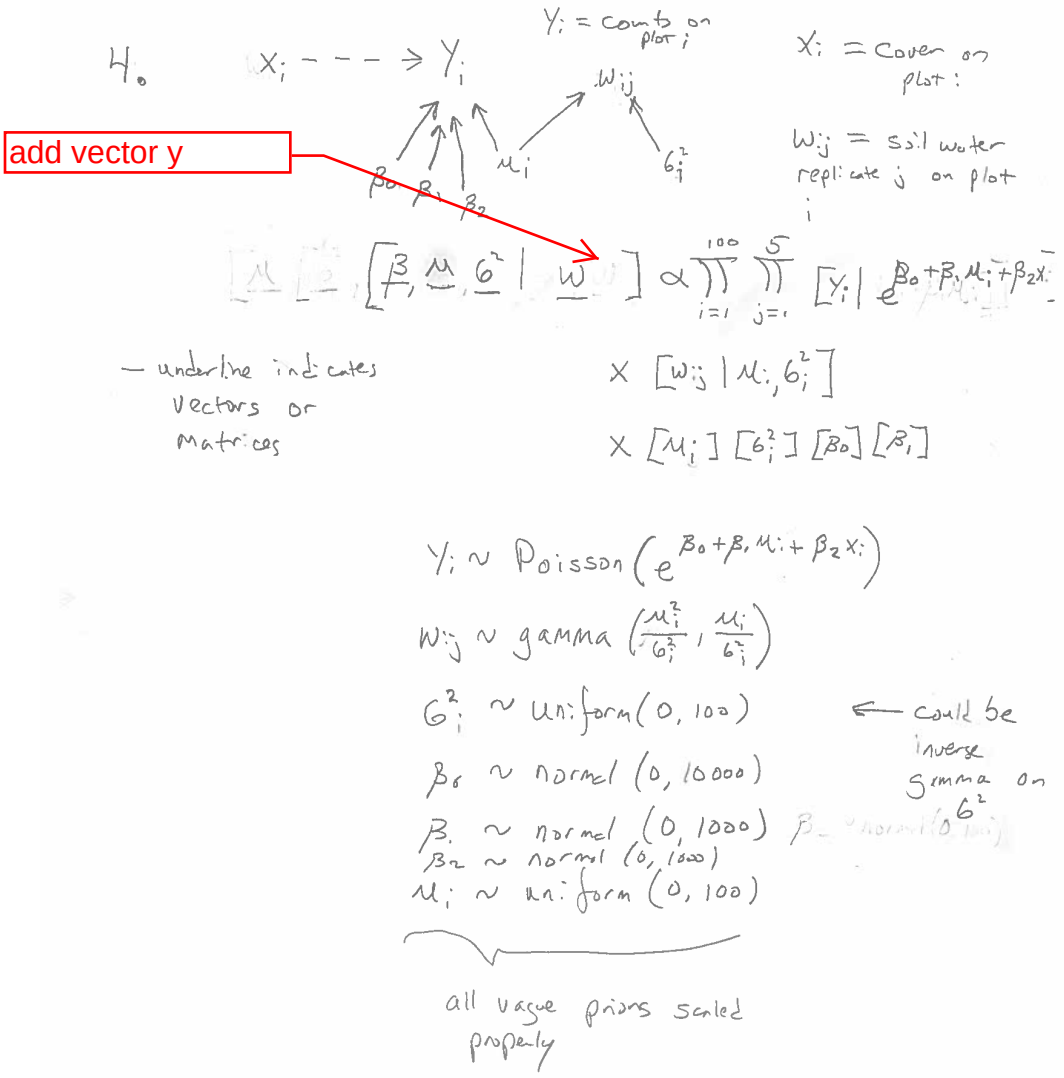
$$\boldsymbol{y}_j \sim \text{multinomial}(n_j, \boldsymbol{\phi})$$

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

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. 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?


H_0
$x_i \dashrightarrow Y_i$
Y_i = counts on plot i
x_i = cover on plot i
w_{ij}
w_{ij} = soil water replicate j on plot i
β_0 β_1 β_2
μ_i
σ_i^2
add vector y
$[\beta, \underline{\mu}, \underline{\sigma^2} \mid \underline{w}] \propto \prod_{i=1}^{100} \prod_{j=1}^{5} [Y_i \mid e^{\beta_0 + \beta_1 \mu_i + \beta_2 x_i}]$
— underline indicates vectors or matrices
$\times [w_{ij} \mid \mu_i, \sigma_i^2]$
$\times [\mu_i] [\sigma_i^2] [\beta_0] [\beta_1]$
$Y_i \sim \text{Poisson}(e^{\beta_0 + \beta_1 \mu_i + \beta_2 x_i})$
$w_{ij} \sim \text{gamma}\left(\frac{\mu_i^2}{\sigma_i^2}, \frac{\mu_i}{\sigma_i^2}\right)$
$\sigma_i^2 \sim \text{uniform}(0, 100)$
← could be inverse gamma on σ^2
$\beta_0 \sim \text{normal}(0, 10000)$
$\beta_1 \sim \text{normal}(0, 1000)$
$\beta_2 \sim \text{normal}(0, 1000)$
$\mu_i \sim \text{uniform}(0, 100)$
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

Should be z_i !!!

$$h(\boldsymbol{\alpha}, x_i) = \frac{e^{\alpha_o + \alpha_1 z_i}}{1 + e^{\alpha_o + \alpha_1 z_i}} \quad (5)$$

$$x_i \sim \text{beta}\big(m(h(\boldsymbol{\alpha}, z_i), \varsigma^2)\big) \quad (6)$$

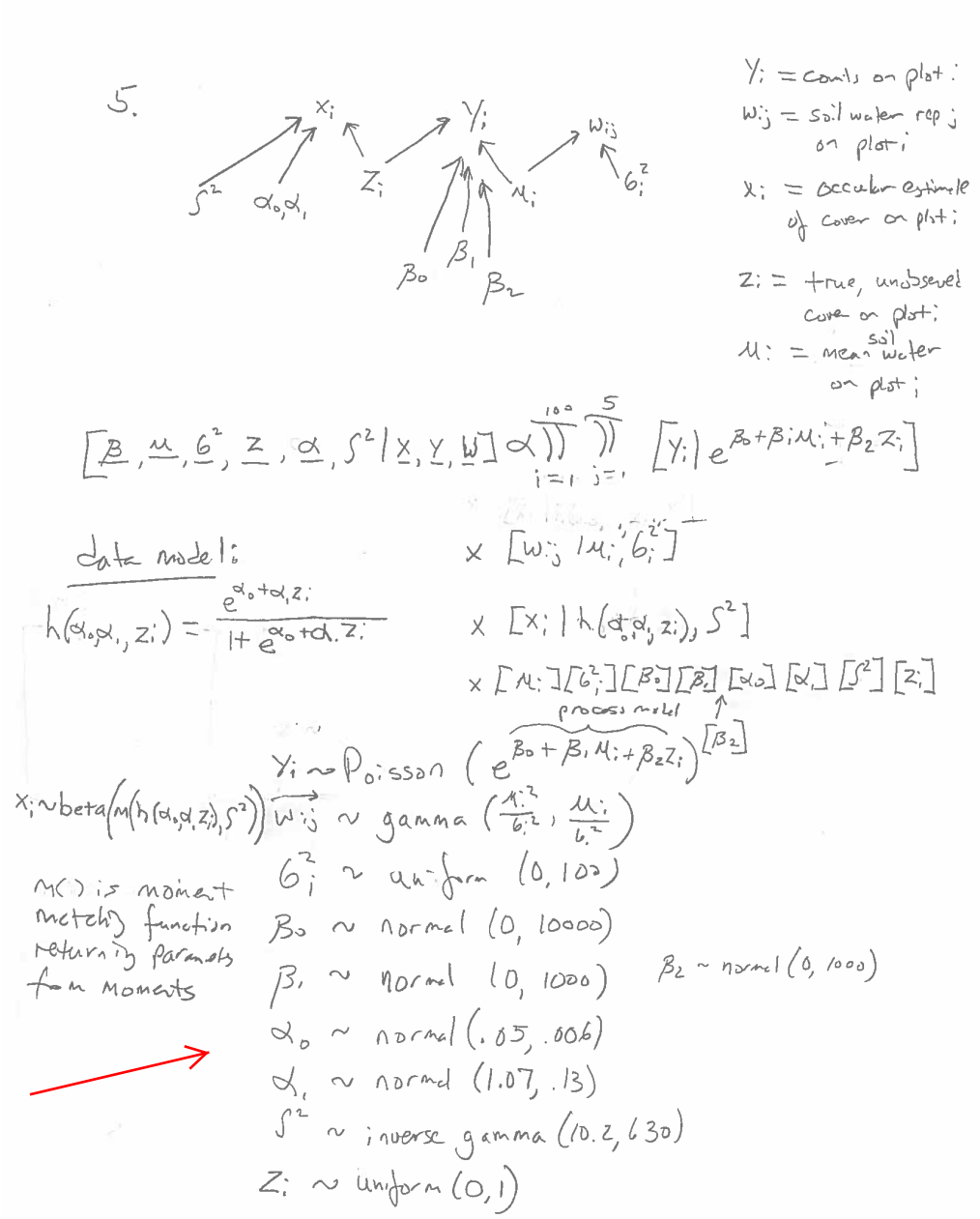
Talk about why we regress observed on true rather than true on observed.

$$\alpha_o \sim \text{normal}(.05,.006) \quad (7)$$

$$\alpha_1 \sim \text{normal}(1.07, .13) \quad (8)$$

$$\varsigma^2 \sim \text{inverse gamma}(10.2, 630) \quad (9)$$

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 α_0, α_1 and ς^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 ef-



What prior is missing here?

fect on establishment?

3. 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. *s*kip 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.