6 Hierarchical Bayesian Models

It is worthwhile to review the key points covered thus far. We started with the first principles rules of probability (sec. 3.2). We used those rules to develop Bayes' theorem (sec. 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 (sec. 3.3). We learned about priors and their influence on the posterior (sec. 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 and (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 (fig. 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 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,

$$\begin{aligned}\Big[\underbrace{\theta_1,\theta_2,z}_{\text{unobserved}} \mid \underbrace{\boldsymbol{y}}_{\text{observed}}\Big] &\propto \underbrace{\big[\theta_1,\theta_2,z,\boldsymbol{y}\big]}_{\text{joint}} &(6.1.1)\\ &\propto \underbrace{\big[\boldsymbol{y}|\theta_1,\theta_2,z\big]}_{\text{likelihood}}\underbrace{[\theta_1]\,[\theta_2]\,[z]}_{\text{priors}} &(6.1.2)\end{aligned}$$

is a simple Bayesian model[2] involving the unobserved quantities θ_1, θ_2, and z, and the observations $\boldsymbol{y}$. It is important to remember that we factor the joint distribution using the rules of probability (sec. 3.3) to obtain the product of the likelihood and priors. The model 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 distribution because we have omitted the denominator of Bayes' theorem, the marginal distribution of the data $\left(\iiint \big[\boldsymbol{y}|\theta_1,\theta_2,z\big]\,[\theta_1]\,[\theta_2]\,[z]\,d\theta_1 d\theta_2 dz\right)$, which is a scalar with a fixed value after we have observed the data. At the risk of getting ahead of ourselves, we are expressing the posterior as proportional to the joint distribution, because this proportionality is all we need to properly develop an algorithm for estimating the parameters and latent state, which we will cover in chapter 7 on the Markov chain Monte Carlo algorithm.

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

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

[1] Including the "knowledge" that little 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.

is hierarchical because we factored $\big[\theta_1,\theta_2,z,y\big]$ to produce $\big[y|\theta_1,z\big]\,[z|\theta_2]\,[\theta_1]\,[\theta_2]$, assuming that 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 next.

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 appears 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 properly factored expressions for joint distributions. Our purpose in this chapter is to emphasize writing mathematical expressions as the proper first step in modeling. For now, we postpone considering how we might implement or evaluate the model. The examples we offer here will be supplemented by

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

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 three themes that recur 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 helpful 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, so it is especially instructive to see how we add detail to 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 to note in the examples is how we partition uncertainty into multiple sources. In particular, we 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):

$$\left[\boldsymbol{\theta}_p, \boldsymbol{\theta}_o, \mathbf{z}|\mathbf{y}\right] \propto \underbrace{[\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}_o]}_{\text{data}}\underbrace{[\mathbf{z}|\boldsymbol{\theta}_p]}_{\text{process}}\underbrace{[\boldsymbol{\theta}_o]\left[\boldsymbol{\theta}_p\right]}_{\text{parameters}}. \tag{6.2.1}$$

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 **z**, a state that is not observable. We relate that state to the observable data (**y**), using a model with a vector of parameters $\boldsymbol{\theta}_o$, including parameters representing uncertainty in our observing system. The behavior of the true state is predicted with a model parameterized by $\boldsymbol{\theta}_p$, including parameters representing stochasticity 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 after Clark (2003a), who studied the effects of individual differences in fecundity on population growth rate of northern 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.

[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.

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 (fig. 6.2.1 A).

Recall that Bayesian networks (fig. 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 represent stochastic relationships among the random variables, and the tails of the arrows specify the parameters defining the distribution of the random variable at the heads of the arrows.

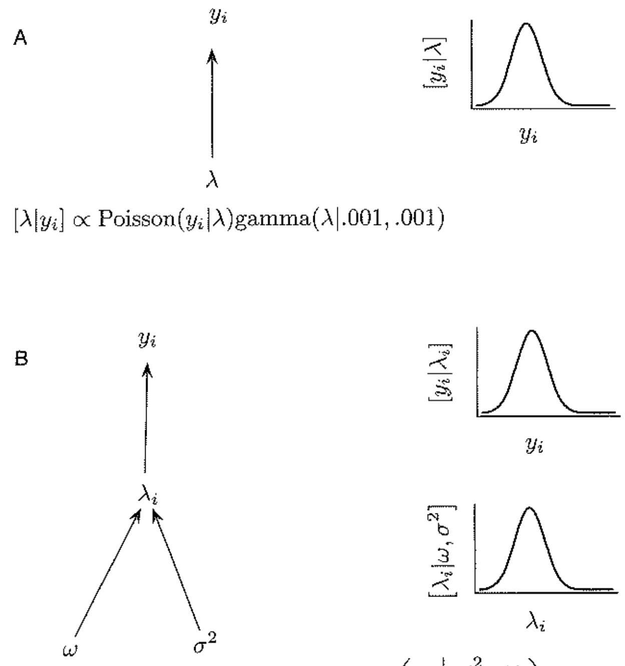
Here is an illustration. Assume we have a single observation (y_i) representing the number of offspring of owl i. We could model average fecundity (λ) using

$$\underbrace{[\lambda|y_i]}_{\text{posterior}} \propto \underbrace{\underbrace{\text{Poisson}\,(y_i|\lambda)}_{\text{likelihood}}\underbrace{\text{gamma}(\lambda|.001, .001)}_{\text{prior}}}_{\text{joint}}. \tag{6.2.2}$$

We will use a Poisson distribution for the likelihood—a logical place to start when modeling count data when we can assume that the variance is approximately equal to the mean.[5] We use a gamma distribution for the prior on λ because it is conjugate to the Poisson. We give numeric arguments to the gamma distribution to make it minimally informative.

Writing the posterior distribution is easy. We simply write 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 is guided by the diagram in figure 6.2.1 A. 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 leading into it is expressed as a prior

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



$$[\lambda|y_i] \propto \text{Poisson}(y_i|\lambda)\text{gamma}(\lambda|.001,.001)$$

$$[\lambda_i,\omega,\sigma^2|y_i] \propto \text{Poisson}(y_i|\lambda_i)\,\text{gamma}\left(\lambda_i \,\middle|\, \frac{\omega^2}{\sigma^2},\frac{\omega}{\sigma^2}\right) \times \text{gamma}(\omega|.001,.001)\,\text{inverse gamma}(\sigma^2|.001,.001)$$

Figure 6.2.1. Bayesian networks for simple (**A**) and hierarchical (**B**) Bayesian models of fecundity of spotted owls assuming a single observation, y_i. 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 Poisson distribution with a mean fecundity (λ) that is the same for all owls. 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 gamma 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 it 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.

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 simple Bayesian models, and your impression is correct. However, these diagrams become more useful in helping visualize and write hierarchical relationships. They are especially helpful (at least for ecologists, if not for statisticians) when there are complex, multilevel 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 of habitats where they establish territories. In this example, we are not trying to determine the causes of individual variation but simply acknowledge that it exists and 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 individual (λ_i) as a random variable drawn from a gamma distribution with mean ω and variance σ^2. We use the diagram in fig. 6.2.1 B as a template to write the posterior and joint distributions:

Moment matching allows posterior to represent mean of the distribution λ's

$$\left[\lambda_i,\omega,\sigma^2|y_i\right] \propto \text{Poisson}(y_i|\lambda_i)\,\text{gamma}\left(\lambda_i \,\middle|\, \frac{\omega^2}{\sigma^2},\frac{\omega}{\sigma^2}\right) \tag{6.2.3}$$

$$\times\ \text{gamma}(\omega|.001,.001)\text{inverse gamma}(\sigma^2|.001,.001). \tag{6.2.4}$$

The likelihood is the same as in our previous model except for 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. The important difference between the simple Bayesian model (eq. 6.2.2) and the hierarchical one (eq. 6.2.3) is the addition of a model for the λ_i; that is, $\lambda_i \sim \text{gamma}(\lambda_i|\frac{\omega^2}{\sigma^2},\frac{\omega}{\sigma^2})$. Assuming that individual owls have fecundities that are drawn from a distribution treats fecundity as a *random effect*, whereas assuming all individuals have the same average fecundity treats fecundity as a *fixed effect* (box 6.2.1). We choose a gamma prior distribution for the population mean fecundity ω because it is continuous and nonnegative. We choose an inverse gamma prior for σ^2 because it is a variance.[6] Both priors have values of parameters chosen to make them minimally informative. No prior is required for λ_i because it occurs on the left-hand side of a conditional—its distribution is determined by the parameters ω and σ^2, which do have priors.

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, pg. 245) offer several examples of inconsistent use of the terms. They recommend dispensing with the use of the term "random effects" altogether and 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 to 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. For example, imagine that you wish to estimate the average aboveground biomass in a grassland. You take a sample of biomass in several 0.25 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 ith plot (y_i) would be

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

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 normal random variable (ϵ_i) to it to represent uncertainty makes no sense, because μ cannot be

(continued)

(Box 6.2.1 *continued*)

negative. Alternatively, the notation $[y_i|\gamma, \beta]$ 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 would somewhat complicate 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, that is, $y_{ij} \sim \text{normal}(\mu_j, \sigma^2)$. When we do this we are treating the μ_j as *fixed* across the locations. (A pooled model would have a single mean regardless of site, i.e., $y_{ij} \sim \text{normal}(\mu, \sigma^2)$.) 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 allow each location to have its own mean biomass drawn from a distribution of means with *hyperparameters*: mean of means equal to α, and variance of means equal to ς^2. Our model then becomes

$$y_{ij} \sim \text{normal}(\mu_j, \sigma_j^2), \tag{6.2.5}$$
$$\mu_j \sim \text{normal}(\alpha, \varsigma^2). \tag{6.2.6}$$

In this case, we are treating the effect of location as random, an effect that 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

$$y_{ij} = \mu_j + \epsilon_{ij}, \tag{6.2.7}$$
$$\mu_j = \alpha + \eta_j, \tag{6.2.8}$$
$$\epsilon_{ij} \sim \text{normal}(0, \sigma_j^2), \tag{6.2.9}$$
$$\eta_j \sim \text{normal}(0, \varsigma^2). \tag{6.2.10}$$

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, such as

$$y_{ij} \sim \text{normal}(\beta_j + \beta_1 x_{ij}, \sigma^2), \tag{6.2.11}$$
$$\beta_j \sim \text{normal}(\mu, \varsigma^2). \tag{6.2.12}$$

[6] Several other distributions (e.g., uniform, inverse gamma, and Cauchy) can be used as priors on variances. However, as you will see later in the book, the inverse gamma is often the distribution of choice because of conjugate relationships (sec. 5.3) for normally distributed random variables. Conjugacy can facilitate model implementation (as we describe in sec. 7.3.2.3). Gamma distributions are used for the inverse of the variance, $1/\sigma^2$, the precision, for the same reason. See Gelman (2006) for a thoughtful discussion of priors for variances in hierarchical models.

Notation that might be puzzling is seen in the parameters for the gamma distribution, $\frac{\omega^2}{\sigma^2}$ and $\frac{\omega}{\sigma^2}$. Where did these come from? The parameters for a gamma distribution are α, the shape, and β, the rate. Recall from section 3.4.4 on moment matching that the mean of the gamma distribution is α/β with variance α/β^2, allowing us to solve for α and β terms of the mean and variance; that is, $\alpha = \omega^2/\sigma^2$, $\beta = \omega/\sigma^2$. The average fecundity for the population is $\omega = \alpha/\beta$.

These clarifications make an important point about drawing Bayesian networks and converting them into mathematical expressions. Recall 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., fig. 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 the model as

$$[\boldsymbol{\lambda}, \alpha, \beta|\mathbf{y}] \propto \prod_{i=1}^{n} \text{Poisson}(y_i|\lambda_i)\ \text{gamma}(\lambda_i|\alpha, \beta) \\ \times \text{gamma}(\alpha|.001, .001)\,\text{gamma}(\beta|.001, .001). \qquad (6.2.13)$$

The point is that Bayesian networks are thinking tools—graphical aids for properly writing 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 embellish the example from Clark (2003) 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 midlife before declining (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 ($x_i = \widetilde{x}_i - x_{0,i}$) where $\widetilde{x}_i$ is the chronological age of the ith individual, $x_{0,i}$ is the age of first reproduction. Thus, an animal is $x_i = 0$ when it first reproduces. Defining age this way makes it convenient to interpret the intercept.

[7] Actually, most statisticians would prefer this form, because it focuses attention on the parameters.

We now model the process "change in fecundity with age" $g(\alpha, \boldsymbol{\beta}, x_i)$ as

$$g\,(\alpha,\boldsymbol{\beta}, x_i) = \alpha + \beta_1 x_i + \beta_2 x_i^2, \qquad (6.2.14)$$

$$\left[\boldsymbol{\lambda}, \boldsymbol{\beta}, \alpha, \sigma_p^2|\mathbf{y}\right] \\ \propto \prod_{i=1}^{n} \text{Poisson}(y_i|\lambda_i)\text{gamma}\left(\lambda_i \,\middle|\, \frac{g\,(\alpha, \boldsymbol{\beta}, x_i)^2}{\sigma_p^2}, \frac{g\,(\alpha, \boldsymbol{\beta}, x_i)}{\sigma_p^2}\right) \qquad (6.2.15) \\ \times \prod_{j=1}^{2} \text{normal}(\beta_j|0, 100)\ \text{normal}(\alpha|0, 100) \\ \times \text{inverse gamma}(\sigma_p^2|.001, .001),$$

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 the influences that create variation in fecundity beyond the effect of the bird's age. It is also important to see that we have replaced the parameter ω with a model $g\,(\alpha, \boldsymbol{\beta}, x_i)$ that exploits observations on an owl's age and our understanding of the relationship between age and fecundity.

Again, we choose a gamma distribution for λ_i because it is continuous and nonnegative. We could also use other continuous distributions with nonnegative support, for example, the lognormal. The distribution for λ_i could be viewed as a "prior" informed by our process model. We use an inverse gamma distribution for the flat prior on σ_p^2 because it is a variance (but see Gelman 2006). We choose a normal distribution for the β's because they are continuous random variables that can take on any real value. To minimize the information contained in the priors for the β's 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

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

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 $g(\alpha, \boldsymbol{\beta}, x_i)$. Sometimes, the predictor variables *do* appear in the posterior distribution, and we describe these cases in a subsequent example and in box 6.2.2.

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 $g(\alpha, \boldsymbol{\beta}, x_i)$ 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 ith individual, γ is the maximum potential fecundity during the first reproduction, and 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 $g(\alpha_i, \boldsymbol{\beta}, x_i)$, where

$$\alpha_i \sim \text{gamma}\left(\frac{h(\gamma, \nu, u_i)^2}{\varsigma_p^2}, \frac{h(\gamma, \nu, u_i)}{\varsigma_p^2}\right). \tag{6.2.16}$$

The parameter ς_p^2 is the process variance associated with the territory model, including all the influences on an individual's fecundity at first reproduction that are not determined by territory size.[9]

We can now see the relationship between this model and the general template we outlined in equation 6.2.1:

$$[\boldsymbol{\theta}_p, \boldsymbol{\theta}_o, \mathbf{z}|\mathbf{y}] \propto \underbrace{[\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}_o]}_{\text{data}} \underbrace{[\mathbf{z}|\boldsymbol{\theta}_p]}_{\text{process}} \underbrace{[\boldsymbol{\theta}_o]\,[\boldsymbol{\theta}_p]}_{\text{parameters}} \tag{6.2.17}$$

$$\underbrace{[y_i|z_i, \boldsymbol{\theta}_o]}_{\text{data}} = \text{Poisson}(y_i|\lambda_i)$$

$$\underbrace{[z_i|\boldsymbol{\theta}_p]}_{\text{process}} = \text{gamma}\left(\lambda_i \,\middle|\, \frac{g(\alpha_i, \boldsymbol{\beta}, x_i)^2}{\sigma_p^2}, \frac{g(\alpha_i, \boldsymbol{\beta}, x_i)}{\sigma_p^2}\right)$$

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

[9] We point out that this is the first time you have seen the symbol ς, which is called *variant sigma*. We will use it often in the remainder of the book when we use more than one variance term in a model.

$$\underbrace{[\boldsymbol{\theta}_o]\,[\boldsymbol{\theta}_p]\,[\gamma][\nu]}_{\text{parameters}} = \prod_{j=1}^{2} \text{normal}(\beta_j|0, 100)\ \text{gamma}(\gamma|.001, .001)$$

$$\times \text{gamma}(\nu|.001, .001)\ \text{inverse gamma}(\sigma_p^2|.001, .001)$$

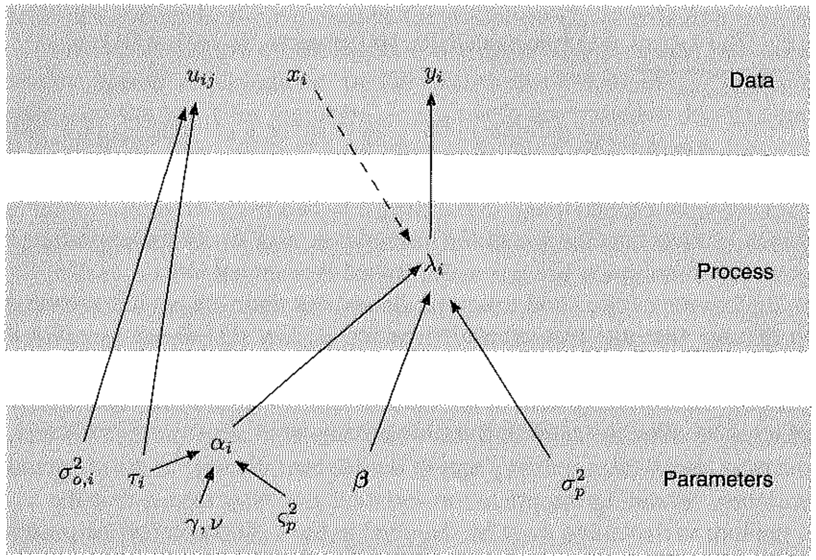
$$\times \text{inverse gamma}(\varsigma_p^2|.001, .001).$$

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 life, 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 $j = 1, \ldots, 3$ observations of territory size for each bird. We can now think of an observation of territory size as a random variable arising from $[u_{ij}|\tau_i, \sigma_{o,i}^2]$, where τ_i is the true, unobserved territory size of the ith bird, and $\sigma_{o,i}^2$ is the observation variance. Modeling the predictor variables this way means that the u_{ij} *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.

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.[10] In our first hierarchical model (eq. 6.2.13), 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 (eq. 6.2.15), 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 (eq. 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 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 nonlinear. In the fourth model (fig. 6.2.2),

[10] 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, but doing so would probably require repeated observations of the fecundity of each individual.



$$
\begin{aligned}
g(\alpha_i,\beta,x_i) &= \alpha_i + \beta_1 x_i + \beta_2 x_i^2 \\
h(\gamma,\nu,\tau_i) &= \gamma e^{-\nu\tau_i} \\
[\boldsymbol{\lambda},\boldsymbol{\alpha},\boldsymbol{\beta},\gamma,\nu,\boldsymbol{\tau},\boldsymbol{\sigma}^2_o,\sigma^2_p,\varsigma^2_p|\mathbf{y},\mathbf{u}] &\propto \prod_{i=1}^{n} \text{Poisson}\,(y_i|\lambda_i)\,\text{gamma}\left(\lambda_i \,\middle|\, \frac{g(\alpha_i,\beta,x_i)^2}{\sigma^2_p}, \frac{g\,(\alpha_i,\beta,x_i)}{\sigma^2_p}\right) \\
&\times \text{gamma}\left(\alpha_i \,\middle|\, \frac{h(\gamma,\nu,\tau_i)^2}{\varsigma^2_p}, \frac{h(\gamma,\nu,\tau_i)}{\varsigma^2_p}\right) \\
&\times \prod_{j=1}^{3} \text{gamma}\left(u_{ij} \,\middle|\, \frac{\tau_i^2}{\sigma^2_{o,i}}, \frac{\tau_i}{\sigma^2_{o,i}}\right) \text{gamma}\,(\tau_i|.001,.001) \\
&\times \text{inverse gamma}\,(\sigma^2_{o,i}|.001,.001)\,\text{gamma}\,(\nu|.001,.001) \\
&\times \text{gamma}\,(\gamma|.001,.001) \prod_{k=1}^{2} \text{normal}\,(\beta_k|0,100) \\
&\times \text{inverse gamma}\,(\sigma^2_p|.001,.001) \\
&\times \text{inverse gamma}\,(\varsigma^2_p|.001,.001)
\end{aligned}
$$

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 of an individual (λ_i) is modeled as a quadratic function of the owl's age (x_i) with parameters α_i, β_1, β_2. We assume age is known. Variation in the λ_i not captured by the model is represented by σ^2_p. We assume that the parameter α_i, the fecundity of owl i at first reproduction, decreases exponentially with increasing territory size τ_i. Observations of territory size (u_{ij}) arise from a distribution with mean τ_i and observation variance $\sigma^2_{o,i}$. 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^2_p. 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.

we add uncertainty in observations of territory size. The observed territory size is a random variable arising from a distribution governed by the true territory size (τ_i) and measured observation variance ($\sigma^2_{o,i}$).

We must deal with two parts of equation 6.2.1[illegible] that might be confusing. First are the sources of uncertainty. We have an explicit parameter for the process variance (σ^2_p), but there doesn't appear to be a parameter controlling variance in the observations. Does that mean we assume there is no observation variance? The answer is no. Remember that the variance of the Poisson distribution is the same as the mean, so the observation variance is implicit in the likelihood—variance that is highly constrained. We also raise a caution here, which we treat in more detail in section 6.3. The fecundity model lacks replication at the level of the individual; that is, we observe only a single fecundity for each owl. This means we would probably not be able to separate observation variance from process variance if the distribution for the likelihood were less constrained than the Poisson. We discuss this important issue more fully in section 6.3.

Box 6.2.2 When Are Predictor Variables Included in the Posterior Distribution?

A common error in writing expressions for the posterior and joint distribution is to include predictor variables, that is, the **x**, on the right-hand side of the conditioning in the posterior distribution, when we assume (rightly or wrongly) that the **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.

(continued)

(Box 6.2.2 *continued*)

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, 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 drop the $g()$ wrapper to make the factoring more clear. Consider the *wrong* expression for the posterior and joint distributions,

$$\left[\boldsymbol{\theta}, \sigma_p^2 | y_i, x_i\right] \propto \left[y_i, \boldsymbol{\theta}, \sigma_p^2, x_i\right], \tag{6.2.18}$$

$$\left[\boldsymbol{\theta}, \sigma_p^2 | y_i, x_i\right] \propto \left[y_i | \boldsymbol{\theta}, \sigma_p^2, x_i\right] [\boldsymbol{\theta}] \left[\sigma_p^2\right] [x_i], \tag{6.2.19}$$

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

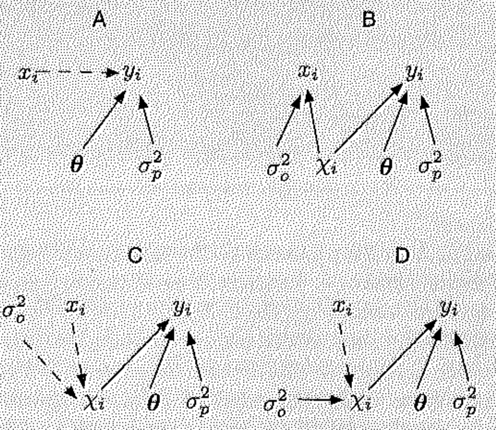
$$\left[\boldsymbol{\theta}, \sigma_p^2 | y_i\right] \propto \left[y_i, \boldsymbol{\theta}, \sigma_p^2\right], \tag{6.2.20}$$

$$\left[\boldsymbol{\theta}, \sigma_p^2 | y_i\right] \propto \left[y_i | \boldsymbol{\theta}, \sigma_p^2, \right] [\boldsymbol{\theta}] \left[\sigma_p^2\right], \tag{6.2.21}$$

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

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

to highlight the deterministic model $g(\boldsymbol{\theta}, x_i)$.



(continued)

(Box 6.2.2 *continued*)

Sometimes, we treat the predictor variables as random variables because we wish 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 $g(\boldsymbol{\theta}, \chi_i)$. As shown in panel B, we now have an expression for the posterior that factors correctly:

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

One more point bears mentioning. Models for predictor variables that take the form $\left[\chi_i | x_i, \sigma_o^2\right]$ 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). In this case, the expression for the posterior and joint distributions is

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

Again, the deterministic model is $g(\boldsymbol{\theta}, \chi_i)$. Note that there is no 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,

$$\left[\boldsymbol{\theta}, \sigma_p^2, \sigma_o^2, \chi_i | y_i\right] \propto \left[y_i | \boldsymbol{\theta}, \sigma_p^2, \chi_i\right] \left[\chi_i | x_i, \sigma_o^2\right] [\boldsymbol{\theta}] \left[\sigma_p^2\right] \left[\sigma_o^2\right]. \tag{6.2.24}$$

6.2.2 Multilevel Models: Controls on Nitrous Oxide Emissions from Agricultural Soils

Data in ecological research are often collected at multiple scales or levels of organization in nested designs (fig. 6.2.3). "Group" is a catchall term for the upper level in many different types of nested hierarchies. Groups can logically be formed by populations, locations, species, treatments, life stages, and individual studies. Measurements are taken within groups on individual organisms, plots, species, time periods, and so on. Measurements may also be taken on the groups themselves, that is, covariates that apply at