Chapter 1

Introduction to Spatial Point Patterns

The goal of this chapter is to provide a brief overview of spatial data leading to the focus on spatial point patterns, the primary subject of this book. This is taken up in Section 1.1. Furthermore, since the inferential framework for the book is Bayesian, we review the basics of Bayesian inference in Section 1.2. In particular, since the Bayesian revolution has been driven by hierarchical modeling, we devote Section 1.3 to elaboration of this rich modeling structure. Finally, since Gaussian processes, along with associated covariance functions underlie much spatial modeling, we briefly visit these ideas in Section 1.4.

1.1. What are spatial data?

There has been a revolution in interest in spatial data over the last twenty years. Increasingly, researchers are recognizing the value of recording spatial information in the data they collect – taking into account the locations of events as well as variables recorded at those locations altogether can provide a richer understanding of a complex process. Specifically, researchers in diverse areas such as climatology, ecology, environmental exposure, public health, and real estate marketing are increasingly faced with the task of analyzing data that are highly multivariate, with many important predictors and response variables, geographically referenced, and often presented as maps. It is anticipated that there is dependence between locations of points as well as between measurements taken at these point, necessitating the introduction of spatial correlation. Often, the data are collected spatially over time, leading to so-called spatio-temporal data, needing temporal correlation, as in longitudinal or other time series structures. This then leads to possible space-time dependence. Does spatial dependence change/evolve over time? If we have a time series at each location, does temporal dependence vary over space?

The focus of this monograph is on formal generative modeling using stochastic specifications that could actually produce the data you have observed and are trying to explain. Generative models are attractive for considering complex processes – for incorporating features to provide behaviors of the process that you seek to emulate. Evidently, for processes yielding geo-referenced data, these ideas certainly apply. Generative modeling enables us to think hierarchically, to specify modeling in stages, and to reflect, in some sense, the actual functioning of the process. It enables us to incorporate all sources of information about the process, incorporating this information at suitable levels of the modeling. Additionally, if such models are fitted within a Bayesian framework, full posterior inference is available with exact inference regarding uncertainty under the models. All of this motivates development of the basics of hierarchical modeling and data analysis for complex spatial (and spatio-temporal) data sets.

As a simple example, in an epidemiological investigation we might wish to analyze lung, breast, colorectal, and cervical cancer rates by county and year in a

particular state. Risk factors which we seek to connect to these rates, such age, race, smoking, mammography, and other screening and staging information, are also available at some, possibly different, spatial scale. As a second example, in a meteorological investigation, we might wish to analyze temperature and precipitation data with hourly or daily measurements at a network of monitoring stations, with a mean surface that reflects elevation, or perhaps a trend in elevation.

However, most appropriate for the framing of this monograph, we may be interested in the point pattern of locations for, say, in an ecological setting, two different species, e.g., juniper trees and pine trees. We have geo-coded locations for each of the trees and a label indicating which species along with environmental features to help explain species distributions, possibly collected over time in order to see change, evolution, diffusion of the patterns.

As a first step, one may be interested in displaying the data collected. However, more value emerges if one has interest in carrying out statistical *inference* tasks, such as modeling of trends and correlation structures, estimation of underlying model parameters, hypothesis testing (or *comparison* of competing models), and prediction at unobserved times or locations. One might seek to employ regression specifications to explain spatial response. Returning to the above, we might conceptualize a process model specification of the form

$$[\text{data}|\text{process},\text{parameters}][\text{process}|\text{parameters}][\text{parameters}].$$

Here, the bracket notation specifies distributions, in particular probability density or mass functions.

The first stage distribution brings in the data revealing how it is connected to the process, or suitable levels of the process. The first stage parameters reflect our ability to propose a *parametric* specification for this relationship but that we do not know the proposed relationship explicitly. The second stage distribution enables us to provide a suitable stochastic description of the process, or at least a description of features of the process, that we seek to learn about but can not observe directly. Again, this description will have a probabilistic form but will involve parameters that are unknown. Finally, the third stage collects all of the unknown parameters and, with a Bayesian lens, requires a prior distribution specification. The form also reveals a generative model: unknowns are chosen at random, then a realization of the process arises at random, and finally, given the unknowns and the process realization, a sample of data is realized.

Do not let this rather innocuous looking hierarchical form underestimate its richness. The data can be of arbitrary type, multivariate, and collected over space and over time. The process specification can range from fairly simple to quite complex, perhaps introducing an uncountable number of unknowns, spatial dependence, and dynamics. We return to this form of specification throughout the monograph, with general elaboration in Section 1.3.

In an expository sense, it is generally asserted that spatial data arises in three different flavors:
(i) point-referenced (or geostatistical) data, where $Y(\mathbf{s})$ is a random vector at a selected location $\mathbf{s} \in \mathbb{R}^r$ and $\mathbf{s}$ varies continuously over D, a fixed subset of $\mathbb{R}^r$, with $r = 2$, perhaps 3 if we add a temporal index to the data. Figure 1.1 shows an example of geostatistical data, capturing counts of hemlocks at 142 selected stands which, at the spatial scale shown, are viewed as points.
(ii) areal data, where D is again a fixed region (of regular or irregular shape), but now partitioned into a finite number of areal units with well-defined boundaries

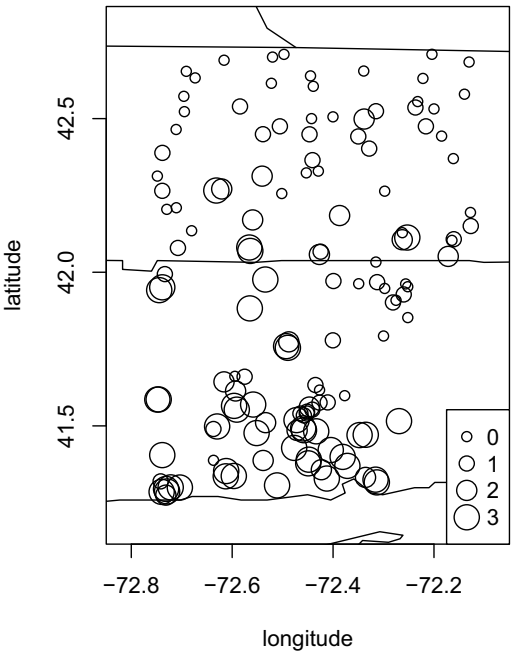

FIG 1.1. *Assessment of hemlock woolly adelgid in 142 eastern hemlock stands across Massachusetts and Connecticut in 1997-1999 [182]. The size of the circle denotes ordinal abundance, where 0 indicates the species was not present and 3 denotes very abundant.*

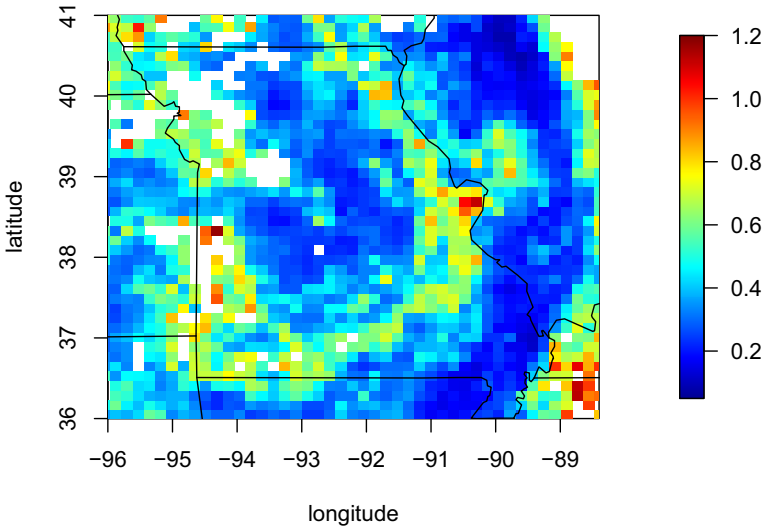

FIG 1.2. *Aerosol optical thickness observed by VIIRS Satellite on July 3, 2013 across the state of Missouri at 12km resolution [180]. White squares denote missing observations.*

and observations are associated with the areal units; we refer to this as discrete spatial data. Figure 1.2 shows an example of areal unit data gathered at grid cell scale from remotely sensed satellite images. The value associated with a grid cell is the aerosol optical thickness associated with the cell.

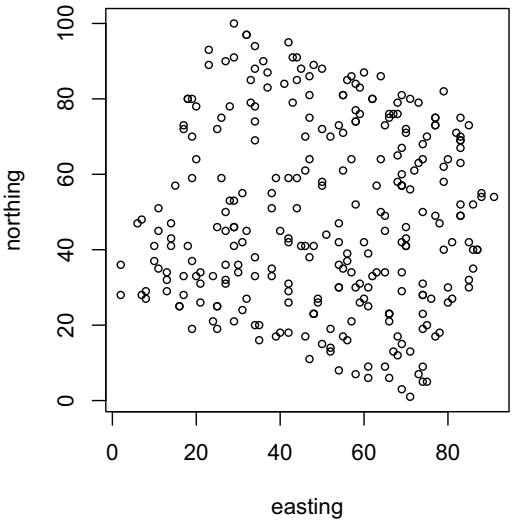

FIG 1.3. *Locations of trees (dbh > 12.7cm) in a forest stand at the Coweeta Hydrologic Laboratory in the southern Appalachians [42, 181].*

(iii) point pattern data, where now the set of locations in D are themselves random; its index set gives the locations of random events that are the spatial point pattern. One can assign $Y(\mathbf{s}) = 1$ for all $\mathbf{s} \in D$ (indicating occurrence of the event), clarifying that only a finite number of events occur in D, and that there are an uncountable number of 0's. Figure 1.3 shows a point pattern of trees, i.e., the random number and locations, in a forest stand.

Our emphasis here is on the third type of data. As an extension, we might assign labels to the points (producing a marked point pattern process). Such marks may be discrete, e.g., species type as above, or continuous, e.g., a size measurement associated with the point. Introduction of marks forces us to think about whether we model the distribution of the marks and then provide a model for the points given the mark or whether we model the pattern of points and then assign a random mark to each point. The distributional specifications are quite different according to the direction of conditioning and a more thorough discussion of this issue is taken up in Section 2.3.3.

1.2. Principles of Bayesian inference

The reader is surely familiar with Bayes' Theorem. In its simplest form

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{1.1}$$

It is natural to ask how this elementary probability law became an inference paradigm (much less a controversial one!)? Suppose we move to random variables in (1.1) obtaining

$$P(X \in A|Y \in B) = \frac{P(Y \in B|X \in A)P(X \in A)}{P(Y \in B)}. \tag{1.2}$$

Then, it is a simple step to move (1.2) to densities which generate probabilities,

$$f(x|y) = \frac{f(y|x)f(x)}{f(y)}. \tag{1.3}$$

Finally, letting $\mathbf{Y}$ denote what you observe and replacing X with $\boldsymbol{\theta}$ denoting what you don't know (didn't or couldn't observe) we arrive at

$$f(\boldsymbol{\theta}|\mathbf{Y}) = \frac{f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{f(\mathbf{Y})} \tag{1.4}$$

or

$$f(\boldsymbol{\theta}|\mathbf{Y}) \propto f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}). \tag{1.5}$$

Now, we have a natural inference paradigm. *You can infer about what you don't know given what you have observed.* This demands direct comparison with the classical inference approach using a sampling distribution (usually approximate) for a statistic T, $T(\mathbf{Y})$ given $\boldsymbol{\theta}$. *You imagine what you might observe given what you don't know.* It seems evidently more sensible (and certainly scientifically more appropriate) to follow the former path which motivates our following of the Bayesian path for this entire monograph.

Returning to (1.4), we really are just thinking about two ways of writing a joint distribution

$$f(\mathbf{Y}, \boldsymbol{\theta}) = f(\mathbf{Y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\mathbf{Y})f(\mathbf{Y}). \tag{1.6}$$

The first conditional/marginal form is generative, comprised of the likelihood (aleatory), $f(\mathbf{Y}|\boldsymbol{\theta})$ and the prior (epistemic), $\pi(\boldsymbol{\theta})$. The second conditional/marginal form is inferential, the posterior, $f(\boldsymbol{\theta}|\mathbf{Y})$ and the marginal distribution of the data, $f(\mathbf{Y})$, for model checking.

The benefits of fitting data using models in a Bayesian framework are clear. One obtains an entire posterior distribution for an unknown rather than perhaps a point estimate and an asymptotic variance, as with usual classical inference (except in cases too simple to be of current interest). Formally, posterior inference for parameters is obtained from $f(\boldsymbol{\theta}|\mathbf{Y})$. Posterior predictive inference for a new observation, Y_0 arises from $f(Y_0|\mathbf{Y}) = \int f(Y_0|\mathbf{Y}, \boldsymbol{\theta})f(\boldsymbol{\theta}|\mathbf{Y})d\boldsymbol{\theta}$.

The primary criticism of the Bayesian paradigm comes from the need to specify priors for all unknowns and the inherent subjectivity that this entails. How can science be carried out objectively if prior specifications are needed? With two different choices of priors yielding two different sets of inference, which one are we to believe, if either? While no rebuttal can be completely convincing, we can argue that there is also subjectivity in the specification of the likelihood, i.e., of the data generating mechanism. Why shouldn't there be comparable concern regarding different choices of the likelihood? Moreover, since no models are correct but some are useful (an adaptation of an old bromide from George Box), parameters are not *real*; they are artifacts of a model. Then, certainly there is no "true" value of a parameter. Perhaps the best we can do is to assume that unknowns are random just like the data are assumed to be random. Finally, we can (and should) do some sensitivity analysis to the prior specification. Often, with a fairly large dataset and fairly weak priors specified at the last hierarchical stage, there may be little sensitivity.

We will not spend much time here on prior specifications. There is a large world regarding improper priors, objective priors, reference priors, etc. [21] which is be-

yond the scope here[1]. Rather, we always will use proper priors since this ensures a proper joint distribution for the data and the unknowns; we don't have to concern ourselves with checking whether the induced posterior distributions are proper. We will make these priors *noninformative*, that is, weak or vague. We will be more specific in the ensuing applications. However, for regression coefficients, we can always use normal priors with large variances where *large* is usually clear in terms of the impact of extreme coefficient values on the resulting regression. Similarly, priors on variances can usually be taken to be inverse gamma distributions, denoted by $IG(a, b)$ where say $a = 1$ (implying no mean or variance) or 2 (implying a mean but no variance). Centering often is not consequential and a crude choice induced from the variation of the data will typically be satisfactory. Priors for range or decay parameters in covariance functions (developed in Section 1.4) can be obtained from the size of the region of interest D. (See [16] for further discussion on priors in spatial settings.)

Returning to the formulation

$$[\text{data}|\text{process,parameters}][\text{process}|\text{parameters}][\text{parameters}],$$

we can write down a simple version as $f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})$. Here, we can think of $\boldsymbol{\theta}$ as capturing the process and the first stage parameters, with $\boldsymbol{\lambda}$ capturing parameters in the process model. Again, since $\boldsymbol{\lambda}$ will not be known, a second stage (hyperprior) distribution $h(\boldsymbol{\lambda})$ will be required. Therefore, we have

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\boldsymbol{\lambda}}{\int f(\mathbf{y}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{\lambda})h(\boldsymbol{\lambda})\, d\boldsymbol{\theta} d\boldsymbol{\lambda}}\,. \tag{1.7}$$

Alternatively, we might replace $\boldsymbol{\lambda}$ in $p(\boldsymbol{\theta}|\mathbf{y}, \boldsymbol{\lambda})$ by an estimate $\hat{\boldsymbol{\lambda}}$; this is called *empirical* Bayes analysis. These days there is little interest in the Bayesian community in empirical Bayes analysis (though see [59]). Pretending that $\boldsymbol{\lambda}$ is "known" rather than averaging over the uncertainty associated with it leads to underestimation of uncertainty across the model. Moreover, using modern Bayesian model fitting techniques [77], there is no need to estimate the hyperparameters.

Returning to (1.4), $p(\boldsymbol{\theta}|\mathbf{y}) \neq \pi(\boldsymbol{\theta})$. This is referred to as Bayesian learning, the change in the posterior distribution compared with the prior. We can extend this to so-called Bayesian updating, often referred to as "crossing bridges as you come to them." In particular, it simplifies sequential data collection. As the simplest version, suppose Y_1 and Y_2 are independent given θ. Then, the joint model is $p(y_2|\theta)p(y_1|\theta)\pi(\theta) \propto p(y_2|\theta)\pi(\theta|y_1)$, i.e., Y_1 updates $\pi(\theta)$ to $\pi(\theta|y_1)$ before Y_2 arrives. In different words, the posterior, $\pi(\theta|y_1)$ becomes the prior to use with the new data Y_2. This updating notion works much more generally – for more than two updates, for updating in blocks, and for both dependent and independent data.

Next, we make some remarks regarding posterior inference, in particular the benefit of having a full distribution to work with. With regard to measures of centrality, you can select whatever seems appropriate - posterior mean, median, or mode. For uncertainty, you can use ranges or variances. For confidence statements, you can use credible intervals. In fact, these are really probability statements about the unknown rather than the frequentist version which arises from pivoting a probability statement from the sampling distribution of some statistic. You can also make more general statements regarding the quantiles for an unknown or the tail behavior of an unknown, e.g., the probability that it exceeds some specified threshold.

[1]Improper priors change the role of the prior from providing a generative model to providing an inference device satisfying some optimality criterion.

Furthermore, hypothesis testing is enriched in the Bayesian setting and may be more suitably referred to as model comparison. The point here is that customary classical hypothesis testing considers a null hypothesis nested within a fuller parameter set and has little to offer for non-nested hypothesis testing. The Bayesian framework enables such model comparison employing Bayes factors relying on "weight of evidence" to distinguish them [23].

More importantly, in a world of complex hierarchical modeling, model comparison is much more demanding than traditional hypothesis testing. Extending this argument below, we argue that model comparison should be done in the data space using predictive distributions and not in the parameter space, since observations are real and parameters are artifacts of the model.

Expanding this a bit, since the hierarchical Bayesian framework is so *liberating* (Section 1.3), we often explore many models. Again, the familiar adage, "All models are wrong but some models are useful" applies, necessitating assessment of adequacy of models and comparison of models. Indeed, with computational tools to fit Bayesian models [77], we face the issue of "overfitting" (more often than underfitting). That is, we often specify models that are richer than the data can support, or are capable of explaining.

We offer some initial words regarding model adequacy and model comparison within the Bayesian framework. Returning to (1.4), the quantity $f(\mathbf{Y})$ is used to assess model adequacy. That is, $f(\mathbf{Y})$ is the marginal density of the data under the specified model and $f(\mathbf{Y}_{obs})$ is the density ordinate at the observed data. In principle, a large value would support adequacy of the model. However, there are serious challenges with using $f(\mathbf{Y}_{obs})$. First, the marginal density is difficult to compute, requiring integration over $\boldsymbol{\theta}$, a high dimensional integral for most hierarchical models of interest. Second, even if we can compute it, as a value for a high dimensional density, say over thousands or more observations (certainly common these days) it is difficult to calibrate it. In different words, model adequacy requires an *absolute criterion*. Does the model meet certain performance standards with regard to such a criterion? We pursue this issue further below.

In the world of complex multi-level modeling, while we may be able to discard some models as inadequate, there will still be many models that are adequate so that we need comparison criteria. Fortunately, these criteria are *relative*; they order models, enabling a choice of best model within the collection under investigation.

Formal Bayesian model comparison can be developed following Bayes rule. Say we have models M_1, M_2, M_k with prior probability of being correct, $p_1, p_2, \ldots, p_k$. Then, with data $\mathbf{Y}$,

$$P(M_j|\mathbf{Y}) = \frac{P(\mathbf{Y}|M_j)p_j}{\sum_{j=1}^{k} P(\mathbf{Y}|M_j)p_j}. \tag{1.8}$$

Calculating $P(\mathbf{Y}|M_j) = \int P(\mathbf{Y}|\boldsymbol{\theta}_j; M_j)\pi(\boldsymbol{\theta}_j|M_j)d\boldsymbol{\theta}_j$ can be challenging due to the integration over $\boldsymbol{\theta}_j$. Moreover, since none of the models are *true*, what do the p_j's mean? Where would they come from? Should they be equally likely? Should we reward smaller, more parsimonious specification? Additionally, where does the set of k models come from? In practice, model development is evolutionary which does not fit into this formal paradigm and, evidently, can contaminate probabilistic assessment of model selection. So, we do not pursue this formal approach further.

Instead, we turn to model selection criteria. There is an enormous literature on model such criteria [23, 40] and there will never be agreement on a "best" model criterion. The choice of criterion greatly depends on the utility for a model

in a particular setting. A further concern is that such criteria reduce a model to a single number. This may be unsatisfying when complex multi-level models are being considered. We might wish to employ multiple criteria (though aggregating them to enable comparison takes us back to our single number challenge).

Arguably, the most important issue is whether we evaluate models in the parameter space or in predictive space. That is, we have a posterior distribution for the parameters which leads to a posterior distribution for the likelihood; we have a posterior predictive distribution for an observation or for a set of observations. Since the parameter space varies with the model and, in fact, according to marginalization, the parameters in the likelihood can vary within a given model, we avoid criteria which operate in the parameter space. We only consider model comparison (and model adequacy) in predictive space. This leads us to the idea of holding out data and doing cross-validation. That is, making the choice of a fitting or training dataset and a test or validation dataset, along with possible replication of this activity, so-called k-fold cross-validation.

Returning to model adequacy/checking, working in predictive space, we immediately come to the decision between prior predictive checks [50] and posterior predictive checks [79]. The approach here is to generate samples under the model and compare them with the observed data in some fashion. If there is adequate agreement, then the model will be declared adequate.

The posterior predictive approach says generate $\mathbf{Y}_{rep}$ from $f(\mathbf{Y}_{rep}|\text{model}, \mathbf{Y}_{obs}) = \int f(\mathbf{Y}_{rep}|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\mathbf{Y}_{obs})d\boldsymbol{\theta}$. The prior predictive approach says generate $\mathbf{Y}_{rep}$ from $f(\mathbf{Y}_{rep}|\text{model}) = \int f(\mathbf{Y}_{rep}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}$. In either case, we introduce a discrepancy function $D(\mathbf{Y}_{rep}, \mathbf{Y}_{obs})$ and then consider its posterior distribution or its prior distribution. With a hierarchical model, we can introduce second stage (latent) variables to consider first stage or second stage discrepancies.

What is the debate? Should we generate samples under the actual model (the prior predictive approach) when you know that the prior is weak and not a realistic distribution to sample parameters from? Should we generate samples with a distribution for $\boldsymbol{\theta}$ that you are more comfortable with (the posterior predictive approach) but is not the generative model for the data? In this regard, posterior predictive model adequacy uses the data twice - once to obtain the posterior and then again in the discrepancy function. As a result posterior predictive checks tend to be less critical of the model. Prior predictive checks have the flavor of Monte Carlo tests, comparing a feature of the observed data with a distribution of that feature under the model. That is, compute a function of the observed data, $T(\mathbf{Y}_{obs})$ and compare it with a set $T(\mathbf{Y}_{rep,b}), b = 1, 2, \ldots, B$ generated under the model.

This relates to the question of whether we should we implement such checks in-sample vs. out-of-sample, relating to the cross-validation idea above. If we can implement posterior predictive checks, employing training data for obtaining the posterior and hold-out data for checking, then we mitigate concerns regarding using the data twice. As we will see in subsequent chapters, with some point pattern models it is not always possible to hold out data. Holding out data will change the nature of interactions between points.

A model adequacy check we primarily use is empirical vs. nominal coverage of predictive intervals. For a variable of interest, say $T(\mathbf{Y})$, first obtain the posterior predictive distribution, $f(T(\mathbf{Y})|\mathbf{Y}_{obs})$ under the model and the data. Then, determine whether $T(\mathbf{Y}_{obs})$ is in or out of the 90% predictive interval of $f(T(\mathbf{Y})|\mathbf{Y}_{obs})$. Do this for many choices of T, compute the empirical coverage (the proportion of times the predictive interval contained the observed value), and compare with the nominal coverage. We note that the posterior predictive distributions will not

usually be available explicitly. However, customary Bayesian model fitting is done with simulation methods, e.g., Markov chain Monte Carlo and Gibbs sampling [77], providing posterior samples of parameters which in turn provide posterior predictive samples. For any given statistic T, these samples provide posterior predictive realizations for that statistic and can be used to create predictive intervals at a specified level of coverage.

Most important is the interpretation that goes with the comparison of empirical vs. nominal coverage. If the empirical coverage is much smaller than the nominal, then the model is not performing well; it is underestimating uncertainty and therefore will not be satisfactory for inference. If the empirical coverage exceeds the nominal coverage, perhaps was 100%, we would also not be happy. Intervals are wider than they need to be, meaning we have too much uncertainty, perhaps are overfitting, and introducing variance inflation.

Returning to model comparison, since we propose to work in predictive space, we walk away from familiar criteria such as AIC, BIC, DIC, and the posterior log likelihood, $\pi(L(\boldsymbol{\theta};\mathbf{Y}_{obs})|\mathbf{Y}_{obs})$ [77]. Instead, we consider criteria in predictive space. A first version employs a penalized posterior predictive loss criterion [73] which attempts to penalize for model complexity. We need a loss function that rewards goodness of fit to the observed data as well as predictive performance for new or replicate data. We adopt a *balanced* loss function. Illustratively, for squared error loss, we obtain $D_k = \frac{k}{k+1}G + P$ where $G = \sum_l (E(Y_{l,new}|\mathbf{y}) - y_{l,obs})^2$ and $P = \sum_l \mathrm{Var}(Y_{l,new}|\mathbf{y})$. Here, G is a goodness of fit term, P is a penalty term, and k provides weighting of the terms. Usually model comparison is not sensitive to the choice of k. Small values of D_k are preferred but the magnitudes of G and P are useful as well. This criterion can be employed both in-sample and out-of-sample. Again, with sampling based model fitting, the posterior predictive mean and variance are readily computed.

Criteria we will consider in the sequel are the following: (i) predictive mean square error (PMSE) - $\sum_{\ell=1}^{L}(E(Y_\ell|\mathbf{Y}_{obs}) - Y_{\ell,obs})^2$, (ii) predictive mean absolute error (PMAE), replacing square with absolute value, (iii) average length of predictive intervals. We might use alternative loss functions for observations that are not continuous, e.g., for counts a common loss function is $(\text{pred} - \text{obs})^2/\text{pred}$.

An attractive criterion which has emerged from the probabilistic forecasting literature and provides a proper scoring rule [85] is the ranked probability score (RPS) for counts, and, with continuous observations, the continuous ranked probability score (CRPS). The intention here is to compare an entire (predictive) distribution with an observation, rather than comparing a feature (e.g., mean) of the predictive distribution to the observation. The idea is that the more concentrated the predictive distribution is around the held out observation the better.

For any continuous distribution/cdf F, CRPS $\equiv \int (F(y) - 1(y > y_{obs}))^2 dy$. For the RPS, we have a discrete distribution and replace the integral with a sum. CRPS is challenging to compute explicitly but a very useful alternative form, under the posterior predictive distribution for Y_ℓ, is CRPS$= E|Y_\ell - Y_{\ell,obs}| - \frac{1}{2}E|Y_\ell - Y_m|$. Averaging is done over different $Y_{\ell,obs}$. Here, cross-validation is appropriate and, in this form, small values are preferred. Posterior predictive samples enable convenient Monte Carlo integration for these expectations.

Finally, we note that for the spatial point pattern modeling that is our focus, we will elaborate specific versions of the foregoing tools for model assessment within the Bayesian framework in Chapter 2.

1.3. Hierarchical modeling

In the 21st century we are experiencing a dramatically changing statistical landscape. We are witnessing remarkable growth in data collection, with datasets now of enormous size, terabytes to petabytes. We are also witnessing a change toward examination of observational data, rather than being restricted to carefully-collected, experimentally designed data. Furthermore, we are studying increasingly complex systems using such data, requiring synthesis of multiple sources of information (empirical, theoretical, physical, etc.), and necessitating the development of multi-level models. The general hierarchical framework which we have alluded to above, [data|process,parameters][process|parameters][parameters], is intended to reflect these dramatic changes, albeit in a simple expression. What it really conveys is the need for stochastic modeling – sophisticated modeling – that can incorporate behaviors we seek to emulate involving uncertainty, nonlinearity, scale, dependence, etc.

The role of the statistician is evolving in this landscape to that of an integral participant in team-based research: a participant in the framing of the questions to be investigated, the determination of data needs to investigate these questions, the development of models to examine these questions, the development of strategies to fit these models, and the analysis and summarization of the resultant inference under these specifications. These are exciting times, offering an exciting new world for modern statistics. The range of applications runs the scientific gamut, e.g., biomedical and health sciences, economics and finance, environment and ecology, engineering and natural science, political and social science.

Again, hierarchical modeling has taken over the landscape in contemporary stochastic modeling. We use this subsection to attempt a partial elaboration of the rich opportunities encompassed in hierarchical modeling. In this regard, though analysis of such modeling can be attempted through non-Bayesian approaches, the Bayesian paradigm (as elaborated in the previous subsection) enables exact inference and proper uncertainty assessment within the given specification.

With the revolution in hierarchical modeling and the objective of fitting within a Bayesian framework has come an enormous revolution in Bayesian computing. Approaches that have emerged over the past thirty years include importance sampling, Markov chain Monte Carlo and Gibbs sampling, along with sequential importance sampling, particle filters and particle learning, and now, the emergence of integrated nested Laplace approximation (INLA), approximate Bayesian computation (ABC), and variational Bayes methods. We do not have the space here to review this large literature. Rather, with a focus on Bayesian inference for spatial point patterns, in Chapter 4, we review the computational strategies appropriate for fitting various models to such data.

1.3.1. What are hierarchical models?

"Hierarchical model" is a broad term that refers to a wide range of model specifications. In particular, here is a partial list which likely includes at least one modeling scenario that the reader will have encountered:

- Multilevel models
- Random effects models
- Random coefficient models
- Variance-component models
- Mixed effects models
- Latent variable models
- Missing data models
- State space models.

The key feature is that hierarchical models are statistical models providing a formal framework for analysis with a complexity of structure that matches the system being studied. Four important concepts are associated with such models:
(i) Modeling data with a complex structure - There is a large range of *nested* structures that can be handled routinely using hierarchical models, e.g. pupils nested classes, classes nested in schools or houses nested in neighborhoods, neighborhoods nested within metropolitan areas.
(ii) Modeling heterogeneity - Standard regression hierarchical models allow for heterogeneity of variance as well as modeling of variances at multiple levels, e.g., variability in house prices can vary from neighborhood to neighborhood as well as from house to house within a neighborhood.
(iii) Modeling dependent data - Capturing potentially complex dependencies in the outcome over time, over space, over context, e.g. house prices within a neighborhood tend to be similar, environmental exposures tend to be similar at locations near each other and at times close to each other.
(iv) Modeling contextuality - Introducing micro and macro relations, e.g., individual house prices depend on individual property characteristics as well as on neighborhood characteristics. Regression coefficients can be attached at appropriate scales.

While there is by now a rich array of techniques for fitting Bayesian models, the simulation based approach incorporated into Gibbs sampling and MCMC is ideally suited to fitting such models. More precisely, the overarching *building block* is the notion of latent variables, e.g., random effects, missing data, labels. These variables introduce unobservable process features which will be of interest, as well as facilitate model fitting. That is, for fitting, Gibbs sampling loops become natural - update other parameters given the values of the latent variables and then update the latent variables given the values of the other parameters.

To illustrate the structure, we consider the standard hierarchical linear model:

$$\text{First stage} : \mathbf{Y}|\mathbf{X}, \boldsymbol{\beta} \sim N(\mathbf{X}\boldsymbol{\beta}, \Sigma_{\mathbf{Y}})$$

$$\text{Second stage} : \boldsymbol{\beta}|\mathbf{Z}, \boldsymbol{\alpha} \sim N(\mathbf{Z}\boldsymbol{\alpha}, \Sigma_{\boldsymbol{\beta}})$$

$$\text{Third stage} : \boldsymbol{\alpha} \sim N(\boldsymbol{\alpha}_0, \Sigma_{\boldsymbol{\alpha}}).$$

We typically specify vague Gaussian priors for the regression parameters and inverse Gamma or inverse Wishart priors for the variances or covariance matrices. Fitting within the Bayesian framework becomes routine. Due to the conjugacy, every updating step within a Gibbs sampler is a standard distribution - normal, inverse Gamma, or inverse Wishart. We have what might be referred to as a *vanilla* Gibbs sampler

If we replace the Gaussian first stage model with an exponential family distribution model (adopting a suitable link function), we have a hierarchical generalized linear model. Now, conjugacy between the first and second stages is lost. Metropolis-Hastings updating would likely be used with adaptive tuning of the acceptance rates [77].

1.3.2. A collection of examples

Conditionally independent hierarchical models

Early hierarchical modeling work began with conditionally independent hierarchical models (CIHMs) at Carnegie Mellon University in the 1980s using Laplace approximation [199]. Being implemented through Gaussian approximations, it preceded

the use of Gibbs sampling and MCMC as Bayesian computation tools. Notably, it is now enjoying a revival through the recent development of integrated nested Laplace approximation (INLA) [176].

The CIHM takes the basic form $\Pi_i[\mathbf{Y}_i|\boldsymbol{\theta}_i]\Pi_i[\boldsymbol{\theta}_i|\boldsymbol{\eta}][\boldsymbol{\eta}]$. Exchangeable $\boldsymbol{\theta}_i$ are assumed. If $\boldsymbol{\eta}$ is fixed, we are fitting separate models for each i. With unknown $\boldsymbol{\eta}$, we add a hyperprior for $\boldsymbol{\eta}$. Now, the models across i are linked; now $\boldsymbol{\eta}$ is informed by each i. More importantly, we now bring in shrinkage or borrowing strength with regard to inference across the $\boldsymbol{\theta}_i$'s. This is an attractive feature both for smoothing and for expected loss under various loss functions. Further development of the CIHM included the hierarchical GLM as well as natural extension to autoregressive moving average (ARMA) time series models. Illustratively, we might have

$$Y_{it} = \mathbf{x}_{it}^T\boldsymbol{\beta}_i + \sum_j \phi_{ij}Y_{i,t-j} + \sum_k \theta_{ik}\epsilon_{i,t-k} + \epsilon_{it}. \tag{1.9}$$

At the second stage, we might adopt exchangeable $\boldsymbol{\beta}_i$, $\boldsymbol{\phi}_i$, $\boldsymbol{\theta}_i$. Then, we could add a vague Gaussian prior on $\boldsymbol{\beta}$, with constrained priors on the ϕ's and θ's (to ensure stationarity), and finally, $\epsilon_{it} \sim N(0, \sigma^2)$.

Random effects models

Random effects are introduced under both Bayesian and frequentist modeling, customarily as normal random variables with an associated variance which is referred to as a variance component. These effects can be at different levels of the modeling but usually assumed exchangeable, in fact, independent and identically distributed (i.i.d.). A typical linear version with i.i.d. effects takes the form

$$Y_{ij} = \mathbf{x}_{ij}^T\boldsymbol{\beta} + \phi_i + \epsilon_{ij}. \tag{1.10}$$

At the second stage, $\boldsymbol{\beta}$ has a Gaussian prior while the ϕ_i are i.i.d. $\sim N(0, \sigma_\phi^2)$. The ϵ_{ij} are i.i.d. $\sim N(0, \sigma_\epsilon^2)$. The variance components, σ_ϕ^2 and σ_ϵ^2, become the third stage hyperparameters. Care is required with the prior specifications for σ_ϕ^2, σ_ϵ^2. It is important to avoid an $IG(\epsilon, \epsilon)$ specification where ϵ is very small since such priors are nearly improper and produce posteriors that are nearly improper, resulting in poorly behaved MCMC model fitting. A protective recommendation is an $IG(1, b)$ or $IG(2, b)$. Nowadays, we are seeing random effects with structured dependence in, e.g., dynamic, spatial and spatio-temporal models [16, 213]. In the context of point pattern models, we introduce these random effects in log Gaussian Cox processes (Section 2.3).

Missing data and imputation

In collecting information on, e.g., individuals, we typically gather vectors of data. Often, one or more of the components is missing. Similarly, when we collect data from monitoring stations often observations are missing. We don't want to be restricted to analyzing only the complete data cases; we don't want to discard the information for the partially observed individuals. In order to use the individuals with missing data, we must *complete* them by doing so-called imputation [61]. There are many ad hoc imputation techniques for filling in missing data. However, we would prefer model-based imputation, i.e., filling in the missingness under the proposed model. Such imputation recognizes that the missing data must be treated

as random variables under the model specification. Importantly, the uncertainty in the imputation propagates to uncertainty in subsequent inference about other unknowns in the model. Fully model-based imputation in the Bayesian setting results in latent variables and Gibbs looping. That is, we iterate between updating the missing data given the model parameters and then update the model parameters given the full data (imputed missing data as well as the observed data). In the Bayesian setting, we extend the Expectation-Maximization (EM) algorithm [49] to provide full posterior inference.

As a simple example, suppose $\mathbf{Y}_i \sim N(\boldsymbol{\mu}_i, \Sigma)$ (the components of $\boldsymbol{\mu}_i$ may have regression forms). Assume some components of some of the $\mathbf{Y}_i$'s are missing. Again, we can use Gibbs sampling to perform the imputation: update the missing data given values for parameters and then update the parameters given values for the missing data along with the observed data. Another standard example considers missing categorical counts (say, from aggregation) within a multinomial model where the multinomial cell probabilities might be modeled using some version of a multivariate logit model. Here, we need to impute/sample missing cell counts under a multinomial model with sum constraints, the constraints arising from the observed aggregated counts over the missing cell counts. After imputation of the cell counts we have a complete set of counts, then we sample all of the parameters in the multinomial model.

Latent variables

Again, latent variables are at the heart of most hierarchical modeling. We can envision latent variables beyond random effects or missing data. A customary version is a hierarchical specification of the form $[\mathbf{Y}|\mathbf{Z}][\mathbf{Z}|\boldsymbol{\theta}][\boldsymbol{\theta}]$. Here, Y's are observed, Z's are latent and the "regression" modeling is moved to the second stage.

As an elementary example, suppose $Y_i \sim \text{Bernoulli}(p(\mathbf{x}_i))$. Let $\Phi^{-1}(p(\mathbf{x}_i)) = \mathbf{x}_i^T\boldsymbol{\beta}$ with a prior on $\boldsymbol{\beta}$. It is awkward to sample $\boldsymbol{\beta}$ using the likelihood in this form. So, following ideas in [4], instead, we introduce $Z_i \sim N(\mathbf{x}_i^T\boldsymbol{\beta}, 1)$. Immediately, $P(Y_i = 1) = \Phi(\mathbf{x}_i^T\boldsymbol{\beta}) = 1 - \Phi(-\mathbf{x}_i^T\boldsymbol{\beta}) = P(Z_i \geq 0)$. Now, we have a routine Gibbs sampler: update the Z's given $\boldsymbol{\beta}, \mathbf{y}$ (sampling from a truncated normal). Then, update $\boldsymbol{\beta}$ given the Z's and $\mathbf{y}$ (the usual conjugate normal updating). This example can be elaborated to include ordinal categorical data where random cut points on $\mathbb{R}^1$ are used to define the categories and latent Gaussian variables are converted to ordinal categorical observations using these cut points.

Errors in variables models

Errors in variables models offer another latent variables setting. In this context, the usual objective is to learn about the relationship between say response Y and predictor X. Unfortunately, X is not or can not be observed. Rather, we observe W instead of X. W may be a version of X, subject to measurement error, i.e., W may be X_{obs} while X may be X_{true}. Alternatively, W may be a variable (variables) that play the role of a surrogate for X.

Conceptually, we may propose a model for $W|X$, referred to as a measurement error model or a model for $X|W$ referred to as a Berkson model [35, 72]. In fact, we could imagine a further errors in variables component. Perhaps we only observe Z, a surrogate for Y. Altogether, we obtain a hierarchical model with latent X's,

possibly Y's. For the measurement error case we have

$$\Pi_i[Z_i|Y_i,\gamma][Y_i|X_i,\beta][W_i|X_i,\delta][X_i|\alpha] \tag{1.11}$$

while for the Berkson case we have

$$\Pi_i[Z_i|Y_i,\gamma][Y_i|X_i,\beta][X_i|W_i,\delta]. \tag{1.12}$$

Note that, for the Berkson case we do not have to model the W_i's. Usually, we have some *validation* data to help to inform about the components of the specification.

Perhaps most remarkable is that, with a full Bayesian specification, we can learn about the relationship between Y and X without ever observing X (and, possibly, without observing Y as well). Though there may be high uncertainty in our ability to learn about this relationship, it does reveal the inferential power of hierarchical specifications.

Mixture models

Mixture models are now widely used due to their flexibility for distributional shapes and their representation of a population in terms of unidentified groups. What we envision here is a setting where we anticipate latent groups within the population we are sampling from but we do not know the group membership for the observations. So, this is different from sampling a population and labeling the individuals by say sex, or race, or ethnicity.

There is a rich literature on mixture models [e.g., 136, 200], parametric and nonparametric, incorporating discrete (finite, countable) or continuous mixing. Here, we consider the most basic finite mixture version

$$\mathbf{Y} \sim \sum_{l=1}^{L} p_l f_l(\mathbf{Y}|\boldsymbol{\theta}_l). \tag{1.13}$$

where the p_l are non-negative and sum to 1 and the f_l are a collection of parametric density functions over the same domain. Often, the f_l are normal densities, whence, we have a normal mixture model.

If L is specified and we observe vectors (perhaps scalars) $\mathbf{Y}_i, i = 1, 2, \ldots, n$, then we envision a latent *label*, L_i, for each $\mathbf{Y}_i$. That is, if $L_i = l$, then $\mathbf{Y}_i \sim f_l(\mathbf{Y}|\boldsymbol{\theta}_l)$. With the labeling variables, the hierarchical model becomes

$$\Pi_i[\mathbf{Y}_i|L_i,\boldsymbol{\theta}][\Pi_i[L_i|\{p_l\}][\boldsymbol{\theta}][\{p_l\}]. \tag{1.14}$$

The prior for p_l's might make them equally likely. Again, Gibbs sampling is routine with obvious Gibbs looping. We update $\boldsymbol{\theta}, \{p_l\}$ given the L's and the data. Then, we update the L_i's given $\boldsymbol{\theta}, \{p_l\}$, and the data. We sample each label from an associated L-valued discrete distribution, i.e. a multinomial trial. If L is unknown we will need to add a prior specification for it. Now, model dimension changes with L. Model fitting options here are reversible jump MCMC [91] or model choice over a set of L's. It is evident that identifiability of the parameters is a challenge. For example, with $L = 2$ and model $pf_1 + (1-p)f_2$, unless we restrict $p < .5$, we can not identify f_1 and f_2.

Revisiting random effects

Consider the setting of individual level longitudinal data say with interest in growth curves. A customary strategy is to model individual level curves centered around a population level curve. We are interested in the population level curve to see *average* behavior of the process. We are interested in the individual level curves, for example, to prescribe *individual* level treatment.

If Y_{ij} is the jth measurement for the ith individual, let

$$Y_{ij} = g(\mathbf{x}_{ij}, \mathbf{z}_i, \boldsymbol{\beta}_i) + \epsilon_{ij} \tag{1.15}$$

where $\epsilon_{ij} \sim N(0, \sigma_i^2)$. The form for g depends upon the application but it need not be linear.

At second stage, we set $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \eta_i$ where the η_i have mean 0 (or perhaps replace $\boldsymbol{\beta}$ with a regression in the $\mathbf{z}_i$). Then, the $\boldsymbol{\beta}_i$ (or the η_i) are the random effects. They provide the individual curves with $\boldsymbol{\beta}$ providing the global curve. We see that this specification can be viewed as a CIHM. Learning with regard to any individual curve will borrow strength from the information about the other curves.

We now offer an important remark. Again, hierarchical models usually introduce latent variables in addition to parameters. Recalling our general hierarchical specification, at times these will be variables associated with the process, e.g., true environmental exposures. However, often they are introduced either to facilitate computation or explanation. This raises the opportunity to introduce the latent variables at the first stage or at the second stage. At the first stage, they imply that the observations are a function of them; at the second stage, they imply that they are explaining the mean of the function.

To be more explicit, consider the simplest example. Suppose the data, Y_i's are Bernoulli trials and suppose the latent Z_i's are normal variables. In the first case, we set say $Y_i = g(Z_i) = 1(Z_i \geq 0)$. In the second case, we set $E(Y_i) = P(Y_i = 1) = P(Z_i \geq 0)$, a probit model. We return to this example in Section 5.3.

Another example is to handle positive random variables using the Tobit, e.g., $Y_i = \max(0, Z_i)$ vs. $Z_i^* = \max(0, Z_i)$ and $E(Y_i) = Z_i^*$. Other possibilities include Poisson, ordinal categorical data, and compositional data. The point is that neither modeling specification is right or wrong. Rather, it is a modeling decision which requires deciding whether you want to use the latent variables to deterministically yield the data or to have them provide a probability distribution for the data.

We conclude this subsection with some caveats associated with hierarchical modeling. Hierarchical models offer an extremely powerful modeling tool but it is easy to abuse them. First, while laying out the stages sequentially can be useful in terms of process specification, it is usually technically very challenging, and typically analytically intractable, to see the impact of changes in the structural specifications on the resultant posterior distributions. Next, with Gibbs sampling and MCMC as very capable model fitting tools, proposed models often grow very big. We often specify models which are too large for the data to support, meaning we are overfitting the data. This leads to challenges in the model fitting in terms of identifiability of parameters. It results in poorly behaved MCMC fitting due to multiple modes in the posterior space which can be hard to find, difficult to assign the appropriate posterior mass to, and potentially difficult to interpret. Additionally, bigger models usually are built with the introduction of more random effects - at different levels, perhaps with dependence in, e.g., space and/or time. These random effects provide so much model flexibility that they will tend to annihilate the coefficients of the

fixed effects which typically enter the modeling linearly. This results in damaging the explanatory capability of the modeling (though it often improves the predictive capability!). In summary, hierarchical models must be handled with care.

1.4. Gaussian processes

Gaussian processes play a crucial role in spatial modeling. They provide extremely flexible specifications for introducing spatial and spatio-temporal dependence. More precisely, they play a primary role in geostatistical modeling and have a role in spatial point pattern modeling as well. In particular, they supply random effects with spatially structured dependence. These effects enable local adjustment to regression modeling at locations using spatially referenced regressors. They specify a spatial surface as a realization of a stochastic process over a region of interest. Moreover, despite the terminology of "Gaussian" process, they offer what would be more appropriately referred to as a nonparametric model for these random effects in that they deliver an uncountable number of random variables, one at each location in the study domain. In this regard, they differ from spatial surfaces provided by splines [51]. Spline surfaces are often viewed as *nonparametric* specifications but, in fact, they employ a specified number of basis functions and so, only introduce a finite set of coefficients with regard to these functions in order to create the desired spline surface.

Formally, a process $Y(\mathbf{s})$ is said to be Gaussian, i.e., a Gaussian process, a GP, if, for any $n \geq 1$ and any set of sites $\{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, $\mathbf{Y} = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n))$ has a multivariate normal distribution. How do we create these multivariate normal distributions? We specify a mean function $\mu(\mathbf{s})$ and a "valid" covariance function (see below) $C(\mathbf{s}, \mathbf{s}') \equiv \mathtt{cov}(Y(\mathbf{s}), Y(\mathbf{s}'))$. Then, $\mathbf{Y} \sim N(\boldsymbol{\mu}, \Sigma)$ where $\boldsymbol{\mu}_i = \mu(\mathbf{s}_i)$ and $\Sigma_{ij} = C(\mathbf{s}_i, \mathbf{s}_j)$. Avoiding technical details, this recipe for providing finite dimensional distributions satisfies the Kolmogorov consistency conditions [152], ensuring that the stochastic process is determined.

The mean function is usually some form of regression specification although when we employ a GP as a random effects model, we set the mean function be 0 over the entire domain of interest. The covariance function is specified through a few parameters say $\boldsymbol{\theta}$, so we have $\Sigma(\boldsymbol{\theta})$ providing *structured* dependence.

Why do we love GPs? After all, there are other distributional families which, in principle, could be used to provide the required finite dimensional distribution. Restriction to Gaussian processes enables several advantages. Here is a list of these advantages:

(i) Gaussian processes offer convenient specification since the mean function and the covariance function determine all finite dimensional distributions.

(ii) Gaussian processes have convenient distribution theory since joint, marginal, and conditional distributions are all immediately obtained from standard results given the mean and covariance structure.

(iii) With hierarchical modeling, a Gaussian process assumption for spatial random effects at the second stage of the model aligns with the way independent random effects with variance components are customarily introduced in the foregoing linear or generalized linear mixed models.

(iv) Technically, with Gaussian processes and stationary models, strong stationarity, $f(Y(\mathbf{s}_1+\mathbf{h}), Y(\mathbf{s}_2+\mathbf{h}), \ldots, Y(\mathbf{s}_n+\mathbf{h})) = f(Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n))$ for arbitrary n, $\mathbf{h}$, and sites, is equivalent to weak stationarity, $\mathrm{cov}(Y(\mathbf{s}+\mathbf{h}, Y(\mathbf{s})) = C(\mathbf{h})$ for arbitrary $\mathbf{s}$ and $\mathbf{h}$.

TABLE 1.1
Common isotropic covariance functions.

Model	Covariance function, $C(\|\|\mathbf{h}\|\|)$
Spherical	$C(\|\|\mathbf{h}\|\|) = \begin{cases} 0 & \text{if } \|\|\mathbf{h}\|\| \geq 1/\phi \\ \sigma^2\left[1 - \frac{3}{2}\phi\|\|\mathbf{h}\|\| + \frac{1}{2}(\phi\|\|\mathbf{h}\|\|)^3\right] & \text{if } 0 < \|\|\mathbf{h}\|\| \leq 1/\phi \end{cases}$
Exponential	$C(\|\|\mathbf{h}\|\|) = \begin{cases} \sigma^2 \exp(-\phi\|\|\mathbf{h}\|\|) & \text{if } \|\|\mathbf{h}\|\| > 0 \end{cases}$
Powered exponential	$C(\|\|\mathbf{h}\|\|) = \begin{cases} \sigma^2 \exp(-\|\phi\|\|\mathbf{h}\|\|^p) & \text{if } \|\|\mathbf{h}\|\| > 0 \end{cases}$
Gaussian	$C(\|\|\mathbf{h}\|\|) = \begin{cases} \sigma^2 \exp(-\phi^2\|\|\mathbf{h}\|\|^2) & \text{if } \|\|\mathbf{h}\|\| > 0 \end{cases}$
Matérn at $\nu = 3/2$	$C(\|\|\mathbf{h}\|\|) = \begin{cases} \sigma^2 (1 + \phi\|\|\mathbf{h}\|\|) \exp(-\phi\|\|\mathbf{h}\|\|) & \text{if } \|\|\mathbf{h}\|\| > 0 \end{cases}$

(v) It is difficult to criticize a Gaussian assumption. We have $\mathbf{Y} = (Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n))$, a single realization from an n-dimensional distribution. With a sample size of one, can we criticize any multivariate distributional specification?

Strictly speaking this last assertion is not quite true with a Gaussian process model. That is, the joint distribution is a multivariate normal with mean, say, 0, and a covariance matrix that is a parametric function of the parameters in the covariance function. As n grows large, the effective sample size will also grow. By linear transformation, $\Sigma^{-\frac{1}{2}}$ we can obtain a set of uncorrelated variables through which the adequacy of the normal assumption might be studied. However, the difficulty is that we don't know the parametrized matrix, $\Sigma^{-\frac{1}{2}}$. We would have to use the data to estimate it and, with an estimated $\hat{\Sigma}^{-\frac{1}{2}}$, we don't produce uncorrelated variables.

Returning to the concept of a valid covariance or, say up to a scaling, a valid correlation function, the challenge is to provide a function where, for all n and all $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n$, the resulting covariance matrix is positive definite. It will not suffice to write down an arbitrary function on $[-1, 1]$; we need a positive definite function [66]. A covariance function is said to be *stationary* if $\text{cov}(Y(\mathbf{s} + \mathbf{h}), Y(\mathbf{s}) = C(\mathbf{h})$ for all $\mathbf{s}$ and $\mathbf{h}$. A covariance function is said to be *isotropic* if $\text{cov}(Y(\mathbf{s}+\mathbf{h}), Y(\mathbf{s}) = C(||\mathbf{h}||)$, i.e., a function of the length of $\mathbf{h}$, for all $\mathbf{s}$ and $\mathbf{h}$. Isotropy is a very strong assumption, implying that dependence has no directionality. It is almost surely never true in practice; it is hoped that, adjusted for a suitable mean function, it will be adequate as a random effects model.

Isotropic covariance functions are widely used in practice. Some common choices are included in Table 1.1. The exponential function is most frequently used due to its convenient functional form and ease of interpretability. For instance, to define a *range*, we might choose the distance such that $e^{-\phi d} = .05$, i.e., the effective range would become the distance at which dependence is deemed to be negligible, i.e., beyond which correlation is less than .05. It is easy to calculate that, with the exponential correlation function, the range is essentially $3/\phi$, facilitating prior specification for ϕ.

The Matérn correlation function is seeing increasing usage [16]. It takes the form

$$C(t) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}||\mathbf{h}||\phi)^\nu K_\nu(2\sqrt{(}\nu)||\mathbf{h}||\phi) \quad \text{if } ||\mathbf{h}|| > 0$$

where K_ν is the modified Bessel function of order ν (computationally tractable in `C/C++` or `geoR`) It introduces ν, a smoothness parameter, where $\nu = 1/2 \Rightarrow$ exponential; $\nu \to \infty \Rightarrow$ Gaussian; $\nu = 3/2 \Rightarrow$ the convenient closed form above. The smoothness idea, in two-dimensions, asserts that the greatest integer in ν indicates the number of times process realizations will be mean-square differentiable [189]. We have the very powerful idea that the smoothness of a random realization of an

uncountable number of random variables, i.e., of a stochastic process surface over a region, is determined by the covariance function driving that realization.

We conclude this chapter with a few more words on covariance functions. As remarked above, to be a valid covariance function the function must be positive definite. Whether a function is positive definite or not can depend upon dimension. In any event, C is a valid covariance functions if and only if it is the characteristic function of a symmetric about 0 random variable (Bochner's Theorem) [120], i.e., $c(\mathbf{h}) = \int \cos(\mathbf{w}^T\mathbf{h})G(d\mathbf{w})$. Once we think in terms of characteristic functions, we immediately think of Fourier transforms using the 1-to-1 correspondence. In turn, this leads to spectral distributions and spectral densities, working with dependence structure in the spectral domain and frequencies rather than distances. Further development is beyond our scope here but see the book of [189] for a useful development.

Finally, there are simple ways of constructing valid covariance functions from familiar ones, e.g., those above by using properties of characteristic functions. For example, we can multiply valid covariance functions (this corresponds to summing independent random variables), we can mix covariance functions (this corresponds to mixing distributions), and we can convolve covariance functions (if c_1 and c_2 are valid then $c_{12}(\mathbf{s}) = \int c_1(\mathbf{s}-\mathbf{u})c_2(\mathbf{u})d\mathbf{u}$ is valid).