CHAPTER 2

Statistical Preliminaries

In this chapter, we expand on some of the ideas presented in Chapter 1, as well as present some statistical results needed for the rest of the book. We give an overview of how Statistics and Science have related to each other in the past, and we give a viewpoint for how Statistical Science will evolve in the twenty-first century. We are deliberately broad and not overly technical in this chapter, for readers who have not yet had a lot of exposure to Statistics. Here we address general issues to help explain the statistical modeling and inference decisions made in the spatial, temporal, and spatio-temporal contexts of the subsequent chapters.

Several explanations of terminology are needed before we start. Uncertainty in data, processes, or parameters means there will be uncertainty in conclusions. Statisticians call this drawing of conclusions in the presence of uncertainty, *statistical inference* (or just *inference*); in this book, inference will be either *estimation* of fixed but unknown parameters, or *prediction* of unknown random quantities. (Notice that "forecasting," namely concluding something about the future, is a special case of "prediction.")

The terms *normal distribution* and *Gaussian distribution* are synonymous. In this book, we prefer the latter and use the expression $Z \sim Gau(\mu, \sigma^2)$ to denote a random variable Z whose probability distribution is Gaussian (i.e., normal) with mean μ and variance σ^2; it is equivalent to the expression $Z \sim N(\mu, \sigma^2)$, which one might see in other books or articles. The random vector $\mathbf{Z} \equiv (Z_1, \ldots, Z_m)'$ is an m-dimensional column vector, where the symbol "$'$" means *transpose*. Then $\mathbf{Z} \sim Gau(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes an m-dimensional Gaussian distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. The covariance matrix $\boldsymbol{\Sigma}$ (sometimes called a variance matrix or a variance–covariance matrix) is a symmetric, positive-definite (occasionally nonnegative-definite) $m \times m$ matrix whose (i, j)th entry is $\text{cov}(Z_i, Z_j)$; $i, j = 1, \ldots, m$.

Statistics for Spatio-Temporal Data, by Noel Cressie and Christopher K. Wikle

Let $[A]$ denote the probability distribution of the random quantity A. For example, the expression, $Z \sim Gau(\mu, \sigma^2)$, is equivalently written as

$$[Z] = \{(2\pi\sigma^2)^{-1/2} \exp[-(1/2)(z-\mu)^2/\sigma^2] : z \in \mathbb{R}\},$$

and $\mathbf{Z} \sim Gau(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\mathbf{Z} = (Z_1, \ldots, Z_m)'$, is written as

$$[\mathbf{Z}] = \{(2\pi)^{-m/2}|\Sigma|^{-1/2} \exp[-(1/2)(\mathbf{Z}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{Z}-\boldsymbol{\mu})] : \mathbf{z} \in \mathbb{R}^m\}.$$

If $g(A)$ is a well defined random quantity for some function $g(\cdot)$, then its *expectation*, $E(g(A))$, is equivalently written as $E(g(A)) = \int g(A)[A]\,dA$ in the continuous case, and is written as $E(g(A)) = \Sigma g(A)[A]$ in the discrete case.

Furthermore, let $[A|B]$ denote the conditional distribution of the random quantity A, conditional on specifying a particular value of (the random quantity) B. This is also referred to as the conditional distribution of A given B. For example, the expression, $\mathbf{Z}|\mathbf{Y} = \mathbf{y} \sim Gau(\mathbf{y}, \sigma^2\mathbf{I})$, is equivalently written as

$$[\mathbf{Z}|\mathbf{Y}] = \{(2\pi\sigma^2)^{-m/2} \exp[-(1/2)(\mathbf{z}-\mathbf{y})'(\mathbf{z}-\mathbf{y})/\sigma^2] : \mathbf{z} \in \mathbb{R}^m\},$$

where it is understood that on the left-hand side, $\mathbf{Y} = \mathbf{y}$.

Let the spatial domain of interest be $D_s \subset \mathbb{R}^d$, a subset of d-dimensional Euclidean space, and let the temporal domain of interest be $D_t \subset \mathbb{R}^1$. The spatial index $\mathbf{s}$ (a d-dimensional vector) and the temporal index t (a real number) can vary continuously or discretely over their respective domains, D_s and D_t. This book is concerned, amongst other things, with models for spatio-temporal random processes. When t varies continuously, we write the generic process as $\{Y(\mathbf{s}; t) : \mathbf{s} \in D_s, t \in D_t\}$. To follow the usual notational convention for time series, when t varies *discretely*, we write instead $\{Y_t(\mathbf{s}) : \mathbf{s} \in D_s, t \in D_t\}$.

Spatial Description and Temporal Dynamics: A Simple Example

The best way to compare space and time in our statistical context is to consider a simple example, where we let $d = 1$, $D_s = \{s_0, s_0 + \Delta, \ldots, s_0 + 24\Delta\}$, and $D_t = \{0, 1, 2, \ldots\}$. Think about D_s as being an east–west transect of regular spacing Δ in a field of wild prairie grass, where the observations on the process $\{Y_{t_0}(s_0), \ldots, Y_{t_0}(s_0 + 24\Delta)\}$ are nondestructive-biomass measurements taken at 25 equally spaced spatial locations $\{s_0, \ldots, s_{24}\} \equiv \{s_0, \ldots, s_0 + 24\Delta\}$, at a fixed point in time ($t = t_0$): 3 pm on a given day in the middle of a given spring. Compare the spatial process $\{Y_{t_0}(s_0), \ldots, Y_{t_0}(s_0 + 24\Delta)\}$ to the temporal process $\{Y_t(s_0) : t = t_0, \ldots, t_0 + 24\}$ of nondestructive-biomass measurements taken at the fixed spatial location $s = s_0$, at 3 pm for each of 25 consecutive days in the middle of the same spring.

Define the *spatial process* at the fixed time point t_0 to be

$$\mathbf{Y}_{t_0} \equiv (Y_{t_0}(s_0), \ldots, Y_{t_0}(s_0 + 24\Delta))',$$

and define the *temporal process* at fixed spatial location s_0 to be

$$\mathbf{Y}(s_0) \equiv (Y_{t_0}(s_0), \ldots, Y_{t_0+24}(s_0))'.$$

By comparing spatial statistical models for $\mathbf{Y}_{t_0}$ and time series models for $\mathbf{Y}(s_0)$. we can see to what extent space is modeled differently from time.

A simple departure from independence for a *spatial process* is nearest-neighbor dependence expressed through conditional distributions. Assume, for $i = 1, \ldots, 23$, the Gaussian (conditional) distribution

$$\begin{aligned} &Y_{t_0}(s_i) \mid \{Y_{t_0}(s_j)\colon j \neq i\} \\ &\quad \sim Gau((\phi_{t_0}/(1+\phi_{t_0}^2))\{Y_{t_0}(s_{i-1}) + Y_{t_0}(s_{i+1})\}, \sigma_{t_0}^2/(1+\phi_{t_0}^2)), \end{aligned} \tag{2.1}$$

where recall that $s_i \equiv s_0 + i\Delta$; $i = 0, \ldots, 24$. On the edges of the transect, assume

$$Y_{t_0}(s_0)|\{Y_{t_0}(s_j)\colon j \neq 0\} \sim Gau(\phi_{t_0} Y_{t_0}(s_1), \sigma_{t_0}^2),$$

$$Y_{t_0}(s_{24})|\{Y_{t_0}(s_j)\colon j \neq 24\} \sim Gau(\phi_{t_0} Y_{t_0}(s_{23}), \sigma_{t_0}^2).$$

In (2.1), assume that the *spatial-dependence parameter* ϕ_{t_0} satisfies $|\phi_{t_0}| \leq 1$. It can be shown that $E(\mathbf{Y}_{t_0}) = \mathbf{0}$, and the correlation between nearest neighbors is (Section 4.2)

$$\operatorname{corr}(Y_{t_0}(s_i), Y_{t_0}(s_{i-1})) = \phi_{t_0}, \qquad i = 1, \ldots, 24.$$

A simple departure from independence for a *time series* is a first-order autoregressive process. Assume that

$$Y_t(s_0) = \phi(s_0) Y_{t-1}(s_0) + \varepsilon_t, \qquad t = t_0 + 1, \ldots, t_0 + 24, \tag{2.2}$$

where ε_t is independent of $Y_{t-1}(s_0)$, and $\varepsilon_t \sim \ ind.Gau(0, \sigma^2(s_0))$, for $t = t_0, t_0 + 1, \ldots, t_0 + 24$. To initialize the process, assume

$$Y_{t_0}(s_0) \sim Gau(0, \sigma^2(s_0)/(1 - \phi(s_0)^2)).$$

In (2.2), we assume that the temporal-dependence parameter $\phi(s_0)$ satisfies $|\phi(s_0)| < 1$. It can be shown that $E(\mathbf{Y}(s_0)) = \mathbf{0}$ and the correlation between two adjacent time points is (Section 3.4.3)

$$\operatorname{corr}(Y_{t-1}(s_0), Y_t(s_0)) = \phi(s_0), \qquad t = t_0 + 1, \ldots, t_0 + 24.$$

The process (2.2) is *dynamical* in that it shows how current values are related mechanistically to past values. Generally, the dependence of current values on

past values is expressed probabilistically, and (2.2) has an equivalent probabilistic expression:

$$Y_t(s_0)|Y_{t-1}(s_0), \ldots, Y_{t_0}(s_0) \sim Gau(\phi(s_0)Y_{t-1}(s_0), \sigma^2(s_0)).$$

Such time series models are sometimes referred to as causal (Section 3.4.3).

Notice the similarity between the spatial process (2.1) and the time series (2.2). Both are Gaussian, zero mean; and if $\phi_{t_0} = \phi(s_0)$, they imply the *same* correlation between adjacent random variables. In fact, as we show in Section 4.2, if $\phi_{t_0} = \phi(s_0)$, then the processes are probabilistically identical! However, the spatial process (2.1) looks east and west for dependence, in contrast to the time series (2.2), which looks to the past. The example has a cautionary aspect. Clearly, a *description* of the properties of spatial or temporal statistical dependence of the model through just moments or even joint probability distributions can completely miss the genesis of the statistical dependence, such as the *dynamical* structure given by (2.2).

Now, when it comes to considering space and time together, we believe that (whenever possible) the temporal dependence should be expressed dynamically, based on physical/chemical/biological/economic/etc. theory, since here the etiology of the phenomenon is clearest. In a contribution to the Statistics literature that was well ahead of its time, Hotelling (1927) gave various statistical analyses based on stochastic differential equations (albeit without a spatial dimension). Our approach contrasts to that of some others, where time is treated as an extra (although different) dimension, and descriptive expressions of spatial dependencies are modified to account for the temporal dimension; see Section 6.1 for further discussion of the two approaches.

2.1 CONDITIONAL PROBABILITIES AND HIERARCHICAL MODELING (HM)

There is a very general way to express uncertainties coming from different sources, through an approach known as hierarchical (statistical) modeling. Chapter 1 gives a discussion of the HM approach through the introduction of a data model and a process model, where the uncertainties are expressed in terms of *conditional probabilities*. This book is about Statistics for spatio-temporal data, and the quantities we are interested in could be as complicated as spatio-temporal stochastic processes of random variables, random vectors, or random sets.

The conditional-probability distributions specified in the hierarchical model (also abbreviated as HM) typically depend on unknown parameters. If a parameter model is included in the HM, in order to express probabilistically the uncertainty on the parameters, the HM is called a *Bayesian Hierarchical Model* (BHM); see Chapter 1. An alternative approach to specifying a prior distribution is to assume that the parameters are fixed and to estimate them using the data; they are then substituted into the data model and the process model as

if they were known. The result is an *Empirical Hierarchical Model* (EHM); also see Chapter 1. We note here (and later in the chapter) that it is possible to put prior distributions on some parameters and to estimate others. In this book, we typically use the term "prior distribution" to be synonymous with "parameter model." However, we recognize that prior information goes into all three components of the hierarchical model. Traditionally, Bayesians have considered what we call the process and parameter distributions to make up the prior distribution. We do not disagree with this but simply prefer to make a distinction between the process and parameters whenever possible.

Consider three generic quantities of interest, Z, Y, and θ, in the HM; for expository purposes we often consider these simply to be random variables. Think of Z as data, Y as a (hidden) process that we wish to predict, and θ as unknown parameters. In a realistic example where Z, Y, and θ are more complicated random quantities, say for spatial statistical mapping of a region's air quality in a given week, Z might be 100-dimensional, Y might be 1000-dimensional, and θ might be five-dimensional. Based on Z, we wish to make inference on Y and θ. That is, in a BHM, we wish to predict both Y and θ; and in an EHM, we wish to predict Y and to estimate/predict θ.

We now give some basic results from probability theory. Recall the notation $[A]$ and $[A|B]$ for marginal and conditional probability distributions, respectively. Then the joint distribution of A and B can be written as

$$[A, B] = [A|B][B], \tag{2.3}$$

and the law of total probability can be written as

$$[A] = \int [A|B][B]\, dB, \tag{2.4}$$

where recall that $\int g(B)[B]\, dB$ denotes the expectation (either an integral or a summation in the case where B is a discrete random quantity) of some function $g(B)$ of B. Finally, in terms of this notation, *Bayes' Theorem* (Bayes, 1763) can be written as

$$[B|A] = \frac{[A|B][B]}{\int [A|B][B]\, dB} = \frac{[A|B][B]}{[A]}. \tag{2.5}$$

2.1.1 Bayesian Hierarchical Modeling (BHM)

The basic representation of a BHM is obtained by splitting up the model into three levels (Berliner, 1996):

Data model:	$[Z\|Y, \theta]$
Process model:	$[Y\|\theta]$
Parameter model:	$[\theta]$.

Note that sometimes we write $[Z|Y, \theta_D]$ and $[Y|\theta_P]$ to emphasize the data-model parameters θ_D and the process-model parameters θ_P. Then $\theta = \{\theta_D, \theta_P\}$, and the parameter model is $[\theta_D, \theta_P]$.

Now the joint distribution can be decomposed recursively. From (2.3), we have

$$\begin{aligned}[Z, Y, \theta] &= [Z, Y|\theta][\theta] \\ &= [Z|Y, \theta][Y|\theta][\theta],\end{aligned} \tag{2.6}$$

which is simply a product of the data model, the process model, and the parameter model. A special case would be where $\theta = \theta_0$, known, and $[\theta]$ concentrates all its probability at θ_0.

Bayes' Theorem gives the conditional distribution of Y and θ, given the data Z, which is typically called the *posterior distribution*. From (2.5), we obtain

$$\begin{aligned}[Y, \theta|Z] &= \frac{[Z|Y, \theta][Y, \theta]}{\int\int [Z|Y, \theta][Y, \theta]\, dY\, d\theta} \\ &= \frac{[Z|Y, \theta][Y|\theta][\theta]}{\int\int [Z|Y, \theta][Y|\theta][\theta]\, dY\, d\theta} \\ &= \frac{[Z|Y, \theta][Y|\theta][\theta]}{[Z]}.\end{aligned} \tag{2.7}$$

Within the framework of Bayesian decision theory, all inference on Y and θ in the BHM depends on this distribution.

Suppose that the data come in two "bursts," $Z^{(1)}$ followed by $Z^{(2)}$. After the first burst of data, $Z^{(1)}$, the posterior distribution is

$$[Y, \theta|Z^{(1)}] = [Y|\theta, Z^{(1)}][\theta|Z^{(1)}]; \tag{2.8}$$

think of the two probability distributions on the right-hand side of (2.8) as the "updated" process model and the "updated" parameter model, respectively. The updated probability distributions represent *scientific learning* about Y and θ, respectively. From (2.7) the posterior distribution is proportional to

$$[Z^{(1)}|Y, \theta][Y|\theta][\theta].$$

Now, after the second burst of data, $Z^{(2)}$, the posterior distribution should be recalculated:

$$\begin{aligned}[Y, \theta|Z^{(1)}, Z^{(2)}] &= \frac{[Z^{(1)}, Z^{(2)}|Y, \theta][Y, \theta]}{[Z^{(1)}, Z^{(2)}]} \\ &= \frac{[Z^{(2)}|Y, \theta, Z^{(1)}][Y, \theta|Z^{(1)}]}{[Z^{(2)}|Z^{(1)}]},\end{aligned} \tag{2.9}$$

which shows how "today's posterior becomes tomorrow's prior." Substituting (2.8) into (2.9) shows that the posterior distribution is proportional to

$$[Z^{(2)}|Y, \theta, Z^{(1)}][Y|\theta, Z^{(1)}][\theta|Z^{(1)}].$$

This expression shows that the posterior distribution is proportional to the product of the updated data model, the updated process model, and the updated parameter model. This is the essence of the sequential implementation given in Section 8.1.1. Furthermore, it is often the case that, given the process Y, $Z^{(1)}$ will not affect the conditional distribution of Y. That is, the first term in this expression often simplies to $[Z^{(2)}|Y, \theta]$.

Since data can come sequentially, these simple calculations are very relevant to how scientists can ascend the knowledge pyramid (see Chapter 1). That is, Bayes' Theorem allows knowledge to be continually improved in a coherent manner.

The numerator in (2.7) is a straightforward product of the individual components of the BHM, but a major problem usually arises when calculating the denominator. The denominator is the normalizing constant that ensures that the posterior distribution has total probability equal to 1. (Because the posterior distribution is conditional on Z, in fact the normalizing "constant" depends on Z.)

When Y and θ are each random variables, the integral in the denominator of (2.7),

$$[Z] = \int\int [Z|Y, \theta][Y|\theta][\theta]\, dY\, d\theta,$$

is only two-dimensional and usually quite easy to calculate using numerical quadrature. However, spatio-temporal BHMs can often yield integrals that are of dimensions on the order of thousands (e.g., Wikle, Berliner, and Cressie, 1998). In the last 20 years, computational breakthroughs have been made so that rather than calculating the posterior distribution analytically or numerically, one can often *simulate* from it (Section 2.3). These computational methods, including Markov chain Monte Carlo (MCMC) and importance sampling (IS), have brought HM into the panoply of many statisticians, including those concerned with modeling spatio-temporal data.

2.1.2 Empirical Hierarchical Modeling (EHM)

The following two-level model also qualifies to be called a HM:

Data model: $[Z|Y, \theta]$
Process model: $[Y|\theta]$,

where it is assumed that the parameter θ is *fixed*, but unknown. Formally, one could still consider a third level, but where the parameter model $[\theta]$ concentrates all its probability at the fixed θ. Recall that sometimes we emphasize

the data-model parameters as θ_D and the process model parameters as θ_P by writing the two-level model as $[Z|Y, \theta_D]$, $[Y|\theta_P]$, and $\theta = \{\theta_D, \theta_P\}$.

In an EHM, all probability distributions are conditional on θ. Inference on Y depends on the distribution

$$[Y|Z, \theta] = \frac{[Z|Y, \theta][Y|\theta]}{[Z|\theta]}, \tag{2.10}$$

where $[Z|\theta] = \int [Z|Y, \theta][Y|\theta]dY$. Equation (2.10) is sometimes called the *predictive distribution*, but in this book we take some license and continue to call it the posterior distribution. The difference between (2.7) and (2.10) is clear, and which one is used as the posterior distribution depends on the type of HM fitted. Notice that the integral in the denominator of (2.10) is lower dimensional than that in (2.7), but it could still be of a dimension on the order of thousands. The "Empirical" part of the EHM arises from the practice of replacing (2.10) with $[Y|Z, \hat{\theta}]$, where $\hat{\theta}$ is an *estimator* of θ (i.e., depends only on the data Z). It is also possible that θ is estimated from an independent study.

Importantly, (2.10) does not require explicit specification of a prior distribution for the parameters, a task that some statisticians are reluctant to do. It can also be the case that (2.10) is faster to compute than (2.7). The price of not specifying uncertainty in the parameter θ is that inferences on Y are generally too liberal, since a simple substitution of $\hat{\theta}$ for θ does not account for the extra variability associated with the estimation of θ (e.g., Carlin and Louis, 2000, Chapter 4; Kang, Liu, and Cressie, 2009). This can result in misleading inferences for say g(Y), where g(·) is a nonlinear functional (Ghosh, 1992; Stern and Cressie, 1999). Second-order adjustments to these inferences on Y to account for the variability in $\hat{\theta}$ are available for simple cases (e.g., Rao, 2003, Section 6.2).

We have already noted that some parameters might have a prior distribution specified for them and some might be assumed fixed but unknown and estimated. This is the case in the example that follows, where an EHM was used to search for a missing nuclear submarine.

2.1.3 Search for the USS Scorpion

In late May 1968, for reasons that are still unclear, the USS *Scorpion* (SSN-589), a nuclear submarine, was lost at sea as it was returning to its naval base at Norfolk, Virginia. An official search for the vessel was started in early June of 1968, in an area of the Atlantic Ocean approximately 400 miles southwest of the Azores (Richardson and Stone, 1971). The search was complicated by the remoteness and extreme depth of this part of the ocean, the lack of certainty as to the location of the *Scorpion* when it (presumably) went down, and the cause of its sinking. Because of success in using Bayesian statistical methods to find a hydrogen bomb lost in the Mediterranean Sea in 1966, scientists at the U.S. Naval Research Laboratory (NRL) implemented a hierarchical statistical search procedure to help find the *Scorpion*. This procedure is conceptually simple (but can have practical challenges) and provides an introduction to the power of HM.

The idea is first to define a spatial "grid" and to propose a first guess (prior) of the probabilities that the object in question (here, the *Scorpion*) is located in each of the grid boxes. The prior probabilities suggest which grid box to search first. If the object is not found in that box, the prior probabilities are then updated (yielding the posterior), and the process is repeated until the object is found. This procedure suggests a fairly simple HM.

Assume that the domain of interest is a part of the ocean floor we call D_s, which is made up of n spatial areas (i.e., grid boxes). Let $Y_i = 1$ if the submarine is in the ith grid box, and let $Y_i = 0$ if it is not; $i = 1, \ldots, n$. Now, it is critical to recognize that when a grid box is searched, there is a chance that, even if the submarine is present, it will not be detected. So, let $Z_i = 1$ if the submarine is found in the ith grid box, and $Z_i = 0$ if not. Borrowing from the terminology associated with occupancy modeling in Ecology (e.g., Royle and Dorazio, 2008, p. 100), we distinguish between the *detection probability*,

$$p_i = \Pr(Z_i = 1 | Y_i = 1), \qquad i = 1, \ldots, n,$$

which is a conditional probability, and the *occurrence probability*,

$$\pi_i = \Pr(Y_i = 1), \qquad i = 1, \ldots, n.$$

This suggests an HM of the following form:

$$\begin{aligned} &\text{Data model:} && Z_i | Y_i \sim \textit{ind. } \textit{Ber}(Y_i p_i), && i = 1, \ldots, n \\ &\text{Process model:} && Y_i \sim \textit{ind. } \textit{Ber}(\pi_i), && i = 1, \ldots, n, \end{aligned}$$

where, for now, we assume that $\{p_i\}$ and $\{\pi_i\}$ are known or (more realistically) determined by expert opinion; and $Ber(p)$ denotes the Bernoulli distribution of a binary random variable, where p is the probability of obtaining a "1." Note that this HM suggests that if the submarine is not in the ith grid box (i.e., $Y_i = 0$), then $Z_i = 0$. If it is in the ith grid box, Z_i follows a Bernoulli distribution with (detection) probability p_i.

Now, assume that the ith grid box is searched and the submarine is not found (i.e., $Z_i = 0$). In this case, the probability that the submarine is in the ith grid box (i.e., $Y_i = 1$) is updated using Bayes' Theorem. This yields the posterior probability,

$$\begin{aligned} &\Pr(Y_i = 1 | Z_i = 0) \\ &\quad = \frac{\Pr(Z_i = 0 | Y_i = 1)\Pr(Y_i = 1)}{\Pr(Z_i = 0)} \\ &\quad = \frac{\Pr(Z_i = 0 | Y_i = 1)\Pr(Y_i = 1)}{\Pr(Z_i = 0 | Y_i = 1)\Pr(Y_i = 1) + \Pr(Z_i = 0 | Y_i = 0)\Pr(Y_i = 0)} \\ &\quad = \frac{(1 - p_i)\pi_i}{(1 - p_i)\pi_i + (1)(1 - \pi_i)} \\ &\quad = \frac{(1 - p_i)\pi_i}{1 - p_i\pi_i}, \end{aligned}$$

where we assume that there are no false-positive detections (i.e., $\Pr(Z_i = 0|Y_i = 0) = 1$). Note that the posterior probability of the submarine being in the ith grid box is *less* than or equal to the prior probability π_i, given that the submarine was not observed there.

If the submarine is not detected in the ith grid box, then this should also affect the posterior probability in the other grid boxes. For example, consider the jth grid box, where $j \neq i$. Then,

$$\Pr(Y_j = 1|Z_i = 0) = \frac{\Pr(Z_i = 0|Y_j = 1)\Pr(Y_j = 1)}{\Pr(Z_i = 0)}$$

$$= \frac{\pi_j}{1 - p_i\pi_i},$$

where we further assume that there is only one submarine and, hence, $\Pr(Z_i = 0|Y_j = 1) = 1$; $j \neq i$. Thus, the posterior probability of the submarine being in the jth grid box is *greater* than or equal to the prior probability π_j, given that the submarine is not detected in a different (i.e., the ith) grid box. These new posterior probabilities would then become the prior probabilities in a sequential procedure that would help determine which grid box should be searched next. Recall that "today's posterior becomes tomorrow's prior."

In practice, the parameters $\{p_i\}$ and $\{\pi_i\}$ need to be estimated (EHM), or other information could be used to specify them with more or less uncertainty (BHM). In the actual search for the *Scorpion*, the probabilities $\{\pi_i\}$ were determined from subjective expert opinions and physical/scientific information. NRL scientists Dr. J.P. Craven and Dr. F.A. Andrews considered nine plausible scenarios associated with the (presumed) accident that led to the disappearance of the *Scorpion*. For each of these, they assigned weights, based in part on "the views and opinions of Navy operating personnel as well as the analysis of specialists in many scientific areas" (Richardson and Stone, 1971, p. 144). Basically, for each scenario, "... the movement of the submarine ... [was] then simulated with random numbers drawn as required to represent the uncertainties in course, speed, and position, at the time the emergency occurred, as well as other variables" (Richardson and Stone, 1971, p. 144–145). In the search, a 20×20 grid (i.e., $n = 400$) was considered, and the probabilities $\{\pi_i : i = 1, \ldots, 400\}$ were calculated and assigned. The assignment was in no way uniform; in fact, two grid boxes accounted for 23% of the total probability.

A great deal of scientific and engineering expertise went into determining the detection probabilities, $\{p_i\}$, as well. There were multiple instruments used in the search, all with different degrees of detectibility, and operational procedures actually used to conduct the search had to be factored in (Richardson and Stone, 1971). On October 28, 1968, the *Scorpion* was located and its location was within 260 yards of the edge of the grid box with the highest initial prior probability! Unfortunately, it sank in very deep water, its hull was crushed, and all 99 officers and crew were lost.

2.1.4 "Classical" Statistical Modeling

While it might seem unusual, we use "classical" as an adjective here for both frequentist and Bayesian modeling. The HM introduces data Z, process Y, and parameters θ; however, the "classical" model found in the work of Fisher (e.g., Fisher, 1935) has only data Z and parameters θ, as does the contribution of Bayes and many who followed him (e.g., Press, 1989). "Classical" frequentists base their inference on the *likelihood*, $[Z|\theta]$. "Classical" Bayesians base their inferences on the posterior distribution, $[\theta|Z]$, which requires *both* a likelihood and a *prior* (i.e., $[\theta]$) to be specified. Both classical approaches miss the fundamental importance of modeling the process Y, where the Physics/Chemistry/Biology/Economics/etc. typically resides.

To be sure, Statistics has played an important role in Science, but often using blunt instruments, like correlation and regression analyses. Without Y being made explicit in statistical models, Science has often chosen its own statistical path, since it is on Y that scientific theories are postulated. Scientists also know that parameters θ are important; these might be starting values, or boundary conditions, or diffusion constants, or etc. In what follows, we give a deliberately simplistic description of how a scientist might view the role of Statistics, although we note that this is changing fast. One of the goals of this book is to accelerate this change.

Scientific experiments produce data Z, and variability in the data is generally recognized. One approach in Science to analyzing the data (to support, refine, or refute a scientific theory) has been to "smooth" them first. Consider the smoother f and write

$$\tilde{Y} = f(Z).$$

The scientist might then assume that any (random) variability has been *removed* and that $\tilde{Y}$ can now be considered the true process with no uncertainty. Less extreme would be to assume that $\tilde{Y}$ and the true process Y are "close." The scientist might then fit a model for Y using the "data" $\tilde{Y}$. If the model for Y is $[Y|\theta_P]$, namely a process model that recognizes uncertainty, the scientist might use "classical" Statistics to fit $[Y|\theta_P]$ to $\tilde{Y}$. Science alert! Control of uncertainty has been lost. While the approach just described can work sometimes, typically when the "signal" is strong, it also has the potential to declare the presence of a "signal" when it may simply be the result of chance fluctuations.

Given the data are to be smoothed, it should be recognized that they are often a combination of raw observations and algorithmic manipulation. The statistician might write instead

$$\tilde{Z} = f(Z), \tag{2.11}$$

where the notation $\tilde{Z}$ in (2.11) is used to suggest a fundamental divide between the two approaches.

Now, a hierarchical statistical model can be fitted using the data $\tilde{Z}$, where the data model $[\tilde{Z}|Y, \theta_D]$ recognizes any remaining uncertainty in $\tilde{Z}$. Inference is based on the posterior distribution, which we choose to show here for an EHM:

$$[Y|\tilde{Z}, \theta_D, \theta_P] \propto [\tilde{Z}|Y, \theta_D][Y|\theta_P]. \tag{2.12}$$

At first glance, it may seem that thinking of the smoothed data as $\tilde{Z}$ in (2.11) (rather than as $\tilde{Y}$) is inconsequential, but we believe that this notational prop guides us (statisticians and scientists, alike) through our "uncertainty audit."

While the picture painted above is simplistic, it does illustrate that Statistics has often failed to establish its proper presence in Science. Scientific interest is in Y, and if a classical frequentist statistician were to fit the scientific model $[Y|\theta_P]$ directly to $\tilde{Z}$, it should be done through the (marginal) model,

$$[\tilde{Z}|\theta_D, \theta_P] = \int [\tilde{Z}|Y, \theta_D][Y|\theta_P]dY.$$

That is, the classical frequentist should integrate out Y; alas, if there is no recognition of Y in the first place, the model chosen to be fitted, $[\tilde{Z}|\theta]$, may be difficult to interpret scientifically or, worse yet, may be inappropriately interpreted.

In this book, we would like to take Science on a path where original observations Z are used as much as possible, where smoothed data are thought of as $\tilde{Z}$, where uncertainties are captured in a HM using conditional probabilities, and where inference is based on the posterior distribution, such as (2.12) or its BHM version. This is the sharp statistical tool needed for scientists to ascend the knowledge pyramid (Chapter 1).

2.1.5 Hierarchical Statistical Modeling

The real power of the HM becomes apparent when dependencies become complicated. Each component distribution can be decomposed further if necessary, and they may be simplified with modeling assumptions. In what follows, we give examples and a discussion of HM's strengths and limitations, based on the presentation found in Cressie et al. (2009).

Data Model

Say we are interested in a process Y for which we have several different data sets, Z_1, Z_2, Z_3, all of which measure the process Y with uncertainty, and perhaps they are measured at different spatial or temporal scales. It is often possible in such cases to make the following data-modeling assumption:

$$[Z_1, Z_2, Z_3|Y, \theta_D] = [Z_1|Y, \theta_{D,1}][Z_2|Y, \theta_{D,2}][Z_3|Y, \theta_{D,3}],$$

where the *data-model parameters* θ_D are here given by $\{\theta_{D,1}, \theta_{D,2}, \theta_{D,3}\}$. That is, we might assume that the different datasets are independent, conditional upon the true process. Although such an assumption must be justified, it is often plausible and provides a very convenient approach for synthesizing various types of observations. The parameters in each of the component distributions can accommodate changes of resolution and alignment, as well as different measurement-error characteristics (e.g., Gelfand, Zhu, and Carlin, 2001; Wikle and Berliner, 2005). This is discussed further in Sections 4.1.3 and 7.1.2.

It is also the case that different datasets that correspond to different processes (or parameters) can be combined. For example, assume we have two datasets, Z_1 and Z_2, corresponding to the processes Y_1 and Y_2, respectively, where $Y = (Y_1, Y_2)$. We can often make use of conditional-independence assumptions, as follows:

$$[Z_1, Z_2|Y, \theta_D] = [Z_1|Y_1, \theta_{D,1}][Z_2|Y_2, \theta_{D,2}],$$

where the data-model parameters are here given by $\theta_D = \{\theta_{D,1}, \theta_{D,2}\}$.

Process Model

Decomposition of process-model distributions can also be considered. For example, consider Y to be made up of two subprocesses, Y_1 and Y_2, as in the example just above. We can often make use of conditional-probability modeling in this context:

$$[Y_1, Y_2|\theta_P] = [Y_1|Y_2, \theta_P][Y_2|\theta_P],$$

where θ_P are the *process-model parameters*. That is, it is often possible to simplify the joint interaction of two components of the process by using conditional probabilities.

A well known example of this occurs in a time series model with Markov assumptions (Section 3.4.3). Specifically, consider the time series, $Y_1, Y_2, \ldots, Y_{T-1}, Y_T$; it is often very difficult to specify the joint distribution of all of these random variables. However, consider the conditional-probability factorization:

$$[Y_T, \ldots, Y_1] = [Y_T|Y_{T-1}, \ldots, Y_1][Y_{T-1}|Y_{T-2}, \ldots, Y_1]$$
$$\ldots [Y_2|Y_1][Y_1].$$

This factorization could be simplified, for example, with a first-order Markov assumption, in which the conditional-probability distribution of the process at time t, conditioned on the process at times prior to t, is only dependent on the most recent time. That is, the *first-order Markov* assumption is

$$[Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_1] = [Y_t|Y_{t-1}].$$

Then, for first-order Markov processes, the joint distribution is

$$[Y_T, Y_{T-1}, \ldots, Y_1] = [Y_T|Y_{T-1}][Y_{T-1}|Y_{T-2}]\cdots[Y_2|Y_1][Y_1]$$
$$= [Y_1]\prod_{i=2}^{T}[Y_i|Y_{i-1}]. \tag{2.13}$$

Notice that the right-hand side now depends only on univariate and bivariate probability distributions, since

$$[Y_i|Y_{i-1}] = [Y_i, Y_{i-1}]/[Y_i].$$

The property (2.13) plays an important role in the updating mechanism that underlies the Kalman filter (Kalman, 1960), which Meinhold and Singpurwalla (1983) demonstrate is the result of optimal prediction of the hidden time series $\{Y_t : t = 1, 2, \ldots\}$ in a temporal HM. When Y_t is a random spatial process, spatio-temporal HMs can be built that result in a spatio-temporal Kalman filter (Huang and Cressie, 1996; Wikle and Cressie, 1999; see also Sections 8.2.1 and 9.2).

Again, these process-model decompositions are attractive because they make use of modeling assumptions that are scientifically plausible. In that way, very complicated joint-probability distributions can be modeled by relatively simple conditional-probability specifications. Often, in the process model, deterministic models can be reformulated with stochastic components to account for scientific uncertainty. For example, Wikle (2003b) uses a reaction-diffusion partial differential equation to motivate a process model for invasive species.

Of course, it is also possible that the process model can be factorized further or can be simplified from scientifically based modeling assumptions. For example, Wikle (2003b) specifies the distribution of spatially explicit diffusion-coefficient parameters, conditional on habitat covariates and spatial random fields. The latter processes could then be modeled too (conditional on parameters), to incorporate sublevels into the process model. HMs can have many levels, so long as there is scientific insight or data upon which to base the decomposition. When there is no longer scientific insight, the BHM typically invokes "noninformative" priors as part of the parameter model.

EHM- and BHM-Based Inferences

Now, if one agrees that the HM approach is reasonable, there remains the question of how to do inference in this setting. Certainly, modeling the uncertainty can be carried out at the first two levels: $[Z|Y, \theta_D]$ and $[Y|\theta_P]$, and the parameters $\theta = \{\theta_D, \theta_P\}$ could be considered fixed but unknown. We have seen already that this is the EHM approach. The classical linear mixed model can be thought of as an EHM (e.g., Demidenko, 2004, Chapter 3). In addition, spatial prediction (kriging) fits into this framework (e.g., Cressie, 1988), as do sequential time series methods such as the Kalman filter (e.g., Meinhold and Singpurwalla, 1983; see also Section 8.2.1).

Depending on the complexity of the data model and the process model, it is often possible to use classical statistical estimation approaches to obtain estimates of the parameters θ. Common approaches for estimation in this EHM context include maximum likelihood estimation, the Expectation-Maximization (EM) algorithm, conditional-likelihood and pseudo-likelihood methods, and estimating equations (e.g., Demidenko, 2004), some of which are illustrated in Chapter 8. Although the EHM approach does not explicitly recognize the uncertainty in estimating the parameters, that uncertainty can often be accounted for by resampling and bootstrap procedures (e.g., Stoffer and Wall, 1991; Efron and Tibshirani, 1993; Lahiri, 2003). Bootstrapping (including the parametric bootstrap) in the spatial and temporal context is discussed by Cressie (1993, pp. 492–497). In simple cases, the uncertainty can be accounted for approximately, using statistical perturbation arguments (e.g., Rao, 2003, Section 6.2).

A coherent approach for inference in complicated HMs is to account for the uncertainty in parameter values by adding a parameter model, $[\theta_D, \theta_P]$. We have already seen that this is the BHM approach; recall that inference is based on the posterior distribution of process and parameters, given the data:

$$[Y, \theta_D, \theta_P | Z] \propto [Z | Y, \theta_D][Y | \theta_P][\theta_D, \theta_P].$$

As discussed at the beginning of this section, it is seldom possible to obtain an analytical expression for this posterior distribution. The discovery that *Markov chain Monte Carlo (MCMC)* computational methods could be used to simulate from general BHMs (Gelfand and Smith, 1990) revolutionized Statistical Science and made ever-more-complicated modeling scenarios possible. These computational aspects are discussed initially in Section 2.3 and for spatio-temporal dynamical models in Chapters 8 and 9.

Where, When, and then Why

The problem of determining a causative relationship in a process model can be expressed in terms of conditional probabilities. If Y_1 is a phenomenon that could directly affect Y_2 through a physical/chemical/biological/economic/etc. mechanism, and $[Y_2|Y_1]$ changes as Y_1 changes, then Y_1 is a candidate to be a cause of Y_2.

However, even the best of theories can miss an important factor (F, say), which might damp down the relationship or yield a negative dependence where there was originally a positive one. We called this Simpson's Paradox and F a "lurking variable" in Chapter 1.

For the simple process, $Y = (Y_1, Y_2, F)$, the process model can be written as

$$\begin{aligned}[Y] &= [Y_1, Y_2 | F][F] \\ &= [Y_2 | Y_1, F][Y_1 | F][F].\end{aligned}$$

Thus, the question, "Why does Y_2 behave as it does?" can be answered, "Because Y_1 causes that behavior, and the *type* of dependence is governed

by the factor F." If we mistakenly focused on $[Y_2|Y_1]$ instead of $[Y_2|Y_1, F]$, we may infer incorrect dependencies. When F represents "level of spatial aggregation," these spurious inferences have been called the *ecological fallacy* (Robinson, 1950); a spatial-statistical interpretation of this can be found in Cressie (1996), and it was called change-of-support in Chapter 1.

More generally, many factors have spatial and temporal variability; hence, space and time can act as a *proxy* for F, should the process model fail to account for it. In other words, modeling spatio-temporal variability, along with good experimental design (see Chapter 1), can get us closer to Science's holy grail, namely *causation*.

Additional Remarks

There are many challenges associated with building (Bayesian) hierarchical statistical models and then carrying out valid inferences. A historical criticism of Bayesian methods is that they require "subjective" specification of prior information on the parameters. Of course, there is also subjectivity in the specification of the likelihood in classical marginal probability models. In fact, a broader perspective is that there is subjectivity involved with the specification of *all* model components: data models, process models, and parameter models. However, it isn't always clear what "subjective" means in this context. For example, it might be "subjective" to use deterministic relationships to motivate a stochastic model, such as for tropical winds (e.g., Wikle et al., 2001), yet the science upon which such a model is based comes from Newton's laws of motion! Thus, we believe that it is not helpful to try to classify probability distributions that determine the statistical model, as subjective or objective. Better questions to ask are about the sensitivity of inferences to model choices and whether such choices make sense scientifically.

Given that a modeler brings so much information to the table when developing models, the conditional-probability framework presented earlier can be used to recognize that this information, say I, is part of what is involved in the conditioning. For a BHM, we have

$$[Y, \theta_D, \theta_P|Z, I] \propto [Z|Y, \theta_D, I][Y|\theta_P, I][\theta_D, \theta_P|I].$$

A major challenge in this paradigm is, to the extent possible, appreciation of the importance of this information, I. It is often the case that a team of researchers at the table have a collective "I" that is better quantified and more appropriate than any individual's "I."

In the HM approach, there are certainly cases where models have to be simplified due to practical concerns. Perhaps the computational issues in a given formulation are limiting, which usually leads to a modification of the model. Such practical concerns are not limited to BHM-based or EHM-based inferences, but they concern all statistical inferences in complicated modeling scenarios.

Lastly, it must be recognized that complicated models often take a life of their own; it is the responsibility of the research team to remember the

Goldilocks Principle discussed in Chapter 1 and to keep uppermost the goal of converting data and information to knowledge. To temper the tendency to fit ever-more-complicated models, there are model-selection criteria that could be invoked (e.g., AIC, BIC, DIC, etc.), which concentrate on the twin pillars of *predictability* and *parsimony* (e.g., Spiegelhalter et al., 2002). But they do not address the third pillar, namely *scientific interpretability* (i.e., knowledge). Our approach is to choose, where possible, statistical models based on this third pillar, while not ignoring the other two. As a consequence, we have placed less emphasis on statistical-model-selection criteria in this book.

2.2 INFERENCE AND DIAGNOSTICS

A dataset can be thought of as a window through which knowledge can be obtained, enough to infer answers to the "why" question. What stops a scientist from truly *deducing* answers, rather than inferring them, is the ubiquitous presence of *uncertainty*. Statistics accounts for the uncertainty with a statistical model. Within that framework, statisticians are constantly looking for optimal procedures and then for ways to quantify the uncertainty in those optimal procedures.

In contrast to Statistics, Data Mining does not systematically seek optimality nor quantification of uncertainty. The data miner is often looking for parts of the dataset that are unusual, whereas the statistician is often looking for comprehensive structures that describe the data. Sometimes, the statistician will hold back observations that are unusual, for further study; the data miner tends to focus in on those unusual observations.

The HM approach, when implemented properly, can account for unusual observations in Z through a heterogeneous process Y. In the absence of an HM approach, classical statistical approaches can struggle to account for unusual observations. For example, suppose that precipitation data for a city, say Z_1, shows several extreme values over time. An HM could handle this heterogeneity by modeling the precipitation process, Y_1, to depend on Y_2, say the aerosol content of the atmosphere. That is,

$$[Y_1, Y_2|\theta] = [Y_1|Y_2, \theta][Y_2|\theta].$$

Each component on the right-hand side might be a simple conditional-probability distribution, but their product can yield a complex model. Then the "unusualness" in Z_1 is manifested through the marginal-probability model,

$$[Z_1|\theta] = \int\int [Z_1|Y_1, \theta][Y_1|Y_2, \theta][Y_2|\theta]dY_1dY_2.$$

Extreme rainfall values might be due to atmospheric aerosol content on those days. A more classical statistical approach might delete the outliers and fit a time series model or attempt a regression model for $[Z_1|Z_2, \theta]$, where Z_2 are

aerosol data, without making the processes Y_1 and Y_2 explicit. The HM gives us a structure to look for reasons for outliers.

A data-mining approach might "tease out" parts of Z_1 that are unusual, often using *ad hoc* (i.e., nonoptimal) methods to find them and often not being able to say what "unusual" means in any consistent way. The particular dataset under study is often the focus. Perhaps a data miner studying the rainfall data Z_1 will think of heterogeneity due to Y_2, perhaps not. Perhaps the "unusual" parts of Z_1 are not actually unusual, and the data miner has found something that is there simply due to chance fluctuations in the data or in the process. We believe that Data Mining has an important place in exploratory data analysis and gives the scientist ideas about how to model the phenomenon (i.e., process) being studied. However, we believe that answering the "why" question (i.e., inference) involves accounting for uncertainty in a coherent way. Statistical modeling and, in particular, HM fulfills this requirement.

Generally speaking, the data model is well defined by the scientist's data-gathering protocol. However, the process model is where the underlying scientific theories, and their uncertainties, are characterized in the form of $[Y|\theta_P]$. Such a theory is often complex, requiring sublevels of conditional-probability models to be defined; examples are given in Chapter 9. These sublevels consist of simpler conditional-probability distributions whose product yields $[Y|\theta_P]$. Consequently, classes of probability distributions developed in the twentieth century, but deemed too simple to model the datasets of the twenty-first century, might in fact have use as simple conditional-probability components of the HM. We shall briefly discuss some of the ideas behind these simple spatial and temporal models.

In the absence of detailed scientific information, simple probability models can provide an initial description of spatio-temporal variability. *Stationarity* in space (time) is a fundamental notion that formally says that a spatial (temporal) process' statistical properties are invariant under translation. Informally, this says that the spatial (temporal) behavior of the process is statistically identical anywhere in D_s (D_t).

Isotropy in space is defined as invariance under rotation about a given spatial location. Consider a "simple" probability distribution for the spatial process Y that is both stationary and isotropic (discussed more fully in Section 4.1). For example, consider the Gaussian process $\{Y(\mathbf{s}): \mathbf{s} \in D_s\}$ whose mean is μ, a constant, and whose *covariance* for any two locations $\mathbf{s} = (s_1, \ldots s_d)'$ and $\mathbf{x} = (x_1, \ldots, x_d)'$ is *stationary*,

$$\operatorname{cov}(Y(\mathbf{s}), Y(\mathbf{x})) = C_Y(\mathbf{s} - \mathbf{x}),$$

and *isotropic*. That is, $C_Y(\mathbf{s} - \mathbf{x})$ is in fact a function of $\|\mathbf{s} - \mathbf{x}\|$, where $\|\mathbf{s} - \mathbf{x}\| \equiv \{(s_1 - x_1)^2 + \cdots + (s_d - x_d)^2\}^{1/2}$. Then it can be shown that the (Gaussian) *process* $Y(\cdot)$ is stationary and isotropic in $\mathbb{R}^d$.

In fact, there are two forms of stationarity in space (and in time). *Strongly stationary* spatial processes have the property that any finite-dimensional joint

distribution at spatial locations $\{\mathbf{s}_i\}$ is identical to the finite-dimensional joint distribution at the lagged locations, $\{\mathbf{s}_i + \mathbf{h}\}$, where $\mathbf{h} = (h_1, \ldots, h_d)'$. For example, the spatial Gaussian process given just above is strongly stationary, since Gaussian processes are defined uniquely by their means and covariances. There is an obvious, analogous definition of strongly stationary temporal processes, where the spatial lag $\mathbf{h}$ is replaced with the temporal lag τ.

The terms *weakly stationary* and *second-order stationary* in space (time) are synonymous. The process $Y(\cdot)$ is second-order stationary if it has finite variance, constant mean, and covariance $C_Y(\cdot)$ that depends only on spatial lag $\mathbf{h}$ (or temporal lag τ). Hence, the class of second-order stationary processes contains the class of strongly stationary processes, provided that $C_Y(\mathbf{0}) = \sigma_Y^2 < \infty$. Notice that the Gaussian process given just above is not only strongly stationary but also second-order stationary.

Ergodic processes (in space or time) have the property that expectations of functionals of $Y(\cdot)$ are well approximated by spatial or temporal averages, respectively, of the same functional. Ergodic processes are a subset of strongly stationary processes (Cressie, 1993, pp. 53–58).

The covariance function $C_Y(\cdot)$ of a second-order stationary process (in space or time) has a Fourier representation in terms of sines and cosines. The Fourier transform of $C_Y(\cdot)$ defines the (power) spectrum, or spectral density, which is an equivalent way to represent the dependence in the covariance function. Equally, the process itself can be represented as a linear combination of Fourier basis functions (sines and cosines at various frequencies). The frequencies that have large squared coefficients most likely have a physical explanation. Thus, the spectral density is another way to characterize (spatial or temporal) dependence in a second-order stationary process.

Underlying these notions is the concept of information content, and whether more information on parts of Y allows us to decrease our uncertainty. This is a complex issue that has been discussed by Cressie (1993, Section 5.8) using the terminology *infill asymptotics* and *increasing-domain asymptotics*. Lahiri et al. (1999) use *mixed-domain asymptotics*, where the information content is increased by simultaneously allowing the sampling region to grow bigger and the distance between sampling locations to become smaller. In the HM approach, there is rarely need to do asymptotic statistical analysis, but it is still an important notion because, from one extreme point of view, Statistics for spatial, temporal, and spatio-temporal data appears to be *impossible*! The data Z almost never involve replication; Nature does not "do it over" under the same conditions.

Sometimes observations repeated over time, $Z_1, \ldots, Z_T$, can be thought of as comparable and provide a type of pseudo-replication. We saw in Section 2.1 how such data quite often depend on (recent) past values; the first-order Markov assumption is one commonly used model of this temporal dependence. In that case, pseudo-replicates, $Z_1, \ldots, Z_T$, do not have the same information content as a sample of *iid* observations; Cressie (1993, Chapter 1) discussed this from

the point of view of "equivalent number of independent observations." Nevertheless, for temporal data, the information content usually grows as $T \to \infty$, in spite of the temporal dependence. Cressie (1993, Section 5.8) called this increasing-domain asymptotics and contrasted it with infill asymptotics (sometimes referred to as fixed-domain asymptotics) that is often more appropriate for spatial data.

Consider the observations Z of a spatial process Y: From an extreme point of view, it is a sample of size 1, from which we wish to make inference! However, this is *not impossible* if there is spatial-dependence structure in the statistical model. We have seen above how that structure might be stationarity and/or isotropy; in an HM it appears as parametric specification of $[Z|Y, \theta_D]$ and $[Y|\theta_P]$. In this new century of very large Z and high-dimensional Y (e.g., spatio-temporal statistical analyses of climate data), invoking stationarity or isotropy assumptions, even within a parametric model, may be inappropriate. We need to look for other ways to avoid tackling an impossible problem, and dimension reduction has proved to be particularly successful; see Sections 3.5.2, 4.1.4, and 7.1.3, and the review by Wikle (2010a). By keeping in mind the notion of information content in Z, one can try to avoid specifying an HM for which inference on unknown quantities can only be weak at best.

2.2.1 Optimal Prediction

Statistical methodology is centered around the notion of optimal inference. Optimal estimators and optimal predictors represent a gold standard, which the collective talents of the Statistics profession have investigated over the last 100 years. These investigations could be classified according to type of inference (Bayesian, frequentist, fiducial, etc.), criteria (bias, variance, efficiency), non/semi/parametric distributional assumptions, exact/approximate/asymptotic optimality, quantification of the optimal estimators'/predictors' uncertainty, and so forth. In this book, our emphasis is on prediction of random quantities Y (and θ) in an HM. Estimation of θ is needed for an EHM approach, and we shall leave it as understood that optimal estimators $\widehat{\theta}$ are chosen, to the extent possible; optimal estimation is treated, for example, in great detail in Lehmann (1983).

In what follows, we shall illustrate how to derive optimal predictors of the random quantity Y from the data Z. For the purpose of this presentation, θ will be assumed fixed but unknown (EHM); however, we note that the approach is extendable to the case where both Y and θ are to be predicted (BHM). In the following paragraphs, it is understood that all distributions are conditional on θ.

Let $L(\widehat{Y}, Y)$ denote a loss function that quantifies the consequence of using the predictor $\widehat{Y} = a(Z)$ to make inference on the true value Y. The *optimal predictor* (e.g., De Groot, 1970) is the $a^*(Z)$ that minimizes $E(L(a(Z), Y))$ with respect to $a(\cdot)$. Consequently, if the loss function L changes, the optimal predictor $a^*(Z)$ changes.

We note that the problem of optimal prediction could be reformulated in terms of a utility function and its maximization with respect to $a(\cdot)$. Loss can

be thought of as a negative utility, and hence they give equivalent predictors. It can be shown (e.g., Ferguson, 1967, Chapter 2) that $a^*(\cdot) \in \mathcal{A}$ is the predictor that minimizes the posterior expected loss, where $\mathcal{A}$ is the class of possible predictors. That is, $a^*(\cdot)$ satisfies

$$E\{L(a^*(Z), Y)|Z\} \leq E\{L(a(Z), Y)|Z\}, \qquad \text{for all } a \in \mathcal{A}. \tag{2.14}$$

The derivations below assume that the *predictand* Y and the *predictor* $\widehat{Y}$ are random variables; see, for example, Gotway and Cressie (1993) for derivations when Y and $\widehat{Y}$ are random vectors. We shall consider two examples based on two important loss functions, *squared-error loss* and *0–1 loss*.

Assume squared-error loss:

$$L(\widehat{Y}, Y) = (Y - \widehat{Y})^2,$$

and write $\widehat{Y} = a(Z)$. Then to find the optimal predictor $a^*(\cdot)$,

$$E\{(Y - a(Z))^2|Z\}$$

is minimized with respect to $a(\cdot)$.

In the calculation below, we demonstrate that $a^*(Z) \equiv E(Y|Z)$ is the optimal predictor in an HM. For any $a(\cdot)$, the posterior expected loss is

$$\begin{aligned} E\{(Y - a(Z))^2|Z\} = E\{(Y - a^*(Z))^2|Z\} + E\{(a^*(Z) - a(Z))^2|Z\} \\ + 2E\{(a^*(Z) - a(Z))(Y - a^*(Z))|Z\}. \end{aligned}$$

Now the last term on the right-hand side is zero, since $E\{(Y - a^*(Z))|Z\} = E(Y|Z) - a^*(Z) = 0$. Hence,

$$E\{(Y - a(Z))^2|Z\} \geq E\{(Y - a^*(Z))^2|Z\},$$

which implies that

$$a^*(Z) \equiv E(Y|Z) = \textit{mean of } [Y|Z],$$

is the optimal predictor of Z. We see in Section 4.1 that this optimal predictor is the basis of *kriging*, and we see in Section 8.2 that it is the basis of *smoothing*, *filtering*, and *forecasting*.

One way to quantify the uncertainty in any predictor $a(\cdot)$ is through the expected loss, $E\{L(a(Z), Y)\}$, where here the expectation, $E\{\cdot\}$, is taken over the joint distribution of Y and Z. In the case of squared-error loss, this yields the mean squared prediction error (MSPE),

$$E\{(Y - a(Z))^2\}.$$

For the optimal predictor we have $a^*(Z) = E(Y|Z)$, and hence the *minimized MSPE* is

$$E\{(Y - E(Y|Z))^2\} = E\{\text{var}(Y|Z)\}.$$

The MSPE is the basic quantification of uncertainty in kriging (Chapter 4) and in Kalman smoothing/filtering/forecasting (Chapter 7).

In spite of its optimality, $E(Y|Z)$ has a major weakness. Suppose it is known that Y is binary; that is, $Y \in \{0, 1\}$. However, the optimal predictor is *not* binary: $E(Y|Z) \in [0, 1]$. That is, the optimal predictor takes values *between* 0 and 1, which Y can *never* take. This weakness is apparent when Y is an array of pixels from a black-and-white image, and Z is a noisy gray-scale version of Y. The optimal predictor, $E(Y|Z)$, gives a "smoother" image than Z, as is desired, but it is *not* black-and-white, as is the true image Y. One remedy is to change the loss function. The *0–1 loss function* is often used in statistical image processing.

Assume 0–1 loss:

$$L(\widehat{Y}, Y) = \begin{cases} 0 & \text{if } \widehat{Y} = Y \\ 1 & \text{if } \widehat{Y} \neq Y \end{cases}$$
$$= 1 - I(\widehat{Y} = Y),$$

where $\widehat{Y} = a(Z)$ and $I(\cdot)$ is the indicator function. Then, to find the optimal predictor $a^*(\cdot)$, it is equivalent to minimize

$$E\{1 - I(a(z) = Y)|Z\},$$

with respect to $a(\cdot)$. This is easily seen to be equivalent to *maximizing* $\Pr(Y = a(Z)|Z)$. In other words, the optimal predictor $a^*(Z)$ satisfies

$$[Y|Z]|_{Y = a^*(Z)} \geq [Y|Z],$$

where $[Y|Z]$ is the posterior probability density (or mass function). That is, the optimal predictor is

$$a^*(Z) = \textit{mode of } [Y|Z].$$

This is in contrast to squared-error loss, where the optimal predictor is the *mean* of $[Y|Z]$.

Notice that the weakness of the (posterior) mean is avoided with the (posterior) mode, since it *is* one (actually, the most likely) of the possible values of Y. However, the 0–1 loss function is in some sense uncompromising; in statistical image processing, even one pixel that is incorrect causes a loss of 1, and the loss remains 1 no matter how many additional pixels are incorrect. Finally, there is no agreed-upon measure of uncertainty for the mode of the posterior distribution.

2.2.2 Diagnostics

An HM is formulated from the parametric assumptions $[Z|Y,\theta]$ (data model) and $[Y|\theta]$ (process model). The posterior distribution $[Y|Z,\theta]$ (from the EHM) or $[\theta|Z]$ (from the BHM) is the basis of inference on the random quantities of interest. But, does the HM fit properly?

To answer this question in a complete way would go beyond the scope of this chapter, but we shall present several generic diagnostic procedures to check the fit of an HM. As with all diagnoses, they have the *potential* to eliminate bad theories, but it is quite possible that two models (one good, one bad) could give similar diagnostic results. The aphorism "absence of evidence is not always evidence of absence" should be kept in mind.

Two obvious diagnostic procedures are *validation* and *cross-validation*. Validation involves splitting the data into two parts, $Z = (Z_{obs}, Z_{val})$, where $Z_{val} \equiv \{Z_{val,i}\}$ will be held back for validation. Implicitly, the process is also split into $Y = (Y_{obs}, Y_{val})$. Then the HM is fitted from data Z_{obs}, resulting in the posterior distribution, $[Y|Z_{obs}]$. From this, $[Y_{val}|Z_{obs}]$ is obtained by marginalization and, from this, (optimal) predictors $\hat{Y}_{val} \equiv \{\hat{Y}_{val,i}\}$ are produced.

Furthermore, under the (usual) assumption that Z_{obs} and Z_{val} are conditionally independent given Y, we obtain

$$[Z_{val}|Z_{obs}] = \int [Z_{val}|Y_{val}][Y_{val}|Z_{obs}]dY_{val}.$$

Then simulating Y^*_{val} from $[Y_{val}|Z_{obs}]$, followed by Z^*_{val} from $[Z_{val}|Y^*_{val}]$, yields a simulation from $[Z_{val}|Z_{obs}]$. From this, (optimal) predictors $\hat{Z}_{val} \equiv \{\hat{Z}_{val,i}\}$ are produced. The validation procedure involves "comparing" $\hat{Z}_{val}$ to Z_{val} to see how "close" they are. That measure of closeness is often the empirical mean-squared prediction error,

$$\underset{i}{\text{ave}}\{(Z_{val,i} - \hat{Z}_{val,i})^2\}.$$

A less direct, but commonly used, measure is obtained by replacing $\hat{Z}_{val,i}$ above with $\hat{Y}_{val,i}$.

Cross-validation involves writing $Z = (Z_1, \ldots, Z_m)$; using suggestive notation, we write $Y = (Y_1, \ldots, Y_m, Y_{m+1}, \ldots, Y_n)$, which accounts for the $(n - m)$ missing data. For $i = 1, \ldots, m$, Z_i is deleted, and the posterior distribution $[Y|Z^{(-i)}]$ is obtained, where $Z^{(-i)} \equiv (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_m)$. In the same manner as for validation, the (optimal) predictors, $\hat{Z}_i^{(-i)}$ and $\hat{Y}_i^{(-i)}$, of Z_i and Y_i, respectively, are produced. Then $\hat{Z}_i^{(-i)}$, or less directly $\hat{Y}_i^{(-i)}$, is compared to Z_i; $i = 1, \ldots, m$. The way the m comparisons are made, individually or *en masse*, is a choice that determines the type of cross-validation diagnostic procedure; for example, Cressie (1993, pp. 101–104) discusses this (in the spatial context) for kriging.

We have already mentioned in Section 2.1 the problem of model selection and the criteria known as AIC, BIC, and DIC (e.g., Spiegelhalter et al., 2002). Another common approach to model selection in a Bayesian statistical context is based on Bayes factors; for example, see the review by Kass and Raftery (1995). In a broad sense, these represent diagnostic procedures also, as does the less-formal strategy for breaking links in the graphical models presented in Section 2.4. In the rest of this section, a diagnostic procedure is presented that is well suited to the HM approach and to prediction as the form of inference. It is based on the *posterior predictive distribution*, and it is at its best when a single model is being fitted. It is similar to a classical significance-testing approach to testing hypotheses, in the sense that a specific alternative model is not specified.

The goal in diagnostics for model fitting is to determine whether the observed data are representative of the type of data expected under the model. In an HM, this can be assessed as follows (presented here for a BHM): Let Z_{rep} denote an independent replicate of the data with the same (unknown) values Y and θ that produced the data Z. The *posterior predictive distribution* (Rubin, 1984; Gelman, Meng, and Stern, 1996) of Z_{rep} is defined as

$$[Z_{rep}|Z] = \int\int [Z_{rep}|Y, \theta, Z][Y, \theta|Z]\, dY\, d\theta$$

$$= \int\int [Z_{rep}|Y, \theta][Y, \theta|Z]\, dY\, d\theta, \tag{2.15}$$

where the second equality reflects the (conditional) independence of the replicate Z_{rep} produced by the HM. We note that there is another type of diagnostic based on the prior predictive distribution (Box, 1980), $[Z_{rep}] = \int [Z_{rep}|Y, \theta][Y, \theta]\, dY\, d\theta$, but it was developed at a time when simulating from the posterior distribution was usually prohibitive. As we see below, (2.15) involves replicating the entire dataset from a proposed HM and applying the results to rather general discrepancy measures. Other posterior predictive approaches can be found in West (1986) and Gelfand, Dey, and Chang (1992) in settings where, respectively, data are arriving sequentially and cross-validation is being implemented on the existing dataset without the benefit of replicating it.

Gelman, Meng, and Stern (1996) advocate basing inference on replicates from (2.15), as follows: Use the computational devices discussed in Section 2.3 to draw a sample $\{(Y^{(\ell)}, \theta^{(\ell)}): \ell = 1, \ldots, L\}$ from the posterior distribution, $[Y, \theta|Z]$. Then for $\ell = 1, \ldots, L$, $Z_{rep}^{(\ell)}$ is obtained by simulating from the data model, $[Z|Y^{(\ell)}, \theta^{(\ell)}]$. This is also mentioned briefly in Gelfand (1996), although without guidance on how the $\{Z_{rep}^{(\ell)}: \ell = 1, \ldots, L\}$ would be used in a diagnostic procedure.

If the BHM given by the three levels, $[Z|Y, \theta]$, $[Y|\theta]$, and $[\theta]$, is an appropriate model, then the replicates, $\{Z_{rep}^{(1)}, Z_{rep}^{(2)}, \ldots, Z_{rep}^{(L)}\}$, should "look like"

the data Z. It is the formalization of this idea that yields *posterior predictive diagnostics*.

To diagnose whether the model fits, we introduce $T(Z; Y, \theta)$ as a "discrepancy measure" that is intended to capture the goodness-of-fit of the model to the data. For example, T may be an overall measure of fit, or it may be a measure designed to tell whether a particular source of variability is adequately addressed by the model. Importantly, T is not restricted to take the form of a test statistic, since it can depend on (Y, θ) as well as on Z. The fit of the model is diagnosed with respect to T by comparing the posterior distribution of $T(Z; Y, \theta)$ to the posterior predictive reference distribution of $T(Z_{rep}; Y, \theta)$.

The joint posterior distribution of $T(Z_{rep}; Y, \theta)$ and $T(Z; Y, \theta)$ can be studied empirically. Recall that $\{(Y^{(\ell)}, \theta^{(\ell)}): \ell = 1, \ldots, L\}$ are L samples from the posterior distribution $[Y, \theta|Z]$. Then, for example, the L pairs of values $\{T(Z_{rep}^{(\ell)}; Y^{(\ell)}, \theta^{(\ell)}), T(Z; Y^{(\ell)}, \theta^{(\ell)}): \ell = 1, \ldots, L\}$ could be displayed in a scatter plot. If the points on the scatter plot are far removed from the 45-degree line, then a lack of model fit would be diagnosed.

One summary of the joint distribution is the *posterior predictive p-value*,

$$\Pr(T(Z_{rep}; Y, \theta) \geq T(Z; Y, \theta)|Z),$$

where the probability is calculated over $[Z_{rep}, Y, \theta|Z]$, a distribution that is easy to compute empirically from posterior samples, $\{(Y^{(\ell)}, \theta^{(\ell)}): \ell = 1, \ldots, L\}$, whose computation is discussed in Section 2.3. Very small posterior predictive p-values would result in rejection of the current model. More moderate values may cast doubt on the model, but whether the model is rejected or not may depend on the ultimate purpose for which it will be used.

One class of discrepancy measures are omnibus measures of fit. An example from this class is the discrepancy based on the usual chi-squared goodness-of-fit measure,

$$T(Z; Y, \theta) = \sum_{i=1}^{m} \frac{(Z_i - E(Z_i|Y, \theta))^2}{\text{var}(Z_i|Y, \theta)}, \tag{2.16}$$

where here the data Z consist of the m individual observations, $Z_1, \ldots, Z_m$, that make up Z. Notice the similarity to classical goodness-of-fit testing when the statistical model is not hierarchical and is simply $[Z|\theta]$. In that case, there is no "Y" in (2.16), and θ is typically replaced with an estimator $\widehat{\theta}$ (e.g., the maximum likelihood estimator). For these classical test statistics, there are analytical results establishing their asymptotic distributions as central chi-squared under the null hypothesis of the test. In contrast, the posterior predictive distribution provides a suitable reference distribution for any T and any sample size m, since it is based on (simulating from) the posterior distribution in a BHM.

In the EHM, the posterior distribution is $[Y|Z, \theta]$, and the posterior predictive distribution is

$$[Z_{rep}|Z, \theta] = \int [Z_{rep}|Y, \theta][Y|Z, \theta]\, dY.$$

When a posterior predictive diagnostic is computed, an estimator $\widehat{\theta}$ is substituted for θ throughout the implementation. In the context of an EHM, the discrepancy measure focuses on the appropriateness of the data model and the process model.

To enhance the sensitivity of posterior predictive diagnostics and to make posterior predictive p-values uniformly distributed under the assumed model, Bayarri and Berger (2000) developed a related approach based on posterior distributions that condition on only part of the information in the data. However, their approach requires more calculation and can be difficult to apply for complex HMs. Stern and Cressie (2000) combine the ideas of cross-validation and posterior predictive distributions to obtain diagnostics for spatial BHMs on a spatial lattice.

2.3 COMPUTATION OF THE POSTERIOR DISTRIBUTION

Statistics tackles uncertainty directly through the use of (conditional) probability distributions. Let X be a random quantity with *joint* probability distribution $[X]$. For example, suppose $X = (X_1, \ldots, X_n)$ and $\{X_i\}$ are iid $Gau(\mu, \sigma^2)$. Then

$$[X] = \prod_{i=1}^{n} \{(2\pi\sigma^2)^{-1/2} \exp[(-1/2)(x_i - \mu)^2/\sigma^2] : x_i \in \mathbb{R}\}$$

$$= \{(2\pi\sigma^2)^{-n/2} \exp[(-1/2) \sum_{i=1}^{n} (x_i - \mu)^2/\sigma^2] : \mathbf{x} \in \mathbb{R}^n\},$$

where $\mathbf{x} \equiv (x_1, \ldots, x_n)'$. That is, in this case, there is an analytical expression for the joint probability distribution, $[X]$. It would also be possible to calculate all the moments of X_i, $\{E(X_i^k) : k = 1, 2, \ldots\}$, and cross-moments are easy too because of the independence between $\{X_i\}$; for example,

$$E(X_1^2 X_2 X_3^4) = E(X_1^2)E(X_2)E(X_3^4).$$

But even in this ideal case, there is one expression that is not available analytically:

$$\Pr(X_1 \leq x) = \int_{-\infty}^{x} [X_1]\, dX_1, \qquad x \in \mathbb{R}.$$

However, as any student in an introductory Statistics class will verify, there are tables available to look up this probability for any x on their exam.

This uneven availability of analytical expressions for densities, moments, sampling distributions, p-values, posterior distributions, and so on, in the twentieth century, was arguably a factor that held back the application of Statistics to complex scientific problems. However, the computing revolution has led to a statistical-modeling revolution! Consider data $Z \equiv (Z_1, \ldots, Z_m)$. Then the core model of the twentieth century,

$$Z_1, \ldots, Z_m \; iid \; Dist(\theta),$$

where $Dist(\theta)$ is a generic parametric probability distribution with unknown parameters θ, has now been replaced with the core model of the twenty-first century,

$$[Z|Y, \theta],$$

$$[Y|\theta],$$

which is an HM made up of two basic levels of conditional-probability distributions. Recall that $Z = (Z_1, \ldots, Z_m)$, and think of $Y = (Y_1, \ldots, Y_n)$, where generally m and n are different. Even though the posterior distribution,

$$[Y|Z, \theta] = \frac{[Z|Y, \theta][Y, \theta]}{[Z|\theta]},$$

is not generally analytically tractable (due to the denominator, $[Z|\theta]$), there are now ways to simulate from $[Y|Z, \theta]$.

2.3.1 Simulation-Based Inference

To make the exposition simple, we assume that X is a single random variable and that we wish to evaluate the moment:

$$E(g(X)) = \int g(X)[X]\,dX, \tag{2.17}$$

where for mathematical correctness, it must be assumed that the moment exists. An example of (2.17) is the cumulative distribution function (CDF):

$$F(x) = \Pr(X \leq x) = E(I(X \leq x)),$$

where recall that $I(\cdot)$ is the indicator function ($I(A)$ equals 1 if A is true, and it equals 0 if not). The mean of X is the moment $E(X)$, the variance of X is $\text{var}(X) = E(X^2) - (E(X))^2$, and so forth. Indeed, every summary of $[X]$ can be formulated as a moment, $E(g(X))$.

Assume for the moment that $X^{(1)}, \ldots, X^{(L)}$ are simulated according to a Monte Carlo procedure that produces *iid* realizations from $[X]$. By the strong law of large numbers (e.g., Loève, 1977, p. 25), $E(g(X))$ can be approximated by

$$\widehat{E}(g(X)) \equiv (1/L) \sum_{\ell=1}^{L} g(X^{(\ell)}), \tag{2.18}$$

where the approximation improves as L increases. We can even characterize the goodness of the approximation using mathematical statistical results such as Berry–Esseen bounds and the central limit theorem (e.g., Loève, 1977, Chapters V and VI). With cheap computing, the use of (2.18) with large L, to approximate (2.17), can save a lot of agonizing analytical derivations for various $g(\cdot)$. For example, approximations to the examples given above are straightforward to compute.

$$\begin{aligned} &\text{CDF:} && \widehat{F}(x) \equiv \widehat{E}(I(X \le x)), \\ &\text{mean:} && \widehat{E}(X), \\ &\text{variance:} && \widehat{\text{var}}(X) \equiv \widehat{E}(X^2) - (\widehat{E}(X))^2. \end{aligned}$$

In a BHM, obtaining the posterior distribution $[Y, \theta | Z]$ is vital for all (optimal) inference, but it is usually not available in an analytical form. It is fundamentally a joint distribution of (Y, θ), albeit conditional on the data Z, and hence *simulation* from this distribution would allow statistical inference to proceed. Geman and Geman (1984) saw this clearly, where their setting was a spatial HM for image data. Gelfand and Smith (1990) brought the idea to the general statistical community in a landmark paper that was the beginning of the HM's ascendancy.

Consider the spatial context, and suppose X is an image made up of 256×256 pixels. Then $[X]$ is a 65,536-dimensional joint distribution. Now a simulation, $X^{(1)}, \ldots, X^{(L)}$, from $[X]$ may result in, say, $L = 10{,}000$ images. How does this result in more knowledge? The answer is that Statistics is concerned with *optimal* inference. From Section 2.2, the goal is to choose *one* image that minimizes the posterior *expected* loss, say, and to accompany that image with a measure of its uncertainty; see the introduction to Chapter 4. That is, statistical inference is typically focused on summaries and properties of the posterior distribution. For example, squared-error loss leads to the *mean* of the posterior distribution (Section 2.2.1).

Stepping back for a moment, the emphasis in Statistical Science has moved from being able to derive *analytical* and/or asymptotic expressions for $[X]$ and its properties, to being able to *simulate* from $[X]$. Analytical calculations are still important because from them come deep understanding and sharp focus; however, the benefits of simulation are enormous.

2.3.2 Markov Chain Monte Carlo (MCMC)

Simulation from the posterior distribution of an HM can usually be formulated in terms of simulation from an aperiodic and irreducible Markov chain. Under the assumptions of aperiodicity and irreducibility, the Markov chain's *stationary distribution* exists, and the general intent is to make it the much-sought-after posterior distribution. If it is, then after an initial "burn-in" period, the sequence of simulations from the Markov chain are samples from that stationary distribution. *Markov chain Monte Carlo (MCMC)* is the generic algorithm that ensures the Markov chain's stationary distribution *is* the posterior distribution of the associated HM.

There is one result from probability theory that is needed in order to tie down this program we have described, of model building and inference from HMs based on MCMC. Suppose that realizations $X^{(1)}, \ldots, X^{(L)}$ are successive realizations from an aperiodic, irreducible Markov chain, following a "burn-in" period. Hence, $X^{(1)}, \ldots, X^{(L)}$ are identically distributed according to the Markov chain's stationary distribution, but they are *not* independent. Then from the Ergodic Theorem (for which regularity conditions have to be checked; see Loève, 1978, Chapter X), with probability 1, we obtain

$$(1/L)\sum_{\ell=1}^{L} g(X^{(\ell)}) \to E(g(X)),$$

as $L \to \infty$. Hence, (2.18) is still an appropriate approximation of (2.17), and (after "burn-in") the MCMC can be continued until L is very large, sometimes on the order of 10^6. By examining the sample variance of the realizations $\{X^{(\ell)}: \ell = 1, \ldots, L\}$ when L is very large, the influence of the Markov dependence on the "effective" size of L can be determined. If storage of MCMC output is a problem, the chain can be "thinned" to reduce the within-sample autocorrelation. These and many other considerations are discussed in Robert and Casella (2004).

The MCMC algorithm is iterative, but it does not converge in the usual, numerical-analysis sense of converging to a solution. It converges in the statistical sense that each realization $X^{(\ell)}$ is distributed according to the stationary distribution of the Markov chain.

Before we describe a very simple MCMC algorithm for a BHM, we would like to emphasize again that the goal is simulation from a joint distribution. MCMC is one of a number of possibilities that include importance sampling, rejection sampling, Laplace approximation, perfect sampling, slice sampling, and so on (e.g., Robert and Casella, 2004).

The easiest MCMC to describe is the *Gibbs sampler*, described here for the BHM. Notice that in the notation of Section 2.1.1, X becomes (Y, θ_P, θ_D) and $[X]$ becomes the conditional-probability distribution, $[Y, \theta_P, \theta_D | Z]$. To sample from this using the Gibbs sampler, simulate successively from the steps:

$$[Y|\theta_P, \theta_D, Z],$$

$$[\theta_P|Y, \theta_D, Z],$$

$$[\theta_D|Y, \theta_P, Z],$$

and repeat; at each step, the latest values obtained from the previous steps are used in the conditioning arguments. This defines a *Markov chain* whose stationary distribution is the posterior distribution, $[Y, \theta_P, \theta_D|Z]$.

The conditional distributions in the Gibbs sampler above, are commonly referred to as the *full conditional distributions*. When one of these is difficult to simulate from because it can only be calculated up to a normalizing constant, the simulation in that step can be performed using a Metropolis-type simulation (e.g., Tierney, 1994; Robert and Casella, 2004). For example, consider the first step and suppose that $[Y|\theta_P, \theta_D, Z]$ is given by the density:

$$f(\cdot|\theta_P, \theta_D, Z)/\int f(y|\theta_P, \theta_D, Z)\,dy,$$

where f is known analytically, but the integral is not. Let Y_{cur} be the *current* value of Y and suppose that Y_{sim} is a simulated random quantity from a distribution centered at Y_{cur} (and satisfying a symmetry property given, e.g., by Robert and Casella, 2004, p. 271). Define the next value of the Markov chain, Y_{nex}, as follows:

$$Y_{nex} \equiv \begin{cases} Y_{sim} \text{ with probability, } \min\{1, f(Y_{sim})/f(Y_{cur})\}, \\ Y_{cur} \text{ with probability, } 1 - \min\{1, f(Y_{sim})/f(Y_{cur})\}. \end{cases}$$

Then Y_{nex} is the updated value of Y (given θ_P, θ_D, Z) in that step of the Gibbs sampler.

The Metropolis algorithm (and its commonly used variant, the Metropolis–Hastings algorithm) can slow up the MCMC procedure; the acceptance probability for Y_{sim} has to be chosen carefully (e.g., 25–30% acceptance rate), so considerable effort is made to avoid using it in the Gibbs sampler. This has perhaps led to (conditional) distributional choices when building the HM that more suit the computational efficiencies than the scientific mechanisms. Clearly, this is a balancing act, and such choices should not have a major impact on the results (see Section 2.5). MCMC algorithms for spatio-temporal HMs are described in Chapter 8.

We now give a similarly simple example for an EHM. In this case, a computational algorithm (e.g., MCMC, Kalman filter) is built to simulate from the posterior distribution, $[Y|Z, \theta]$. Since θ is fixed but unknown, the "empirical" part of the EHM is to substitute in an estimate $\widehat{\theta}$ for θ to simulate from $[Y|Z, \widehat{\theta}]$. In the following paragraph, we give a simple description of the EM estimate $\widehat{\theta}$.

In the HM context, the EM algorithm defined below provides an estimator of θ that attempts to maximize the likelihood, $[Z|\theta]$. It does so by defining

the *complete likelihood*, $[Z, Y|\theta]$, which is equal to $[Z|Y, \theta][Y|\theta]$, the product of the data model and the process model. Dempster, Laird, and Rubin (1977) presented the algorithm in terms of successive iterations of an E-step and an M-step. The ℓth iteration is given by the following:

E-step: Calculate $E(\ln[Z, Y|\theta]|Z = Z_{obs}, \widehat{\theta}^{(\ell-1)}) \equiv q(\theta|\widehat{\theta}^{(\ell-1)})$.
M-step: Find the θ that maximizes $q(\theta|\widehat{\theta}^{(\ell-1)})$; call this $\widehat{\theta}^{(\ell)}$.

With a starting value $\widehat{\theta}^{(0)}$ and repetition of the EM steps, for $\ell = 1, 2, \ldots$, we choose $\widehat{\theta}$ to be the value of $\widehat{\theta}^{(\ell)}$ that satisfies pre-specified convergence criteria. Section 8.3.1 gives an implementation in the spatio-temporal context.

2.3.3 Summaries of the Posterior Distribution

In the case of a BHM, the final result is a simulation, $\{(Y^{(\ell)}, \theta^{(\ell)}): \ell = 1, \ldots, L\}$, from the posterior distribution $[Y, \theta|Z]$. Hence, the optimal predictor $E(Y|Z)$ can be approximated by $\widehat{E}(Y|Z) \equiv (1/L)\sum_{\ell=1}^{L} Y^{(\ell)}$. A measure of its uncertainty, the posterior variance $\text{var}(Y|Z)$, can be approximated by $\widehat{\text{var}}(Y|Z)$. Another measure of uncertainty is the *Bayesian credible interval*,

$$C_{1-\alpha} \equiv \{y: f(y|Z) \geq k_{1-\alpha}\},$$

where $f(y|Z) \equiv dF(y|Z)/dy$, and $k_{1-\alpha}$ is chosen such that

$$\int_{C_{1-\alpha}} f(y|Z)\, dy = 1 - \alpha, \qquad 0 \leq \alpha < 1.$$

Notice that as $\alpha \to 1$, we obtain the posterior mode. Then $\widehat{C}_{1-\alpha}$, the MCMC approximation to $C_{1-\alpha}$, is obtained by replacing $F(\cdot|Z)$ with $\widehat{F}(\cdot|Z)$ above and defining $\widehat{f}(\cdot|Z)$ as the (smoothed) histogram obtained from the CDF $\widehat{F}(\cdot|Z)$. Similar inferences could be obtained for θ.

Notice the advantage of simulating from the joint posterior distribution, $[Y, \theta|Z]$. If one of the quantities (e.g., θ) is not needed, the simulated $\{Y^{(1)}, \ldots, Y^{(L)}\}$ are, *by definition*, distributed according to the marginalized posterior distribution, $\int [Y, \theta|Z]\, d\theta$. Although this might be a difficult integral to obtain analytically, the "law of the unconscious statistician" simply allows us to approximate any moment, $E(g(Y)|Z)$, with $\widehat{E}(g(Y)|Z)$.

2.3.4 Additional Remarks

There is software available to fit HMs using MCMC, the most notable one being BUGS (Spiegelhalter et al., 2010). For the problems we encounter in our research, we have almost exclusively developed our own algorithms and written our own code. This book's emphasis is on the statistical-modeling aspects, but we would like to say that our lesser emphasis here on computations

does not reflect our regard for its importance. For example, when developing MCMC algorithms to simulate from the posterior distribution, we take great care in developing full conditional distributions, in looking for ways to increase statistical efficiency (e.g., Rao–Blackwellization) and computational efficiency (e.g., block updating), and in diagnosing the convergence of the Markov chain to its stationary distribution. [More details on MCMC algorithms can be found in the excellent books of Gilks, Richardson, and Spiegelhalter (1996), Gelman et al. (2003), and Robert and Casella (2004).]

We close this section by pointing out that when doing Statistics for temporal, spatial, and spatio-temporal data, the dimensions of Z, Y, and θ can grow fast as the HMs become more complex. The problem becomes particularly acute in the spatio-temporal domain, where the high spatial dimension is combined with a desire to filter the spatial fields as more data arrive. This has led to considerable research in the area of sequential Monte Carlo methods, in particular *particle filtering* (e.g., Doucet, de Freitas, and Gordon, 2001; Andrieu, Doucet, and Holenstein, 2010). We discuss this further in Section 8.4.5.

2.4 GRAPHICAL REPRESENTATIONS OF STATISTICAL DEPENDENCIES

Statistical independence is a property of probability distributions. Two random quantities A and B are (statistically) independent if $[A, B] = [A][B]$. In other words, to know the joint distribution, it is enough to know the marginal distributions and to put the joint distribution equal to their product. *Statistical dependence* is the absence of statistical independence. According to this definition, there are myriad ways that two or more random quantities could be (statistically) dependent. A key component of statistical modeling of complex phenomena is to specify the dependence structure. One way to visualize dependence structures is through a graph.

2.4.1 Directed and Undirected Graphs

Graphs are made up of nodes and edges. If two random quantities A and B are statistically dependent, we could equivalently represent this dependence by a node with label "A," a node with label "B," and an edge (or line) connecting the two nodes. If it is thought that the dependence is causal, namely A causes B, it is sensible to write the dependence in the joint distribution $[A, B]$ as in (2.3):

$$[A, B] = [B|A][A],$$

and the edge between A and B becomes a *directed edge* (or arrow) from A to B. (Of course, the decomposition, $[A, B] = [A|B][B]$, is equally true, but from a modeling point of view it is not a meaningful way to write the joint probability.)

An *undirected graph* consists of two or more *nodes* (appropriately labeled) and *edges* (lines) joining the nodes, where a typical node is joined to only some of the other nodes. A *directed graph* also has nodes, and all edges are replaced with *directed edges* (arrows), where the direction of the arrow is chosen based on the idea that something happening at the arrow-head node has been "caused" by something happening at the arrow-tail node. Figure 2.1 shows both an undirected graph (Figure 2.1a) and a directed graph (Figure 2.1b) for the simple case of two nodes.

An HM can be represented as a directed graph. Figure 2.2a shows the initial node θ, with directed edges from it to nodes Y and Z. There is also a directed edge from node Y to node Z. That is, dependencies in an HM are expressed through the conditional-probability distributions, $[Z|Y, \theta]$ and $[Y|\theta]$. If we split θ up into θ_D and θ_P, so that the BHM is

$$[Z|Y, \theta] = [Z|Y, \theta_D],$$

$$[Y|\theta] = [Y|\theta_P],$$

$$[\theta_D, \theta_P] = [\theta_D][\theta_P],$$

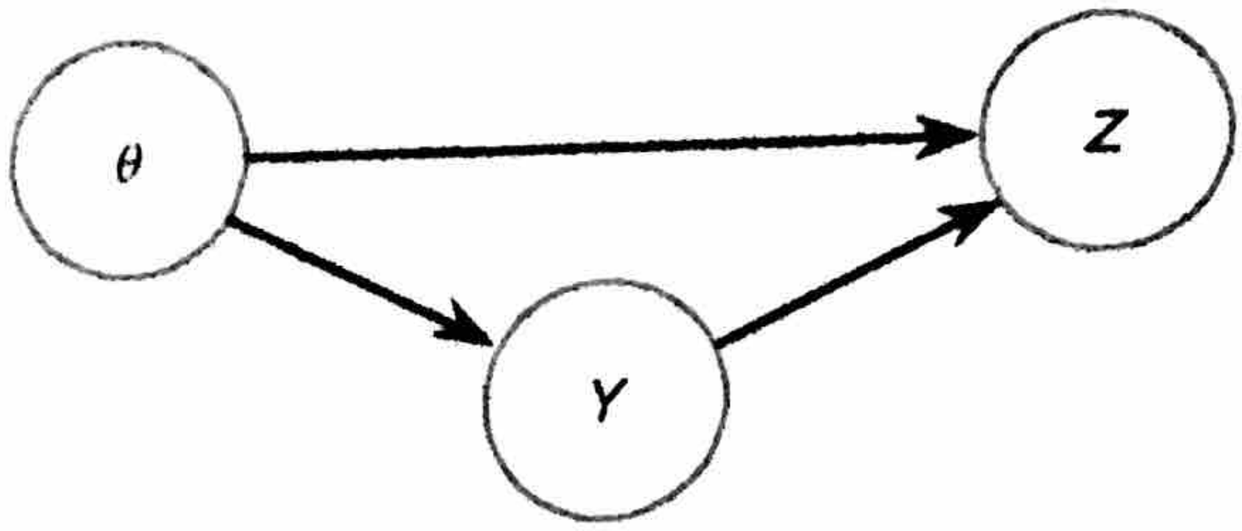
then the independence of θ_D and θ_P yields the directed graph in Figure 2.2b (showing the two "root" nodes θ_D and θ_P with no edge between them). Here, the directed edges from θ_D and Y both go to Z, and the directed edge from θ_P goes to Y. This type of HM is often found in the hierarchical-statistical-modeling literature (e.g., Gelman et al., 2003).

Both undirected and directed graphs can be used to show statistical dependence between random variables, the difference being how that dependence is

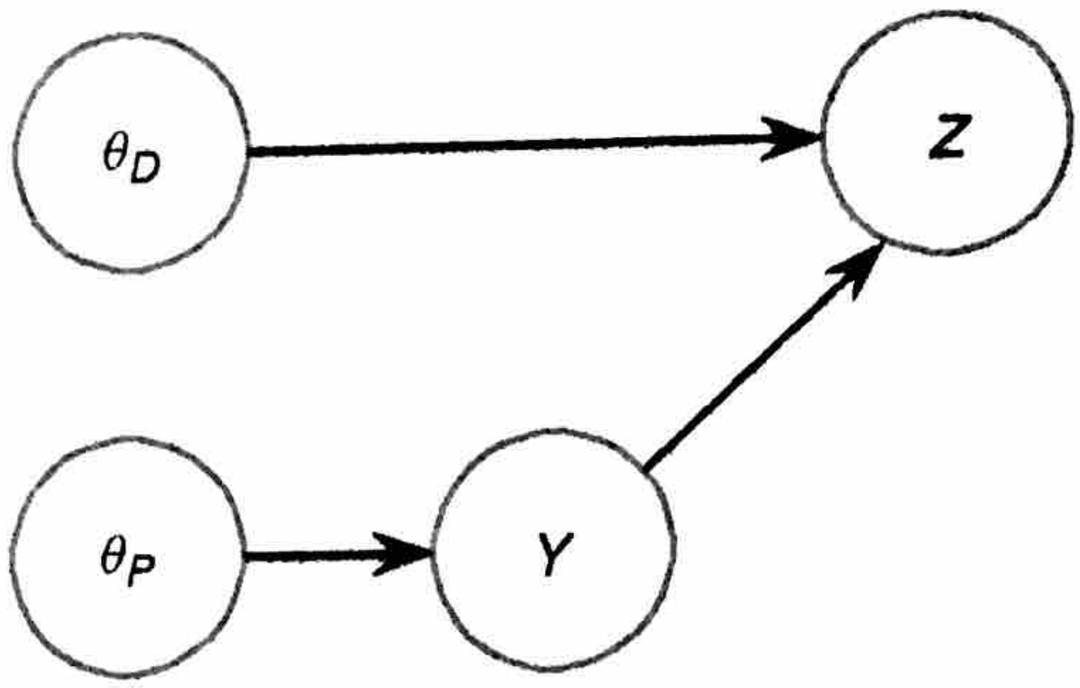
(a)

(b)

Figure 2.1 Graphs with nodes $\{A, B\}$, where A and B are random quantities. (**a**) Undirected graph showing dependence between A and B, illustrated by an edge (*line*); no line between A and B would illustrate independence. (**b**) Directed graph showing that B depends on A causally, as illustrated by the directed edge (*arrow*) from A to B.



(a)



(b)

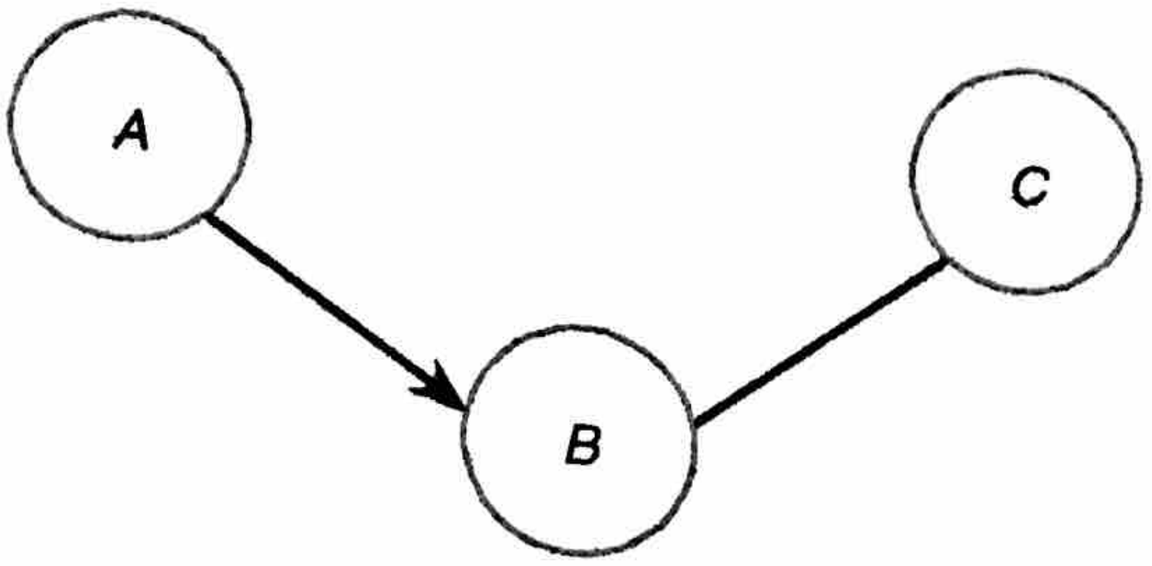
Figure 2.2 Hierarchical model represented as a directed graph. (**a**) Generic model showing that Z is (causally) dependent on Y and θ and that Y is (causally) dependent on θ. (**b**) The node θ is separated into two nodes θ_D and θ_P, and θ_D and θ_P are assumed independent (no edge between θ_D and θ_P).

expressed in the statistical model. Graphs can have a combination of undirected and directed edges, and these have been called *chain graphs* (e.g., Lauritzen, 1996, p. 7). Figure 2.3a shows a three-node chain graph where B depends causally on A, but there is a noncausal type of dependence (undirected edge) between B and C. Notice that a directed graph can be obtained by combining nodes B and C and leaving node A where it is; the result is a two-node directed graph shown in Figure 2.3b, deduced from Figure 2.3a. However, Figure 2.3a cannot be recovered from the graph in Figure 2.3b, which illustrates an important point. When building dependence structures, the statistical modeler should try to use as *many* nodes as there are random quantities of interest, but with as *few* edges as there are dependencies between the quantities.

2.4.2 Conditional Independence

Whether the edges in Figure 2.3a are directed or not is immaterial to the next calculation. Assume that there is no edge between A and C, and consider

$$\begin{aligned}[A, C|B] &= [A, B, C]/[B] \\ &= [A|B, C][B|C][C]/[B] \\ &= [A|B][B|C][C]/[B],\end{aligned}$$



(a)

(b)

Figure 2.3 (**a**) Chain graph. (**b**) Directed graph resulting from (**a**), but showing the dependence structure less precisely.

where we assume the denominator $[B]$ is positive (to keep technicalities to a minimum); the last equality is due to the assumed lack of edge between nodes A and C (see Figure 2.3a). Then from Bayes' Theorem given by (2.5), here applied to B and C,

$$[A, C|B] = [A|B][C|B]. \tag{2.19}$$

The relation (2.19) is remarkable. The lack of an edge between A and C in Figure 2.3a is expressed in (2.19) in terms of *conditional* independence of A and C given B.

How can conditional independencies be read from a (undirected or directed) graph? The answer is surprisingly simple (e.g., Lauritzen, 1996). If the conditioning nodes (here B) are removed from the graph, and that breaks apart the graph into unconnected graphs, then the random quantities associated with the two graphs (here A and C) are conditionally independent, conditional on those at the removed nodes (here B).

Intuitively, the conditional independence relation (2.19) should imply

$$[A|B, C] = [A|B]; \tag{2.20}$$

that is, conditional on B, also knowing C should be immaterial to the probability distribution of A. The proof of (2.20) is straightforward:

$$[A|B, C] = [A, C|B]/[C|B],$$

where we assume that the denominator $[C|B]$ is positive (again to keep technicalities to a minimum). Hence, from (2.19), we obtain

$$[A|B, C] = [A|B].$$

It is also straightforward to show that (2.20) implies (2.19). An example of this in a BHM can be deduced from Figure 2.2b:

$$[Z|Y, \theta_D, \theta_P] = [Z|Y, \theta_D]. \tag{2.21}$$

A more complicated example comes from our wish to diagnose (see Section 2.2) whether or not there is a directed edge from Y_1 to Y_2 in the process $Y = (Y_1, Y_2)$. If there is no edge, then $[Y_2|Y_1, \theta_P] = [Y_2|\theta_P]$.

For simplicity of exposition, we use notation that does not show conditioning on the parameters, but they should be thought of as being present. Associated with Y_1 and Y_2 are data Z_1 and Z_2, respectively. The directed graph is

$$Z_1 \rightarrow Y_1 \rightarrow Y_2 \leftarrow Z_2. \tag{2.22}$$

For example, Craigmile et al. (2009) consider Y_1 to be presence of heavy metals in soils, ambient air, and water supply (regional environment), and consider Y_2 to be presence of the same heavy metals in the household environment. In that study, it was of interest to determine whether data Z_1 on the regional environment could help one "learn" what is happening in the household environment. It is not hard to see that if $[Y_2|Y_1] = [Y_2]$ (i.e., no edge), then

$$[Y_2|Z_2, Z_1] = [Y_2|Z_2].$$

That is, there is no learning about Y_2 from Z_1 beyond what can be learned from Z_2. Hence, a diagnosis that the directed edge from Y_1 to Y_2 is needed would come from seeing changes in the posterior of Y_2, depending on whether data Z_2 or data (Z_1, Z_2) are used. Notice that it would be equally appropriate to look for changes in $[Y_1|Z_1]$ in comparison to $[Y_1|Z_1, Z_2]$.

While graphs are a wonderful way to visualize the presence or absence of conditional dependence, they are silent on the actual form of the conditional probability distribution. Thus, the graphical representation should be viewed as a template upon which conditional-probability models are overlaid.

2.4.3 Graphical Models for Temporal, Spatial, and Spatio-Temporal Processes

Time is one-dimensional and ordered, while space is d-dimensional and has no natural ordering. Therefore, directed graphs are natural for time (Section 3.4.6), and undirected graphs are natural for space (Section 4.2.2). Then, chain graphs (i.e., graphs with directed *and* undirected edges) would provide a natural template for expressing spatio-temporal dependencies; see Section 6.4.2.

2.5 DATA/MODEL/COMPUTING COMPROMISES

The modern statistical approach to modeling is fundamentally about capturing uncertainty with HMs and the conditional-probability distributions upon which they are based. While this provides a framework for inference, actually building and implementing an HM involves a number of compromises. First, there is the notion of the data/model compromise. Imagine trying to gain knowledge about a complicated process. With enough noiseless observations at all appropriate temporal and spatial scales, one could in principle re-create the underlying dependence relationships that are inherent in the process. That is, with a fairly generic model structure, the true underlying scientific mechanism that controls the process could be inferred. However, it is never possible to obtain *noiseless* observations, and a process can never be observed at *all* temporal and spatial scales. Thus, inclusion of both a data model and a process model is a way to formalize the data/model compromise and, depending on the nature of the data that is actually present, the process and parameter models can be chosen to show more or less structure. The HM is very flexible, since it allows inclusion of established scientific properties of the process (which to some extent were themselves inferred from historical data analyses). And, since inference for the HM is based on the posterior distribution, this knowledge can compensate for lack of data. However, there is a flip side to this. It might be the case that an established scientific theory suggests a very complex model with multiple levels in the HM (Wikle and Hooten, 2010). When the available data are not rich enough to learn about the processes and parameters of such a model, then a *practical* lack of identifiability (or, lack of learning) may inhibit successfullly fitting the complex HM.

In addition, the practical matter of implementation leads to the computing/model compromise. While advances in statistical computation have allowed HMs to flourish in the last couple of decades, these computational algorithms can be finicky to "tune" and can be quite time consuming to implement. In many cases, one has to recognize the difference between what one *wants to do* and what one *can do*, when it comes to implementation. This can lead to simpler model structure than what is truly desired. The tension between modeling and computing is healthy if one starts with what one wants to do (answer the Science question) and "dials back" to what one can do. Consequently, as new generations of computing technology emerge, harder problems with fewer compromises can be solved.

This book emphasizes modeling, but our experience with implementation and inference for scientific problems has taught us that the "art of compromise" is ever present. In Chapter 9, examples are given that illustrate different aspects of modeling, implementation, and inference.