Chapter 6

Hierarchical Bayesian Models

It is worthwhile to review the key points covered thus far. We started with the first principles rules of probability (Section 3.2). We used those rules to develop Bayes theorem (Section 5.1) and to show how we can factor joint distributions of observed and unobserved quantities into parts based on our knowledge of conditioning and independence (Section 3.3). We learned about priors and their influence on the posterior (Section 5.4).

We now apply what we have learned to ecological examples of hierarchical Bayesian models. These models offer unusually revealing and broadly useful routes to insight because they allow us to decompose complex, high-dimensional problems into parts that can be thought about and analyzed individually. We can use the same approach for virtually any problem, regardless of its particular features.

This chapter has two objectives: 1) To explain hierarchical models and how they differ from simple Bayesian models; 2) To illustrate building hierarchical models using mathematically correct expressions. We illustrate the first two sets of steps in the general modeling process that we introduced in the Preface (Figure 0.0.1 A, B).

We begin with the definition of hierarchical models. Next we introduce four, general classes of hierarchical models that have broad application in ecology. These classes can be used individually or in combination to attack virtually any research problem. We use examples to show how to draw Bayesian networks that portray stochastic relationships between observed and unobserved quantities. We show how to use the network drawings as a guide for writing out posterior and joint distributions.

6.1 What is a hierarchical model?

A statistical model is Bayesian if the unobserved quantities we seek to understand are random variables whose probability distributions are estimated from observations and prior knowledge.[1] Recall from Chapter 5 that a Bayesian model is simple if it represents the joint distribution of those random variables as the product of the likelihood multiplied by the prior distributions. For example,

$$\left[\underbrace{\theta_1, \theta_2, z}_{\text{unobserved}} \mid \underbrace{y}_{\text{observed}}\right] \propto \underbrace{[\theta_1, \theta_2, z, y]}_{\text{joint}} \tag{6.1.1}$$

$$\propto \underbrace{[y|\theta_1, \theta_2, z]}_{\text{likelihood}} \underbrace{[\theta_1]\,[\theta_2]\,[z]}_{\text{priors}} \tag{6.1.2}$$

is a simple Bayesian model of the unobserved quantities θ_1, θ_2, and z, and the observations y.[2] It is important to remember that we factor the joint distribution using the rules of probability (Section 3.3), to obtain the product of the likelihood and priors. It is not hierarchical because there is no conditioning beyond the dependence of the data on the unobserved quantities. This means that every quantity that appears on the right hand side of the conditioning symbol in the likelihood is found in a prior. The posterior distribution is proportional to the joint because we have the omitted denominator of Bayes theorem, the marginal distribution of the data $\left(\int_{\boldsymbol{\theta},z} [y|\theta_1, \theta_2, z]\,[\theta_1]\,[\theta_2]\,[z]\,d\theta_1 d\theta_2 dz\right)$, which is a scalar after we have observed the data. At the risk of getting ahead of ourselves, we are expressing the posterior as being proportional to the joint distribution because this proportionality is all we need to do to properly develop an algorithm for estimating the parameters and latent state, which we will cover in the Chapter on the Markov chain Monte Carlo algorithm (Chapter 7).

A Bayesian model is *hierarchical* whenever we use probability rules for factoring (Section 3.3) to express the joint distribution as a product of conditional distributions. For example,

$$\begin{aligned}[\theta_1, \theta_2, z|y] &\propto [\theta_1, \theta_2, z, y] \\ &\propto [y|\theta_1, z]\,[z|\theta_2]\,[\theta_1]\,[\theta_2]\end{aligned} \tag{6.1.3}$$

is hierarchical because we have factored $[\theta_1, \theta_2, z, y]$ to become $[y|\theta_1, z]\,[z|\theta_2]\,[\theta_1]\,[\theta_2]$, assuming that

[1] Including the "knowledge" that nothing is known.

[2] Strictly speaking this assumes that θ_1, θ_2, and z are independent *a priori*. This is a common assumption in Bayesian models. Inference is rarely sensitive to this assumption.

θ_1 and θ_2 are independent *a priori*. We can quickly see that the model is hierarchical because the unobserved quantity z appears on the right hand side of the "|" in the likelihood $[y|\theta_1, z]$ and on the left hand side of the "|" in the distribution $[z|\theta_2]$. Note that there is no prior distribution[3] for z because it is conditional on a quantity for which there *is* a prior distribution, θ_2. The factoring of joint distributions into products of conditional distributions is not arbitrary, but rather is based on our knowledge of an ecological process, how we observe it, and the assumptions we can use to simplify it, as we illustrate below.

6.2 Example hierarchical models

Hierarchical models are most often applied in ecological research to deal with four commonly encountered challenges:

1. Representing variation among individuals arising, for example, from genetics, location, or experience.

2. Studying phenomena operating at more than one spatial scale or level of ecological organization.

3. Estimating uncertainty that arises from modeling a process as well as uncertainty that results from imperfect observations of the process.

4. Understanding changes in states of ecological systems that cannot be observed directly. These states arise from "hidden" processes.

These broad challenges are not mutually exclusive; more than one often arises within the same investigation. Hierarchical models can be used to create a robust and flexible framework for analysis that is capable of meeting these challenges as they arise.

In the following examples we illustrate different types of hierarchical models. At the same time we show how to graphically represent relationships between observed and unobserved quantities in Bayesian networks, also called directed acyclic graphs (DAGs), a concept introduced in Section 3.3. Bayesian networks form a template for writing out properly factored expressions for joint distributions. Our purpose in this chapter is to emphasize writing out mathematical expressions

[3] Some would call $[z|\theta_2]$ a hierarchical prior and $[\theta]$ a hyper prior, but this perspective is somewhat unconventional.

as the proper first step in modeling. For now, we will avoid considering how we might implement or evaluate the model, which will come later. The examples we offer here will be supplemented by worked problems in model building in Part III, problems that will challenge you to diagram and write models.

As you read the following sections, it will be especially useful to notice four themes that reoccur in the examples. The first theme is the one-to-one relationship between diagrams of stochastic relationships and the mathematical expressions for the posterior and joint distributions. This is a critical insight. Next, it will be useful to see how we compose stochastic models by combining deterministic functions with probability distributions. Hierarchical models are often developed by substituting a model for a parameter – it is especially instructive to see how we add detail models and exploit additional explanatory data by "modeling parameters." This process illustrates how models of high dimension can be composed, even though the examples here are relatively simple. The final crosscutting theme in the examples is how we partition uncertainty into multiple sources. In particular, we will often use a particular factoring of the joint distribution first proposed by Berliner (1996) and later elaborated by Wikle (2003); Clark (2005); Cressie et al. (2009), and Wikle et al. (2013),

$$[\theta_p, \theta_d, \mathbf{z}|y] \propto \underbrace{[y|\mathbf{z}, \theta_d]}_{\text{Data}} \underbrace{[\mathbf{z}|\theta_p]}_{\text{Process}} \underbrace{[\theta_d]\,[\theta_p]}_{\text{Parameters}} . \tag{6.2.1}$$

We decompose the joint distribution this way because it represents such a broad range of problems in ecological research. There is a "true" ecological state of interest $\mathbf{z}$, a state that is not observable. We relate that state to the observable data, $\mathbf{y}$, using a model with a vector of parameters θ_d, including parameters representing uncertainty in our observing system. The behavior of the true state is predicted with a model parameterized by θ_p including parameters representing stochastically in the process.[4] This model represents our hypothesis about how an ecological process works.

6.2.1 Understanding individual variation: fecundity of spotted owls

Understanding variation in processes caused by variation among individual organisms forms a central challenge in population and community ecology. Our first example is fashioned from Clark (2003a) who studied the effects of individual differences in fecundity on population growth rate of northern

[4] You may wonder, "Where's the x? What happened to observations of predictor variables?" Suspend disbelief for a moment. We will deal with this question in the next section.

spotted owls, *Strix occidentalis caurina*. In this example, we are interested in estimating the average number of offspring annually produced by each breeding female, that is, their average fecundities, as well as the average fecundity for the population.

A simple Bayesian model requires the assumption that all owls have the same average fecundity. This means that variation among individuals occurs from year to year because fecundity is a random variable, but a sample of many years would have the same average reproductive output for all individuals. We can represent these ideas in a Bayesian network (Figure 6.2.1 A).

Recall that Bayesian networks (Figure 3.3.1), are drawings that depict probability distributions graphically. These drawings are particularly useful for showing the dependencies in hierarchical models. The nodes in the diagrams represent random variables; solid arrows in the diagrams represent stochastic relationships among the random variables. The tails of the arrows specify the parameters defining the distribution of the random variable at the heads of the arrows.

Here is an example. Assume we have an observation (y_i) representing the number of offspring of owl i . Before it is observed, the y_i arises from a probability distribution with a mean λ and variance σ_o^2. We can use the diagram (Figure 6.2.1 A) as a guide to formulate the simple Bayesian model,

$$\underbrace{[\lambda, \sigma_o^2 | y_i]}_{\text{posterior}} \propto \underbrace{[y_i | \lambda, \sigma_o^2]\,[\lambda]\,[\sigma_o^2]}_{\text{joint}}. \tag{6.2.2}$$

Writing out the posterior distribution is easy. We simply write down a distribution with the unobserved quantities on the left hand side of the "|" and the observed quantities on the right hand side. Composing expressions for the joint distribution (i.e., the likelihood multiplied by the priors) is guided by the diagram – the nodes at the heads appear on the left hand side of a "|" and the nodes at the tails of the arrows appear on the right hand side. Any node at the tail of an arrow that does not have an arrow coming into it is expressed as a prior distribution. The prior distributions must have numeric arguments for their parameters. Because the parameters of priors are constant (i.e., they are not random variables) they do not appear as nodes in the diagram. Remember, nodes represent random variables.

It may strike you that diagrams are superfluous when you are writing down simple Bayesian models and your impression is correct. However, these diagrams become more useful in helping us visualize and write down hierarchical relationships. They are especially helpful (at least for

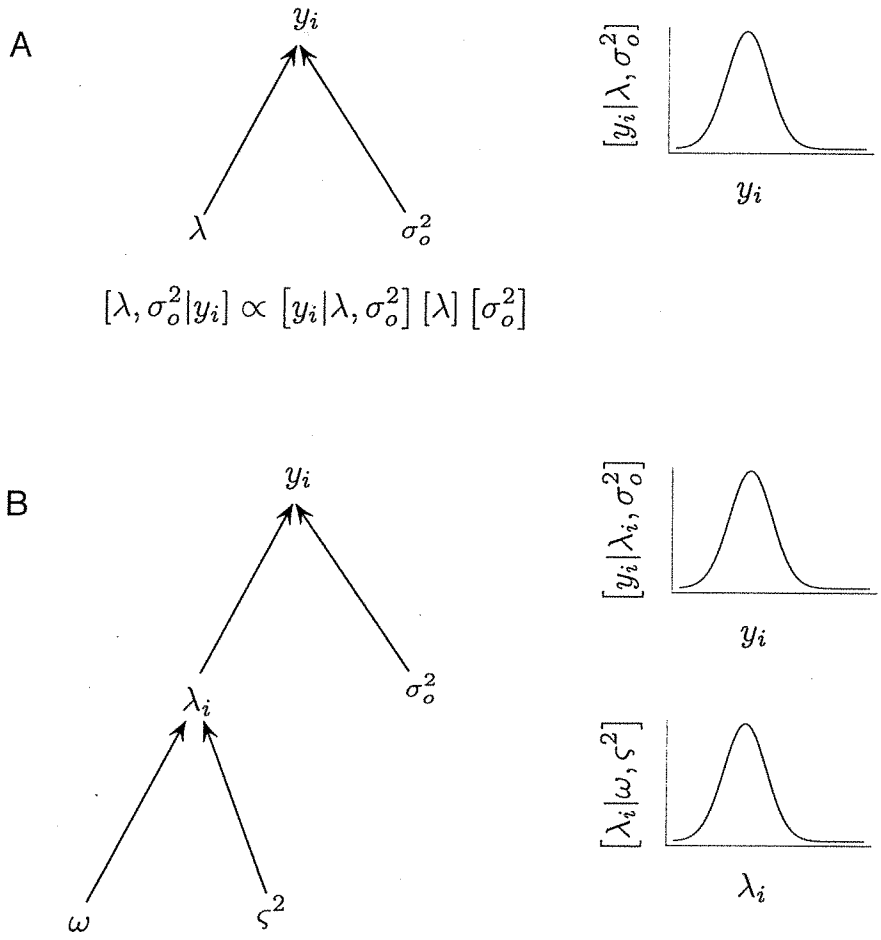


Figure 6.2.1: Bayesian networks for simple (**A**) and hierarchical (**B**) Bayesian models of fecundity of spotted owls. There are only two levels in the simple model (**A**) because the joint distribution is a product of the likelihood and the priors. In this case, we assume the data arise from a mean fecundity (λ) that is the same for all owls. The only variation is due to sampling, represented by σ_o^2. There are three levels in the hierarchical model (B) because the joint distribution is a product of two conditional distributions and the priors. In this case, we assume that each owl has its own average fecundity (λ_i) that is drawn from a distribution with mean ω and variance ς^2. Note the correspondence between the heads of arrows and random variables on the left hand side of conditioning symbols in the joint distribution and the tails of arrows and random variables on the right hand side of conditioning symbols. Any random variable at the tail of an arrow without an arrow leading into requires a prior distribution. The equations and the diagrams represent distributions (right column) where the heads of the arrows are the random variables shown on the x-axis and the tails of the arrows are the moments (or the parameters) that define the distributions.

ecologists, if not statisticians) when there are complex, multi-level relationships among observed and unobserved quantities, as we will soon see.

We now model the case where *each* owl has its *own* mean fecundity. Variation in average fecundity among individuals might occur because of differences in genetics or age or variation in the quality habitats where they establish territories. In this example, we are not trying to determine the causes of individual variation, but simply to acknowledge that it exists and to include it in our model. This is a key idea.

Consider a network with an additional level in the hierarchy (Figure 6.2.1 B). We now treat the average fecundity of each of individual (λ_i) as a random variable drawn from a distribution mean ω and variance ς^2. The observation y_i comes from a distribution of the annual fecundities of each individual. Note the subscript on λ_i indicating that each individual has a fecundity – the observations for owl i will vary from year to year, but over the long term the observations on owl i will average λ_i . Assuming individual owls have fecundities that are drawn from a distribution treats fecundity as a *random effect* where as assuming all individuals have the same average fecundity treats fecundity as a *fixed effect* (Box 6.2.1).

Box 6.2.1 Random effects

The terms *random effect* and *fixed effect* are used in the scientific literature in ways that can be confusing. Gelman and Hill (2009, page 245) offer several examples of inconsistent use of the terms random and fixed effects. They recommend dispensing with the use of the term "random effects" all together, replacing it with group level effects. This is a sensible suggestion because all "effects" are considered to be random variables in the Bayesian framework. However, "random effects" is widely used, sometimes pertaining to individuals rather than groups. We will use the term later in the book and explain it here.

In Bayesian hierarchical modeling, random effects are used to describe variation that occurs beyond variation that arises from sampling alone. Here is an example. Imagine that you want to estimate the average aboveground biomass in a grassland. You take a sample of biomass in several .25 m^2 plots. If the biomass is randomly distributed across the area you sample, then a

reasonable way to model the variation in the biomass in the i^{th} plot (y_i) would be

$$\begin{aligned} y_i &= \mu + \epsilon_i \\ \epsilon_i &\sim \text{normal}(0, \sigma^2), \end{aligned}$$

which is the same as

$$y_i \sim \text{normal}(\mu, \sigma^2),$$

where μ is the mean biomass per plot and σ^2 is the variance among plots. We generally prefer the latter notation because not all variation is additive. If a random variable like μ is strictly positive, then adding a random variable to it to represent uncertainty makes no sense because it cannot have a mean of 0. Alternatively, the notation $[y_i|\gamma, \beta]$is a probability mass or probability density function and γ and β are parameters works for any random variable, regardless of its support. We are using a normal distribution for clarity here, but because biomass is strictly positive, a better choice might be lognormal or gamma. However, this somewhat complicates the example, so to keep things simple and familiar, we chose the normal.

Now imagine that you sampled at five different locations, indexed by j. If we treat location as a *fixed effect*, our model doesn't change because we assume that the variation is due entirely to sampling, i.e., $y_{ij} \sim \text{normal}\left(\mu, \sigma^2\right)$. When we do this we are treating the μ as *fixed* across the locations. Alternatively, we might more reasonably assume that there are differences in productivity among sites arising from any number of different sources – soil type, depth to the water table, topography, level of herbivory and so on. In this case, we would allow each location to have its own mean biomass drawn from a distribution of means with *hyper-parameters* mean of means equal to α and variance of means equal to ς^2. Our model then becomes

$$y_{ij} \sim \text{normal}(\mu_j, \sigma_j^2) \tag{6.2.3}$$

$$\mu_j \sim \text{normal}(\alpha, \varsigma^2). \tag{6.2.4}$$

In this case, we are treating the effect of location as random – it varies randomly according to sources of variation that we acknowledge exist but that we are not attempting to explain. You

will also see this written as

$$y_{ij} = \mu_j + \epsilon_{ij} \tag{6.2.5}$$

$$\mu_j = \alpha + \eta_j \tag{6.2.6}$$

$$\epsilon_{ij} \sim \text{normal}\left(0, \sigma_j^2\right) \tag{6.2.7}$$

$$\eta_j \sim \text{normal}\left(0, \varsigma^2\right). \tag{6.2.8}$$

We have used the problem of estimating a mean to illustrate random effects, but the same idea applies to any parameter in any model. For example, a common use of random effects is to allow the intercepts of regressions to vary by location or some other grouping variable, e.g.,

$$y_{ij} \sim \text{normal}\left(\beta_j + \beta_1 x_{ij}, \sigma^2\right) \tag{6.2.9}$$

$$\beta_j \sim \text{normal}\left(\mu, \varsigma^2\right). \tag{6.2.10}$$

We used the diagram (Figure 6.2.1 B) as a template to write down the posterior and joint distributions,

$$\underbrace{\left[\lambda_i, \sigma_o^2, \omega, \varsigma^2 | y_i\right]}_{\text{posterior}} \propto \underbrace{\left[y_i | \lambda_i, \sigma_o^2\right] \left[\lambda_i, | \omega, \varsigma^2\right] \left[\sigma_o^2\right] \left[\omega\right] \left[\varsigma^2\right]}_{\text{joint}}. \tag{6.2.11}$$

Again, it is important to see the relationship between equation 6.2.11 and the Bayesian network representing the relationships the equation embodies (Figure 6.2.1 B).

We have come a long way toward writing out our complete model but we are not finished. We need to choose appropriate probability distributions for each of the random variables described with the bracket notation and we need to think about how to represent multiple observations (i.e., a vector, $\mathbf{y}$). Recall that the support for the random variable and its dispersion guide our choice of a distribution to represent the random variable. The random variable y_i is discrete – a female produces individual offspring – so the y_i are integers. If we assume that the variance in the y_i is approximately the same as the mean, then a logical choice for modeling the y_i is the Poisson

distribution,[5]

$$y_i \sim \text{Poisson}(\lambda_i).$$

The average fecundities (λ_i) are continuous, non-negative random variables, so a gamma distribution is a logical choice to model them. Moreover, ω and ς^2 are also continuous and strictly non-negative, so we use a gamma distribution for their prior distributions. Vague priors for these parameters are

$$\omega \sim \text{gamma}(.001,.001) \tag{6.2.12}$$

$$\varsigma^2 \sim \text{gamma}\,(.001,.001)\,. \tag{6.2.13}$$

These are reasonable choices if we have no previous knowledge of the distribution of the λ_i. To assemble our full model, we use the full data set ($\mathbf{y}$) consisting of n observations,

$$\begin{aligned}[\boldsymbol{\lambda}, \omega, \varsigma^2|\mathbf{y}] \propto & \prod_{i=1}^{n} \text{Poisson}\,(y_i|\lambda_i)\,\text{gamma}\left(\lambda_i \,\middle|\, \frac{\omega^2}{\varsigma^2}, \frac{\omega}{\varsigma^2}\right) \times \\ & \text{gamma}(\omega|.001,.001)\text{gamma}\left(\varsigma^2|.001,.001\right). \end{aligned} \tag{6.2.14}$$

Remember that taking a product across the individual likelihoods to estimate the total likelihood requires the assumption that the observations are independent.

There are two potential sources of confusion here, both of which are instructive. First, what happened to σ_o^2? In our original Bayesian network, the distribution of fecundities was governed by a mean and a variance–which makes sense because all random variables are drawn from distributions and distributions have means and variances.[6] However, there is no σ_o^2 in our hierarchical model (equation 6.2.14). Actually, there *is* a variance for y_i in equation 6.2.14 – recall that in the Poisson distribution the variance and the mean are assumed to be equal.

The second result that might be puzzling is seen in the parameters for the gamma distribution, $\frac{\omega^2}{\varsigma^2}$ and $\frac{\omega}{\varsigma^2}$. Where did these come from? The parameters for a gamma distribution are α, the shape, and β, the rate. Recall from the section on moment matching the mean of the gamma distribution is $\frac{\alpha}{\beta}$ with variance $\frac{\alpha}{\beta^2}$ (3.4.4), allowing us to solve for α and β terms of the means and variance, i.e.,

[5] If this assumption doesn't hold, then the negative binomial distribution would be a better choice. Later (8.1), we will learn methods to evaluate the assumptions we make in choosing distributions.

[6] There are exceptions to this generality. For example the means and variance of the Cauchy distribution are not defined. However, all of distributions we will use in this book have means and variances. See Appendix tables A.1 and A.2.

$\alpha = \frac{\omega^2}{\varsigma^2}, \beta = \frac{\omega}{\varsigma^2}$. The average fecundity for the population $\omega = \frac{\alpha}{\beta}$.

These clarifications make an important point about drawing Bayesian networks and converting them into mathematical expressions. Remember that the heads of arrows in Bayesian networks are random variables governed by a distribution defined by the parameters at the tails of the arrows (i.e., Figure 3.3.1). Thus, it is possible to define these distributions in terms of means and variances or in terms of parameters. It follows that it would have been perfectly correct[7] to write out the model as

$$[\lambda, \alpha, \beta|\mathbf{y}] \propto \prod_{i=1}^{n} \text{Poisson}\left(y_i|\lambda_i\right) \text{gamma}\left(\lambda_i|\alpha, \beta\right) \times \text{gamma}(\alpha|.001, .001)\text{gamma}\left(\beta|.001, .001\right). \quad (6.2.15)$$

The point is that Bayesian networks are thinking tools – graphical aids for properly writing out models. In some cases it will be most helpful to think about stochastic relationships in terms of the moments of distributions; in other cases it will be more useful to think in terms of parameters. Moment matching allows these approaches to be interchangeable. We can be flexible in our use of tools.

We now extend this example to illustrate how we might add parameters and explanatory observations (i.e., covariates) to our model to explain variation among individuals in fecundity. Reproductive success for many species of vertebrates rises to a peak during mid-life before declines (Part and Forslund, 1996; Hamel et al., 2012) as individuals grow old. Thus, it might be reasonable to model λ_i as a quadratic function of "reproductive age," defined as time after the animal is capable of reproduction, $x_i = \widetilde{x}_i - x_i^0$ where $\widetilde{x}_i$ is the chronological age of the i^{th} individual, x_i^0 is the age of first reproduction. Thus, an animal is $x_i = 0$ when it first reproduces. Defining age this way makes for a convenient interpretation of the intercept.

We now model of the process "change in fecundity with age" $g\left(\alpha, \beta, x_i\right)$ as,

[7] Actually, most statisticians would prefer this.

$$
\begin{aligned}
g\left(\alpha,\beta,x_i\right) &= \alpha+\beta_1 x_i+\beta_2 x_i^2 \\
\left[\lambda,\beta,\alpha,\sigma_p^2|\mathrm{y}\right] &\propto \prod_{i=1}^{n}\mathrm{Poisson}\left(y_i|\lambda_i\right)\mathrm{gamma}\left(\lambda_i\left|\frac{g\left(\alpha,\beta,x_i\right)^2}{\sigma_p^2},\frac{g\left(\alpha,\beta,x_i\right)}{\sigma_p^2}\right.\right)\times \\
&\quad \prod_{j=1}^{2}\mathrm{normal}\left(\beta_i|0,100\right)\mathrm{normal}(\alpha|0,100)\mathrm{gamma}(\sigma_p^2|.001,.001)
\end{aligned}
\tag{6.2.16}
$$

where α is the average reproduction of an owl at reproductive age 0, β_1 and β_2 are parameters that control the change in fecundity with age, and σ_p^2 is process variance. It is important to understand that process variance includes all of the influences that create variation in fecundity beyond the effect of the bird's age. It is important to see that we have replaced the parameter ω with a model $g\left(\alpha,\beta,x_i\right)$ that exploits observations on an owl's age and our understanding of the relationship between age and fecundity.

Again, we choose gamma distributions for λ_i and σ_p^2 because they are continuous and strictly positive. The distribution for λ_i could be viewed as a "prior" informed by our process model. Parameters of the other gamma distributions are chosen to be weakly informative. We choose a normal distribution for the βs because they are continuous random variables that can take on any real value. To minimize the information contained in the priors for the βs we center them on 0 and assign a variance that is very large relative to their values.

You might reasonably ask, "Why doesn't the data set $\mathbf{x}$ appear in the posterior distribution in the same way that $\mathbf{y}$ does? After all, both are observed quantities." The short answer is this. The $\mathbf{x}$ are not treated as random variables in this formulation. You are right that both the $\mathbf{x}$ and the $\mathbf{y}$ are observed, but in this case, we are assuming that the $\mathbf{x}$ data are observed *perfectly*[8], while the data $\mathbf{y}$ are random variables. This means the $\mathbf{x}$ are known, fixed quantities, treated no differently than the constant π in the normal distribution. They are not random variables and hence, they should not appear in the expression for the posterior distribution which, by definition, is composed of random variables. The predictor variables correctly appear as arguments to the deterministic function $g\left(\alpha,\beta,x_i\right)$. There are cases when the predictor variables *do* appear in the posterior distribution and we will describe these cases in a subsequent example and in Box 6.2.2.

[8]You may recall the assumption of conventional linear regression, customarily but often wrongly ignored by ecologists, that the predictor variables are measure without error.

What else might influence fecundity? We might reasonably hypothesize that the fecundity of each owl at reproductive age 0 should increase with decreasing territory size (e.g., Elbroch and Wittmer, 2012), which is to say that territory size shifts the curve $g(\alpha, \beta, x_i)$ up or down. A reasonable deterministic model of this process is $h(\gamma, \nu, u_i) = \gamma e^{-\nu u_i}$, where u_i is the observed area of the territory of the i^{th} individual; γ is the maximum potential fecundity during the first reproduction; ν controls the decline in mean fecundity that occurs as territory area increases. We can include this process in our model by allowing each individual to have a different intercept in the "change in fecundity with age" model $g(\alpha_i, \beta, x_i)$ where

$$\alpha_i \sim \text{gamma}\left(\frac{h(\gamma, \nu, u_i)^2}{\varsigma_p^2}, \frac{h(\gamma, \nu, u_i)}{\varsigma_p^2}\right). \tag{6.2.17}$$

The parameter ς_p^2 is the process variance associated with the territory model, including all of the influences on an individuals fecundity at first reproduction that are not determined by territory size.

We can now see the relationship between this model and the general template we outlined above (equation 6.2.1),

$$[\theta_p, \theta_d, \mathbf{z}|y] \propto \underbrace{[y|\mathbf{z}, \theta_d]}_{\text{Data}} \underbrace{[\mathbf{z}|\theta_p]}_{\text{Process}} \underbrace{[\theta_d][\theta_p]}_{\text{Parameters}} \tag{6.2.18}$$

$$\underbrace{[y_i|z_i, \theta_d]}_{\text{data}} = \text{Poisson}(y_i|\lambda_i) \tag{6.2.19}$$

$$\underbrace{[z_i|\theta_p]}_{\text{process}} = \text{gamma}\left(\lambda_i \left| \frac{(g(\alpha_i, \beta, x_i))^2}{\sigma_p^2}, \frac{g(\alpha_i, \beta, x_i)}{\sigma_p^2}\right.\right) \times \tag{6.2.20}$$

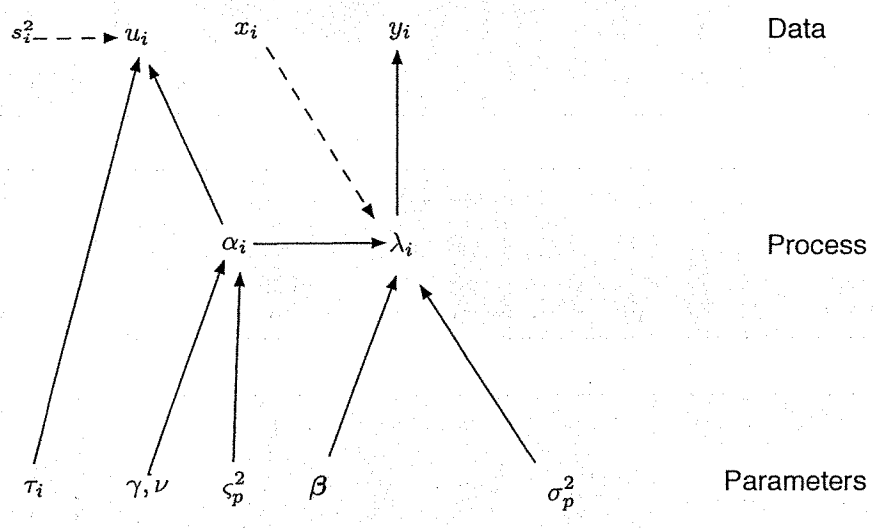
$$\text{gamma}\left(\alpha_i \left| \frac{h(\gamma, \nu, u_i)^2}{\varsigma_p^2}, \frac{h(\gamma, \nu, u_i)}{\varsigma_p^2}\right.\right)$$

$$\underbrace{[\theta_d][\theta_p]}_{\text{Parameters}} = \prod_{j=1}^{2} \text{normal}(\beta_i|0, 100)\,\text{gamma}(\sigma_p^2|.001, .001) \times \tag{6.2.21}$$

$$\text{gamma}\left(\varsigma^2|.001, .001\right).$$

Again, notice that the predictor variables $\mathbf{x}$ and $\mathbf{u}$ do not appear in the posterior distribution because we assumed they are known.

Birds were marked and followed throughout their lives, so it is reasonable to assume that age was measured perfectly. But this is not a reasonable assumption for territory size. Assume that we have data on the variance in the estimate of territory size for each bird, s_i^2. We can now think of an observation of territory size as a random variable arising from $\left[u_i|\tau_i, s_i^2\right]$ where τ_i is the true, unobserved territory size and s_i^2 is the measured observation variance. Modeling the predictor variables this way means that the u_i *is* a random variable and must be included in the expression for the posterior distribution. The full model predicting owl fecundity is shown as a Bayesian network and an expression for the posterior and joint distributions in Figure 6.2.2. We provide general guidance on when to include predictor variables in posterior distributions in Box 6.2.2.



$$
\begin{aligned}
g(\alpha_i, \beta, x_i) &= \alpha_i + \beta_1 x_i + \beta_2 x_i^2 \\
h(\gamma, \nu, \tau_i) &= \gamma e^{-\nu \tau_i} \\
[\boldsymbol{\lambda}, \beta, \alpha, \sigma_p^2, \varsigma_p^2 | \mathbf{y}, \mathbf{u}] &\propto \prod_{i=1}^{n} \text{Poisson}(y_i|\lambda_i)\, \text{gamma}\left(\lambda_i \,\middle|\, \frac{g(\alpha_i,\beta,x_i)^2}{\sigma_p^2}, \frac{g(\alpha_i,\beta,x_i)}{\sigma_p^2}\right) \times \\
&\quad \text{gamma}\left(\alpha_i \,\middle|\, \frac{h(\gamma,\nu,\tau_i)^2}{\varsigma_p^2}, \frac{h(\gamma,\nu,\tau_i)}{\varsigma_p^2}\right) \times \\
&\quad \text{gamma}\left(u_i \,\middle|\, \frac{\tau_i^2}{s_i^2}, \frac{\tau_i}{s_i^2}\right) \text{gamma}(\tau_i|.001,.001) \times \\
&\quad \prod_{j=1}^{2} \text{normal}(\beta_i|0,100) \times \\
&\quad \text{gamma}(\sigma_p^2|.001,.001)\text{gamma}(\varsigma^2|.001,.001)
\end{aligned}
$$

Figure 6.2.2: Hierarchical model of fecundity of spotted owls. Relationships between random variables are shown with solid arrows; deterministic relationships are shown with dashed arrows. The observation of fecundity of each owl y_i is a random variable controlled by its average fecundity (λ_i) and sampling variation resulting from the particular year the owl was sampled. The average fecundity individual λ_i is modeled as a quadratic function of the owl's age (x_i) with parameters $\alpha_i, \beta_1, \beta_2$. We assume age is known. Variation in the λ_i not captured by the model is represented by σ_p^2. We assume that the parameter α_i, the fecundity of owl i at first reproduction, decreases exponentially with increasing territory size u_i, which is measured with error captured by the known observation variance s_i^2. The rate of decrease in α_i is controlled by the parameter ν. The maximum possible value of α_i is γ, which occurs at territory size of 0. Variation in the α_i not represented in the exponential model is represented by ς_p^2. The expressions for the posterior and joint distributions of the unobserved and observed are shown at the bottom of the figure. Note the correspondence between the diagram and the expression for the joint distribution. Quantities at the heads of the solid arrows are on the left hand side of the conditioning symbols. Quantities at the tails are on the right hand side. Quantities at the tails of solid arrows with no arrow leading into them must have prior distributions with numeric arguments. The quantities at the tails of dashed arrows are treated as known.

It is useful to think about the relationships between the equations we used to construct the model and to consider where uncertainty arises. We have observations of a process that includes sampling error in our estimates of the fecundity of individual owls.[9] In our first model (equation 6.2.14), we have a single term for uncertainty that arises in the process of reproduction because different owls have different mean fecundities resulting from differences in age, location, genetics, and all other sources of variation. In our second model (equation 6.2.16), we seek to reduce that uncertainty about the process by including additional knowledge – the age of each owl – and by using a model that explains variation in fecundity in a biologically sensible way. In our third model (equation 6.2.17), we seek to reduce uncertainty further by modeling the average reproduction at reproductive age 0, the intercept in the effects of age model, as a function of territory size. We include all of the variation in the true, average fecundity that is not explained by our model in the stochastic terms σ_p^2 and ς_p^2. It is important to understand that our deterministic models $g()$ and $h()$ could have taken any functional form, linear or non-linear. In the fourth model (Figure 6.2.2) we add uncertainty in observations of territory size. The observed territory size a random variable arising from a distribution governed by the true territory size τ_i and measured observation variance, s^2, treated as known. If we had information on the distribution of the observation variance, we could have treated it as a random variable.

Box 6.2.2 When are predictor variables included in the posterior distribution?

A common error in writing out expressions for the posterior and joint distribution is to include predictor variables, i.e., the $\mathbf{x}$, on the right hand side of the conditioning in the posterior distribution when we assume (rightly or wrongly) that the $\mathbf{x}$ are measured without error – they are *known*. Hence, they are not random variables and should not be included in the posterior distribution. It is fine that they are arguments to a deterministic function representing an ecological process, but if we include them in the posterior distribution then the factoring of the joint distribution doesn't work out in a sensible way.

Consider a simple example. We have a deterministic model $g(\boldsymbol{\theta}, x_i)$, the output of which gives the mean of a response, y_i. Variation in y_i occurs because our model omits many influences,

[9] Remember, the observation variance in this case equals the mean. We could use a different distribution, for example a negative binomial, if we wanted to estimate the observation variance separately.

which we quantify with process variance σ_p^2. We assume the y_i are measured perfectly, but we nonetheless treat them as random variables because of the uncertainty about the process that our model fails to capture. We will drop the $g(\,)$ wrapper to make the factoring more clear. Consider the **wrong** expression for the posterior and joint distribution,

$$\left[\theta, \sigma_p^2 | y_i, x_i\right] \propto \left[y_i, \theta, \sigma_p^2, x_i\right] \tag{6.2.22}$$

$$\left[\theta, \sigma_p^2 | y_i, x_i\right] \propto \left[y_i | \theta, \sigma_p^2, x_i\right] \left[\theta\right] \left[\sigma_p^2\right] \left[x_i\right], \tag{6.2.23}$$

which is obviously incorrect because we require a prior on the known value of the observation x_i. The **correct** expression is

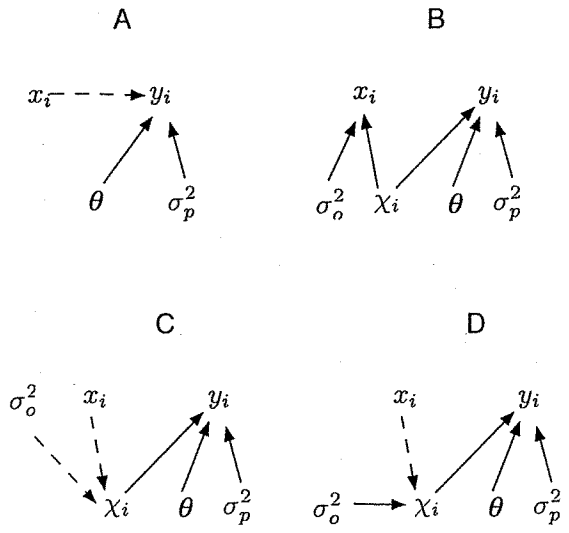
$$\left[\theta, \sigma_p^2 | y_i\right] \propto \left[y_i, \theta, \sigma_p^2\right] \tag{6.2.24}$$

$$\left[\theta, \sigma_p^2 | y_i\right] \propto \left[y_i | \theta, \sigma_p^2, \right] \left[\theta\right] \left[\sigma_p^2\right], \tag{6.2.25}$$

as illustrated in panel **A** of the diagram below. The x_i are implicitly part of these expressions as shown in the Bayesian network. It would also be correct to write

$$\left[\theta, \sigma_p^2 | y_i\right] \propto \left[y_i | g\left(\theta, x_i\right), \sigma_p^2, \right] \left[\theta\right] \left[\sigma_p^2\right],$$

to highlight the deterministic model $g\left(\theta, x_i\right)$.



There are cases where we treat the predictor variables as random variables because we want to model errors in observing them. If we assume that the observations of the predictor variable are imperfect, then we might model them arising from a distribution $[x_i|\chi_i, \sigma_o^2]$ where χ_i is the true, unobserved value of x_i and σ_o^2 represents uncertainty in the observation process. Our deterministic model is now $g(\theta, \chi_i)$. As shown in panel **B**, we now have an expression for the posterior that factors correctly,

$$[\theta, \sigma_p^2, \chi_i, \sigma_o^2|y_i, x_i] \propto [y_i|\theta, \sigma_p^2, \chi_i] [x_i|\chi_i, \sigma_o^2] [\theta] [\sigma_p^2] [\sigma_o^2] [\chi_i]. \quad (6.2.26)$$

One more point bears mentioning. Models for predictor variables that take the form $[\chi_i|x_i, \sigma_o^2]$ are sometimes seen in the scientific literature. These models portray the true, unobserved value of the predictor variable as a random variable determined by the *known* observation and *known* observation variance (panel **C**, above). In this case, the expression for the posterior and joint is

$$[\theta, \sigma_p^2, \chi_i|y_i] \propto [y_i|\theta, \sigma_p^2, \chi_i] [\chi_i|x_i, \sigma_o^2] [\theta] [\sigma_p^2]. \quad (6.2.27)$$

Again, the deterministic model is $g(\theta, \chi_i)$. Note that there is no longer on a prior on χ_i because it is seen on both sides of a conditional symbol. Also note that x_i and σ_o^2 are no longer seen in the expression for the posterior because they are not random variables. Hence, they do not require a prior. We think a better way to do this would be to treat σ_o^2 as a random variable informed by a strong prior developed in calibration studies (panel **D**), in which case

$$[\theta, \sigma_p^2, \sigma_o^2, \chi_i|y_i] \propto [y_i|\theta, \sigma_p^2, \chi_i] [\chi_i|x_i, \sigma_o^2] [\theta] [\sigma_p^2] [\sigma_o^2]. \quad (6.2.28)$$

6.2.2 Multi-level models: controls on nitrous oxide emissions from agricultural soils

Data in ecological research are often collected at multiple scales or levels of organization in designs that are nested (Figure 6.2.3). "Group" is a catchall term for the upper level in many different types of nested hierarchies. Groups could logically be formed by populations, locations, species, treatments, life stages, and individual studies. We have measurements within groups on individ-

ual organisms, plots, species, time periods, and so on. We also have measurements on the groups themselves, that is covariates that apply at the upper level of organization or spatial scale. Multilevel models represent the way that a quantity of interest responds to the combined influence of observations taken at the group level and within the group.

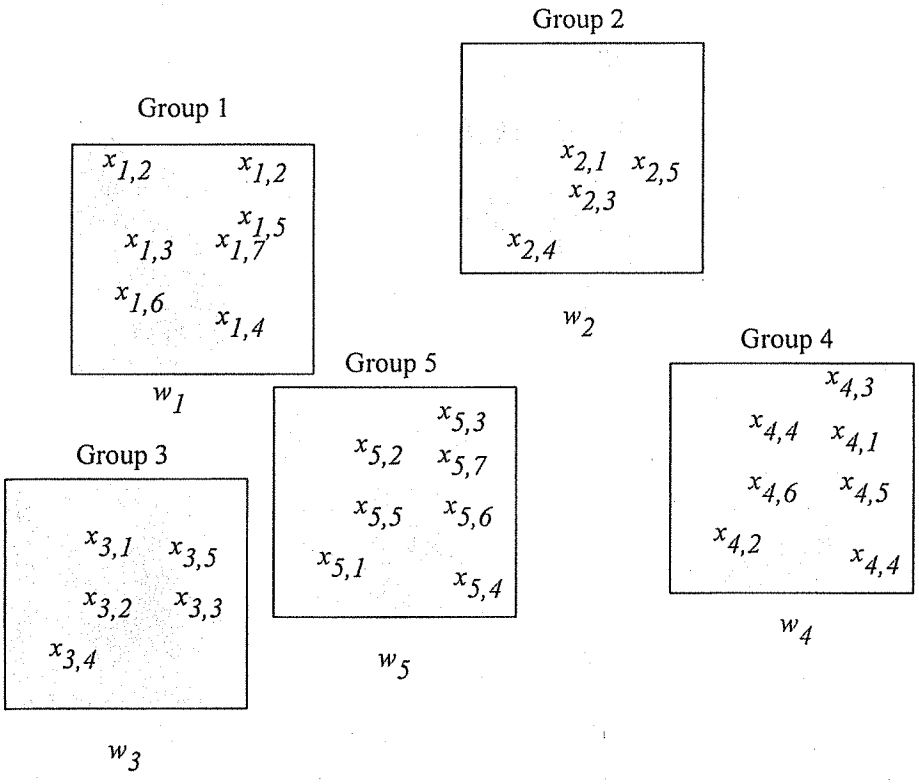


Figure 6.2.3: Observations of states and processes in ecology are often made at different scales of time, space, or level of organization. We can think of the upper level as a "group" with associated observations w_j on the j^{th} group. We also have observations within each group, x_{ij} where the notation ij means the i^{th} observation in group j. In this illustration there are variable numbers of observations within groups (i.e., the design is unbalanced) and $j = 1, \ldots, 5$ observations for groups. Note that when the number of observations is unbalanced, as it is here, product symbols in likelihoods for observations within groups must have an upper index appropriate for the number of observations, e.g., n_j.

Here, we modify the hierarchical example developed by Qian et al. (2010) (using the data of Carey (2007)) to illustrate a multi-level model. Nitrous oxide, a greenhouse gas roughly 300 times more potent than carbon dioxide in forcing atmospheric warming, is emitted when nitrogen is added to the soil in synthetic fertilizers. Carey (2007) conducted a meta-analysis of effects of nitrogen fertilizer addition ($gN \cdot ha^{-1} \cdot d^{-1}$), reviewing 164 studies. In this example, studies occurred at different locations, forming the group level in the hierarchy. Soil carbon content ($g \cdot C \cdot g^{-1}$organic matter) was measured as a group level covariate that was assumed to be measured without error. Replicated observations of N_2O emission, also assumed to be measured without error, were paired

with measurements of fertilizer addition (kgN· ha^{-1}). The type of fertilizer was also studied, but we choose to omit this effect to simplify the example. There were a total of 1085 observations across all of the studies.

We could model the observations of N_2O emission as

$$g(\alpha_j, \beta, x_{ij}) = \alpha_j + \beta x_{ij} \quad (6.2.29)$$

$$[\alpha_j, \beta, \sigma_j^2, \varsigma^2 | y_{ij}] \propto [y_{ij} | g(\alpha_j, \beta, x_{ij}), \sigma_j^2] [\alpha_j | \mu, \varsigma^2] [\beta] [\mu] [\sigma_j^2] [\varsigma^2] \quad (6.2.30)$$

where y_{ij} is the i^{th}observation of N_2O emissions in study j and x_{ij} is a paired measurement of fertilizer addition. The model $g(\alpha_j, \beta, x_{ij})$ represents the hypothesis that emissions increase in direct proportion to fertilizer additions. The intercept α_j varies among studies as a random variable drawn from distribution with parameters μ and ς^2. The fact that we explicitly represent variation among studies using the distribution of the α_j is what sets this analysis apart from conventional, single level regression that could be done separately for each of the 164 individual sites or by pooling all of the data across sites to estimate a single intercept and slope. The σ_j^2 represents the uncertainty about N_2O emissions that comes from sampling variation within a study and the ς^2 represents the uncertainty that arises as a result of variation among studies. An advantage of this hierarchical approach is know as *borrowing strength*, which means that estimates of the intercepts from locations will small datasets are made more precise by studies with larger datasets (Box 6.2.3).

Box 6.2.3 What does "borrowing strength" mean?

You'll often read the phrase "borrowing strength" in papers that use Bayesian hierarchical models. In this context, borrowing strength refers to the sharing of information among unknowns in Bayesian models. For example, consider the situation where a researcher measures leaf area index (LAI) for each plant in a set of five plots. Suppose that the numbers of plants in each plot are 10, 12, 8, 10, and 2. Clearly the last plot will carry less information about plot-level LAI because of its smaller sample size. A classical Bayesian remedy for this small sample situation is to specify a hierarchical model to help learn about plot-level mean LAI. In this case, let y_{ij} be the LAI measurement for plant i $(i = 1, \ldots, n_j)$ on plot j $(j = 1, \ldots, J)$ and the plot-level mean LAI be z_j. A complete Bayesian model could be formulated as

$$\begin{aligned} y_{ij} &\sim \text{normal}(z_j, \sigma_y^2) , && (6.2.31) \\ z_j &\sim \text{normal}(\mu, \sigma_z^2) , && (6.2.32) \\ \mu &\sim \text{normal}(\mu_0, \sigma_0^2) , && (6.2.33) \\ \sigma_z^2 &\sim \text{inverse gamma}(\alpha_z, \beta_z) , && (6.2.34) \\ \sigma_y^2 &\sim \text{inverse gamma}(\alpha_y, \beta_y) . && (6.2.35) \end{aligned}$$

At the risk of getting ahead of ourselves, the full-conditional distribution for the mean of plot with the small sample size z_5, is $[z_5|\cdot]$, where the notation reads "the distribution of z_5 conditional on the data and other parameters that influence its value." (We cover full-conditionals in detail in Section 7.3.2.1). Thus, the distribution of z_5, will contain two terms:

$$[z_5|\cdot] \propto \prod_{i=1}^{n_5} \text{normal}(y_{i5}|z_5, \sigma_y^2)\text{normal}(z_5|\mu, \sigma_z^2) , \quad (6.2.36)$$

one term containing the portion of data collected at plot 5 and a second term that depends on the mean of plot-level means and an associated variance component.

In fitting the model, all of the data (not just the data for plot 5) will help estimate μ and σ_z^2. We can think of $\text{normal}(z_5|\mu, \sigma_z^2)$ as a "prior" for z_5. The information contained in this second term will lead to a posterior for z_5 that is more precise than the resulting posterior from

a model where σ_z^2 is assumed to be known and vague *a priori*.

In a sense, the information about μ and σ_z^2 from the rest of the plots, will "shrink" z_5 for plot 5 back to the appropriate distribution of plot-level means. The plots with more data will provide more information about the proper amount of dispersion in the distribution of plot-level means which, in turn, provides more information about plots with smaller sample size. Therefore, the variance of z_5 will be smaller than it would have been if we had simply used a vague prior for all z_j.

This concept of borrowing strength is not really unique to Bayesian statistics, as it can be interpreted as a random effect in the model. However, the Bayesian perspective of these types of random effects is particularly clear and rigorous. We will return to the general concept of "shrinkage" in later chapters when we describe statistical regularization and its many benefits, including model selection.

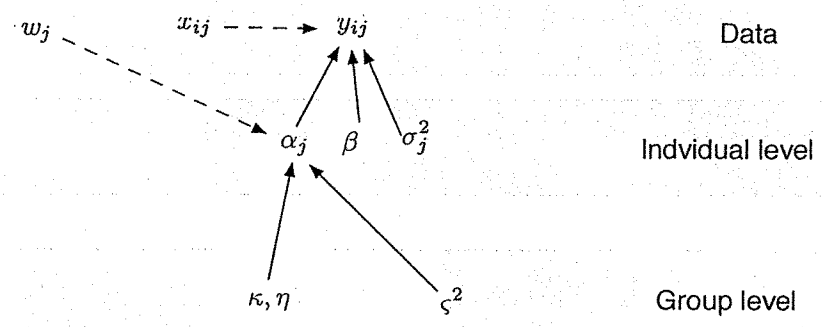
We now seek to explain some of the variation among sites using the observation of soil carbon content taken at the group level (similar to the example above, Section 6.2.1, where we modeled fecundity as a function of the covariate, age). Instead of simply estimating an average value for the intercept (equation 6.2.30), we instead model the intercept for each study as a linear function of the observations of soil carbon at the group level, that is, the α_j are predicted using the model $h(\kappa, \eta, w_j) = \kappa + \eta w_j$ (Figure 6.2.4). Choosing lognormal distributions[10] for the distribution of the data and the intercepts makes sense because they are continuous and non-negative

$$y_{ij} \sim \text{lognormal}\left(\log\left(\alpha_j + \beta_{x_{ij}}\right), \sigma_j^2\right) \tag{6.2.37}$$

$$\alpha_j \sim \text{lognormal}\left(\log\left(\kappa + \eta w_j\right), \varsigma^2\right). \tag{6.2.38}$$

The parameters in these expressions might require some review. Let μ be the first parameter of a lognormal distribution. The median of the lognormal distribution equals e^{μ}, so log(median) equals μ. Thus, if our deterministic model predicts the median of the posterior distribution, we take its log to obtain the first parameter. The second parameter is the variance of the random variable on the log scale.

[10] Gamma distributions could also be used.



$$
\begin{aligned}
g(\alpha_j, \beta, x_{ij}) &= \alpha_j + \beta x_{ij} \\
h(\kappa, \eta, w_j) &= \kappa + \eta w_j \\
[\alpha, \beta, \sigma, \kappa, \eta, \varsigma | \mathrm{y}] &\propto \prod_{j=1}^{164} \prod_{i=1}^{n_j} \text{lognormal}\left(y_{ij} | \log\left(g(\alpha_j, \beta, x_{ij})\right), \sigma_j^2\right) \times \\
&\quad \text{lognormal}\left(\alpha_j | \log\left(h(\kappa, \eta, w_j)\right), \varsigma^2\right) \times \\
&\quad \text{inverse gamma}\left(\sigma_j^2 | .001, .001\right) \times \\
&\quad \text{normal}\left(\beta | 0, 1000\right) \text{gamma}\left(\kappa | .001, .001\right) \times \\
&\quad \text{normal}(\eta | 0, 1000) \text{inverse gamma}(\varsigma^2 | .001, .001)
\end{aligned}
$$

Figure 6.2.4: Bayesian network and posterior and joint distribution for the meta-analysis of effects of fertilizer on nitrous oxide (N_2O) omissions. The y_{ij} are observations of N_2O emissions accumulated from 164 different studies. The y_{ij} are modeled as a linear function of the level of fertilizer added within a given study, x_{ij} where j indexes study and the subscript ij indicates the i^{th} observation within study j. The n_j are the number of observations from study j. Intercepts in the individual level model (α_j) varied among groups (i.e., studies) as a function of soil carbon content (w_j). Uncertainty within individual studies (σ_j^2) was allowed to vary among studies. Lognormal distributions were chosen for y_{ij} and α_{ij} because both quantities must be non-zero. Inverse gamma priors were chosen for the variances (σ_j^2, ς^2) because an inverse gamma distribution is a conjugate for the variance of lognormal distribution assuming the mean is known (A.3).

We are not limited to modeling the intercept at the group level; we could also allow the slopes to vary among sites or allow both intercepts and the slopes to vary. See Gelman and Hill (2009, page 376) for details. Moreover, we emphasize that our choice of linear models to represent the process of N_2O emission is no way mandatory – we could use any function form that makes biological sense.[11]

6.2.3 Hidden processes: effects of predation by tree snakes on lizard populations

Ecologists often want to answer questions about the state of a system that changes over time or space. Many of the states that we strive to understand cannot be observed directly but instead arise from processes that are "hidden" (e.g., Newman et al., 2006; Tavecchia et al., 2009; Liberg et al., 2012; Gimenez et al., 2012). We must make inferences about these unobservable states and hidden processes from the behavior of quantities that we *can* observe. We will refer to unobservable states as *latent*. Here, we illustrate an especially valuable use of Bayesian hierarchical models: estimating latent states and how they change over time, space, and in response to perturbation.

Campbell et al. (2011) tested the hypothesis that exothermic predators influence the abundance of their exothermic prey using predation by brown tree snakes (*Boiga irregularis*) on lizards on the island of Guam as a model system. This research offers an especially useful example because it illustrates how hierarchical models can be used to analyze designed, manipulative experiments. The research team observed lizard abundance on four, 1 ha plots, which we will index by $m = 1, \ldots 4$. All tree snakes were removed from two of the plots and the two were left as controls with ambient levels of snake abundance. Lizards were counted on five transects within each plot (indexed by i). Counts were repeated seven times on each transect (on different days, indexed by t) within each of six monitoring periods. We will omit the monitoring period dimension of the experiment, which means that time (t) refers strictly to repeated measures of transects.

The research team needed to estimate the true, unobserved lizard abundance on each transect based on counts along the transect. A key problem in this kind of research is that counting all individuals is virtually impossible because some lizards that are present inevitably escape detection. The mismatch between what we are able to observe and the true state we want to understand requires building a model of the data, a way to estimate the probability that a lizard is observed (ϕ) on a transect at a given time conditional on it being present. A sensible approach for modeling

[11] It might make *more* sense, for example, to model the intercept as an asymptotic function of soil carbon, something like $h(\kappa, \eta, w_i) = \frac{\kappa w_i}{\eta + w_i}$.

2298 the data would start with

$$y_{itm} \sim \text{binomial}\left(z_{im}, \phi\right), \tag{6.2.39}$$

$$\phi \sim \text{beta}\left(1, 1\right), \tag{6.2.40}$$

2299 which simply says that the observations of the number of lizards counted on transect i observed at
2300 time t on plot m can be represented as a random variable y_{itm} drawn from a binomial distribution
2301 where z_{im} is the known true, unobserved number of individuals on the i^{th} transect of plot m and ϕ
2302 is the probability of observing an individual. Thus, z_{im} represents the number of "trials" on transect
2303 i in plot m, that is, the number of lizards that were present and might be found; y_{itm} is the number
2304 of "successes," the number of lizards that were found, and ϕ is the probability of a success on a
2305 single trial, the probability that we would observe a lizard if it were present (Royle, 2004). By
2306 taking replicate observations on the transect, and assuming for the moment that z_{im} is known, we
2307 can estimate ϕ on the back of a napkin using a beta-binomial conjugate prior relationship (i.e., as
2308 shown in Section 5.3). The full expression for the posterior and joint distributions is

$$[\phi|\mathbf{y}] = \prod_{i=1}^{5}\prod_{t=1}^{7}\prod_{m=1}^{4} \text{binomial}\left(y_{itm}|z_{im}, \phi\right) \text{beta}\left(\phi|1, 1\right) \tag{6.2.41}$$

2309 Recall that we do not include $\mathbf{z}$ in the posterior distribution because at this point, we are assuming
2310 it is known.

2311 Our initial model (equation 6.2.39) requires the assumption that there is a *single* detection
2312 probability applying identically to all time periods and all transects within a plot in much the same
2313 way that we assumed that all owls had the same average fecundity (Section 6.2.1). We could improve
2314 on the model by allowing each transect and time to have its own detection probability, reasoning
2315 that observability is likely to vary over time and space. Transects might differ in the availability
2316 of lizard hiding places and observation times might include different temperatures that affect lizard
2317 activity levels, both of which would alter lizard exposure to the observer. Campbell et al. (2011)

included this variability in the data model using

$$y_{itm} \sim \text{binomial}\left(z_{im}, \phi_{itm}\right) \tag{6.2.42}$$

$$\text{logit}\left(\phi_{itm}\right) = \alpha_0 + \alpha_{1,itm} \tag{6.2.43}$$

$$\alpha_{1,itm} \sim \text{normal}\left(0, \sigma^2_{\alpha_1}\right) \tag{6.2.44}$$

where α_0 is the overall, mean probability of detection of lizards on the plot, and α_{itm} is a *random effect* (Box 6.2.1) of transect and observation time. Thus, $\alpha_{1,itm}$ represents the variation in detection probability that arises from differences among transects and sampling occasions on plot m.

It might be useful to think of equation 6.2.42 as an "intercept only" (i.e., α_0) linear model, making it analogous to examples developed above that had group-level intercepts (Sections 6.2.2 and 6.2.4). As before, we are allowing the "groups" time and transect to have their own intercepts drawn from a distribution with mean inverse logit (α_0). There is variation around this mean probability of detection created by random variation among transects and observation times, variation that is portrayed by the parameters α_{1itm}. Notice that we are not modeling how this variation arises as we might do if we had covariates, say on temperature or vegetation cover. Instead, we simply acknowledge that the variation exists and include it using the random effect terms.

It might appear on first glance that the model (equations 6.2.42 - 6.2.44) contains only a single source of uncertainty, the random effect $\alpha_{1,itm}$ and the associated parameter controlling its distribution $\sigma^2_{\alpha_1}$. This flies in the face of the idea that random effects are included in models to capture uncertainty that extends beyond sampling variation. However, remember that there is a sampling variability included in the binomial likelihood (6.2.42); the variance of which is $n\phi\left(1-\phi\right)$ (i.e., Section 3.4.3.1). So, sampling variability is implicit in the binomial and hence there are two sources of uncertainty in this observation model – uncertainty arising from sampling and uncertainty arising because the ability to detect lizards varies among times and transects.

It is important to understand that we could achieve the exact same meaning with slightly different notation,

$$y_{itm} \sim \text{binomial}\left(z_{im}, \phi_{itm}\right) \tag{6.2.45}$$

$$\text{logit}\left(\phi_{itm}\right) \sim \text{normal}\left(\mu_\phi, \sigma^2_\phi\right). \tag{6.2.46}$$

Notice that in the first case (equations 6.2.42 - 6.2.44) we a mean 0 random effect ($\alpha_{1,itm}$) to an overall mean. In the second case (equations 6.2.45 - 6.2.44), we model random effects random as random draws α_{itm} from a distribution with mean μ_α. It is important to see that these are algebraically identical because both types of notation are widely used in the literature.

A third equally correct alternative for this model would be:

$$y_{itm} \sim \text{binomial}\left(z_{im}, \phi_{itm}\right) \tag{6.2.47}$$

$$\phi_{itm} \sim \text{beta}\left(\phi_{itm}|\alpha, \beta\right). \tag{6.2.48}$$

In this case, we not transforming ϕ_{itm} to cause it to take on values that could range from $-\infty$ to $+\infty$ appropriate for the normal distribution, but rather are choosing a distribution appropriate for a random variable that can take on values in the interval 0 to 1. The mean of this distribution, you will recall from Section 3.4.4, is $\frac{\alpha}{\alpha+\beta}$.

Up to now, we have assumed that the true number of lizards on a transect was known, which of course it is not. We now develop a model of processes controlling the true, unobserved lizard abundance each transect, z_{im}. Thus, z_{im} is now a random variable. The main interest in this study is the variation of numbers of lizards among plots, particularly the variation that is contributed by the snake-removal treatment. So, now we model the means of the four plots (two of which had predators removed) using:

$$z_{im} \sim \text{Poisson}\left(e^{\beta_{0,m}+\beta_1 x_{im}}\right) \tag{6.2.49}$$

$$\beta_{0,m} \sim \text{normal}(0, 100)$$

$$\beta_1 \sim \text{normal}(0, 100) \tag{6.2.50}$$

where x_{im} is an indicator variable equal to 1 if plot m had snakes removed and 0 otherwise. It is entirely reasonable to assume that the x_{im} are measured without error because treatments were assigned by the investigators. In this model, $\beta_{0,m}$ is the mean abundance of lizards on plot m. Note the average abundance on transect i within plot m is determined by this mean abundance as modified by the treatment effect, β_1. Campbell et al. (2011) made repeated observations at the transect scale but mean abundance, i.e., $\beta_{0,m}$ was modeled at the plot scale, recognizing that plots

were the experimental units.

The Bayesian network for the relationships between the knowns and unknowns and the expression for the posterior and joint distributions is shown in Figure 6.2.5. The process models (equations 6.2.49 - 6.2.50) and data model (equation 6.2.42 - 6.2.44) are linked by latent state z_{im}, the true average number of lizards on transect i within plot m (Figure 6.2.5). This example shows how hierarchical models can be used in designed experiments to represent variation at multiple scales, to estimate unobservable quantities, to estimate the effect of a treatment, and to properly account for uncertainty arising from multiple sources.

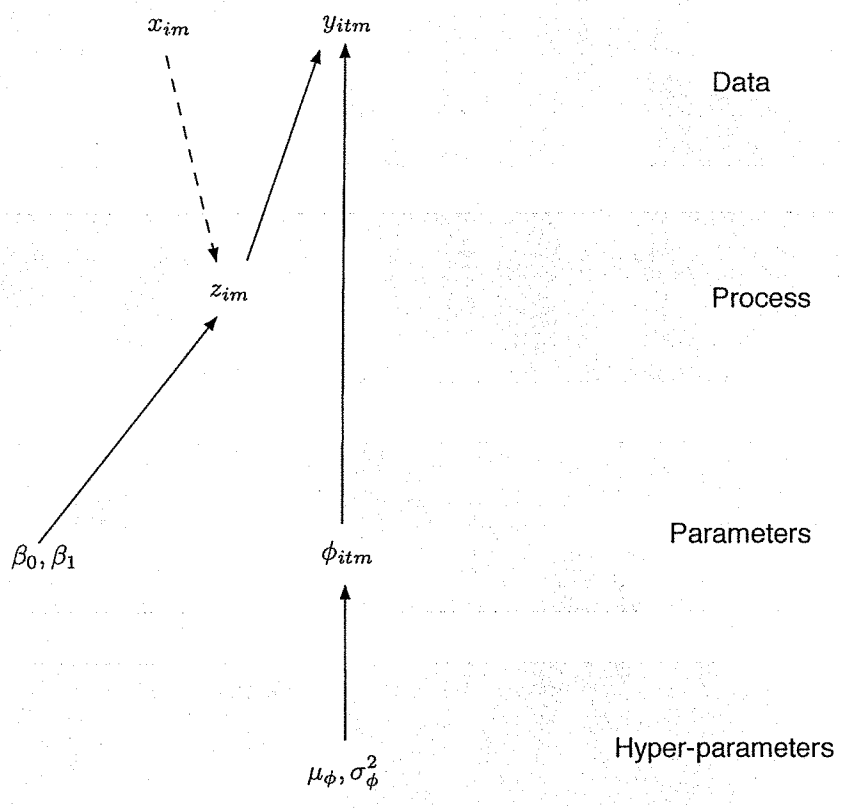
6.2.4 Multi-level models: functional traits of Neotropical trees mediate effects of light and size on growth

Relationships between traits of individuals that control physiological and reproductive function are known to shape demographic rates of populations. The relationship between species functional traits and demography has provided fundamental insight into the compromises underlying life-history strategies (Westoby et al., 2002; Westoby and Wright, 2006; van Kleunen et al., 2010). In this example, we feature the work of Rüger et al. (2012) who used a Bayesian hierarchical model to reveal how functional traits modify the influence of light and size on the growth rate of species of Neotropical trees.

The growth rate of individual trees within each species was modeled as a power function of light availability ($x_{1,ij}$) and tree diameter at breast height ($x_{2,ij}$), where i indexes individuals and j indexes species. These were assumed to be measured without error. Thus, species were treated as a group-level variable and there were measurements of individual growth responses and covariates for individuals within each species. The true growth rate (λ_{ij}) of individual i within species j was modeled using a log transformation to linearize the power function,

$$\begin{aligned} g\left(\boldsymbol{\beta}, \mathbf{x}_{ij}\right) &= \beta_{0,j} + \beta_{1,j} \log\left(x_{1,ij}\right) + \beta_{2,j} \log(x_{2,ij}) \\ \lambda_{ij} &\sim \text{lognormal}\left(g\left(\boldsymbol{\beta}, \mathbf{x}_{ij}\right), \sigma^2_{p,j}\right) \end{aligned} \tag{6.2.51}$$

where $\sigma^2_{p,j}$ represents the process variance for species j. Another way to look at this that might be more clear is to use the untransformed model, i.e. the power function for growth, $g\left(\boldsymbol{\beta}, \mathbf{x}_{ij}\right) =$



$$
\begin{aligned}
[\phi, z, \beta, \mu_\phi, \sigma^2_\phi | y] \quad \propto \quad & \prod_{i=1}^{5} \prod_{t=1}^{7} \prod_{m=1}^{4} \text{binomial}(y_{itm} | z_{im}, \phi_{itm}) \times \\
& \text{normal}(\text{logit}(\phi_{itm}) | \mu_\phi, \sigma^2_\phi) \times \\
& \text{Poisson}\left(z_{im} | e^{\beta_0 + \beta_1 x_{im}}\right) \times \\
& \text{normal}(\beta_0 | 0, 1000) \text{normal}(\beta_1 | 0, 1000) \times \\
& \text{normal}(\mu_\phi | 0, 1000) \times \\
& \text{inverse gamma}(\sigma^2_\phi | .001, .001)
\end{aligned}
$$

Figure 6.2.5: Bayesian network for model of effects of snake predation on lizards modified from Campbell III et al. (2011). The data (y_{itm}) are the number of lizards observed on transect i at time t on plot m. The true number of lizards on transect i of plot m is z_{im}. The parameter ϕ_{itm} is the probability of detecting a lizard that is truly present. We model the logit of these probabilities as draws from a normal distribution with mean μ_ϕ and variance σ^2_ϕ. The model $e^{\beta_0 + \beta_1 x_{im}}$ represents the effect of removing predators on the true number of lizards on a transect.

$e^{\beta_0} x_{1,ij}^{\beta_1} x_{2,ij}^{\beta_2}$ in which case the true growth rate is $\lambda_{ij} \sim \text{lognormal}\left(\log\left(g\left(\boldsymbol{\beta}, \mathbf{x}_{ij}\right)\right), \sigma_{p,j}^2\right)$ where now $\sigma_{p,j}^2$ is the variance of $\log\left(\lambda_{ij}\right)$.

Functional traits of species (wood density, maximum height, leaf area, seed mass, leaf mass per area and leaf nutrient content) comprising the data vector $\mathbf{w}_j$ were used as group level covariates to model the β coefficients (Figure 6.2.6) in the individual level model using vectors of group level parameters, $\boldsymbol{\alpha}, \boldsymbol{\gamma}$, and $\boldsymbol{\eta}$,

$$\beta_0 \sim \text{normal}\left(\alpha_0 + w_j'\boldsymbol{\alpha}, \varsigma_{\beta_0}^2\right) \tag{6.2.52}$$

$$\beta_1 \sim \text{normal}\left(\gamma_0 + \mathbf{w}_j'\boldsymbol{\gamma}, \varsigma_{\beta_1}^2\right) \tag{6.2.53}$$

$$\beta_2 \sim \text{normal}\left(\eta_0 + \mathbf{w}_j'\boldsymbol{\eta}, \varsigma_{\beta_2}^2\right). \tag{6.2.54}$$

The process model (equation 6.2.51) was related hierarchically to observations of tree growth (Figure 6.2.6 Data level) with a normal mixture model (see Section 3.4.5) to reflect uncertainty in the observations,

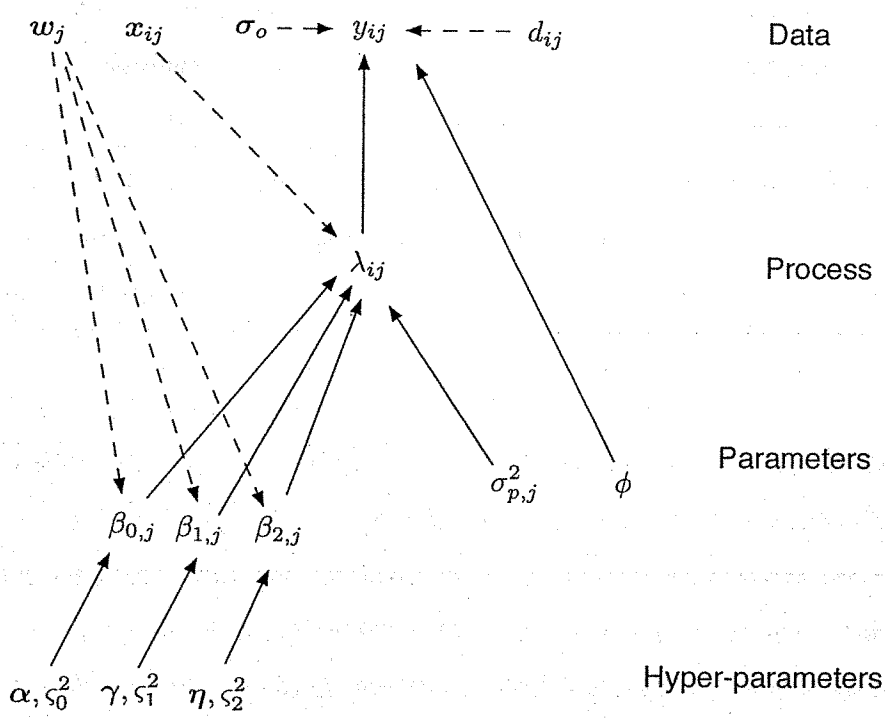
$$[y_{ij}|\lambda_{ij}, \phi] = (1-\phi) \cdot \text{normal}\left(\lambda_{ij}, \frac{\sigma_{o,1}}{d_{ij}}\right) + \phi \cdot \text{normal}\left(\lambda_{ij}, \frac{\sigma_{o,2}}{d_{ij}}\right). \tag{6.2.55}$$

This is a novel approach because it breaks the observation uncertainty into two parts. Small routine errors caused by a slightly different placement of the calipers or tape measure cause size dependent uncertainty represented in the standard deviation $\sigma_{o,1}$, which was assumed to be known from calibration studies. Large, size-independent errors were caused by missing a decimal place or recording a number with the wrong tree. These uncertainties were represented in $\sigma_{o,2}$, also assumed to be known. The two standard deviations were divided by the number of days elapsed between the two height measurements of the tree (d_{ij}) used to calculate the observed annual growth rate (y_{ij}) assuming that the magnitude of errors was proportional to the time between measurements.[12]

The full expression for the joint and posterior distributions (Figure 6.2.6) assembles the model of processes governing the true, unobserved growth of individual trees (equation 6.2.51), the model explaining species effects on the true growth rate using functional traits (equation 6.2.52) and a model of the data relating the true growth rate to the observations of growth rate (equation 6.2.55).

[12]It should be clear to you why d_{ij} and σ_o do not appear in the likelihood. It is because they are treated as known. If this is not clear, see Box 6.2.1.

2407 This example illustration shows how hierarchical models can be used to deal with two common
2408 problems in ecology. We seek to understand how characteristics of species, in this case their func-
2409 tional traits, modify responses of individuals to their local environments. The ecological process we
2410 seek to understand is itself hierarchical. We also require a hierarchical model because we are model-
2411 ing an underlying process (tree growth) based on observations that are not a perfect representation
2412 of the process (diameter at breast height). As in the owl fecundity example, we need to separate
2413 the uncertainty that arises from imperfect observations from uncertainty created by the failure of
2414 our model to represent the process.



$$
\begin{aligned}
[y_{ij}|\lambda_{ij},\phi] &= (1-\phi)\cdot \text{normal}\left(y_{ij}|\lambda_{ij},\frac{\sigma_{o,1}}{d_{ij}}\right) + \phi \cdot \text{normal}\left(y_{ij}|\lambda_{ij},\frac{\sigma^2_{o,2}}{d_{ij}}\right) \\
g\left(\beta,\mathbf{x}_{ij}\right) &= e^{\beta_0} x_{1,ij}^{\beta_1} x_{2,ij}^{\beta_2} \\
[\beta,\sigma^2_p,\varsigma^2,\alpha,\gamma,\eta|y] &\propto \prod_{j=1}^{171}\prod_{i=1}^{n_j}[y_{ij}|\lambda_{ij},\phi] \times \\
&\quad \text{lognormal}(\lambda_{ij}|\log(g\left(\beta,\mathbf{x}_{ij}\right)),\sigma^2_{p,j}) \times \\
&\quad \text{gamma}\left(\sigma^2_{p,j}|.001,.001\right) \times \\
&\quad \text{normal}(\beta_0|\alpha_0+\alpha' w_j,\varsigma_0^2) \times \\
&\quad \text{normal}(\beta_1|\gamma_0+\gamma' w_j,\varsigma_1^2) \times \\
&\quad \text{normal}(\beta_2|\eta_0+\eta' w_j,\varsigma_2^2) \times \\
&\quad \prod_{q=0}^{6}\text{normal}(\alpha_q|0,1000)\text{normal}(\gamma_q|0,1000)\text{normal}(\eta_q|0,1000)) \times \\
&\quad \prod_{m=0}^{2}\text{inverse gamma}(\varsigma^2_m|.001,.001) \\
&\quad \text{beta}(\phi|1,1)
\end{aligned}
$$

Figure 6.2.6: Hierarchical model of controls on growth of tropical trees (Rüger et al., 2012). Observations (y_{ij}) of the growth rate of an individual (λ_i) were made with error. A mixture model $\left[y_{ij}|\lambda_{ij},\sigma^2_o,\phi,d_{ij}\right]$ was used to account for two types of error: mistakes made in measurement of individuals that depended on tree size and size-independent mistakes resulting from errors like missing a decimal place or recording a number with the wrong tree. The true growth rate was modeled using a power function with individual traits (diameter at breast height and light availability) as predictor variables. The intercept of the power function was modeled as a linear function of six species traits, wood density, maximum height, leaf area, seed mass, leaf mass per area, and leaf nutrient content. Definitions of symbols are given in the text.

6.2.5 Multiple states, multiple types of data

We now cover an important topic, the use of multiple types of data to estimate parameters and latent states. Ecologists are accustomed to datasets that contain multiple covariates, but the idea of multiple responses may be unfamiliar. In Chapter 2 we introduced the idea that models of relatively high dimension may be needed to represent complex ecological relationships, for example, the functioning of ecosystems or the dynamics of populations and communities. In these cases the need to represent interactions and composite forces motivates multiple parameters and states. Models that have many parameters and that predict more than one latent state will usually require multiple types of data. Otherwise, their unknowns will not be identifiable.

We first show how to use multiple data sources in a general way before offering a more specific, example. Assume you have a model $g(\theta_p, \mathbf{x})$ that predicts the central tendency of L latent states[13], $z_1, .., z_L$. These states might be estimates of fluxes of different molecular forms of carbon and nitrogen from soils. The states might be census of individuals of different species in a community and independently obtained estimates of their proportions in the community. They might be temperature, pH, turbidity, and salinity of an estuary. You have w vectors of data on these states, $\mathbf{y}_1, ..., \mathbf{y}_w$. We have a data model $h(\theta_d, \mathbf{z})$ that relates the observations to the true value of the latent state.

You might reasonably choose a multivariate distribution for the likelihood, something like:

$$\mathbf{y}_i \sim \text{multivariate normal}\left(h\left(\theta_d, \mathbf{z}_i\right), \Sigma_y\right) \tag{6.2.56}$$

$$\mathbf{z}_i \sim \text{multivariate normal}\left(g\left(\theta_p, \mathbf{x}_i\right), \Sigma_z\right) \tag{6.2.57}$$

were the Σ are covariance matrices. In this case, we are modeling the observations and the true states as vectors following a multivariate normal distribution. The elements of $\mathbf{y}_i$ and $\mathbf{z}_i$ can be correlated and can have their own individual variance terms. There is nothing wrong with this approach as long as we can plausibly assume that the stochasticity in the latent state and in the observations can be represented using the normal distribution.

The assumption of normality for all states and responses constrains our options. Often we want to understand quantities of interest that have different support – some are strictly positive, some

[13]To keep our notation compact we are using a static model as an example here, but bear in mind that the model could just as easily be a dynamic simulation model predicting the central tendency of L states $z_{1,t}, ..., z_{L,t}$ at time t using model $g(\theta, \mathbf{x}_{t-1}, \mathbf{z}_{t-1})$. Our example will use this type of model.

are 0 or 1, some are 0 to 1, some are numbers of individuals in categories.[14] Assuming variance that is constant with the mean, as is the case with the normal distribution, may not be reasonable. We often require a more flexible approach.

A great strength of Bayes (or maximum likelihood for that matter) is that we can combine independent data *sets* in the same way we combine indedpenent individual *observations* within a dataset – by taking their products. This aggregation is justified by the rules of probability (Section 3.3). Combining data sets using the independent product rule allows us to choose probability distributions that are appropriate for the support of each random variable, that is, each observation or each unobserved state. It follows that a general expression for the posterior and joint distributions exploiting multiple datasets is

$$\left[\theta_d, \theta_p, \sigma_o^2, \sigma_p^2, \mathbf{z}|\mathbf{y}_1, ...\mathbf{y}_L\right] \propto \prod_{l=1}^{w}\prod_{i=1}^{n_l}\left[y_{li}|h\left(z_{li}, \theta_d\right), \sigma_o^2\right]_l \left[z_{li}|g\left(\theta_p, \mathbf{x}_i\right), \sigma_p^2\right] \times \left[\theta_d, \theta_p, \sigma_o^2, \sigma_p^2\right] \quad (6.2.58)$$

where i indexes individual observations and states. Note that there is an l subscript on the likelihood, which indicates that different distributions can be used as needed for the different datasets, realizing, of course, that we may need some judicious moment matching to transform the means and variances into the proper parameters. The main point here is to recognize the dual products, one product taken over datasets and likelihoods (indexed by l) and the other one over individual observations within a dataset (indexed by i). Thus, we have used multiple types of data by multiplying the total likelihood of each dataset.

We now make equation 6.2.58 more specific by offering a hypothetical example based on age- or stage-structured population modeling (Caswell, 1988), a widely used approach for modeling dynamics of populations of animals and plants. Presume we are interested in an organism whose life history can be described by m stages (or age classes) in the state vector $\mathbf{z}_t$ that contains the true number of individuals in each stage at time t. Avoiding weedy details, assume we have an appropriately composed projection matrix $\mathbf{A}$ containing fertilities and survival probabilities. We will notate the survival and fertility parameters in $\mathbf{A}$ collectively as θ so that our deterministic

[14] An additional complication arises when the $\mathbf{y}$ vectors differ in length.

model is $g(\theta, \mathbf{z}_{t-1}) = \mathbf{A}\mathbf{z}_{t-1}$ and our stochastic model of the process is

$$\log(\mathbf{z}_t) \sim \text{multivariate normal}\left(\log\left(g(\theta, \mathbf{z}_{t-1})\right), \sigma_p^2\mathbf{I}\right) \tag{6.2.59}$$

where $\mathbf{I}$ is an $m \times m$ matrix with ones on the diagonal. The log transform means that we are portraying each of the i stages at time t as a lognormally distributed random variable with mean $\mathbf{a}_i' z_{i,t}$ (the inner product[15] of the i^{th} row of matrix $\mathbf{A}$ with the state vector $\mathbf{z}_t$) and process variance σ_p^2. We could allow for different process variances for each stage and for their covariance by explicitly specifying a covariance matrix, but we want to keep things simple for this example.

Now assume we have a dataset $\mathbf{y}_1$ containing total census of individuals in the population at T times. We also have independent data on the number of individuals in each stage in the population, $\mathbf{Y}_2$, obtained by sampling a subset of the population and classifying each individual into an appropriate category. The second data set is a matrix because it is composed of a vector of classification counts for each of the T time points. We assume for simplicity that there are no gaps in the two time series of data, but missing data could be modeled if necessary. A reasonable expression for the posterior and joint distribution is

$$\begin{aligned}
[\theta, \mathbf{z}|\mathbf{y}_1, \mathbf{Y}_2] \propto & \underbrace{\prod_{t=2}^{T} \text{Poisson}\left(y_{1,t} \,\middle|\, \sum_{i=1}^{m} z_{it}\right)}_{\text{likelihood for census data}} \times \\
& \underbrace{\text{multinomial}\left(\mathbf{y}_{2,t} \,\middle|\, \sum y_{2,it}, \frac{\mathbf{z}_t}{\sum_{i=1}^{m} z_{it}}\right)}_{\text{likelihood for classification data}} \times \\
& \underbrace{\text{multivariate normal}\left(\mathbf{z}_t|\log\left(g(\theta, \mathbf{z}_{t-1})\right), \sigma_p^2\mathbf{I}\right)}_{\text{process model}} \times \\
& \text{appropriate priors for } \theta, \mathbf{z}_1, \text{ and } \sigma_p^2.
\end{aligned} \tag{6.2.60}$$

There are several points worth emphasizing here. First look at the likelihood for the census data. The estimate of the true, unobserved population size is the sum over the m stages, which forms our estimate of the mean of the distribution of the random variable, y_t, the number of individuals counted at time t. Next focus on the likelihood for the classification data. A vector of proportional

[15] If we have two vectors $\mathbf{u}$ and $\mathbf{v}$ with three elements each, then their inner product is $\mathbf{u}'\mathbf{v} = u_1v_1 + u_2v_2 + u_3v_3$. This is also called the dot product.

contributions of each stage to the total population is $\frac{z_t}{\sum_{i=1}^{m} z_{it}}$, where elements of this vector are the number in each stage divided by the total. This vector forms the second parameter of the multinomial likelihood. The first argument is simply the total number of individuals classified.

You may be wondering, "What happened to the observation variance in equation 6.2.58?" Once again, we need to remember the relationship between parameters and moments (Section 3.4.4). The variance of the Poisson is the same as its single parameter, the mean. This variance reflects the idea that if we censused the population on different days or under different conditions, we would obtain different counts simply because of sampling error.[16] In the case of the multinomial, the observation variance for the estimate of the number of individuals in category $y_{2,it}$ is $n_t p_{it}(1 - p_{it})$ where $n = \sum_{i=1}^{m} y_{2,it}$ and $p_i = \frac{z_{it}}{\sum_{i=1}^{m} z_{it}}$. So the observation variance is implicit in the values of the parameters of the multinomial.

Recall that we calculate the total probability of the data conditional on a parameter (i.e., the total likelihood) from the product of the probabilities of individual observations (equation 4.1.5). This is what we are doing multiplying across the i observations in individual likelihoods above. The products of these total likelihoods for the two datasets gives us the combined probability of the two data sets conditional on the values of latent state z. We not limited to two likelihoods, we might have many. Moreover, although we must assume independence here to keep things simple, non-independent datasets could be used as long as we properly modeled the dependence among them, a topic that is beyond the scope of this book, but that could be tackled after mastering the principles we present.

6.3 When are observation and process variance identifiable?

We have spoken frequently about partitioning uncertainty by separately modeling process variance and observation variance. There are conditions when this is possible and when it is not. To understand this idea we need to introduce the statistical term *identifiability*. For inference on a model to be possible, the parameters in the model must be identifiable, which loosely means that it is possible to learn the true value of this model's parameter(s) conditional on an infinite number of observations. In practice, this means that different values of the parameter(s) must generate

[16] Actually, it would be wise to replicate the census for each t so that we could explicitly estimate the observation variance. We are ignoring the possibility of bias arising from over or undercounting to keep this example simple. See Section 6.2.3 for an example where we account for bias in the counts.

different probability distributions of the observable variables.

Consider a general, hierarchical expression for the posterior and joint distribution of observations and parameters,

$$[\theta, \sigma_p^2, \sigma_o^2|\mathbf{y}] \propto \prod_{i=1}^{n} [y_i|z_i, \sigma_o^2] \left[z_i|g\left(\theta, \mathbf{x}_i\right), \sigma_p^2\right] [\theta, \sigma_p^2, \sigma_o^2] \tag{6.3.1}$$

where $g(\theta, \mathbf{x}_i)$ is a deterministic model of an ecological process, y_i is an observation on the process, σ_p^2 is the process variance, and σ_o^2 is the observation variance that in this case arises purely from sampling.

We can identify σ_p^2 and σ_o^2 if and only if one or more of the following conditions hold:

1. We have replications on the observations for each of the unknown states, i.e.,

$$[\theta, \sigma_p^2, \sigma_o^2|\mathbf{Y}] \propto \prod_{i=1}^{n} \prod_{j=1}^{J} [y_{ij}|z_i, \sigma_o^2] \left[z_i|g\left(\theta, \mathbf{x}_i\right), \sigma_p^2\right] [\theta, \sigma_p^2, \sigma_o^2] \ , \tag{6.3.2}$$

   where j indexes multiple observations for each i. Obtaining replications, of course, requires thoughtful design to achieve, which is a great reason for writing down your model as part of the design of your research. Examples 6.2.2, 6.2.3, and 6.2.4 all had replication at the data level.

2. Your model has strong and differing "structure" in the data and process models. Structure can mean very different distributions for the process and observation models, as in example 6.2.1 above, or it can mean strong and differing spatial or temporal structure in the data or the process, a somewhat advanced topic that is beyond the scope of this book. See Cressie and Wikle (2011) for a full treatment.

3. Your model has strongly informative priors on parameters, particularly the variance components. These could come from previous studies as illustrated in example 6.2.4, again illustrating the value of informative priors (Section 5.4).

2527 To illustrate a model where the two variance components would probably **not** be identifiable,

$$[\theta, \sigma_p^2, \sigma_o^2|\mathbf{y}] \propto \prod_{i=1}^{n} \underbrace{\text{normal}\left(y_i|z_i, \sigma_o^2\right) \text{normal}\left(z_i|g\left(\theta, \mathbf{x}_i\right), \sigma_p^2\right)}_{\text{variances not identifiable}} \times \tag{6.3.3}$$
$$\text{inverse gamma}\left(\sigma_p^2|.001, .001\right) \text{inverse gamma}\left(\sigma_o^2\right) \times$$
$$\text{appropriate priors on } \theta\,.$$

2528 Note that there is no replication and the distributions for the process and the data are the same.

Part II

Implementation