Abstract
The 21st century is witnessing dramatic changes in the way that datasets are being analyzed. Data collection is more and more focused on observational data rather than designed experiments. Datasets are growing increasingly larger. This is particularly true for environmental and ecological data settings. This chapter suggests that a valuable way to infer from such datasets is through formal stochastic model specification. That is, in order to understand and explain the behavior of a complex environmental or ecological process, it
can be most effective to build process-based models. Such models introduce simplification but attempt to capture key features of the process and link the model generatively to the data. This chapter attempts to provide a general overview of such modeling enterprise. Many of the subsequent chapters develop sophisticated examples, fabricated according to their context.

2.1 Introduction

In a handbook on environmental and ecological statistics, the swath of statistical work to review is enormous. Statistical analyses span the spectrum from purely descriptive presentation to sophisticated optimization strategies to multi-level modeling. Over the ensuing chapters, the full range of techniques will be presented for learning about complex environmental and ecological processes from associated data. However, this chapter suggests that a contemporary perspective adopts a fully model based approach.

Arguing that the goal of a statistical analysis is formal inference, we devote this chapter to stochastic modeling, hierarchical modeling (usually handled within a Bayesian framework), identifying sources of modeling error, incorporating uncertainty appropriately, propagating uncertainty, and model assessment - adequacy and comparison. All of these issues will be considered in the subsequent chapters; the goal here is present an overview. Moreover, the amount of material is far more than this chapter can accommodate; each topic could fill more than a chapter. As a result, additional references will be supplied.

Perhaps the most important message to take away is the encouragement of coherent modeling. We prefer to work with models that are coherent in the sense that they are generative; the specification could have yielded the data that were observed. While all models are simplifications of the process under investigation, within environmental science and ecology, when we observe realizations from a process it seems appropriate to ask that our simplified explanation of the process could yield them. This does not assert that there is only one model specification to consider. There can be many plausible generative specifications. This is why we need to assess model adequacy and model choice. Moreover, when model specifications become too complex to enable tractable fitting, we will readily accede to suitable approximation. However, this does not deny an objective of a coherent model.

The format of this chapter is as follows. Section 2 provides a general perspective on stochastic modeling. Section 3 focuses on the Bayesian paradigm. Section 4 opens up the door to hierarchical modeling, the primary theme of the chapter. Section 5 leads us to latent variables, a catchall for many of the unknowns introduced in specifying hierarchical models. Section 6 briefly looks at mixture models while Section 7 takes us back to random effects with linkages to other chapters in this Handbook. Section 8 offers some words on dynamic models, connecting to a subsequent chapter with fuller development. Sections 9 and 10 consider model assessment, adequacy in the former, comparison in the latter. We conclude in Section 11 with a brief summary.

2.2 Stochastic modeling

Again, here we work with probability model specifications. In particular, such a model provides a distributional specification for observed data Y given unknown parameters, θ,
Modeling for environmental and ecological processes

which we write as \( f(y|\theta) \). This expression is almost too general to be useful. What is \( Y \)? For instance, what sorts of variables is \( Y \) composed of - continuous, categorical, ordinal, nominal? Do we have vectors of observations? Do we have observations that are independent or are they dependent? Are they indexed temporally, spatially, or both?

What is \( \theta \)? If we envision a process that generates \( Y \), then \( \theta \) can be envisioned as the process. Practically, the process is typically specified through a finite collection of parameters which we can take as \( \Theta \). Again, this is almost too general to be useful. Moreover, it raises the question of whether parameters are “real”? Since they are features of a model to explain \( Y \) and since the specified model is essentially never the true model, rather an approximation, these parameters need not be real. This does not suggest that it is not useful to infer about these parameters. Such inference can facilitate how covariates/predictors affect responses.

As a result, before continuing discussion of inference for \( \theta \) under \( f(Y|\theta) \), we digress to remark upon parametric inference, usually called estimation vs. predictive inference. The point here is that estimation takes place in the \( \Theta \) space while prediction takes place in the \( Y \) space. Inference about \( \theta \) can not be compared with any truth. On the other hand, inference about \( Y \), perhaps a new \( Y_0 \), takes place in the space of the data and can be compared with observed data, perhaps data held out from fitting the model. Predictive inference has a long history in the Statistics literature [25]. It is playing an increasingly important role in current statistical work with observational, rather than designed, data collection.

In any event, the implication is that, while a model may be used for both estimation and prediction, assessment of model performance, i.e., model adequacy and model comparison, should be carried out in the data space. We will elaborate this theme further in Sections 9 and 10 below.

Model selection is a critical issue in developing explanations for environmental and ecological processes. There is a huge literature on model selection. It is impossible to make a case for any particular criterion as being “best.” Even if one is prepared to specify the utility for a model, there still will not be a mutually agreed upon criterion. We elaborate this a bit further in Section 10 but defer fuller discussion to the literature [13, 17].

A specific model selection challenge is variable selection. When models are developed for explanation, that is, regression models, then, according to application, we can have a rich range of covariates to consider. With regard to environmental and ecological processes, the list can be substantial. For instance, with regard to explaining biodiversity, we have factors like history, habitat heterogeneity, competition, climate and climate variability, productivity, and disturbance. Each of these factors can be captured with a variety of regressors. And, these regressors can interact with each other producing more regressors. Evidently, explanation can be very complicated. Of course, only a subset will be available in a given application but still the selection of variables to include in the model can become an issue.

The need for variable selection arises because with, for example \( k \) predictors, we have \( 2^k \) models. With increasing \( k \), we very quickly exceed the possibility of exploring all of them. Usually, variable selection has, as a general goal, parsimony; we seek simpler explanations. If a model is essentially as complicated as the process, then it may not readily facilitate our understanding of the process. Apart from the size of the model space, variable selection is further complicated by the multicollinearity between variables. In different words, different subsets of variables can perform equally well with regard to explanation. In a high dimensional model space, we can find many regions where model performance is indistinguishable. And, since none of the models are true, it can be difficult to prefer one to another. We often fall back to an initial selection of variables that, from a process perspective, are agreed to be important and then investigate the addition of further variables. Variable selection criteria can also provide a measure of importance for each variable. See Section 9.

Further complications are the following. We can have different distributional specifications for models, different ways to capture stochasticity. This takes us back to model
comparison and may overarch variable selection. We can have multilevel models where explanation can enter at different levels (see Sections 4-8 below). We can have nonlinearities in explanation. That is, usual modeling attempts to explain the mean, perhaps on a transformed scale, and introduces regressors linearly, a generalized linear model [13, 48]. However, with environmental and ecological processes, most relationships between variables are nonlinear. The nature of the nonlinearity, e.g., exponential, asymptotic, sigmoidal, convex, concave, etc. has been gleaned to some extent from various historical experiments. Nonlinear models are much more challenging to fit and have received little attention with regard to variable selection.

Altogether, there is too much to say about variable selection here. We offer some very brief discussion below but, since this issue is not an important focus of this chapter, we are only able to supply some useful references to the literature. Not surprisingly, this literature is, by now, enormous [13, 17], and is an acknowledged challenge across many fields.

Random effects play a critical role in the ensuing modeling ideas. They have a long history in the literature [see, e.g., 56] but in the past twenty years have emerged with a new purpose. That is, traditional random effects modeling was introduced primarily to recognize that in some experiments, individuals selected for treatments or in fact, treatments themselves, e.g., plots, arose randomly from populations. Hence, the effects that they introduced were random rather than fixed in the design and that it was vital to incorporate this source of randomness into the modeling in order to better assess significance of effects. Much of this work played out in designed experiments, typically in analysis of variance or covariance settings.

Again, the past twenty years has seen a dramatic shift toward the analysis of observational data. In this context, the role of random effects has changed dramatically. Rather than being a nuisance because effects could not comfortably assumed fixed, now they have become a valuable modeling tool. They serve as devices for model enrichment, for clarifying uncertainty.

To elaborate a bit, they are often introduced as surrogates for unobservable regressors in a process specification. That is, they flexibly soak up a portion of residual variability from a model with fixed regressors. They can capture dependence, e.g., if we obtain a vector of measurements for a set of individuals, assigning a common random effect to each individual in the modeling makes the measurements associated with that individual dependent. They can capture structured dependence in time, e.g., autoregressive behavior. They can capture dependence in space to reflect stronger association for observations closer to each other in space. In fact, they can capture space-time dependence, allowing space-time interaction in dependence. All of these examples suggest hierarchical structure. That is, the random effects may be introduced at the first stage of the modeling but their distributional specifications will be supplied at a second level of modeling. We illustrate some of these ideas below but they are fleshed out much more fully in the ensuing chapters.

The likelihood is the starting place for essentially all of the modeling in this handbook. Some chapters will take the likelihood and add a prior, leading to Bayesian modeling and inference, the focus of much of the remainder of this chapter. But, at the least, we would want a likelihood which, viewed as a density for the data, becomes a generating mechanism given parameters. It can be anticipated that likelihood-based inference will be a benchmark in formal statistical analysis into the foreseeable future. In this regard, the Likelihood Principle, in its simplest form, asserts that inference should only depend upon the data observed but not on the sampling mechanism used to obtain the data [10]. The process is what it is and does not change because of the way we collect the data\footnote{The Likelihood Principle is not universally supported from a philosophical perspective. See, e.g., [47]}. However, in practice,
fully specifying a stochastic data generation mechanism and then employing the associated likelihood will usually provide inference with good properties.

The likelihood takes the density for the data, say \( f(y|\theta) \), and views it as a function of \( \theta \) for fixed, observed data \( y \). The customary point estimate for \( \theta \) is the maximum likelihood estimate (MLE), the value of \( \theta \) which maximizes the likelihood. The literature on MLE’s is enormous [see, e.g., 41], and, though the concept is older, it has been popular in the Statistics community for more than 100 years, dating to the work of Fisher [4]. Optimality properties for MLE’s under fairly general conditions are well-established. Asymptotic uncertainty is routinely obtained using Fisher information matrices. The EM algorithm [18] is very widely used, extending likelihood inference to missing and incomplete data settings.

However, working with the likelihood in the 21st century brings challenges. Maximization of a function for a high dimensional \( \theta \) is demanding; local optima are common. Inference is limited to a point estimate and an asymptotic covariance matrix, typically adding asymptotic normality for interval estimation. Though we may have rates of convergence, obtained technically, we often can’t employ them usefully in practice. Perhaps, most importantly, likelihood inference is not a good fit with general hierarchical modeling; in fact, the likelihood itself is not uniquely determined (see Section 4). The Bayesian framework much more easily accommodates such models, as we argue for the remainder of this chapter.

### 2.3 Basics of Bayesian inference

We envision that multi-level models will be needed to build effective stochastic models for complex processes. Inference and fitting for such models is most directly handled within a Bayesian framework. So, here, we briefly remind the reader of the basics of Bayesian inference.

To start, how did Bayes’ Theorem, an elementary result in probability, become an inference paradigm and, in fact, a controversial one? Recall the theorem in its simplest form, with two events, \( A \) and \( B \), i.e., \( P(A|B) = \frac{P(B|A)P(A)}{P(B)} \). Suppose we move to events associated with random variables. We obtain \( P(X \in A|Y \in B) = \frac{P(Y \in B|X \in A)}{P(Y \in B)} \). Then, we take the simple step to densities, yielding \( f(x|y) = \frac{f(y|x)f(x)}{f(y)} \).

Finally, letting \( Y \) denote what you observe, i.e., the data in the previous section and replacing \( X \) with \( \theta \) denoting the parameters, as in the previous section, i.e., what you don’t know (didn’t observe), we arrive at

\[
f(\theta|Y) = \frac{f(Y|\theta)\pi(\theta)}{f(Y)} \tag{2.1}
\]
or

\[
f(\theta|Y) \propto f(Y|\theta)\pi(\theta). \tag{2.2}
\]

So, we have an intuitively appealing inference paradigm using the density on the left side: infer about what you don’t know given what you have seen. This contrasts with the classical inference approach, using sampling distributions, \( T(Y) \) given \( \theta \) which asks you to imagine what you might see given what you don’t know.

In fact, we are really just thinking about two ways of writing a joint distribution:

\[
f(Y, \theta) = f(Y|\theta)\pi(\theta) = f(\theta|Y)f(Y). \tag{2.3}
\]

In this specification, we see our friend from the previous section, the model for \( Y \) given
\( \theta \), the likelihood. The specification adds a distribution for \( \theta \), the prior. The result is a generative specification: First a value of \( \theta \) is drawn from \( \pi(\theta) \) and then a realization of the data is achieved by using this \( \theta \) in \( f(y|\theta) \) to obtain a \( Y \).

Conditioning in the reverse order provides so-called posterior inference, again inferring about \( \theta \) given the observed \( Y \). Particularly attractive is that with \( f(\theta|Y) \), we obtain an entire distribution to use for inference about \( \theta \). Apart from familiar measures of centrality and spread, we can obtain quantiles and probabilities of events for \( \theta \). This would seem to be much more satisfying than say the results from a likelihood analysis, i.e., a point estimate, an asymptotic variance, and a confidence interval, typically justified by an appropriate central limit theorem.

Furthermore, prediction is an immediate probabilistic consequence of the specification. That is, \( f(Y_0|Y) \) arises from

\[
f(Y_0|Y) = \int f(Y_0|Y, \theta) f(\theta|Y) d\theta. \tag{2.4}
\]

The first term under the integral becomes \( f(Y_0|\theta) \) if \( Y_0 \) is independent of \( Y \). If not, it would be obtained from the joint distribution, \( f(Y, Y_0|\theta) \) which would be available from the model specification.

### 2.3.1 Priors

The controversial ingredient in the above is \( \pi(\theta) \), the prior. There are two issues here. Should \( \theta \) be viewed as fixed or random? Then, if it is viewed as random, where does \( \pi(\theta) \) come from? The case for assuming \( \theta \) to be random is based upon the idea that, if the model is not true, then there is no true \( \theta \). If we allow \( \theta \) to be random, we incorporate the fact that there is no fixed \( \theta \) into the model. We can more appropriately propagate uncertainty in inference regarding \( \theta \) as well as in prediction of data. We could simply say, naively, that if what we observe is random, shouldn’t what we can not observe also be taken as random.

However, once we assume \( \theta \) to be random, then we have to specify a distribution for it. This is where the controversy lies. If you choose one prior and I choose another, whose inference should be adopted? Is there a role for subjectivity in statistical inference? The argument can be made that the prior enables you to incorporate any information you have about the unknowns in the model specification. The case can be made that we always know something about unknowns, e.g., we always have some idea of what the magnitude of regression coefficients can be because, when applied to regressors, we can see if they provide predictions outside the range of what is observable. Similarly, we can argue that we always know something about variances. If uncertainty is allowed to be too large, again, realizations beyond what is observable will occur. Furthermore, we can (and should) engage in sensitivity analysis to the prior. Typically, the data contribution regarding \( \theta \) will be much stronger than that of the prior so there will be little prior sensitivity.

What does the prior do mathematically? If we look at the product, \( f(Y|\theta)\pi(\theta) \), we see that the prior rescales the likelihood, changing its shape, perhaps its mode, perhaps its spread. What does this mean practically? The prior and the likelihood combine multiplicatively to provide our information about \( \theta \). Intuitively, we arrive at an inferential compromise between the two functions, which becomes the posterior distribution, to provide inference about \( \theta \). In practice, we often use what we call weak, vague, or noninformative priors. These terms themselves are vague and, at times, their impacts are not properly appreciated [40].

In the Bayesian literature there is substantial discussion of objective priors, so-called automatic specifications to remove potential subjectivity. These priors go under various names, e.g., Jeffreys priors, reference priors, g-priors, etc. [8]. For instance, for location parameters, they are just constant over \( R^1 \), for scale parameters, they are the reciprocal of
the parameter or the square of the reciprocal. Typically these priors are improper (i.e., not integrable) raising the possibility that, with an improper joint distribution, we could have an improper posterior [9]. There is also a substantial literature on elicitation [see, e.g., 52, and references therein]. In some settings this can be a valuable exercise but, for most of the hierarchical modeling that is currently employed, there is not much of a role for elicitation. That is, elicitation makes sense for parameters that are, in some sense real. In much of hierarchical modeling, we introduce many parameters for which elicitation would not be feasible.

The fundamental issue here is whether we view the prior as an inference device as opposed to a component of a model specification to provide a generative specification. As suggested above, if you define a parameter, you will almost always know something about it so you can always propose a sensible proper prior and maintain a generative model. In any event, to avoid potentially improper posteriors, perhaps it is best to play it safe and adopt proper priors. A last word here, returning to weak or vague priors, if they are nearly improper, then the posterior may be nearly improper. With usual Markov chain Monte Carlo or Gibbs sampling used for model fitting (see Section 3.3), it may be difficult to see nearly improper posteriors.

### 2.3.2 Posterior inference

Armed with an entire posterior distribution for an unknown, inference becomes nothing more than presenting features of the distribution. We can provide customary measures of centrality like posterior means or medians. We can provide customary measures of uncertainty like posterior variances, standard deviations, ranges. We can provide quantiles and probability statements. In particular, if we provide a predictive interval for an unknown, it is a probability statement for that unknown, a so-called credible interval. It is not a confidence interval created by pivoting some sampling distribution. For a full development of Bayesian inference, see [23]. We defer discussion of Bayesian hypothesis testing, equivalently model comparison, to Section 10.1.

### 2.3.3 Bayesian computation

The foregoing elegance and clarity attached to Bayesian inference ignores the significant computational challenges associated with implementing the inference. The explicit challenge is integration. That is, (2.2) expresses the posterior up to normalizing constant. This constant is needed to calculate probabilities with the this distribution. Further integration is required to calculate any moments of the distribution. In practice, none of these integrations are available explicitly. When \( \theta \) is high dimension, it is not feasible to handle these integrations numerically. The breakthrough that enables feasible Bayesian computation for the rich classes of models we are interested in replaces integration by sampling; to learn about a high dimensional posterior distribution, we just sample it as much as we wish. These days, sampling high dimensional distributions is accomplished, fairly straightforwardly, using Markov chain Monte Carlo (MCMC) and Gibbs sampling [58]. These techniques have become the workhorses of Bayesian computation in the 21st century. See [23] or [58] for full discussions. The remaining material in this chapter as well as the subsequent chapters that fit hierarchical models will assume familiarity with MCMC and Gibbs sampling for model fitting.
2.4 Hierarchical modeling

As noted in the abstract, moving into the second decade of the 21st century, we are wit-
nessing a dramatic paradigm shift in the way that statisticians collaborate with researchers
from other disciplines. Disappearing are the days when the statistician was called in at the
end of a project to provide some routine data analysis and some summary displays. Now the
statistician is an integral player in a research team, helping to formulate hypotheses, iden-
tify data needs, develop suitable stochastic models, and implement fitting of the resulting
challenging models.

As part of this shift, there is increasing attention paid to bigger picture science, to looking
at complex processes with an integrative perspective, to bringing a range of knowledge and
expertise to this effort. Increasingly, we find researchers working with observational data,
less with designed experiments, recognizing that the latter can help inform about the former
but the gathering of such experiments provides only one source of data for learning about
the complex process. Other information sources, empirical, theoretical, physical, etc. should
also be included in the synthesis.

The primary result of all of this is the development of a multi-level stochastic model
which attempts to incorporate the foregoing knowledge, inserting it at various levels of the
modeling, as appropriate. Following the vision of Mark Berliner [11], we imagine a general
three stage hierarchical specification:

\[
\begin{align*}
\text{First stage:} & \ [\text{data} | \text{process}, \text{parameters}] \\
\text{Second stage:} & \ [\text{process} | \text{parameters}] \\
\text{Third stage:} & \ [(\text{hyper})\text{parameters}].
\end{align*}
\]

The simple form of this specification belies its breadth. The process component can
include multiple levels. It can be dynamic, it can be spatial. The data can be conditioned
on whatever aspects of the process are appropriate. The stochastic forms can be multi-
variate, perhaps infinite dimensional with parametric and/or nonparametric specifications.
Moreover, while the focus here is on applications in environment and ecology, the range of
applications to which this generic specification has been applied runs the scientific gamut,
e.g., biomedical and health sciences, economics and finance, engineering and natural science,
political and social science.

In view of the above, hierarchical modeling has taken over the landscape in contem-
porary stochastic modeling. Though analysis of such modeling can be attempted through
nonBayesian approaches [57], working within the Bayesian paradigm enables exact inference
and proper uncertainty assessment (see below) within the given specification.

We can immediately extend the Bayesian framework of the previous section to the hier-
archical setting. Specifically, we add one further level, parametrically. Now, we would have
a prior distribution \( \pi(\theta | \lambda) \), where \( \lambda \) is a vector of hyperparameters. Connecting with (2.5),
we can think of \( \theta \) even more generally as the “process” of interest with some parts known
and some parts unknown. Then, we can write \( f(y|\text{process}, \theta)f(\text{process}, \theta | \lambda)\pi(\theta | \lambda)\pi(\lambda) \),
evidently, a hierarchical specification. If \( \lambda \) is known, the posterior distribution for \( \theta \) is given
by

\[ p(\theta|y, \lambda) = \frac{p(y, \theta|\lambda)}{p(y|\lambda)} = \frac{p(y, \theta|\lambda)}{\int p(y, \theta|\lambda) \, d\theta} = \frac{f(y|\theta)p(\theta|\lambda)}{\int f(y|\theta)p(\theta|\lambda) \, d\theta} = \frac{f(y|\theta)p(\theta|\lambda)}{m(y|\lambda)}. \]

In practice, \( \lambda \) will not be known. A second stage hyperprior distribution \( h(\lambda) \) will be required, as proposed in (2.5), so that

\[ p(\theta|y) = \frac{p(y, \theta)}{p(y)} = \frac{\int f(y|\theta)p(\theta|\lambda)h(\lambda) \, d\lambda}{\int f(y|\theta)p(\theta|\lambda)h(\lambda) \, d\theta \, d\lambda}. \] (2.6)

This is the path that is usually pursued. However, alternatively, we might replace \( \lambda \) in \( p(\theta|y, \lambda) \) by an estimate \( \hat{\lambda} \). This is called empirical Bayes analysis [14]. Because such estimation removes the uncertainty associated with \( \lambda \), uncertainty in overall inference will tend to be underestimated. However, in practice, with models of sufficient complexity, we often implement a bit of empirical Bayes inference, using a bit of data-based fixing of some parameters.

An issue with likelihood based inference for hierarchical models emerges. What is the likelihood? From the multi-level model, we can marginalize over \( \lambda \) to obtain \( p(y|\theta)\pi(\theta) \) where \( \pi(\theta) = \int \pi(\theta|\lambda)h(\lambda) \, d\lambda \), interpreting \( p(y|\theta) \) as the likelihood. Alternatively, we can marginalize over \( \theta \) obtaining \( p(y|\lambda)h(\lambda) \) where \( p(y|\lambda) = \int \theta p(y|\theta)\pi(\theta|\lambda) \, d\theta \). Now the likelihood would become \( p(y|\lambda) \). Which likelihood should we work with? One might argue that we should focus on the parameters of interest. However, it may be more satisfying to fit the full hierarchical model, enabling posterior inference regarding all of \( \theta \) and \( \lambda \).

### 2.4.1 Introducing uncertainty

The generality reflected in [11] enables rich model specification. In fact, as hierarchical modeling has taken over a substantial portion of the statistical modeling landscape, we are finding increasingly challenging examples. With such richness comes the recognition of the opportunity for introducing uncertainty and how to do it appropriately. Again, the adopted stochastic models are only approximations to the complex process so error will always be introduced. The benefit of the transparency implicit in building multi-level models is that it allows us to determine where and how to introduce error. And, once we do, the fact that we are working in a fully probabilistic setting, enables the uncertainty to be properly propagated to the posterior distribution.

What are the various types of uncertainty to consider? Possibilities include: stochastic uncertainty reflecting not just the noise in realizations but also the randomness with regard to where sampling units are; measurement uncertainty (experimental uncertainty), reflecting measurement error in the devices recording the observations; parameter uncertainty, evidently captured in some fashion through prior specifications; model uncertainty, reflecting missing or unobservable information which are part of the process but not accessible; uncertainty in model dimension or functional uncertainty, reflecting the fact that functional forms employed in relating variables in the specifications are only approximate; and multiple model uncertainty, reflecting the interplay between the modeling levels, how specification at one level influences uncertainty at another level. The implications of introducing uncertainty in various ways to achieve various objectives will not be assessable analytically. Rather, a substantial amount of model fitting and model comparison will be required to make such assessments.
2.4.2 Random effects and missing data

Arguably, the utilization of hierarchical models initially blossomed in the context of handling random effects and missing data, using the E-M algorithm [18] for likelihood analysis and Gibbs sampling [30] for fully Bayesian analysis. In this subsection, we offer some elementary remarks, first on random effects, then on missing data.

With regard to random effects, both classical and frequentist modeling specify a stochastic model for these effects, usually assumed to be a normal distribution with an associated variance component. These effects can be introduced at different levels of the modeling but, regardless, in much of the literature, they are assumed to be exchangeable, in fact i.i.d.

More recently, we are seeing random effects with structured dependence in, e.g., dynamic, spatial and spatio-temporal models (see Chapters 4 and 5).

A typical linear version with i.i.d. effects takes the following form. At the first stage:

\[ Y_{ij} = X_{ij}^T \beta + \phi_i + \epsilon_{ij}. \]

At the second stage, \( \beta \) has a Gaussian prior while the \( \phi_i \) are i.i.d. \( \sim N(0, \sigma^2_\phi) \). The \( \epsilon_{ij} \) are i.i.d. \( \sim N(0, \sigma^2_{\epsilon}) \). The variance components become the third stage hyperparameters, i.e., we require prior specifications for \( \sigma^2_{\phi}, \sigma^2_{\epsilon} \). As has been learned over recent years, care is required in these specifications. Recalling our remarks in the previous section on improper priors, they can lead to improper posteriors [9]. The frequently-employed inverse gamma priors, \( IG(\epsilon, \epsilon) \) for small \( \epsilon \) are nearly improper and result in nearly improper posteriors as well as badly behaved MCMC in practice. A protective recommendation is an \( IG(1, b) \) or \( IG(2, b) \). Both are far from improper; the former has no integer moments, the latter has a mean but no variance. Evidently, we can revise the model to have a nonGaussian first stage. Again, care is needed with prior specifications as well as in model fitting.

In collecting information on, e.g., individuals, we often have vectors of data with one or more components of the components missing. It is unattractive to confine ourselves to analyzing only the complete data cases. This may discard too much data and possibly introduce bias with regard to the ones retained. To use the individuals with missing data, we must \textit{complete} them, so-called imputation. There is by now a very substantial literature on imputation [see, e.g., 42]. However, to do a fully model-based imputation in the Bayesian setting results in latent variables (Section 5) and looping with Gibbs samplers, updating parameters given missing data, then updating missing data given parameters. In this sense, the Gibbs sampler extends the EM algorithm to provide full posterior inference rather than an MLE with an asymptotic variance.

As a simple example, consider multivariate normal data, \( Y_i \sim N(\mu_i, \Sigma) \) where the components of \( \mu_i \) may have regression forms in suitable covariates. Some components of some of the \( Y_i \)'s are missing. In order to perform the imputation, we do basic Gibbs sampling: we update \( (\mu_i, \Sigma) \) given values for the missing \( Y \)'s, then we update the missing \( Y \)'s given values for the \( (\mu_i, \Sigma) \). Another standard example considers missing categorical counts within a multinomial model where the multinomial cell probabilities might be modeled using some sort of multivariate logit model [1]. For instance, some categories are aggregated/collapsed so counts for the disaggregated categories are missing. Again, we can envision a looping of the Gibbs sampler: update the parameters given values for all the counts, update the missing counts given values for the parameters.
2.5 Latent variables

Latent variables are at the heart of most hierarchical modeling. Here, we provide examples which suggest they can be envisioned beyond random effects or missing data. Latent variable models customarily result in a hierarchical specification of the form $f(Y|Z)f(Z|\theta)\pi(\theta)$. Here, the $Y$’s are observed, the $Z$’s are latent and the “regression” modeling is shifted to the second stage.

An elementary version of a latent variable model arises with binary data models. In particular, the usual binary response model adopts a logit or probit link function. Illustrating with the probit, suppose $Y_i \sim \text{Bernoulli}(p(X_i))$ (more generally, we can have $Y_i \sim \text{Bi}(n_i,p(X_i))$). Specifically, let $\Phi^{-1}(p(X_i)) = X_i\beta$ with a prior on $\beta$. In fitting this model using MCMC computation, it is awkward to sample $\beta$ using the likelihood in this form. If we introduce $Z_i \sim N(X_i\beta, 1)$ then, immediately, $P(Y_i = 1) = \Phi(X_i\beta) = 1 - \Phi(-X_i\beta) = P(Z_i \geq 0)$. Once we bring in these $Z_i$’s, we achieve a routine Gibbs sampler: update the $Z$’s given $\beta, y$ (this requires sampling from a truncated normal), update $\beta$ given the $Z$’s and $y$ (this is the usual, typically conjugate normal updating). This approach was first articulated in the literature by [3].

It is clear that this approach can readily extend to general ordinal categorical data settings [1]. In particular, for each $i$, the Bernoulli trial is replaced with a multinomial trial. There is still a latent $Z_i$, still following say a Gaussian linear regression. Now the multinomial outcomes are created by introducing cut points along the real line; the intervals determined by the cut points allocate probabilities to each of the multinomial outcomes. The cut points will be random as well, noting that, in order to identify the intercept in the regression, the smallest cut point can be taken to be 0, without loss of generality.

Another routine generalization of this approach accommodates censored or truncated data models. As a simple illustration, suppose we observe a variable with a point mass at 0 and the remainder of its mass spread over $R^+$ or perhaps $(0, 1]$. In the first case, such data arise when studying, for instance, daily precipitation at a location; in the second case when considering, for instance, the proportion of a particular land use classification over a region. Here, with observed $Y_i$’s, we can introduce $Z_i$’s such that $Y_i = g(Z_i), Z_i > 0, Y_i = 0, Z_i \leq 0$, with $g(\cdot)$ a link function from $R^+$ to $R^+$ or perhaps to $(0, 1]$. Then, we could model the $Z_i$’s using a usual Gaussian linear regression.

A further setting for latent variables is change point problems [7]. Frequently, we observe a process over the course of time during which we would like to assess whether some sort of change in regime has occurred. Practically speaking, this requires the notion of a “least” significant change. That is, we may be able to identify even a very small change with enough data but we will find challenging the case where the support for the change has “no change” as a boundary point.

In the change point setting two sampling scenarios can be envisioned. In the first, we have a full set of data. Then we look, retrospectively, to try to find if change(s) occurred. In the second, we look at the data sequentially and we try to identify change(s) as the data collection proceeds. We illustrate with a simple version of the first scenario. Let $f_1(y|\theta_1)$ be the density for i.i.d. observations before the change point, $f_2(y|\theta_2)$ the density after the change point. With data $Y_i, i = 1, 2, ..., n$, let $K$ be the change point indicator, i.e., $K \in \{1, 2, ..., n\}$ where $K = k$ means change at observation $k + 1$; $k = n$ means “no change.” Then, the model is:

\[ f(Y|Z)f(Z|\theta)\pi(\theta) \]

While, in a sense, all variables we can not observe are “missing,” in the previous subsection we took missing to mean some components of the data while here they will be variables different from the data.
L(θ₁, θ₂, k; y) = Π_{i=1}^{k} f_1(y_i|θ_1)Π_{i=k+1}^{n} f_2(y_i|θ_2).
(2.7)

Again, we have a hierarchical model, [y|k, θ₁, θ₂][K = k][θ₁, θ₂]. (Note that we do not include any parameters in the prior for K; with only one change point, we could not hope to learn about such parameters.) With a prior on θ₁, θ₂, K, we have a full model specification. Again, a simple Gibbs sampler emerges for model fitting: update θ’s given k, y (so, we know exactly which observations are assigned to which density); update k given θ’s and y (this is just a discrete distribution, easily sampled). Obvious generalizations would allow the y’s to be dependent, to have order restrictions on θ’s, to imagine multiple change points. Also, in this version, time is discretized to the set of times when the measurements were collected. Extension to continuous time and multiple change point settings is available using point process models [50, 62].

Errors in variables models [24] offer another latent variables setting. Loosely stated, the objective is to learn about the relationship between say Y and X. Unfortunately, X is not observed. Rather, we observe say W instead of X. In some cases, W will be a version of X, subject to measurement error, i.e., W may be X_{obs} while X may be X_{true}. In other cases W may be a variable (variables) that play the role of a surrogate for X. In any event, if we envision a joint distribution for W and X, we may condition in either direction. If we specify a model for W|X we refer to this as a measurement error model [15, 24], imagining W to vary around the true or desired X; if we specify a model for X|W we refer to this as a Berkson model [15]. In fact, we can imagine a further errors in variables component - perhaps we observe Z, a surrogate for Y. Altogether we have a hierarchical model with latent X’s, possibly Y’s. In particular, for the measurement error case, with independent observations, we have:

Π_i f(Z_i|Y_i, γ)f(Y_i|X_i, β)f(W_i|X_i, δ)f(X_i|α) \tag{2.8}

while for the Berkson case we have:

Π_i f(Z_i|Y_i, γ)f(Y_i|X_i, β)f(X_i|W_i, δ) \tag{2.9}

Typically, we will also have some validation data to inform about the components of the specification. We might have some X, Y pairs or perhaps some X, W pairs. It is noteworthy that the measurement error version requires a model (prior) for X while the Berkson model does not. In many applications the former will be more natural. However, in some contexts, model fitting is only feasible with the latter [6]. In any event, what is most remarkable is that, within this hierarchical framework, using a full Bayesian specification, we can learn about the relationship between Y and X without ever observing X (and, possibly, without observing Y as well). This reveals the power of hierarchical modeling but, evidently, what we can learn depends upon the form of what we specify and the data we have.

2.6 Mixture models

Mixture models have now become a staple of modern stochastic modeling [50, 62]. This has arisen on at least two accounts: (i) their flexibility to model unknown distributional shapes and (ii) their intuition in representing a population in terms of groups/clusters that may exist but are unidentified. Mixture models come in several flavors - parametric or nonparametric, incorporating finite, countable or uncountable mixing. In this regard, they
are sometimes referred to as classification problems or discriminant analysis, reflecting a
goal of assigning an individual to a population or assessing whether individuals belong to
the same population.

The most rudimentary finite mixture version takes the form:

\[ Y \sim \sum_{l=1}^{L} p_l f_l(Y|\theta_l). \]  

(2.10)

Often the \( f_l \) are normal densities, whence we obtain a normal mixture. If we assume \( L \) is
specified and we observe \( Y_i, i = 1, 2, \ldots, n \), then what is latent is a label for each \( Y_i \), i.e., an
indicator of which component of the mixture \( Y_i \) was drawn from. These latent labels would
be such that if \( L_i = l \), then \( Y_i \sim f_l(Y|\theta_l) \). Upon introducing these labeling variables, the
resulting hierarchical model becomes:

\[ \Pi_i f(Y_i|L_i, \theta) \Pi_i f(L_i|\{p_l\}) \pi(\theta) \pi(\{p_l\}). \]  

(2.11)

Here, \( \theta \) denotes the collection of \( \theta_l \). Once again, Gibbs sampling is routine to implement. We
create a loop that updates \( \theta, \{p_l\} \) given the \( L \)'s and the data. With observations assigned to
components, this becomes the equivalent of an analysis of variance problem to learn about
the \( \theta \). To update the \( L_i \)'s given \( \theta, \{p_l\} \) and the data requires sampling from an \( L \)-valued
discrete distribution where, for a given \( i \), the mass on the \( l \)'s is determined by the relative
likelihood for the observed \( Y_i \) as well as the prior on the \( p_l \)'s (often a uniform). Hence,
we obtain individual assignment probabilities as well as global assignment weights. Richer
versions introduce covariates into the \( \theta_l \)'s and, possibly, into the \( p_l \)'s [46]. Further challenge
is added if \( L \) is unknown with a prior specification. Now, since the dimension of the model
changes with \( L \) we may attempt reversible jump MCMC [70] to learn about \( L \). A simpler
alternative might be to carry out model choice across a set of \( L \)'s.

It is evident that such mixture models are not identifiable, i.e., the subscripts can be
permuted and the same mixture distribution results. This has led to discussion in the
literature with regard to introducing identifiability constraints or the possibility of fitting
the MCMC, allowing multi-modality in the posterior in the absence of identifiability [46].
The former path seems to be the most widely used, with order constraints on the means,
imposed in some fashion, being the most common choice for achieving identifiability.

Next, we recall that many familiar distributional models can be developed through
continuous mixing, i.e., closed forms are achieved by virtue of conjugacy between the mixed
model and the mixing distribution. Well-known examples include scale mixing of normals to
obtain t-distributions, as well as Poisson-Gamma (equivalently negative binomial) and beta-
binomial models. In some modeling situations we may seek individual level mixing variables,
whence we would introduce such variables as latent quantities. Again, a hierarchical model
arises. An illustration is in the case of outlier detection through the use of suitable individual
level gamma mixing of normals. Here outliers are “detected” through the magnitudes of their
associated mixing variables, i.e., the heavier the posterior tails, the more we are inclined to
classify the observation as an outlier. See, e.g., [27, 63].

2.7 Random effects

Returning to the random effects setting, let us first consider individual level longitudinal
data with interest in explanation through growth curves. A natural specification would
model individual level curves centered around a population level curve. We would need the
22 Handbook of Environmental and Ecological Statistics

population level curve to see average behavior of the process; we need individual level curves in order, for example, to prescribe individual level treatment. With parameters at each level and a third stage of hyperparameters, we again see a hierarchical form.

More precisely, if \( Y_{ij} \) is \( j \)th measurement for \( i \)th individual, let

\[
Y_{ij} = g(X_{ij}, Z_i, \beta_i) + \epsilon_{ij}
\]

where \( \epsilon_{ij} \sim N(0, \sigma^2) \). The form for \( g \) depends upon the application. It is often linear but need not be. We set \( \beta_i = \beta + \eta_i \) where the \( \eta_i \) have mean 0 (or perhaps replace \( \beta \) with a regression in the \( Z_i \)). Then the \( \beta_i \) (or the \( \eta_i \)) are the random effects. They provide the individual curves with \( \beta \) providing the global curve. Learning with regard to any individual curve will borrow strength from the information about the other curves.

Customarily, random effects are modeled using normality. However, they need not be i.i.d. That is, if say the scalar \( \omega_i \) is associated with individual \( i \), we need not insist that the vector, \( \omega \), of \( \omega_i \)’s, be distributed as say, \( \omega \sim N(0, \sigma^2 I) \). We can replace \( \sigma^2 I \) with \( \Sigma(\theta) \) where \( \Sigma(\theta) \) has structured dependence. That is, with say \( n \) individuals, we could not learn about an arbitrary positive definite \( n \times n \) matrix \( \Sigma \) but we could learn about \( \Sigma \) defined as a function of only a few parameters. Structured dependence is at the heart of time series, spatial and spatio-temporal modeling (see Chapters 3 and 5) and is frequently specified through a Gaussian process (GP). For example, in the spatial setting, we envision data in the form \( Y(s_i), i = 1, 2, ..., n \), i.e., \( n \) observations at \( n \) different spatial locations \( s_i \) so \( s_i \in R^2 \). We could imagine three dimensional locations and, more generally, replacing geographic space with say covariate space. This takes us into the growing world of computer models [44, 45, 49]. In any event, with a GP, we need only specify finite dimensional (e.g., \( n \)) joint distributions with the joint dependence determined by a valid covariance function. Customarily, the covariance function assigns stronger association to variables that are closer to each other in geographic space. A common example is the exponential, \( \text{cov}(Y(s), Y(s')) = \sigma^2 \exp(-\phi ||s - s'||) \). With \( n \) locations, \( \Sigma(\theta)_{ij} = \sigma^2 \exp(-\phi ||s_i - s_j||) \).

Data fusion presents another important area for hierarchical modeling (see Chapter 7). The context is frequently spatial or spatio-temporal. For instance, data on environmental exposure can be available from different sources, e.g., monitoring stations, computer models, and satellites. The fusion objective is to use all data sources to infer about the true exposure. We confront misalignment of scales in space, in time. To reconcile the misalignment we may elect to scale up or scale down.

2.8 Dynamic models

Dynamic models have now become a standard formulation for a wide variety of processes, including financial and environmental applications. Alternate names for them in the literature include Kalman filters, state space models and hidden Markov models [33, 51, 59]. They introduce a first stage (or observational model) and then a second stage (or transition model), with third stage hyperparameters. Again, the first stage provides the data model while the second stage provides a latent dynamic process model. See, e.g., [64] for a full development. In particular, there is substantial modeling opportunity at the second stage, allowing evolution of process variables or process parameters, in either case, driven by covariate information. Chapter 4 presents a full development. Here, we just note the hierarchical structure.
Specifically, the basic dynamic model takes the form:

\[ Y_t = g(X_t, \theta) + \epsilon_t, \quad \text{the observation equation} \tag{2.12} \]

with

\[ X_t = h(X_{t-1}; \theta_2) + \eta_t, \quad \text{the transition equation.} \tag{2.13} \]

Evidently, time \( t \) is discrete and we are putting the dynamics in the mean.

We illustrate with a dynamic space-time model. Consider:

Stage 1: Measurement equation

\[ Y(s, t) = \mu(s, t) + \epsilon(s, t); \quad \epsilon(s, t) \overset{\text{ind}}{\sim} N(0, \sigma^2_\epsilon). \]

\[ \mu(s, t) = \mathbf{x}^T(s, t) \tilde{\beta}(s, t). \]

\[ \tilde{\beta}(s, t) = \beta_t + \beta(s, t) \]

with

Stage 2: Transition equation

\[ \beta_t = \beta_{t-1} + \eta_t, \quad \eta_t \overset{\text{ind}}{\sim} N_p(0, \Sigma). \]

\[ \beta(s, t) = \beta(s, t-1) + \mathbf{w}(s, t) \]

where the \( \mathbf{w}(s, t) \) are independent (over \( t \)) innovations of a spatial process [see, e.g., 28].

As noted above, this specification can be connected to a linear Kalman filter [39]. Thus, Bayesian model fitting using the forward filter, backward sample (ffbs) algorithm [16, 23] becomes the customary approach. Again, Chapter 4 explores such models in considerable detail.

Furthermore, Wikle and colleagues [38, 65] have adapted these dynamic forms to the fitting of models motivated by stochastic partial differential equations (SPDE’s). The approach is to discretize time. In particular, there are many interesting ecological diffusions (characterized by SPDE’s) which can be applied to study the behavior over time (and, perhaps space) of: (i) emerging diseases such as avian flu or H1N1 flu; (ii) exotic organisms, e.g., invasive plants and animals; (iii) the evolution of the distribution of size or age of a species; (iv) the dynamics explaining phenomena such as transformation of landscape, deforestation, land use classifications, and urban growth. Our objective for such processes is to forecast likely spread in space and time with associated uncertainty. We anticipate that the evolution will be nonlinear and nonhomogeneous in space and time, driven by explanatory covariates.

2.9 Model adequacy

In a world filled with rich modeling opportunity, it becomes necessary to investigate model adequacy and model comparison. Two brief comments are appropriate here. First, model adequacy is usually viewed as an \textit{absolute} issue. If we develop a criterion, it is usually difficult to calibrate performance under this criterion. How adequate is adequate? The
difference between model adequacy and model comparison is therefore clear. The latter is relative: with a criterion, we can make comparison using this criterion. Second, when we propose complex multi-level models, they are often too big for the data. They often overfit but would rarely be inadequate.

As a result, we tend to treat model adequacy informally, e.g., in the realm of exploration to discard inadequate models, to retain adequate ones for comparison. In this spirit, there is a rich array of diagnostic tools for studying model failures of different sorts. These include investigation of residuals (looking for departure from distributional assumptions), comparison of predicted vs. observed (looking for under or overfitting), empirical Q-Q plots, etc. In fact, more formal goodness of fit tests are well known and have a long history in the literature [35]. However, they are rarely relevant for hierarchical model adequacy.

Within the Bayesian framework, the marginal density ordinate for the observed data is a historical criterion [12]. That is, we plug the data into the marginal density, \( f(Y) \). This criterion is difficult to calculate since it requires marginalizing over \( \theta \) and is also difficult to calibrate for high dimensional data. It is not used with hierarchical models. Variants, such as conditional predictive ordinates [54] have been proposed. These are “leave-one-out” criteria, data deletion approaches which also can be very computationally demanding with large amounts of data.

More common in the current literature are posterior predictive checks, as advocated by [32], and prior predictive checks, as in [19]. Both of these ideas approaches employ discrepancies to capture differences between what is observed and what is predicted. The former uses posterior predictive distributions based upon the observed data and then checks the observations against these predictive distributions. Use of the data twice in this fashion makes this approach insufficiently critical of models. The latter makes comparison using prior predictive distributions, obtained under the model but without using the data. Hence, they are more critical of the actual model specification. They can be most easily implemented using Monte Carlo tests [21], i.e., generating realizations of a feature of the data under the model and comparing with the observed value of the feature.

It can be argued that prior predictive distributions are based upon priors which are generally weak, often not proper, and are only introduced to complete model specifications. Posterior predictive distributions are obtained using posteriors and posteriors are the distributions that will be used for inference. The counterargument is that, if you want to work with generative models (proper priors) and want to criticize such a model, you should criticize what is specified, not what is realized upon applying it to the data.

Finally, we suggest the use of empirical vs. nominal coverage as a convenient model checking strategy. The idea is based upon specifying a hold-out or validation dataset, using the remainder for model fitting. The general approach is discussed further in the next section. Here, we would simply suggest to use the fitted model to obtain posterior predictive credible intervals for for a held out observation and then compare this interval with the actual observation. Doing this for the set of held out points enables comparison between the nominal coverage associated with the interval vs. the observed empirical coverage across the set of hold out observations. Empirical undercoverage suggests that the model is not capturing uncertainty well enough. Empirical overcoverage implies uncertainty is greater than it should be, perhaps the model is overfitting. Again, implementation of this model checking strategy should be done out-of-sample. In-sample checking again suffers from using the data twice.
2.10 Model comparison

As we have argued above, model comparison is often done in the parameter space, leading to penalized likelihood criteria such as AIC, BIC and, with hierarchical models, DIC. AIC and BIC have been widely used in the biology/ecology community [13]. DIC [61] has been advocated within the Bayesian community as a criterion which, attractively, can be implemented directly in the BUGS software for Bayesian model fitting [43].

We suggest that it may be preferable to make comparison in the predictive space using (posterior) predictive distributions. In predictive space, having observations, we can examine quantities like predictive mean square error, predictive coverage, and continuous ranked probability scores. In Section 10.1 we review model comparison in the parameter space while in Section 10.2 we review model selection in predictive space.

2.10.1 Bayesian model comparison

Historically, Bayesian model comparison has been addressed through Bayes factors. We start with a formal application of Bayes’ theorem to this problem. Suppose we have a set of $k$ models under consideration and, a priori, we assume $p_j$ is the prior probability that model $j$ is true, $j = 1, 2, ..., k$. Then, using the theorem, with data $y$ and denoting the $j$th model by $M_j$,

$$
P(M_j|y) = \frac{p_j P(y|M_j)}{\sum_{j=1}^{k} p_j P(y|M_j)}. \tag{2.14}
$$

We would then choose the model with the highest posterior probability. On the surface, this idea is attractive. However, calculating $P(y|M_j) = \int P(y|\theta_j; M_j) \pi(\theta_j|M_j) d\theta_j$ for each $j$ when $\theta_j$ is large can be challenging. Furthermore, where do the $p_j$’s come from? We have repeatedly asserted that none of these models is “true.” Would we assume a prior that they are equally likely? This seems inappropriate since, in application, we would want to reward parsimony or sparseness.

Briefly, we return to the variable selection problem discussed in Section 2, i.e., variable selection in a regression model where parsimony in model dimension, in number of variables, is to be rewarded. Hence, we might place a $Po(\lambda)$ distribution on the number of variables, adding a prior on $\lambda$. In fact, a more useful idea is to consider $p$, a prior inclusion probability for a variable, with exchangeable Bernoulli selection. That is, a priori, $P(\text{variable } X_k \text{ is selected}) = p$, i.e., $P(I_k = 1) = p$, adding a prior on $p$. Then, under a binomial model for the number of variables selected, we seek the posterior probability of $I_k = 1$. Suppose we reparameterize the regression coefficient for $X_k$ to $\theta_k = I_k \beta_k$. Then, we introduce a “spike and slab” prior for $\theta_k$ using the auxiliary variable $I_k$. That is, when $I_k = 1$, $X_k$ is in the model with coefficient $\beta_k$; when $I_k = 0$, $X_k$ is not in the model. There is a growing literature on variable selection starting from these ideas. See, e.g., [53] and further references therein.

Returning to the model selection criterion above, can we avoid specifying the $p_j$’s? Suppose we look at models in pairs (so, really only suitable for a small number of models). For a pair of models $M_0$ and $M_1$, define the Bayes Factor for $M_0$ vs $M_1$ as $B_{01} = \frac{P(Y|M_0)}{P(Y|M_1)}$. This ratio is customarily interpreted as providing a weight of evidence for model $M_0$ relative to $M_1$. The ratio also offers another interpretation. With two models, we have

$$
P(M_0|Y) P(M_1|Y) = P(Y|M_0) P(M_1) \times p_0 \frac{p_0}{p_1}
$$

where $p_1 = 1 - p_0$. That is, the Bayes factor provides the conversion from the prior odds to
the posterior odds without having to provide the prior odds. In practice, the Bayes factor reduces model comparison to, essentially, a hypothesis testing problem. We have two actions with what can be expressed as, essentially, a 0 − 1 loss function.

The Bayes factor requires proper priors for both models since it is arises as a ratio of density ordinates which can only be interpreted when priors are proper. In fact, this reveals a problem with nearly improper priors. Suppose $Y \sim N(\theta, 1)$. $M_0 \equiv H_0 : \theta = 0$, $M_1 \equiv H_A : \theta \neq 0$, i.e., $M_0 \subset M_1$. Then, with a Uniform($-L, L$) prior on $\theta$, $B_{01} = 2\phi(Y)/(\Phi(L - Y) − \Phi(-L - Y))$. The support for $M_0$ tends to $\infty$ as $L$ increases.

Finally, as we commented with the general model selection criterion, calculation of the Bayes factor can be challenging. Marginalization over $\theta$ is required for both the numerator and denominator. In fact, calculation of Bayes factors attracted a fair bit of attention in the literature at the turn of the 21st century, exploring Monte Carlo calculation ideas, bridge and path sampling approaches, etc. [58]. However, Bayes factors are seldom used with hierarchical models.

We mention one last feature of Bayes factors which enables connection to penalized likelihood and therefore to AIC and BIC type model selection criteria. This is the Lindley paradox. Suppose $Y_1, Y_2, ..., Y_n$ i.i.d.$N(\theta, 1)$ with prior $\pi(\theta) = N(0, 1)$. Consider $H_0 : \theta = 0$ vs $H_A : \theta \neq 0$. Then, $B_{01} = N(Y|0, \frac{1}{n})/N(Y|0, n)$. The usual test statistic is $Z = \sqrt{nY}$ in which case

$$B_{01} = \sqrt{n + 1}{\bar{Z}}|e^{-\frac{n}{2}}Z^2/2.$$  

So, regardless of $|Z|$, as $n \to \infty$, $B_{01} \to \infty$, i.e., we choose $M_0$. This implies that the Bayes factor is "too large."

This leads to the issue of penalized likelihood. What do we mean by this? Consider the usual likelihood ratio test which, for $H_0 : \theta \in \Theta_0$ vs. $H_A : \theta \in \Theta - \Theta_0$, takes the form: Reject if

$$\chi^2_{p_2 - p_1} > c.$$  

It can be written more generally for models $M_1$ and $M_2$ such that $M_1 \subset M_2$.

Under weak conditions, when $M_1$ is true, $-2\log \lambda \approx \chi^2_{p_2 - p_1}$ where $p_2$ is the dimension of model $M_2$ and $p_1$ is the dimension of model $M_1$. So, $P(\lambda < c|M_1) = P(-2\log \lambda > c') \approx P(\chi^2_{p_2 - p_1} > c') > 0$. We see that $P(\text{reject } M_1|M_1 \text{ true}) \rightarrow 1$ as $n \to \infty$; $\lambda$, equivalently log$\lambda$ is too small. This leads to proposing a penalty function on $-2\log L(\theta; Y)$, $j = 1, 2$. Here, we have nearly fifty years of literature dating to [2] and [60]. Forms for the penalty include functions of only $p_j$, e.g., $cp_j$. This form of penalty leads to the Akaiake Information Criterion (AIC). Numerous choices of $c$ have been proposed in the literature; $c = 2$ is commonly used. Another choice considers functions of $n$ and $p_j$, e.g., log$np_j$. These are called Bayesian Information Criteria (BIC) in the literature. See [13] for full details. This choice results in the criterion $-2\log \lambda_n + (p_2 - p_1)\log n$. Above, we saw that the Bayes factor was too large and log$\lambda$ is too small. Is there a relationship between them? [26] show that

$$10\log BF = 10\log \lambda_n + 10\log \frac{p_2 - p_1}{2} + O(1).$$  

(2.15)

We note that these criteria really only apply to the case of "fixed" effects modeling where model dimension is clear. With random effects, latent variables, etc., it is no longer clear what model dimension is. So, we can assert that these criteria play a small role in today’s modeling landscape. In this regard, the Deviance Information Criterion (DIC) [61] has been proposed as a version suitable for hierarchical models. Computation is easy, following on from MCMC model fitting. However, examples of negative model dimension, of sensitivity to choice of parametrization, of selecting models marginalized to focus on different parameters,
have raised issues regarding its effectiveness as a criterion. Perhaps the real issue is that the criterion is being applied over parameter space while model comparison with hierarchical models may be better done in predictive space.

### 2.10.2 Model comparison in predictive space

We have suggested above that model comparison may be more appropriately implemented in the space of the data rather than in the space of the parameters. This leads to comparison of the posterior predictive distribution for an observation with the actual observed value of the realization. Expression (4) shows how the predictive distribution for an observation \( Y_0 \) arises. MCMC model fitting provides posterior samples of \( \theta \)'s. If we take a posterior draw \( \theta^* \) and then draw a \( Y^*_0 \) from \( f(Y_0|Y, \theta^*) \), so-called composition sampling [5], then \( Y^*_0 \sim f(Y_0|Y) \). So, we can draw posterior predictive samples one-for-one with posterior parameter samples.

A critical question emerges. Shall we do model comparison in-sample or out-of-sample? In-sample comparison is attractive in the sense that we fit each model under consideration to all of the data. Then, we can choose which observations we want to use in the model comparison. The disadvantage, as with model adequacy, is, in using the data twice, in-sample validation tends to show better model performance than would be seen in practice. Out-of-sample validation [55], also called cross-validation, with randomly selected hold out data, seems preferred. We see how well the models do with data that was not used to fit them. The challenge here is how much data to hold out. We create a training/fitting set and a test/validation set. What proportion of the data should we allocate to each. Customary practice employs 10 – 20% hold out according to the overall size of the dataset. Note that hold out is only used for model comparison. Once a model is selected, the inference results will be presented based upon the full dataset.

[29] proposed a model comparison criterion in the spirit of DIC but in predictive space. It is viewed as an in-sample criterion. Adopting a balanced loss function, they obtained a two component criterion. One component rewards fidelity of prediction relative to the actual observation. The other rewards small posterior predictive uncertainty.

More commonly used these days are criteria like predictive mean square error or predictive mean absolute error, applied out-of-sample. Hold out observations are compared with the associated posterior predictive mean or median, respectively, averaged over the hold out set. A possibly more attractive criterion is the ranked or continuous ranked probability score [34], according to whether the predictive distribution has discrete support (like a species count) or continuous support (like a temperature or ozone level). The objective is to compare the entire predictive distribution to the held out value. Intuitively, the more concentrated the predictive distribution is around the hold out value, the better the model performance.

The proposed measure is the squared integrated distance between the predictive distribution and the degenerate distribution at the observed value,

\[
CRPS(F,y) = \int_{-\infty}^{\infty} (F(u) - 1(u \geq y))^2 du
\]

where \( F \) is the predictive distribution and \( y \) is the observed value. For us, \( Y_0 \) is the observation and \( F \) is the posterior predictive distribution for \( Y_0 \). With a collection of such hold out observations and associated predictive distributions, we would sum the CRPS over these observations to create the model comparison criterion. Recall that, under MCMC model fitting, we will not have \( F \) explicitly but, rather, a sample from \( F \). Fortunately, a convenient
alternative computational form, provided \( F \) has a first moment, is

\[
\text{CRPS}(F, y) = \frac{1}{2} E_F |Y - Y'| + E_F |Y - y| \tag{2.17}
\]

where \( Y \) and \( Y' \) are independent replicates from \( F \). With samples from \( F \), we have immediate Monte Carlo integrations to compute (2.17).

As some last thoughts here, one may question the comfort of taking a very complex multi-level model and reducing it to a single number for model comparison. Perhaps, we might be interested in the local performance of a model. Perhaps, we could choose different models in different parts of the predictive space. Perhaps we could average over models (Bayesian model averaging, [83]). The overarching question asks what is the utility for the model; different models would be preferred under different utilities.

\[\textbf{2.11 Summary}\]

We have argued that hierarchical models provide the stochastic framework within which to develop integrative process models. We have shown, with a variety of examples, that these models typically share a common structure. There is a first stage data model, there is a second stage process model that is latent, i.e., it is endowed with a full model specification but it is unobserved, and a third stage which incorporates prior specifications for all of the remaining parameters in the model. We have noted that, in order to get the uncertainty right, these models should be fitted within the Bayesian framework. We have also noted that fitting of these models typically introduces familiar looping in Gibbs sampling. Hence, it is straightforward to envision how the MCMC model fitting should be implemented. However, according to the size of the dataset and the complexity of the specifications, such model fitting can be very challenging, perhaps infeasible. Indeed, this limitation will become more of a constraint as we continue to seek models which stretch the limits of our computing capabilities. Hence, we imagine a computing future built around simulation based model fitting of these hierarchical forms but incorporating suitable approximation. Thus, the “art” will encompass both specification (with comparison) and approximate fitting (to enable inference and comparison).

\[\textbf{Bibliography}\]


Modeling for environmental and ecological processes


