2
Preliminaries

2.1 Introduction

This chapter outlines the basic considerations in thinking about the design and analysis of computer experiments. This section begins by distinguishing *three* types of variables that can affect the output of a computer code $y(\cdot)$, depending on the phenomenon being modeled. Using this categorization, we identify some possible experimental goals.

The first type of variable that we distinguish is a *control variable*. If the output of the computer experiment is some performance measure of a product or process, then the control variables are those variables that can be set by a engineer or scientist to "control" the product or process. Some authors use the terms *engineering variables* or *manufacturing variables* rather than control variables. We use the generic notation $\boldsymbol{x_c}$ to denote control variables. Control variables are present in physical experiments as well as in many computer experiments.

As examples of control variables, we mention the dimensions b and d of the bullet tip prosthesis illustrated in Figure 1.2 (see Section 1.2.2). Another example is given by Box and Jones (1992) in the context of a hypothetical physical experiment to formulate ("design") the recipe for a cake. The goal was to determine the amounts of three baking variables to produce the best tasting cake: *flour*, *shortening*, and *egg*; hence, these are control variables. The physical experiment considered two additional variables that also affect the taste of the final product: the *time* at which the cake is baked and the *oven temperature*. Both of the latter variables are specified in the baking recipe on the cake box. However, not all bakers follow the box directions exactly and even if they attempt to follow them precisely, ovens can have true temperatures that differ from their nominal settings and timers can be systematically off or be unheard when they ring.

The variables, baking time and oven temperature, are examples of *environmental variables*, a second type of variable that can be present in both computer and physical experiments. In general, environmental variables affect the output $y(\cdot)$ but depend on the specific user or on the environment at the time the item is used. Environmental variables are sometimes called *"noise variables."* We use the notation $\boldsymbol{x_e}$ to denote the vector of environmental variables for a given problem. In practice, we typically regard environmental variables as random with a distribution that is known or unknown. To emphasize situations where we regard the environmental variables as random, we use the notation $\boldsymbol{X_e}$. As an example of a computer experiment where environmental variables occur, in the hip prosthesis design problem of Chang, Williams, Notz, Santner and Bartel (1999) described above, both outputs depended on the *magnitude* and *direction* of the force exerted on the head of the prosthesis. These two variables were patient specific and depended on body mass and activity. They were treated as having a given distribution that was characteristic of a given population.

In addition to control and environmental variables, there is a third category of input variable that sometimes occurs. This third type of input variable describes the uncertainty in the mathematical modeling that relates other inputs to output(s). As an example, O'Hagan, Kennedy and Oakley (1999) consider a model for metabolic processing of U^{235} that involves various rate constants for elementary conversion processes that must be known in order to specify the overall metabolic process. In some cases, such elementary rate constants may have values that are unknown or possibly there is a known (subjective) distribution that describes their values. We call these variables *model* variables and denote them by $\boldsymbol{x}_m$. In a classical statistical setting we would call model variables "model parameters" because we use the results of a physical experiment, the ultimate reality, to estimate their values. Some authors call model variables "tuning parameters."

The following section describes several fundamental goals for computer experiments depending on which types of variables are present and the number of responses that the code produces. For example, if the code produces a single real-valued response that depends on control and environmental variables, then we use the notation $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ to emphasize that the propagation of uncertainty in the environmental variables $\boldsymbol{X}_e$ must be accounted for. In some cases there may be multiple computer codes that produce related responses $y_1(\cdot), \ldots, y_B(\cdot)$ which either represent competing responses or correspond to "better" and "worse" approximations to the response. For example, if there are multiple finite element analysis codes based on greater or fewer node/edge combinations to represent the *same* phenomenon, then one might hope to combine the responses to improve prediction. Another alternative is that $y_1(\cdot)$ represents the primary object of interest while $y_2(\cdot), \ldots, y_L(\cdot)$ represent "related information"; for example, this would be the case if the code produced a response *and* vector of first partial derivatives. A third possibility is when the $y_i(\cdot)$ represent competing objectives; in this case, the goal might to optimize one response subject to minimum performance standards on the remaining ones.

Following the description of experimental goals, we summarize the basic issues in modeling computer output. Then we will be prepared to begin Chapter 3 on the first of the two basic issues considered in this book, that of predicting $y(\cdot)$ at (a new) input $\boldsymbol{x}_0$ based on training data $(\boldsymbol{x}_1, y(\boldsymbol{x}_1)), \ldots, (\boldsymbol{x}_n, y(\boldsymbol{x}_n))$. Chapter 5 will address the second issue, the design problem of choosing the input sites at which the computer should be run.

2.2 Defining the Experimental Goal

2.2.1 Introduction

In this section, we initially consider the case of a single real-valued output $y(\cdot)$ that is to be evaluated at input training sites $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$. We let $\widehat{y}(\boldsymbol{x})$ denote a generic predictor of $y(\boldsymbol{x})$ and consider goals for two types of inputs. In the first type of input, referred to as a *mono-input*, all components of $\boldsymbol{x}$ are either control variables *or* environmental variables *or* model variables. In the second type of input, referred to as a *mixed-input*, $\boldsymbol{x}$ contains at least two of the three different types of input variables: control, environmental, and model. Finally, in Subsection 2.2.4, we outline some typical goals when there are several outputs. In all cases there can be both "local" and "global" goals that may be of interest.

2.2.2 Research Goals for Mono-Inputs Codes

First, suppose that $\boldsymbol{x}$ consists exclusively of control variables, i.e., $\boldsymbol{x} = \boldsymbol{x}_c$. In this case one important objective is to estimate $y(\boldsymbol{x})$ "well" for all $\boldsymbol{x}$ in some domain $\mathcal{D}$. There have been several criteria used to measure the quality of the prediction in an "overall" sense. One appealing intuitive basis for judging the predictor $\widehat{y}(\boldsymbol{x})$ is its *integrated squared error*

$$\int_{\mathcal{D}} \left[\widehat{y}(\boldsymbol{x}) - y(\boldsymbol{x})\right]^2 w(\boldsymbol{x}) d\boldsymbol{x}, \tag{2.2.1}$$

where $w(\boldsymbol{x})$ is a weight function that quantifies the importance of each value in $\mathcal{D}$. For example, $w(\boldsymbol{x}) = 1$ weights all parts of $\mathcal{D}$ equally while $w(\boldsymbol{x}) =$

$I_{\mathcal{A}}(\boldsymbol{x})$, the indicator function of the set $\mathcal{A} \subset \mathcal{D}$, ignores the complement of $\mathcal{A}$ and weights all points in $\mathcal{A}$ equally.

Unfortunately, (2.2.1) cannot be calculated because $y(\boldsymbol{x})$ is unknown. However, later in Chapter 6 we will replace $[\widehat{y}(\boldsymbol{x}) - y(\boldsymbol{x})]^2$ by a posterior mean squared value computed under a certain "prior" model for $y(\boldsymbol{x})$ and obtain a quantity that can be computed (see Section 6.2 for methods of designing computer experiments in such settings).

The problem of predicting $y(\cdot)$ well over a region can be thought of as a global objective. In contrast, more local goals focus on finding "interesting" parts of the input space $\mathcal{D}$. An example of such a goal is to identify (any) $\boldsymbol{x}$, where $y(\boldsymbol{x})$ equals some target value. Suppose

$$\mathcal{L}(t_0) = \{\boldsymbol{x} \in \mathcal{D} | y(\boldsymbol{x}) = t_0\}$$

denotes the "level set" of input values where $y(\cdot)$ attains a target value t_0. Then we wish to determine any input $\boldsymbol{x}$ where $y(\cdot)$ attains the target level, i.e., any $\boldsymbol{x} \in \mathcal{L}(t_0)$. Another example of a local goal is to find extreme values of $y(\cdot)$. Suppose

$$\mathcal{M} = \{\boldsymbol{x} \in \mathcal{D} | y(\boldsymbol{x}) \geq y(\boldsymbol{x}^\star) \text{ for all } \boldsymbol{x}^\star \in \mathcal{D}\} \equiv \arg\max y(\cdot)$$

is the set of all arguments that attain the global maximum of $y(\boldsymbol{x})$. Then an analog of the level set problem is to find a set of inputs that attain the overall maximum, i.e., to determine any $\boldsymbol{x} \in \mathcal{M}$. The problem of finding global optima of computer code output has been the subject of much investigation (Mockus, Tiešis and Žilinskas (1978), Bernardo, Buck, Liu, Nazaret, Sacks and Welch (1992), Mockus, Eddy, Mockus, Mockus and Reklaitis (1997), Jones, Schonlau and Welch (1998), Schonlau, Welch and Jones (1998)).

There is a large amount of literature on mono-input problems when $\boldsymbol{x}$ depends only on environmental variables. Perhaps the most frequently occurring application is when the environmental variables are random inputs with a known distribution and the goal is determine how the variability in the inputs is transmitted through the computer code. In this case we write $\boldsymbol{x} = \boldsymbol{X}_e$ using upper case notation to emphasize that the inputs are to be treated as random variables and the goal is that of finding the distribution of $y(\boldsymbol{X}_e)$. This problem is sometimes called uncertainty analysis (Crick, Hofer, Jones and Haywood (1988), Dandekar and Kirkendall (1993), Helton (1993), O'Hagan and Haylock (1997), and O'Hagan et al. (1999) are examples of such papers). Also in this spirit, McKay, Beckman and Conover (1979) introduced the class of Latin hypercube designs for choosing the training sites $\boldsymbol{X}_e$ at which to evaluate the code when the problem is to estimate the *mean* of the $y(\boldsymbol{X}_e)$ distribution, $E\{y(\boldsymbol{X}_e)\}$. The theoretical study of Latin hypercube designs has established a host of asymptotic and empirical properties of estimators based on them (Stein (1987), Owen (1992a), Owen (1994), Loh (1996), Pebesma and Heuvelink (1999)) and enhancements of such designs (Handcock (1991), Tang (1993), Tang (1994), Ye (1998), Butler (2001)).

The third possibility for mono-input is when $y(\cdot)$ depends only on *model* variables, $\boldsymbol{x} = \boldsymbol{x}_m$. Typically in such a case, the computer code is meant to describe the output of a physical experiment but the mathematical modeling of the phenomenon involves *unknown* parameters, often unknown rate or physical constants. In this situation the most frequently discussed objective in the computer experiments literature is that of *calibration.* Calibration is possible when the results of a physical experiment are available whose response is the physical phenomenon that the computer code is meant to model. The goal is to choose the model variables $\boldsymbol{x}_m$ so that the computer output best matches the output from the physical experiment (examples are Cox, Park and Singer (1996), Craig, Goldstein, Rougier and Seheult (2001), Kennedy and O'Hagan (2001), and the references therein).

2.2.3 Research Goals for Mixed-Inputs

Mixed-inputs can arise from any combination of control, environmental, and model variables. We focus on what is arguably the most interesting of these cases, that of $\boldsymbol{x}$ consisting of both control and environmental variables. In the problems described below, the environmental variables will be assumed to have a known distribution, i.e., $\boldsymbol{x} = (\boldsymbol{x}_c, \boldsymbol{X}_e)$ where $\boldsymbol{X}_e$ has a known distribution. There are related problems for other mixed-input cases.

In this case, for each $\boldsymbol{x}_c$, $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ is a random variable with a distribution that is induced by the distribution of $\boldsymbol{X}_e$. The $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ distribution can change as $\boldsymbol{x}_c$ changes. As discussed above for the mono-input case $\boldsymbol{x}$ $= \boldsymbol{X}_e$, attention is typically focused on some specific aspect of this induced distribution. For example, recall the study of Chang et al. (1999) for designing a hip prosthesis that was introduced in Section 2.1. In their situation, $y(\boldsymbol{x}_c, \boldsymbol{x}_e)$ was the maximum strain at the bone-implant interface; it depended on the engineering variables, $\boldsymbol{x}_c$, that specified the geometry of the device, and on the environmental variables $\boldsymbol{x}_e$, consisting of the force applied to the hip joint and the angle at which it is applied. Chang et al. (1999) considered the problem of finding engineering designs $\boldsymbol{x}_c$ that minimized the *mean strain* where the mean is taken with respect to the environmental variables. Of course, this is equivalent to maximizing the negative of the mean strain and for definiteness, we describe all optimization problems below as those of finding maxima of mean functions.

To describe this, and related goals, in a formal fashion, let

$$\mu(\boldsymbol{x}_c) = E\left\{y(\boldsymbol{x}_c, \boldsymbol{X}_e)\right\} \tag{2.2.2}$$

denote the mean of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ with respect to the distribution of $\boldsymbol{X}_e$. Similarly define (implicitly) the upper alpha quantile of the distribution of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$, denoted by $\xi^\alpha = \xi^\alpha(\boldsymbol{x}_c)$, as

$$P\left\{y(\boldsymbol{x}_c, \boldsymbol{X}_e) \geq \xi^\alpha\right\} = \alpha$$

(assuming for simplicity that there is a unique such upper α quantile). For example, the notation $\xi^{.5}(\boldsymbol{x}_c)$ denotes the *median* of the distribution of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$, which is a natural competitor of the mean, $\mu(\boldsymbol{x}_c)$, when $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ has a skewed distribution.

With this setup, it now possible to describe analogs of the three goals considered above for $y(\boldsymbol{x}_c)$ but now defined for $\mu(\boldsymbol{x}_c)$ (or $\xi^{.5}(\boldsymbol{x}_c)$, say, if the distribution of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ is skewed). Let $\widehat{\mu}(\boldsymbol{x}_c)$ denote a generic predictor of $\mu(\boldsymbol{x}_c)$. The analog of estimating $y(\cdot)$ well over its domain is to estimate $\mu(\boldsymbol{x}_c)$ well over the control variable domain in the sense of minimizing

$$\int \left[\mu(\boldsymbol{x}_c) - \widehat{\mu}(\boldsymbol{x}_c)\right]^2 \, w(\boldsymbol{x}_c) d\boldsymbol{x}_c. \tag{2.2.3}$$

To solve this problem, one must not only choose a particular predictor $\widehat{\mu}(\boldsymbol{x}_c)$ of $\mu(\boldsymbol{x}_c)$, but also the set of input training sites $(\boldsymbol{x}_c, \boldsymbol{x}_e)$ on which to base the predictor. As in the case of (2.2.1), the criterion (2.2.3) cannot be computed, but a Bayesian analog that has a computable mean will be introduced in Chapter 6.

The parallel of the problem of finding a control variable that maximizes $y(\boldsymbol{x}_c)$ is that of determining an $\boldsymbol{x}_c$ that maximizes the mean output $\mu(\boldsymbol{x}_c)$, i.e., finding an $\boldsymbol{x}_c^M$ that satisfies

$$\mu(\boldsymbol{x}_c^M) = \max_{\boldsymbol{x}_c} \mu(\boldsymbol{x}_c). \tag{2.2.4}$$

Similarly, a parallel to the problem of finding $\boldsymbol{x}_c$ to attain target $y(\cdot)$ values is straightforward to formulate for $\mu(\boldsymbol{x}_c)$.

Additional challenges occur in those applications when the distribution of $\boldsymbol{X}_e$ is not known precisely. To illustrate the consequences of such a situation, suppose that $\boldsymbol{x}_c^M$ maximizes $E_{G^N}\{y(\boldsymbol{x}_c, \boldsymbol{X}_e)\}$ for a given *nominal* $\boldsymbol{X}_e$ distribution, G^N. Now suppose, instead, that $G \neq G^N$ is the true $\boldsymbol{X}_e$ distribution. If

$$E_G\{y(\boldsymbol{x}_c^M, \boldsymbol{X}_e)\} \ll \max_{\boldsymbol{x}_c} E_G\{y(\boldsymbol{x}_c, \boldsymbol{X}_e)\}, \tag{2.2.5}$$

then $\boldsymbol{x}_c^M$ is substantially inferior to any $\boldsymbol{x}_c^\star$ that achieves the maximum in the right-hand side of (2.2.5). From this perspective, a control variable $\boldsymbol{x}_c$ can be thought of as being "robust" against misspecification of the $\boldsymbol{X}_e$ distribution if $\boldsymbol{x}_c$ comes close to maximizing the mean over the nominal $\boldsymbol{X}_e$ distribution and $\boldsymbol{x}_c$ is never far from achieving the the maximum on the right-hand side of (2.2.5) for alternative $\boldsymbol{X}_e$ distributions, G. There are several formal methods of defining a robust $\boldsymbol{x}_c$ that heuristically embody this idea.

The classical method of defining a robust $\boldsymbol{x}_c$ is by a minimax approach (Huber (1981)). Given a set $\mathcal{G}$ of possible environmental variable distributions (that includes a "central," nominal distribution G^N), let

$$\mu(\boldsymbol{x}_c, G) = E_G\{y(\boldsymbol{x}_c, \boldsymbol{X}_e)\}$$

denote the mean of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ when $\boldsymbol{X}_e$ has distribution $G \in \mathcal{G}$. Then

$$\min_{G \in \mathcal{G}} \mu(\boldsymbol{x}_c, G)$$

is the smallest mean value for $y(\boldsymbol{x}_c, \cdot)$ that is possible when $\boldsymbol{X}_e$ distributions come from $\mathcal{G}$. We say $\boldsymbol{x}_c^{\mathcal{G}}$ is a $\mathcal{G}$-robust design provided

$$\min_{G \in \mathcal{G}} \mu(\boldsymbol{x}_c^{\mathcal{G}}, G) = \max_{\boldsymbol{x}_c} \min_{G \in \mathcal{G}} \mu(\boldsymbol{x}_c, G).$$

Philosophically, $\mathcal{G}$-robust designs can be criticized because they are *pessimistic*; $\boldsymbol{x}_c^{\mathcal{G}}$ maximizes a worst-case scenario for the mean of $y(\boldsymbol{x}_c)$. In addition, one is faced with the challenge of specifying a meaningful $\mathcal{G}$. Finally, there can be substantial computational problems determining $\mathcal{G}$-robust designs.

An alternative definition, Bayesian in spirit, assumes that it is possible to place a distribution $\pi(\cdot)$ on the $G \in \mathcal{G}$ where $\mathcal{G}$ is the known set of environmental distributions. In the most straightforward case, the distributions in $\mathcal{G}$ can be characterized by a finite vector of parameters $\boldsymbol{\theta}$. Suppose that $\pi(\cdot)$ is a prior density over the $\boldsymbol{\theta}$ values. We define $\boldsymbol{x}_c^{\pi}$ to be $\pi(\cdot)$-robust provided

$$\int \mu(\boldsymbol{x}_c^{\pi}, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} = \max_{\boldsymbol{x}_c} \int \mu(\boldsymbol{x}_c, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta}.$$

A critique of $\pi(\cdot)$-robust designs is that, in addition to the difficulty of specifying a meaningful $\mathcal{G}$, one must also determine a prior $\pi(\cdot)$. However, $\pi(\cdot)$-robust designs are typically easier to compute than $\mathcal{G}$-robust designs.

A third, more heuristic definition of a robust $\boldsymbol{x}_c$ requires only a nominal $\boldsymbol{X}_e$ distribution, and neither a class $\mathcal{G}$ of alternative distributions nor a prior $\pi(\cdot)$ need be specified. This last definition is based on the following observation. Suppose that for a given $\boldsymbol{x}_c$, $y(\boldsymbol{x}_c, \boldsymbol{x}_e)$ is (fairly) "flat" in $\boldsymbol{x}_e$; then the mean of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ will "tend" to be independent of the choice of $\boldsymbol{X}_e$ distribution. Assuming that we desire the mean $\mu(\boldsymbol{x}_c)$ of $y(\cdot)$ under the nominal distribution to be large, a robust $\boldsymbol{x}_c$ maximizes $\mu(\boldsymbol{x}_c)$ among those $\boldsymbol{x}_c$ for which $y(\boldsymbol{x}_c, \boldsymbol{x}_e)$ is flat. We call such an $\boldsymbol{x}_c$ a V-robust design where the V is for "variance" since the constraint can be thought of as a variance of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ (with respect to a heuristically chosen $\boldsymbol{X}_e$ distribution). To define this notion formally, suppose that each component of $\boldsymbol{X}_e$ has a bounded support; the $\boldsymbol{X}_e$ has support on a bounded hyper-rectangle, say $\Pi_i[a_i, b_i]$. Let

$$\sigma^2(\boldsymbol{x}_c) = \frac{1}{\Pi_i[a_i, b_i]} \int y^2(\boldsymbol{x}_c, \boldsymbol{x}_e) d\boldsymbol{x}_e - \left(\frac{1}{\Pi_i[a_i, b_i]} \int y(\boldsymbol{x}_c, \boldsymbol{x}_e) d\boldsymbol{x}_e \right)^2$$

be the "variance" of $y(\boldsymbol{x}_c, \boldsymbol{X}_e)$ with respect to a uniform distribution on $\boldsymbol{X}_e$. We define $\boldsymbol{x}_c^V$ to be M-robust provided $\boldsymbol{x}_c^V$ maximizes

$$\mu(\boldsymbol{x}_c)$$
$$\text{subject to}$$
$$\sigma^2(\boldsymbol{x}_c) \leq M.$$

Here M is an absolute bound on the variability of $y(\boldsymbol{x}_c, \cdot)$. An alternative natural constraint is

$$\sigma^2(\boldsymbol{x}_c) \leq \max_{\boldsymbol{x}_c^\star \in \mathcal{D}} \sigma^2(\boldsymbol{x}_c^\star) \times M,$$

where M is now a relative bound that is < 1. Because $M < 1$, this second formulation has the theoretical advantage that the feasible region is always nonempty whereas in the former specification one may *desire* that the variance be no greater than a certain bound M, but there need not exist control variables $\boldsymbol{x}_c$ that achieve this target value. Using the relative constraint has the computational disadvantage that the maximum variance must be determined. Alternatively, and perhaps more in keeping with the quality control concept of having a "target" mean, we define $\boldsymbol{x}_c^V$ to be V-*robust* if it minimizes $\sigma^2(\boldsymbol{x}_c)$ subject to a constraint on $\mu(\boldsymbol{x}_c)$. Lehman, Santner and Notz (2003) discuss the sequential design of computer experiments to find M-robust and V-robust choices of control variables.

2.2.4 Multiple-Output Experiments

To fix ideas, suppose that $y_1(\cdot), \ldots, y_m(\cdot)$ are the computed outputs. There are at least three different settings that lead to such a situation. First, the outputs can represent multiple codes for the same quantity; for example, Kennedy and O'Hagan (2000) study multiple codes that represent coarser and finer finite element descriptions for the same response.

A second setting that leads to multiple outputs is when the $y_i(\cdot)$ are *competing* responses from *different* codes; in prosthesis design we desire to maximize the strain at the prosthesis–bone interface so that bone resorption does not occur and simultaneously minimize (or at least bound) the side to side "toggling" of the implant. The two objectives, maximizing strain and minimizing toggling, represent competing design goals. A third setting that leads to multiple outputs is when a single code produces $y_i(\cdot)$ that are related to one another. As an example, Morris, Mitchell and Ylvisaker (1993) and Mitchell, Morris and Ylvisaker (1994) consider the estimation of $y(\boldsymbol{x})$ for codes that produce $y(\cdot)$ *and* all its first partial derivatives for each input site $\boldsymbol{x}$. Thus we regard $y_1(\boldsymbol{x}) = y(\boldsymbol{x})$, the original output, and $y_2(\boldsymbol{x})$, $\ldots$, $y_m(\boldsymbol{x})$ as the values of the partial derivatives of $y(\boldsymbol{x})$ with respect to each component of $\boldsymbol{x}$. These derivatives provide auxiliary information that permits more precise estimation of $y(\boldsymbol{x})$ than that based on $y(\cdot)$ alone.

The modeling of multiple $y_i(\cdot)$ depends on which scenario above holds, as do the possible scientific or engineering objectives. For example, when $y_2(\boldsymbol{x})$, $\ldots$, $y_m(\boldsymbol{x})$ represent auxiliary information about $y_1(\boldsymbol{x})$, the goal might simply be to use the additional information to better estimate $y_1(\cdot)$. To continue the example introduced in the previous paragraph, Morris et al. (1993) and Mitchell et al. (1994) show how to model the output from codes that produce a response $y(\cdot)$ and the partial derivatives of $y(\cdot)$. They then use these models to derive (empirical) best linear unbiased estimators of $y(\cdot)$ at new sites $\boldsymbol{x}_0$ based on all the responses. See Section 4.2 for a discussion of modeling multiple responses.

Now consider the scenario where $\boldsymbol{x} = \boldsymbol{x}_c$, $y_1(\cdot)$ is the response of primary interest, and $y_2(\cdot), \ldots, y_m(\cdot)$ are competing objectives. Then we can define a feasible region of $\boldsymbol{x}_c$ values by requiring minimal performance standards for $y_2(\boldsymbol{x}_c), \ldots, y_m(\boldsymbol{x}_c)$. Formally, an analog of the problem of minimizing $y(\cdot)$ is

$$\begin{array}{ccc} \text{minimize} & & y_1(\boldsymbol{x}_c) \\ & \text{subject to} & \\ y_2(\boldsymbol{x}_c) & \geq & M_2 \\ & \vdots & \\ y_m(\boldsymbol{x}_c) & \geq & M_m. \end{array}$$

Here M_i is lower bounds on the performance of $y_i(\cdot)$ that is acceptable. If in addition to control variables, x also contains environmental variables, then we can replace each $y_i(\boldsymbol{x}_c)$ above with $\mu_i(\boldsymbol{x}_c) = E\{y_i(\boldsymbol{x}_c, \boldsymbol{X}_e)\}$. In cases where $\boldsymbol{x} = \boldsymbol{x}_e$, a typical objective is to find the joint distribution of $(y_1(\boldsymbol{X}_e), \ldots, y_m(\boldsymbol{X}_e))$ or, even simpler, that of estimating the mean vector $(E\{y_1(\boldsymbol{X}_e)\}, \ldots, E\{y_m(\boldsymbol{X}_e)\})$.

Lastly, if the $y_i(\cdot)$ represent the outputs of *different* codes of varying accuracy for the same response, then a typical goal is to combine information from the various outputs to better estimate the true response. Specification of this goal depends on identifying the "true" response; we postpone a discussion of this idea until we discuss modeling multiple response output in Section 4.2.

2.3 Modeling Output from Computer Experiments

2.3.1 Introduction

This book uses Bayesian methodology to design and analyze computer experiments. Prior information describing the functional relationship of the input $\boldsymbol{x}$ to the (unknown) output $y(\boldsymbol{x})$ is combined with the information in the training data to predict $y(\cdot)$ at new sites and to accomplish the other goals described in Sections 2.2 and 6.3.

Best linear unbiased prediction, a frequentist methodology, has also been used for prediction of real-valued quantities associated with $y(\cdot)$ and for the calculation of their standard errors (see Section 3.2.3). The conceptual problem with this approach is that the source of randomness that is measured by the standard error, for example, is not easily understood and when specified is often not of interest to the user. For example, one source of randomness that leads to interpretable standard errors is the randomness in the predictor that results from use of a stochastic mechanism to choose the locations of the input data (the "design" of the computer experiment). In this case, the standard error of a predictor of $y(\boldsymbol{x}_0)$ is the variation in the predictor due to the randomly selected training data.

We prefer the Bayesian approach to analyze the data from computer experiments and regard the use of a prior distribution for $y(\cdot)$ as clearer in its intent than the frequentist viewpoint, though not simpler to implement. Computer experiments represent a highly nonparametric setting. Eliciting a prior for the output of a black box code is much more difficult than, say, eliciting the prior for the output of a regression. However, this approach is philosophically more satisfying, for example, in its interpretation of the standard errors that will be specified in Section 4.1—they refer to model uncertainty (given the training data). However, the reader should recognize that reasonable users may disagree about the prior information concerning the input-output function that any particular Bayesian predictor makes (and hence the associated standard error of prediction). Oakley (2002) and Reese, Wilson, Hamada, Martz and Ryan (2000) give advice and case-studies about the formation of prior distributions.

In sum, our attitude toward using the Bayesian approach to problems of the design and analysis of computer experiments is not dogmatic. We *do* attempt to control the characteristics of the functions produced by our priors, but *do not* rigidly believe them. Instead, our goal is to choose flexible priors that are capable of producing many shapes for $y(\cdot)$ and then let the Bayesian machinery allow the data to direct the details of the prediction process.

Sections 2.3.2-2.3.4 will introduce Gaussian random functions and provide the reader with an appreciation for the flexibility of this class of priors. Section 2.3.5 will discuss hierarchical priors based on Gaussian random functions as a method of further enhancing this flexibility.

The final general point we wish to make in this introduction is that computer experiments are not alone in their use of Bayesian prediction methodology to analyze high-dimensional, highly correlated data. Many other scientific fields produce such data, albeit with measurement error. The statistical analyses used in geostatistics (Matheron (1963), Journel and Huijbregts (1979)), environmental statistics and disease mapping (Ripley (1981), Cressie (1993)), global optimization (Mockus et al. (1997)), and statistical learning (Hastie, Tibshirani and Friedman (2001)) are based on the Bayesian philosophy. Hence many of the methodologies discussed in their literatures are also relevant here.

In the following we regard $y(\cdot)$ to be a real-valued function with domain $\mathcal{X}$ where $\mathcal{X}$ is a subset of d-dimensional Euclidean space having positive d-

dimensional volume. We adopt the notation $Y(\cdot)$ to distinguish the random function from its realizations $y(\cdot)$ which are functions. Some authors use the terms "stochastic process" or simply "process" rather than random function and we occasionally also use these terms, although "random function" is the most natural terminology when discussing computer experiments.

Conceptually, a random function should be thought of as a mapping from elements of a sample space of outcomes, say Ω, to a given set of functions, just as random variables are mappings from a set Ω of elementary outcomes to the real numbers. It will occasionally add clarity to our discussion to make this explicit by writing $y(\boldsymbol{x}) = Y(\boldsymbol{x}, \omega)$ to be a *particular* function from $\mathcal{X}$ to $\mathrm{I\!R}^1$, where $\omega \in \Omega$ is a specific element in the sample space. Sometimes we refer to $y(\cdot, \omega)$ as a *draw* from the random function $Y(\cdot)$ or as a *sample path* (in $\mathcal{X}$) of the random function. The introduction of the underlying sample space Ω helps clarify ideas when discussing the smoothness properties of functions drawn from $Y(\cdot)$. In particular, we desire sufficient flexibility in our stochastic model so that, ideally, there is an ω for which $y(\boldsymbol{x}) = Y(\boldsymbol{x}, \omega)$ represents the response to our computer experiment.

We will also consider computer experiments that produce multiple outputs; in such situations we let $\boldsymbol{y}(\boldsymbol{x}) = (y_1(\boldsymbol{x}), \ldots, y_B(\boldsymbol{x}))^\top$ denote the vector of outputs. A typical application that produces multiple outputs is that when the computer code determines not only $y(\boldsymbol{x})$ but also each of the partial derivatives of $y(\boldsymbol{x})$. Then $\boldsymbol{y}(\boldsymbol{x}) = (y(\boldsymbol{x}), \partial y(\boldsymbol{x})/\partial x_1, \ldots\ \partial y(\boldsymbol{x})/\partial x_d)$. In the general multiple output case, we view the random mechanism as associating a vector valued function, $\boldsymbol{y}(\boldsymbol{x}) = \boldsymbol{Y}(\boldsymbol{x}, \omega)$, with each elementary outcome $\omega \in \Omega$. Codes that produce multiple outcomes were introduced in Section 2.2, their modeling will be considered in Section 5.2.1, and the application of these models will be provided in Sections 4.2.3 and 6.3.6.

We begin this overview of stochastic models for generating functions $y(\cdot)$ with the following simple example.

Example 2.1 Suppose that we generate $y(x)$ on $[-1, +1]$ by the mechanism

$$Y(x) = b_0 + b_1 x + b_2 x^2, \tag{2.3.1}$$

where b_0, b_1, and b_2 are independent with $b_i \sim N(0, \sigma_i^2)$ for $i = 1, 2, 3$. Functions drawn from $Y(x)$ are simple to visualize. Every realization $y(\cdot)$ is a quadratic equation ($P\{b_2 = 0\} = 0$) that is symmetric about an axis other than the y-axis (symmetry about the y-axis occurs if and only if $b_1 = 0$ and $P\{b_1 = 0\} = 0$). The quadratic is convex with probability 1/2 and it is concave with probability 1/2 (because $P\{b_2 > 0\} = 1/2 = P\{b_2 < 0\}$). Figure 2.1 illustrates ten outcomes from this random function when $\sigma_0^2 = \sigma_1^2 = \sigma_2^2 = 1.0$.

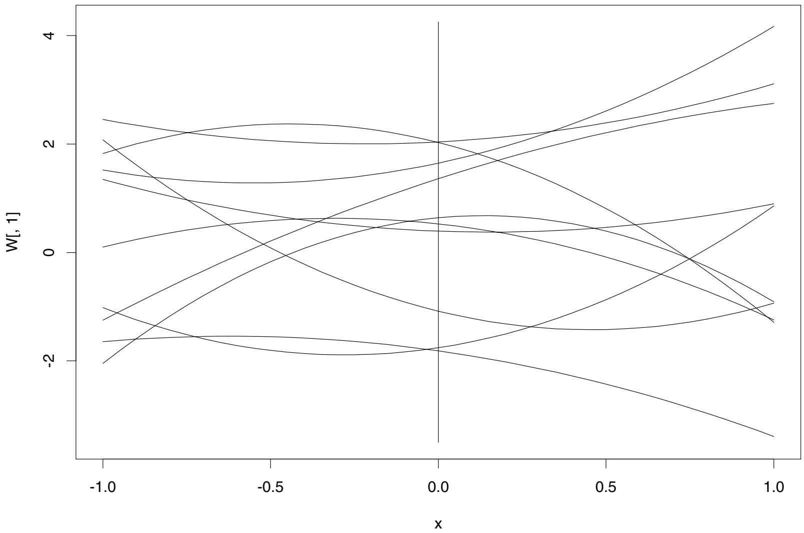

FIGURE 2.1. Ten draws from the random function $Y(x) = b_0 + b_1x + b_2x^2$ on $[-1, +1]$, where b_0, b_1, and b_2 are independent and identically $N(0, 1.0)$ distributed.

For any $x \in [-1, +1]$ the draws from (2.3.1) have mean zero, i.e.,

$$\begin{aligned} E\{Y(x)\} &= E\{b_0 + b_1x + b_2x^2\} \\ &= E\{b_0\} + E\{b_1\} \times x + E\{b_2\} \times x^2 \\ &= 0 + 0 \times x + 0 \times x^2 = 0. \qquad (2.3.2) \end{aligned}$$

Equation (2.3.2) says that for any x, the mean of $Y(x)$ is *zero* over many drawings of the coefficients (b_0, b_1, b_2); this is true because each regression coefficient is independent and centered at the origin so that each regression term is positive and negative with probability 1/2 and thus their sum, $Y(x)$, is also positive and negative with probability 1/2.

For any $x \in [-1, +1]$ the pointwise variance of $Y(x)$ is

$$\begin{aligned} \text{Var}\{Y(x)\} &= E\left\{\left(b_0 + b_1x + b_2x^2\right)\left(b_0 + b_1x + b_2x^2\right)\right\} \\ &= \sigma_0^2 + \sigma_1^2x^2 + \sigma_2^2x^4 \geq 0. \end{aligned}$$

The values of $Y(x_1)$ and $Y(x_2)$ at x_1, $x_2 \in [-1, +1]$ are related, as can be seen from

$$\begin{aligned} \text{Cov}\{Y(x_1), Y(x_2)\} &= E\left\{\left(b_0 + b_1x_1 + b_2x_1^2\right)\left(b_0 + b_1x_2 + b_2x_2^2\right)\right\} \\ &= \sigma_0^2 + \sigma_1^2x_1x_2 + \sigma_2^2x_1^2x_2^2. \qquad (2.3.3) \end{aligned}$$

This covariance can be positive or negative. The sign of the covariance of $Y(x_1)$ and $Y(x_2)$ can intuitively be explained as follows. The covariance formula (2.3.3) is clearly positive for any x_1 and x_2 when both are positive or both are negative. Intuitively this is true because over many drawings of (b_0, b_1, b_2), $Y(x_1)$ and $Y(x_2)$ both tend to be on the same side of the axis of symmetry of the quadratic and thus $Y(x_1)$ and $Y(x_2)$ increase or decrease together. The covariance formula *can* be negative if x_1 and x_2 are on the *opposite* sides of the origin *and* σ_1^2 dominates σ_0^2 and σ_2^2 (algebraically, the middle term in (2.3.3) is negative and can exceed the sum of the other two terms). Intuitively, one circumstance where this occurs is if σ_0^2 is small (meaning the curves tend to go "near" $(0, 0)$), *and* σ_2^2 is small (the curves tend to be linear near the origin), *and* σ_1^2 is large; in this case, the draws fluctuate between those with large positive slopes and those with large negative slopes, implying that $Y(x_1)$ and $Y(x_2)$ tend to have the opposite sign over the draws.

Because linear combinations of a fixed set of independent normal random variables have the multivariate normal distribution, the simple model (2.3.1) for $Y(\cdot)$ satisfies: for each $L > 1$ and any choice of $x_1, \ldots, x_L \in \mathcal{X}$, the vector $(Y(x_1),\ldots, Y(x_L))$ is multivariate normally distributed. (See Appendix B for a review of the multivariate normal distribution.) The $y(\cdot)$ realizations also have several limitations, from the viewpoint of computer experiments. First, the model can *only* produce quadratic draws. Second, the multivariate normal distribution of $(Y(x_1),\ldots, Y(x_L))$ is *degenerate* when $L \geq 4$. In the development below we wish to derive more flexible random functions that retain the computational advantage that $(Y(x_1), \ldots, Y(x_L))$ has the multivariate normal distribution. ■

There are many sources that provide detailed theoretical discussions of random functions, particularly the Gaussian random functions introduced in Sections 2.3.2–2.3.4 (Cramér and Leadbetter (1967), Adler (1981), Adler (1990), and Abrahamsen (1997), for example). It is not our purpose to present a complete account of the theory. Rather, we desire to give an overview of these models, to describe the relationship between the "correlation function" of stationary Gaussian random functions and the smoothness properties of its realizations $y(\boldsymbol{x})$, and to develop intuition about this relationship through a series of examples.

2.3.2 Gaussian Random Function Models

In the computer experiments literature, the most popular models for generating function draws are *Gaussian random functions*, also called the Gaussian stochastic processes. Hence we emphasize these models in this section although, as we will note, some of the concepts that we introduce apply to more general random functions.

Definition Suppose that $\mathcal{X}$ is a fixed subset of $\mathrm{I\!R}^d$ having positive d-dimensional volume. We say that $Y(\boldsymbol{x})$, for $\boldsymbol{x} \in \mathcal{X}$, is a *Gaussian random function* (GRF) provided that for any $L \geq 1$ and any choice of $\boldsymbol{x}_1, \ldots \boldsymbol{x}_L$ in $\mathcal{X}$, the vector $(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_L))$ has a multivariate normal distribution.

Gaussian random functions are determined by their *mean* function, $\mu(\boldsymbol{x}) \equiv E\{Y(\boldsymbol{x})\}$, for $\boldsymbol{x} \in \mathcal{X}$, and by their *covariance* function

$$C^{\star}(\boldsymbol{x}_1, \boldsymbol{x}_2) \equiv \mathrm{Cov}\{Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2)\},$$

for $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathcal{X}$. Some authors call $C^{\star}(\cdot, \cdot)$ the "autocovariance" function to be consistent with the language used in time series analysis.

The $Y(\boldsymbol{x})$ model in Example 2.1 is a GRF. The GRFs that are used in practice are *nonsingular*, which means that for any choice of inputs, the covariance matrix of the associated multivariate normal distribution is nonsingular. Such nonsingular multivariate normal distributions have the advantage that it is easy to compute the conditional distribution of one (or several) of the $Y(\boldsymbol{x}_i)$ variables given the remaining $Y(\boldsymbol{x}_j)$. The prediction methodology used in Section 3.3 requires that these conditional means and conditional variances be known and the predictive distributions of Section 4.1 require that the entire conditional distribution be known. In addition, draws from the most widely used GRFs allow a greater spectrum of shapes than the quadratic equations generated in (2.1). They also permit the modeler to control the smoothness properties of the $y(\boldsymbol{x})$ draws; in most of the scientific applications mentioned above, there is *some* information about the smoothness of $y(\cdot)$, although perhaps only that it is a continuous function of the inputs.

There are two technical concepts that we address briefly before introducing specific GRF models. We wish to make the reader aware of the *practical* difficulties that these two concepts address. The first concept has to do with the fact that our random function models are defined by their *finite-dimensional* distributions while, in the following, we are interested in properties that depend on limiting operations such as assuring that functions drawn from the process have specified smoothness (continuity and differentiability) properties. The continuity and differentiability of $y(\boldsymbol{x})$ as a function of $\boldsymbol{x}$ are *sample path properties*, i.e., they regard $y(\boldsymbol{x}) = Y(\boldsymbol{x}, \omega)$ as a function of $\boldsymbol{x}$ for fixed ω. Thus throughout, we require that our random function models be *separable*, which is a property introduced by Doob (1953) that ensures that the finite-dimensional distributions determine the sample path properties of function draws. Adler (1981) (pages 14-15) states the formal definition of separability and discusses its intuition. For our purposes it suffices to know that given any random function $Y(\cdot)$ on $\mathcal{X}$, there is an equivalent separable random function $Y^s(\cdot)$ on $\mathcal{X}$. The random functions $Y(\cdot)$ and $Y^s(\cdot)$ are *equivalent* provided

$$P\left\{Y(\boldsymbol{x}) = Y^s(\boldsymbol{x})\right\} = 1 \quad \text{for all} \quad \boldsymbol{x} \in \mathcal{X}.$$

We assume throughout (and the proofs of almost sure sample path properties require) that our GRF models have been chosen to be separable.

Our second technical concept concerns a statistical issue. Classical statistical methods make inferences about a population based on a random sample of data from that population. Indeed, the statistical procedure is chosen to have certain sampling characteristics (meaning properties based on repeated sampling from the population). In computer experiments (as well as in most other applications of spatial statistics), we observe $y(\boldsymbol{x}_1)$, ..., $y(\boldsymbol{x}_n)$, where $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_n$ are training data input sites. However, these data are values of a *single* function drawn from a population of functions according to $Y(\cdot)$, i.e., $(y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_n)) = (Y(\boldsymbol{x}_1, \omega), \ldots, Y(\boldsymbol{x}_n, \omega))$. Thus spatial data gives *partial* information about a single function $y(\boldsymbol{x}) = Y(\boldsymbol{x}, \omega)$ rather than a *random sample of functions* drawn according to $Y(\cdot)$. To predict the value of $y(\boldsymbol{x}_{new})$, where $\boldsymbol{x}_{new}$ is a new input site, the process must exhibit some regularity over $\mathcal{X}$. In general, it need not be the case that one can make inference about population quantities which are Ω averages such as the $y(\boldsymbol{x}_{new})$ predictor above, based on a spatial average for a single $\omega \in \Omega$. Process *ergodicity* is the standard property that permits valid statistical inference about that process based on a single draw (for a discussion of this property from a statistical viewpoint, see Cressie (1993), pages 52-58, and the additional references listed there). The technical details of this concept are beyond the scope of this book; we note only that this issue motivates users to restrict attention to GRFs that are (strongly) *stationary* (or homogeneous).

Definition The random function $Y(\cdot)$ is *strongly stationary* provided that for any $\boldsymbol{h} \in \mathbb{R}^d$, any $L \geq 1$, any $\boldsymbol{x}_1$, ..., $\boldsymbol{x}_L$ in $\mathcal{X}$ with $\boldsymbol{x}_1 + \boldsymbol{h}$, ..., $\boldsymbol{x}_L + \boldsymbol{h} \in \mathcal{X}$, it must be the case that $(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_L))$ and $(Y(\boldsymbol{x}_1 + \boldsymbol{h}), \ldots, Y(\boldsymbol{x}_L + \boldsymbol{h}))$ have the *same* distribution.

Notice that this definition is general. When applied to GRFs $Y(\cdot)$, stationarity is equivalent to requiring that $(Y(\boldsymbol{x}_1), \ldots, Y(\boldsymbol{x}_L))$ and $(Y(\boldsymbol{x}_1 + \boldsymbol{h}), \ldots, Y(\boldsymbol{x}_L + \boldsymbol{h}))$ always have the same mean vector and same covariance matrix. In particular, GRFs must have the *same* marginal distribution for all $\boldsymbol{x}$ (taking $L = 1$); their mean and their variance must both be constant. Furthermore, it is not difficult to show that the covariance of a stationary GRF must satisfy

$$\text{Cov}\left\{Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2)\right\} = C\left(\boldsymbol{x}_1 - \boldsymbol{x}_2\right) \qquad (2.3.4)$$

for some function $C(\cdot)$, called the *covariance function* of the process. The equation (2.3.4) means that all pairs of locations $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ having common orientation *and* common inter-point distance will have the same covariance. For example, the pairs of points at the tails and tips of the three arrows in Figure 2.2 all have the same covariance structure (as well as infinitely many other pairs on the two parallel lines depicted in the figure). The (constant) variance of a stationary process can be expressed in terms of its covariance function as $\mathrm{Var}\{Y(\boldsymbol{x})\} = \mathrm{Cov}\{Y(\boldsymbol{x}), Y(\boldsymbol{x})\} = C(\mathbf{0})$.

Technically, the stationarity of a GRF $Y(\boldsymbol{x})$ does not guarantee that $Y(\boldsymbol{x})$ is ergodic but this will be case if $C(\boldsymbol{h}) \to 0$ as $\boldsymbol{h} \to \infty$ and *hence, inference is valid based on data collected from a single sample path* (Adler (1981), page 145). The correlation function examples below satisfy this condition.

An even stronger requirement is that the GRF be invariant under rotations, a property called *isotropy.* A stationary GRF $Y(\cdot)$ can be shown to be isotropic provided

$$\mathrm{Cov}\left\{Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2)\right\} = C\left(\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2\right), \tag{2.3.5}$$

where $\|\boldsymbol{h}\|_2 = \sqrt{\sum_i h_i^2}$ is Euclidean distance. For isotropic models, every pair of points $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ having common inter-point distance must have the same covariance (and correlation) regardless of their orientation (see the right panel of Figure 2.2). For example, for any isotropic GRF, the origin has the same correlation with every point on the unit circle. Isotropic models are usually not useful when component inputs are measured on different scales.

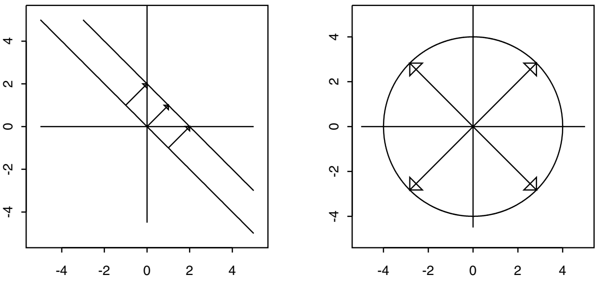

FIGURE 2.2. In the left-hand panel, the tip and tail of each arrow have the same correlation for stationary random functions. In the right-hand panel, all points on the circle have the same correlation for isotropic random functions.

We will occasionally consider random function models that that are nonparametric in that they make only moment requirements on $Y(\cdot)$. The most important such model is that of *second-order stationary.* A random function $Y(\cdot)$ having *constant mean* and *constant variance* is second-order stationary provided its covariance function satisfies (2.3.4).

Despite the arguments given above, stationarity is a substantial restriction and we often require more flexibility in modeling $y(\boldsymbol{x})$. Several approaches have been used in the literature to enhance random function modeling while retaining (some) of the theoretical simplifications that stationarity provides. The most frequently used of these techniques, the one we employ here, is to permit the mean of the stochastic process generating $y(\boldsymbol{x})$ to depend on $\boldsymbol{x}$ in a standard regression manner while assuming the residual variation follows a stationary GRF. The corresponding random function has the form

$$Y(\boldsymbol{x}) = \sum_{j=1}^{p} f_j(\boldsymbol{x})\beta_j + Z(\boldsymbol{x}) = \boldsymbol{f}^\top(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}), \qquad (2.3.6)$$

where $f_1(\cdot), \ldots, f_p(\cdot)$ are *known* regression functions, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is a vector of *unknown* regression coefficients, and $Z(\cdot)$ is a *zero mean* stationary GRF over $\mathcal{X}$. These $Y(\cdot)$ models are, of course, nonstationary.

Most other methods for enhancing $Y(\boldsymbol{x})$ model flexibility have been motivated by environmental applications which often require nonstationary models. While extremely successful in these applications, the nonstationary models introduced in the course of such data analyses have typically been used only in low-dimensional $\boldsymbol{x}$ input settings (two- or three-dimensional space, or three- or four-dimensional space-time applications). Their ability to handle higher dimensional $\boldsymbol{x}$ input cases is untested, although they may well be of use in the analysis of computer experiments. We mention two modeling strategies that have been suggested in the literature.

One method is to generate nonstationary $Y(\boldsymbol{x})$ models from stationary ones by convolving a stationary process with a kernel; Higdon, Swall and Kern (1999) integrate white noise, the spatial analog of a random sample of normal observations, against a Gaussian kernel to produce such a model. In the same spirit, Hass (1995) constructs $Y(\boldsymbol{x})$ models as a moving window over a stationary process. Another approach, introduced by Sampson and Guttorp (1992), is based on deforming the input $\boldsymbol{x}$ of a stationary process to model $Y(\boldsymbol{x})$ (see also Guttorp and Sampson (1994) and Guttorp, Meiring and Sampson (1994)).

2.3.3 The Correlation Function of a Gaussian Random Function Model

To be consistent with the notation introduced in (2.3.6), hereafter we denote the stationary GRF of interest by $Z(\cdot)$. We reiterate that $Z(\cdot)$ has zero mean (by including any overall constant mean value among the regression terms in (2.3.6)). Thus $Z(\cdot)$ is completely determined by its covariance function $C(\cdot)$. In some applications, it is more convenient to separately model the process variance $\sigma_Z^2 = C(\boldsymbol{0})$ and the process correlation function. The *correlation function* of a stationary process $Z(\boldsymbol{x})$ that has finite $\sigma_Z^2 > 0$ and covariance function $C(\cdot)$ is defined to be

$$R(\boldsymbol{h}) = C(\boldsymbol{h})/\sigma_Z^2 \;\text{ for }\; \boldsymbol{h} \in \mathrm{I\!R}^d.$$

The name "correlation function" comes from

$$\begin{aligned}\mathrm{Cor}\{Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2)\} &= \frac{\mathrm{Cov}\{Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2)\}}{\sqrt{\mathrm{Var}\{Z(\boldsymbol{x}_1)\} \times \mathrm{Var}\{Z(\boldsymbol{x}_2)\}}} \\ &= \frac{C(\boldsymbol{x}_1 - \boldsymbol{x}_2)}{\sigma_Z^2} = R(\boldsymbol{x}_1 - \boldsymbol{x}_2).\end{aligned}$$

What properties must valid covariance and correlation functions possess? Assuming that $Z(\boldsymbol{x})$ is nondegenerate, then $C(\mathbf{0})$ $(= \sigma_Z^2) > 0$ while $R(\mathbf{0}) = 1$. Because $\mathrm{Cov}\{Y(\boldsymbol{x}+\boldsymbol{h}), Y(\boldsymbol{x})\} = \mathrm{Cov}\{Y(\boldsymbol{x}), Y(\boldsymbol{x}+\boldsymbol{h})\}$, the covariance and correlation functions of stationary GRFs must be *symmetric about the origin*, i.e.,

$$C(\boldsymbol{h}) = C(-\boldsymbol{h}) \ \text{ and } \ R(\boldsymbol{h}) = R(-\boldsymbol{h}).$$

Both $C(\cdot)$ and $R(\cdot)$ must be *positive semidefinite* functions; stated in terms of $C(\cdot)$, this means that for any $L \geq 1$, and any real numbers $w_1, \ldots, w_L$, and any inputs $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$ in $\mathcal{X}$,

$$\sum_{i=1}^{L} \sum_{j=1}^{L} w_i w_j C(\boldsymbol{x}_i - \boldsymbol{x}_j) \geq 0. \tag{2.3.7}$$

The sum (2.3.7) must be nonnegative because the left-hand side is the variance of $\sum_{i=1}^{L} w_i Y(\boldsymbol{x}_i)$. The covariance function $C(\cdot)$ is *positive definite* provided >0 holds in (2.3.7) for every $(w_1, \ldots, w_L) \neq \mathbf{0}$ (any $L \geq 1$ and any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$ in $\mathcal{X}$).

While every covariance function must satisfy the symmetry and positive semidefinite properties above, these properties do not offer a convenient method for generating valid covariance functions. Rather, what is of greater importance is a characterization of the class of covariance functions because this would allow us to generate valid covariance functions. While a general study of how to determine the form of valid stationary covariance functions is beyond the scope of this book, one answer to this question is relatively simple to state, and we do so next.

As a prelude to identifying this class of covariance functions (and as an introduction to the topic of smoothness which is taken up again in Section 2.3.4), we introduce the concept of mean square (MS) continuity. Mean square properties describe the average performance of the sample paths. For purposes of stating the definitions of MS properties, there is nothing to be gained by restricting attention to GRFs and so we consider general random functions $Y(\cdot)$.

Definition Suppose $Y(\cdot)$ is a stationary process on $\mathcal{X}$ that has finite second moments. We say that $Y(\cdot)$ is *MS continuous* at the point $\boldsymbol{x}_0 \in \mathcal{X}$ provided

$$\lim_{\boldsymbol{x} \to \boldsymbol{x}_0} E\left\{ (Y(\boldsymbol{x}) - Y(\boldsymbol{x}_0))^2 \right\} = 0.$$

The process is *MS continuous on $\mathcal{X}$* provided it is MS continuous at every $\boldsymbol{x}_0 \in \mathcal{X}$.

Suppose $C_Y(\cdot)$ is the covariance function of the stationary process $Y(\cdot)$, then

$$E\left\{(Y(\boldsymbol{x}) - Y(\boldsymbol{x}_0))^2\right\} = 2\left(C_Y\left(\boldsymbol{0}\right) - C_Y\left(\boldsymbol{x} - \boldsymbol{x}_0\right)\right). \tag{2.3.8}$$

The right-hand formula shows that $Y(\cdot)$ is MS continuous at $\boldsymbol{x}_0$ provided $C_Y(\cdot)$ is continuous at the origin—in fact, $Y(\cdot)$ is MS continuous at *every* $\boldsymbol{x}_0 \in \mathcal{X}$ provided $C_Y(\cdot)$ is continuous at the origin. Stated in terms of the correlation function, $C_Y(\boldsymbol{h}) \to C_Y(\boldsymbol{0}) = \sigma_Z^2$ as $\boldsymbol{h} \to \boldsymbol{0}$ is equivalent to

$$R_Y(\boldsymbol{h}) = C_Y(\boldsymbol{h})/\sigma_Z^2 \to 1.0 \ \text{ as } \ \boldsymbol{h} \to \boldsymbol{0}.$$

Continuing our discussion of general random functions $Y(\cdot)$, Bochner (1955) proved that the covariance function of every stationary, MS continuous random function $Y(\cdot)$ on $\mathbb{R}^d$, can be written in the form

$$C_Y(\boldsymbol{h}) = \int_{I\!R^d} \cos(\boldsymbol{h}^\top \boldsymbol{w}) \; dG(\boldsymbol{w}), \tag{2.3.9}$$

where $G(\cdot)$ is positive finite symmetric measure on $\mathbb{R}^d$. In particular, this characterization must hold for the special case of stationary GRFs. (See also the discussions in Cramér and Leadbetter (1967) on page 126, Adler (1981) on page 25, Cressie (1993) on page 84, or Stein (1999) on page 22-25.)

The process variance corresponding to $C_Y(\cdot)$ having the form (2.3.9) is

$$C_Y(\boldsymbol{0}) = \int_{I\!R^d} dG(\boldsymbol{w}) < +\infty$$

which is finite because G is a bounded measure on $\mathbb{R}^d$; $F(\cdot) = G(\cdot)/C_Y(\boldsymbol{0})$ is a symmetric probability distribution, called the *spectral distribution*, corresponding to $C_Y(\cdot)$. The function

$$R_Y(\boldsymbol{h}) = \int_{I\!R^d} \cos(\boldsymbol{h}^\top \boldsymbol{w}) \; dF(\boldsymbol{w}) \tag{2.3.10}$$

is the correlation function corresponding to the spectral distribution $F(\cdot)$. If $F(\cdot)$ has a density $f(\cdot)$, then $f(\cdot)$ is called the *spectral density* corresponding to $R_Y(\cdot)$. In this case

$$R_Y(\boldsymbol{h}) = \int_{I\!R^d} \cos(\boldsymbol{h}^\top \boldsymbol{w}) f(\boldsymbol{w}) \; d\boldsymbol{w}. \tag{2.3.11}$$

The right-hand side of (2.3.11) gives us a method to produce valid correlation functions (and covariance functions)—choose a symmetric density $f(\cdot)$ and evaluate the integral (2.3.11).

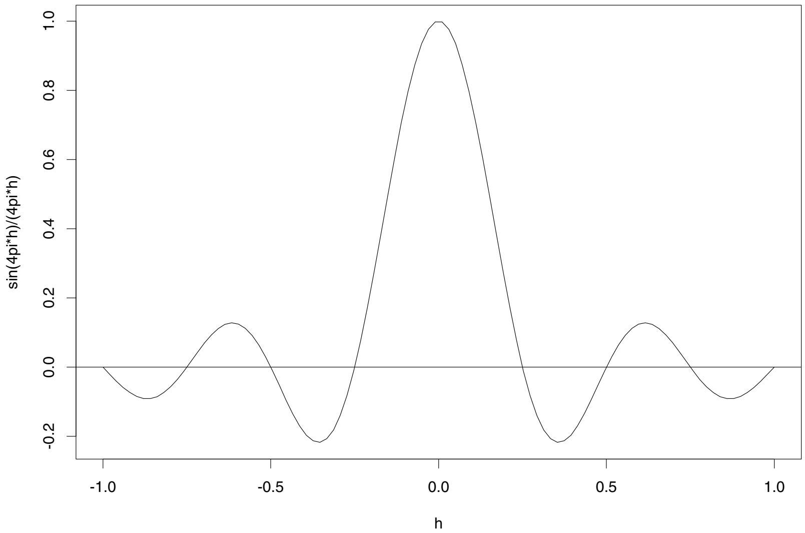

FIGURE 2.3. The correlation function $R(h) = \sin(h/\theta)/(h/\theta)$ for $\theta = 1/4\pi$ over h in $[-1, +1]$

Example 2.2 This first example shows how (2.3.11) can be used to generate valid correlation functions from probability density functions that are symmetric about the origin. Consider the one-dimensional case. Perhaps the simplest choice of one-dimensional density is the uniform density over a symmetric interval which we take to be $(-1/\theta, +1/\theta)$ for a given $\theta > 0$. Thus the spectral density is

$$f(w) = \begin{cases} \theta/2, & -1/\theta < w < 1/\theta \\ 0, & \text{otherwise} \end{cases}$$

and the corresponding correlation function is

$$R(h) = \int_{-1/\theta}^{+1/\theta} \frac{\theta}{2} \cos(hw)\; dw = \begin{cases} \frac{\sin(h/\theta)}{h/\theta}. & h \neq 0 \\ 1, & h = 0 \end{cases} .$$

This correlation has scale parameter θ; Figure 2.3 shows that $R(h)$ can model both positive and negative correlations. ■

Any function $R_Y(\cdot)$ of the form (2.3.10) must satisfy $R_Y(\mathbf{0}) = 1$, must be continuous at $\boldsymbol{h} = \mathbf{0}$, must be symmetric about $\boldsymbol{h} = \mathbf{0}$, and must be positive semidefinite. The first consequence holds because

$$R_Y(\mathbf{0}) = \int_{I\!R^d} \cos(\mathbf{0}^\top \boldsymbol{w})\; dF(\boldsymbol{w}) = \int_{I\!R^d} 1\; dF(\boldsymbol{w}), = 1,$$

where the third equality in the above is true because $F(\cdot)$ is a probability distribution. Continuity follows by an application of the dominated convergence theorem; notice that from the argument following (2.3.8), continuity of $R_Y(\boldsymbol{h})$ at the origin insures that the corresponding process is MS continuous. Symmetry holds because $\cos(-x) = \cos(x)$ for all real x. Positive semidefinite is true because for any $L \geq 1$, any real numbers $w_1, \ldots, w_L$, and any $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_L$ we have

$$\begin{aligned}
&\sum_{i=1}^{L} \sum_{j=1}^{L} w_i w_j R_Y(\boldsymbol{x}_i - \boldsymbol{x}_j) \\
&\quad = \int_{I\!R^d} \sum_{i=1}^{L} \sum_{j=1}^{L} w_i w_j \cos(\boldsymbol{x}_i^\top \boldsymbol{w} - \boldsymbol{x}_j^\top \boldsymbol{w})\ dF(\boldsymbol{w}) \\
&\quad = \int_{I\!R^d} \sum_{i=1}^{L} \sum_{j=1}^{L} w_i w_j \left\{ \cos(\boldsymbol{x}_i^\top \boldsymbol{w}) \cos(\boldsymbol{x}_j^\top \boldsymbol{w}) \right. \\
&\qquad\qquad \left. + \sin(\boldsymbol{x}_i^\top \boldsymbol{w}) \sin(\boldsymbol{x}_j^\top \boldsymbol{w}) \right\}\ dF(\boldsymbol{w}) \\
&\quad = \int_{I\!R^d} \left\{ \left(\sum_{i=1}^{L} w_i \cos(\boldsymbol{x}_i^\top \boldsymbol{w}) \right)^2 + \left(\sum_{i=1}^{L} w_i \sin(\boldsymbol{x}_i^\top \boldsymbol{w}) \right)^2 \right\}\ dF(\boldsymbol{w}) \\
&\quad \geq 0.
\end{aligned}$$

Continuity, symmetry, and positive semidefiniteness also hold for any covariance function $C_Y(\cdot)$ of form (2.3.9).

We conclude by mentioning several additional tools that are extremely useful for "building" covariance and correlation functions given a basic set of such functions. Suppose that $C_1(\cdot)$ and $C_2(\cdot)$ are valid covariance functions. Then their sum and product,

$$C_1(\cdot) + C_2(\cdot) \ \text{ and } \ C_1(\cdot) \times C_2(\cdot),$$

are also valid covariance functions. The sum, $C_1(\cdot) + C_2(\cdot)$, is the covariance of two independent processes, one with covariance function $C_1(\cdot)$ and the other with covariance function $C_2(\cdot)$. Similarly, $C_1(\cdot) \times C_2(\cdot)$ is the covariance function of the product of two independent zero-mean GRFs with covariances $C_1(\cdot)$ and $C_2(\cdot)$, respectively.

The product of two valid correlation functions, $R_1(\cdot)$ and $R_2(\cdot)$, is a valid correlation function, but their sum is not (notice that $R_1(\mathbf{0}) + R_2(\mathbf{0}) = 2$, which is not possible for a correlation function). Correlation functions that are the products of one-dimensional marginal correlation functions are sometimes called *separable* correlation functions (not to be confused with the earlier use of the term separable).

We now introduce two widely–used families of correlation functions that have been used in the literature to specify stationary Gaussian stochastic processes (see also Journel and Huijbregts (1978), Mitchell, Morris and Ylvisaker (1990), Cressie (1993), Vecchia (1988), and Stein (1999)).

Example 2.3 Another familiar choice of a symmetric density that can be used as a spectral density is the normal density. To give a simple form for the resulting correlation function, take the spectral density to be $N(0, 2/\theta^2)$ for $\theta > 0$. Calculation gives

$$\begin{aligned} R(h) &= \int_{-\infty}^{+\infty} \cos(hw) \frac{\theta}{\sqrt{2\pi}\sqrt{2}} \exp\{-w^2\theta^2/4\}\, dw \\ &= \exp\left\{-(h/\theta)^2\right\}. \end{aligned} \tag{2.3.12}$$

This correlation is sometimes called the *Gaussian correlation function* because of its form but the reader should realize that the name is, perhaps, a misnomer. The Gaussian correlation function is a special case of the more general family of correlations called the power exponential correlation family. This family is far and away the most popular family of correlation models in the computer experiments literature. The one-dimensional GRF $Z(x)$ on $x \in \mathbb{R}$ has *power exponential* correlation function provided

$$R(h) = \exp\left\{-|h/\theta|^p\right\} \quad \text{for} \quad h \in \mathbb{R}, \tag{2.3.13}$$

where $\theta > 0$, and $0 < p \leq 2$. In addition to the Gaussian subfamily, the case $p = 1$

$$R(h) = \exp\left\{-(|h|/\theta)\right\}$$

is well-studied. The GRF corresponding to this correlation function is known as the Ornstein-Uhlenbeck process.

For later reference, we note that every power exponential correlation function, $0 < p \leq 2$, is continuous at the origin, and none, except the Gaussian $p = 2$, is differentiable at the origin. In fact, the Gaussian correlation function is infinitely differentiable at the origin.

From the fact that products of correlation functions are also correlation functions,

$$R(\boldsymbol{h}) = \exp\left\{-\sum_{j=1}^{d} |h_j/\theta_j|^{p_j}\right\} \tag{2.3.14}$$

is a d-dimensional separable version of the power exponential correlation function, as is the special case of the product Gaussian family

$$R(\boldsymbol{h}) = \exp\left\{-\sum_{j=1}^{d} (h_j/\theta_j)^2\right\}$$

which has dimension–specific scale parameters. ■

Example 2.4 Suppose that $Z(x)$ is a one-dimensional GRF on $x \in \mathrm{I\!R}$ with correlation function

$$R(h|\theta) = \begin{cases} 1 - 6\left(\frac{h}{\theta}\right)^2 + 6\left(\frac{|h|}{\theta}\right)^3, & |h| \leq \theta/2 \\ 2\left(1 - \frac{|h|}{\theta}\right)^3, & \theta/2 < |h| \leq \theta \\ 0, & \theta < |h| \end{cases}, \tag{2.3.15}$$

where $0 < \theta$ and $h \in \mathrm{I\!R}$. The function $R(h|\theta)$ has two continuous derivatives at $h = 0$ and also at the change point $h = \theta/2$ (see the right column of Figure 2.6). $R(h|\theta)$ assigns zero correlation to inputs x_1 and x_2 that are sufficiently far apart ($|x_1 - x_2| > \theta$). Formally, the spectral density that produces (2.3.15) is proportional to

$$\frac{1}{w^4\theta^3}\left\{72 - 96\cos\left(w\theta/2\right) + 24\cos(w\theta)\right\}.$$

Anticipating Section 3.2 on prediction in computer experiments, the use of (2.3.15) leads to cubic spline interpolating predictors. As in the previous example, we note that

$$R(\boldsymbol{h}|\boldsymbol{\theta}) = \prod_{j=1}^{d} R(h_j|\theta_j)$$

for $\boldsymbol{h} \in \mathrm{I\!R}^d$ is a correlation function that allows each input dimension to have its own scale and thus dimension specific rate at which $Z(\cdot)$ values become uncorrelated. Other one-dimensional cubic correlation functions can be found in Mitchell et al. (1990) and Currin, Mitchell, Morris and Ylvisaker (1991). ■

2.3.4 Using the Correlation Function to Specify a GRF with Given Smoothness Properties

In practice we reduce the choice of a GRF to that of a covariance (or correlation) function whose realizations have desired prior smoothness characteristics. Hence we now turn attention to describing the relationship between the smoothness properties of a stationary GRF, $Z(\cdot)$, and the properties of its covariance function, $C(\cdot)$. To describe this relationship for general processes would require substantial space. By restricting attention to stationary GRFs we can provide a relatively concise overview. See Adler (1990), Abrahamsen (1997), or Stein (1999) for a discussion of these ideas for more general processes and for additional detail concerning the Gaussian process case.

There are several different types of "continuity" and "differentiability" that a process can possess. The definitions differ in their ease of application and the technical simplicity with which they are established. Given a particular property such as continuity at a point or differentiability over an interval, we would like to know that draws from a given random function model $Z(\cdot)$ have that property with probability one. For example, if Q is a property of interest, say continuity at the point $\boldsymbol{x}_0$, then we desire

$$P\left\{\omega : Z(\cdot, \omega) \text{ has property } Q\right\} = 1.$$

We term this *almost sure behavior* of the sample paths.

Section 2.3.3 introduced the widely-used concept of MS continuity. We saw an instance of the general fact that MS properties are relatively simple to prove, although they are not of direct interest in describing sample paths. Below we show that a slight strengthening of the conditions under which MS continuity holds guarantees almost sure continuity.

Recall that in Section 2.3.3 we stated that any stationary random function $Z(\cdot)$ on $\mathcal{X}$ having finite second moments is MS continuous on $\mathcal{X}$ provided that its correlation function is continuous at the origin, i.e., $R(\boldsymbol{h}) \to 1$ as $\boldsymbol{h} \to \mathbf{0}$. GRFs with either the cubic (2.3.15) or the power exponential (2.3.13) correlation functions are examples of such random functions.

Adler (1981) (page 60) shows that for the sample paths of stationary GRFs to be almost surely continuous, one need only add a condition requiring that $R(\boldsymbol{h})$ converge to unity sufficiently fast. For example, a consequence of his Theorem 3.4.1 is that, if $Z(\cdot)$ is a stationary GRF with correlation function $R(\cdot)$ that satisfies

$$1 - R(\boldsymbol{h}) \leq \frac{c}{|\log(\|\boldsymbol{h}\|_2)|^{1+\epsilon}} \quad \text{for all} \quad \|\boldsymbol{h}\|_2 < \delta \qquad (2.3.16)$$

for some $c > 0$, some $\epsilon > 0$, and some $\delta < 1$, then $Z(\cdot)$ has almost surely continuous sample paths. MS continuity requires that $(1 - R(\boldsymbol{h})) \to 0$ as $\boldsymbol{h} \to 0$; the factor $|\log(\|\boldsymbol{h}\|_2)|^{1+\epsilon} \to +\infty$ as $\boldsymbol{h} \to 0$. Thus (2.3.16) holds provided that $1 - R(\boldsymbol{h})$ converges to zero at least as fast as $|\log(\|\boldsymbol{h}\|_2)|^{1+\epsilon}$ diverges to $+\infty$. The product

$$[1 - R(\boldsymbol{h})] \times \left|\log\left(\|\boldsymbol{h}\|_2\right)\right|^{1+\epsilon}$$

is bounded for most correlation functions used in practice. In particular this is true for any power exponential correlation function with $0 < p \leq 2$. One can also use the spectral distribution to give sufficient conditions for almost sure continuity of sample paths. The standard conditions are stated in terms of the finiteness of the moments of the spectral distribution. For example, see Theorem 3.4.3 of Adler (1981) or Sections 9.3 and 9.5 of Cramér and Leadbetter (1967).

Conditions for almost sure continuity of the sample paths of nonstationary GRFs, $Z(\cdot)$, can be similarly expressed in terms of the rate at which

$$E\left\{|Z(\boldsymbol{x}_1) - Z(\boldsymbol{x}_2)|^2\right\}$$

converges to zero as $\|\boldsymbol{x}_1 - \boldsymbol{x}_2\|_2 \rightarrow 0$ (Adler (1981), Theorem 3.4.1).

As for continuity, a concept of mean square differentiability can be defined that describes the mean difference of the usual tangent slopes of a given process and a limiting "derivative process." Instead, here we directly discuss the parallel to almost sure continuity. Consider the individual sample draws $z(\boldsymbol{x}) = Z(\boldsymbol{x}, \omega)$, $\mathcal{X} \subset \mathrm{IR}^d$, corresponding to specific outcomes $\omega \in \Omega$. Suppose that the j^{th} partial derivative of $Z(\boldsymbol{x}, \omega)$ exists for $j = 1, \ldots d$ and $\boldsymbol{x} \in \mathcal{X}$, i.e.,

$$\nabla_j Z(\boldsymbol{x}, \omega) = \lim_{\delta \rightarrow 0} \frac{Z(\boldsymbol{x} + \boldsymbol{e}_j \delta, \omega) - Z(\boldsymbol{x}, \omega)}{\delta}$$

exists where $\boldsymbol{e}_j$ denotes the unit vector in the j^{th} direction. Let

$$\boldsymbol{\nabla} Z(\boldsymbol{x}, \omega) = (\nabla_1 Z(\boldsymbol{x}, \omega), \ldots, \nabla_d Z(\boldsymbol{x}, \omega))$$

denote the vector of partial derivatives of $Z(\boldsymbol{x}, \omega)$, sometimes called the gradient of $Z(\boldsymbol{x}, \omega)$. We will state conditions on the covariance (correlation) function that guarantee that the sample paths are almost surely differentiable. The situation for higher order derivatives can be described in a similar manner, sample pathwise, for each ω.

As motivation for the condition given below, we observe the following heuristic calculation that gives the covariance of the derivative of $Z(\cdot)$. Fix $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ in $\mathcal{X}$, then

$$\begin{aligned}
&\mathrm{Cov}\Big(\tfrac{1}{\delta_1} Z(\boldsymbol{x}_1 + \boldsymbol{e}_j \delta_1) - Z(\boldsymbol{x}_1)\,,\, \tfrac{1}{\delta_2} Z(\boldsymbol{x}_2 + \boldsymbol{e}_j \delta_2) - Z(\boldsymbol{x}_2)\Big) \\
&\quad = \quad \frac{1}{\delta_1 \delta_2} \{C(\boldsymbol{x}_1 - \boldsymbol{x}_2 + \boldsymbol{e}_j(\delta_1 - \delta_2)) - C(\boldsymbol{x}_1 - \boldsymbol{x}_2 + \boldsymbol{e}_j \delta_1) \\
&\quad\quad\quad - C(\boldsymbol{x}_1 - \boldsymbol{x}_2 - \boldsymbol{e}_j \delta_2) + C(\boldsymbol{x}_1 - \boldsymbol{x}_2)\} \\
&\quad \rightarrow \quad - \left.\frac{\partial^2 C(\boldsymbol{h})}{\partial h_j}\right|_{\boldsymbol{h} \,=\, \boldsymbol{x}_1 - \boldsymbol{x}_2}
\end{aligned} \tag{2.3.17}$$

as $\delta_1, \delta_2 \rightarrow 0$ when the second partial derivative of $C(\cdot)$ exists. These calculations motivate the fact that the covariance function of the partial derivatives of $Z(\cdot)$, if they exist, are given by the partial derivatives of $C(\boldsymbol{h})$. Thus it should come as no surprise that to assure that a given Gaussian random field has, almost surely, differentiable draws, the conditions required are on the partial derivatives of the covariance function.

Formally, suppose

$$C_j^{(2)}(\boldsymbol{h}) \equiv \frac{\partial^2 C(\boldsymbol{h})}{\partial h_j^2}$$

exists and is continuous with $C_j^{(2)}(\boldsymbol{0}) \neq 0$; let $R_j^{(2)}(\boldsymbol{h}) \equiv C_j^{(2)}(\boldsymbol{h})/C_j^{(2)}(\boldsymbol{0})$ be the normalized version of $C_j^{(2)}(\cdot)$. Then almost surely $Z(\cdot)$ has j^{th} partial differentiable sample path, denoted $\nabla_j Z(\boldsymbol{x})$, provided $R_j^{(2)}(\cdot)$ satisfies

(2.3.16). In this case $-C_j^{(2)}(\boldsymbol{h})$ is the covariance function and $R_j^{(2)}(\boldsymbol{h})$ is the correlation function of $\nabla_j Z(\boldsymbol{x})$.

Higher order $Z(\cdot)$ derivatives can be iteratively developed in the same way, although a more sophisticated notation must be introduced to describe the higher-order partial derivatives required of $C(\cdot)$. Conditions for nonstationary $Z(\cdot)$ can be determined from almost sure continuity conditions for nonstationary $Z(\cdot)$ (Adler (1981), Chapter 3).

We complete this section by illustrating the effects of changing the covariance parameters on the draws of several stationary GRFs that were introduced earlier and on one important additional family, the Matérn correlation function. In each case, the plot was obtained by linearly joining draws from an appropriate 20 or 40 dimensional multivariate normal distribution; hence the figures give the spirit, if not the detail, of the sample paths from the associated process. The interested reader can gain addition feel for stationary Gaussian processes by using the software of Kozintsev (1999) or Kozintsev and Kedem (2000) for generating two-dimensional Gaussian random fields (see the URL

`http://www.math.umd.edu/~bnk/CLIP/clip.gauss.htm)`

Example 2.3 (Continued–power exponential correlation function) Figures 2.4 and 2.5 show the marginal effects of changing the shape parameter p and the scale parameter θ on the function draws from GRFs over $[0, 1]$ having the power exponential correlation function (2.3.13). *These figures, and those that illustrate the other GRFs that are discussed below, connect 20 points drawn from a multivariate normal distribution having the desired covariance matrix and so illustrate the spirit of the function draws, if not their fine detail.*

For powers $p < 2$, the sample paths are theoretically nondifferentiable and this can be seen in the bottom two panels of Figure 2.4. The sample paths for $p = 2.0$ are infinitely differentiable; the draws in the top panel of Figure 2.4 are very near the process mean of zero for $\theta = 1.0$. As shown in Figure 2.5, the number of local maxima and minima in sample paths is controlled by the scale parameter when $p = 2.0$. Figure 2.5 shows that as the scale parameter θ *increases*, the correlations for each fixed pair of inputs decreases and the sample paths have increasing numbers of local maxima. This is true because the process exhibits less dependence for "near-by" x and thus "wiggles" more like white noise, the case of uncorrelated $Z(\boldsymbol{x})$. As θ *decreases*, the correlation for each pair of inputs increases and, as the correlation approaches unity, the draws become more nearly the constant zero, the process mean. In Figure 2.5 the most extreme case of this phenomenon is shown in the top panel where $(p, \theta) = (2.0, 0.50)$. ■

Example 2.4 (Continued–cubic correlation function) Recall that the cubic correlation (and covariance) function (2.3.15) is twice continuously

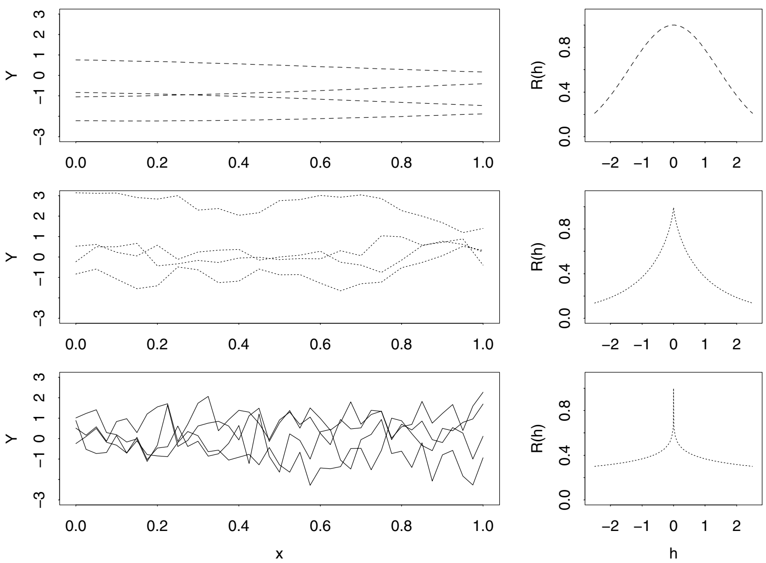

FIGURE 2.4. The Effect of Varying the Power on the Sample Paths of a GRF with a Power Exponential Correlation Function. Four draws from a zero mean, unit variance GRF with the exponential correlation (2.3.13) having fixed $\theta \equiv 1.0$ with $p = 2.0$ (dashed lines), $p = 0.75$ (dotted lines), and $p = 0.20$ (solid lines).

differentiable. Thus draws from a GRF with this correlation structure will be continuous and differentiable. Figure 2.6 shows draws from this process for different θ. As the scale parameter θ *decreases*, the domain where $R(h) = 0$ increases and hence the paths become more like white noise, i.e., having independent and identically distributed Gaussian components. As θ *increases*, the paths tend to become flatter with fewer local maxima and minima. ■

Example 2.5 The Matérn correlation function was introduced by Matérn in his thesis (Matérn (1960) or see the reprint Matérn (1986) and Vecchia (1988) for related work). This model has been used especially to describe the spatial and temporal variability in environmental data (see Rodríguez-Iturbe and Mejía (1974), Handcock and Stein (1993), Handcock and Wallis (1994), and especially Stein (1999)).

From the viewpoint of the spectral representation, the Matérn correlation function arises by choosing the t distribution as the spectral density. Given $\nu > 0$ and $\theta > 0$, use of the t density

$$f(w) = \frac{\Gamma(\nu + 1/2)}{\Gamma(\nu)\sqrt{\pi}} \left(\frac{4\nu}{\theta^2}\right)^{\nu} \frac{1}{\left(w^2 + \frac{4\nu}{\theta^2}\right)^{\nu+1/2}}$$

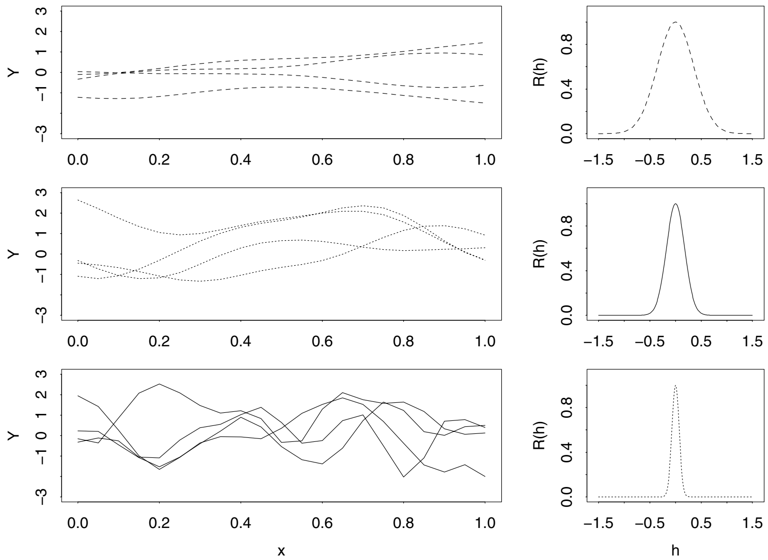

FIGURE 2.5. The Effect of Varying the Scale Parameter on the Sample Paths of a GRF with a Power Exponential Correlation Function. Four draws from a zero mean, unit variance GRF with the exponential correlation function (2.3.12) (having fixed $p = 2.0$) for $\theta = 0.50$ (dashed lines), $\theta = 0.25$ (dotted lines), and $\theta = 0.10$ (solid lines).

in spectral correlation formula (2.3.10) gives the two parameter correlation family

$$R(h) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\,|h|}{\theta} \right)^{\nu} K_{\nu} \left(\frac{2\sqrt{\nu}\,|h|}{\theta} \right), \tag{2.3.18}$$

where $K_\nu(\cdot)$ is the modified Bessel function of order ν. As is usual in the literature, we refer to (2.3.18) as the Matérn correlation function. The parameter θ is clearly a scale parameter for this family. The modified Bessel function arises as the solution of a certain class of ordinary differential equations (Kreyszig (1999)). In general, $K_\nu(t)$ is defined in terms of an infinite power series in t; when ν equals a half integer, i.e., $\nu = n + 1/2$ for $n \in \{0, 1, 2, \ldots\}$, then $K_{n+1/2}(\cdot)$ can be expressed as the finite sum

$$K_{n+1/2}(t) = e^{-t} \sqrt{\frac{\pi}{2t}} \sum_{k=0}^{n} \frac{(n+k)!}{k!\,(n-k)!} \frac{1}{(2t)^k}.$$

The corresponding Matérn correlation function (2.3.18) is

$$e^{-2\sqrt{\nu}|h|/\theta} \left\{ b_0 \left(\frac{|h|}{\theta} \right)^{n} + \, b_1 \left(\frac{|h|}{\theta} \right)^{n-1} + \, b_2 \left(\frac{|h|}{\theta} \right)^{n-2} + \ldots + \, b_n \right\},$$

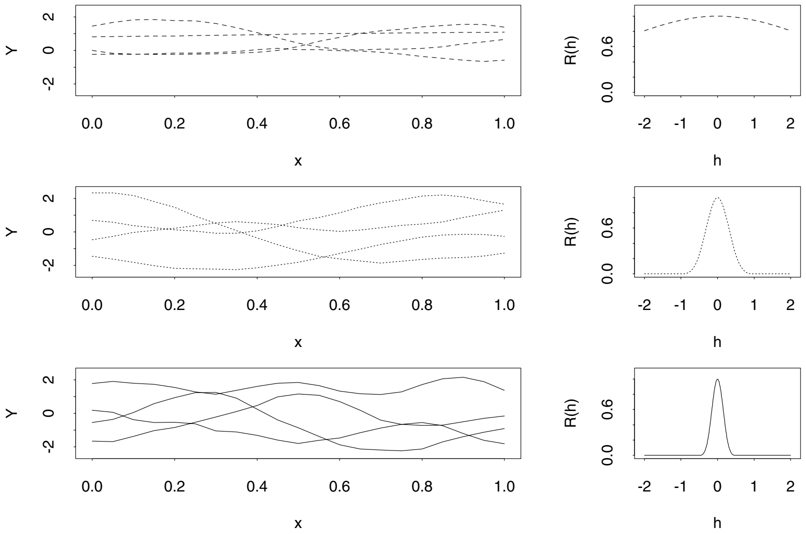

FIGURE 2.6. The Effect of Varying the Scale Parameter on the Sample Paths of a GRF with a Cubic Correlation Function. Four draws from a zero mean, unit variance GRF with the cubic correlation function (2.3.15) for $\theta = 0.5$ (solid lines), $\theta = 1.0$ (dotted lines), and $\theta = 10.0$ (dashed lines). The corresponding correlation function is plotted to the right of each set of sample paths.

where the coefficients are given by

$$b_j = \frac{\sqrt{\pi}\ \nu^{(n-j)/2}}{4^j \Gamma(\nu)}\ \frac{(n+j)!}{j!\ (n-j)!}$$

for $j = 0, 1, \ldots$ where $\nu = n + 1/2$; the b_j depend on ν but not θ. For example, when $n = 0$ ($\nu = 1/2$),

$$K_{1/2}(t) = \sqrt{\pi} e^{-t}/\sqrt{2t} \ \text{ and so } \ R(h) = e^{-\sqrt{2}|h|/\theta},$$

which is a special case of the power exponential correlation function with $p = 1$ that was introduced earlier. Similarly, $R(h) \rightarrow e^{-(h/\theta)^2}$ as $\nu \rightarrow \infty$ so that this class of correlations includes the Gaussian correlation function in the limit.

The smoothness of functions drawn from a GRF with Matérn correlation depends on ν. Let $\lceil \nu \rceil$ denote the integer ceiling of ν, i.e., the smallest integer that is greater than or equal to ν. For example, $\lceil 3.2 \rceil = 4$ and $\lceil 3 \rceil = 3$. Then functions drawn from a GRF having the Matérn correlation have almost surely continuously differentiable sample draws of order $(\lceil \nu \rceil - 1)$. Thus we refer to ν as the smoothness parameter of the Matérn family (see Cramér and Leadbetter (1967)).

Products of the one-dimensional Matérn correlation function can be useful for modeling d-dimensional input responses. In this case, the family might include dimension specific scale parameters and a common smoothness parameter,

$$R(\boldsymbol{h}) = \prod_{i=1}^{d} \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\,|h_i|}{\theta_i} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}\,|h_i|}{\theta_i} \right),$$

or dimension specific scale and smoothness parameters.

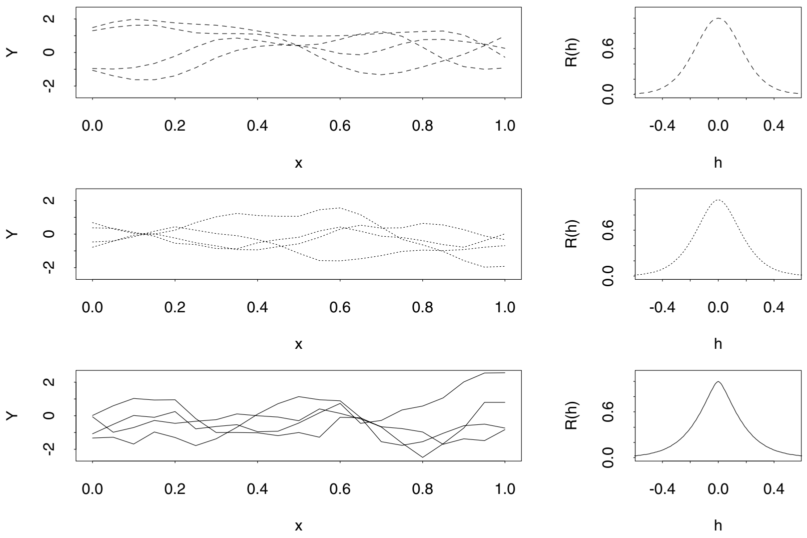

FIGURE 2.7. The Effect of Varying the ν Parameter on the Sample Paths of a GRF with Matérn Correlation Function. Four draws from a zero mean, unit variance GRF with the Matérn correlation function (2.3.18) (having fixed $\theta = 0.25$) for $\nu = 1$ (solid lines), $\nu = 2.5$ (dotted lines), and $\nu = 5$ (dashed lines).

We conclude by displaying sets of function draws from one-dimensional GRFs on $[0, 1]$ having different Matérn correlation functions to illustrate the effect of changing the scale and shape parameters.

Figure 2.7 fixes the scale parameter at $\theta = 0.25$ and varies $\nu \in \{1, 2.5, 5\}$. The draws clearly show the increase in smoothness as ν increases. As a practical matter, it is difficult for most observers to distinguish sample paths having 3 or 4 continuous derivatives from those that are infinitely differentiable. In contrast, Figure 2.8 fixes the smoothness parameter at $\nu = 4$ and varies $\theta \in \{0.01, 0.25, 2.0\}$. For fixed ν and $0 < h < 1.0$, the scaled range of $|h|/\theta$ varies substantially for different θ; $|h|/\theta$ ranges from

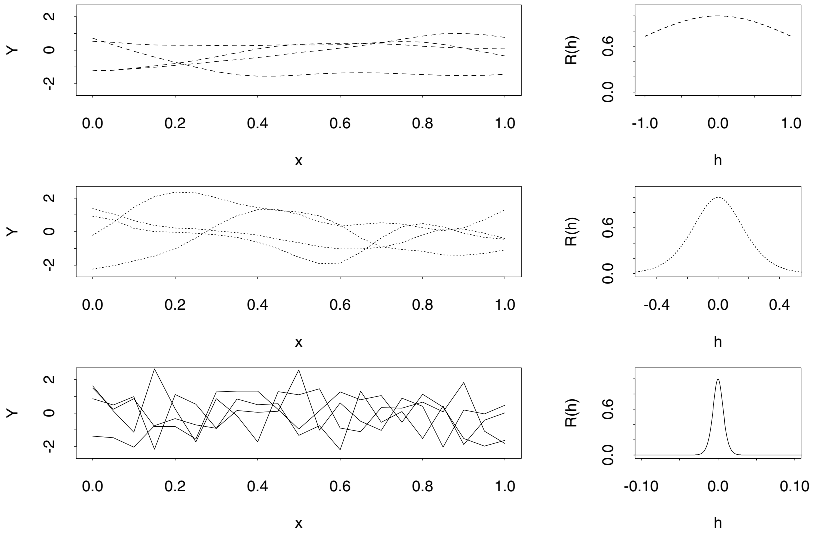

FIGURE 2.8. The Effect of Varying the Scale Parameter on the Sample Paths of a GRF with Matérn Correlation Function. Four draws from a zero mean, unit variance GRF with the Matérn correlation function (2.3.18) (having fixed $\nu = 4$) for $\theta = 0.01$ (solid lines), $\theta = 0.25$ (dotted lines), and $\theta = 2.0$ (dashed lines).

0.0 to 100 for $\theta = 0.01$ while this ratio only varies over 0.0 to 0.5 for $\theta = 2.0$. Notice that we use different h ranges for plotting $R(h)$ in Figure 2.8 to better illustrate the character of the correlation function near the origin. As θ increases, the correlation function of any two fixed points decreases (to zero) and hence the sample paths "look" more like white noise. Thus the bottom panel of this figure plots a process with many more local maxima and minima than does the top panel. ■

2.3.5 Hierarchical Gaussian Random Field Models

While the examples above can provide guidance about the choice of a specific GRF prior for $y(\cdot)$, it will often be the case that the user will not be prepared to specify every detail of the GRF prior. For example, it will often be difficult to specify the correlation function of the GRF. A flexible alternative to the complete specification of a GRF is to use a *hierarchical* GRF prior model for $Y(\cdot)$. To describe this model, suppose that

$$Y(\boldsymbol{x}) = \sum_{j=1}^{p} f_j(\boldsymbol{x})\beta_j + Z(\boldsymbol{x}) = \boldsymbol{f}^\top(\boldsymbol{x})\boldsymbol{\beta} + Z(\boldsymbol{x}),$$

where $Z(\cdot)$ is a Gaussian random field with zero mean, variance σ_Z^2, and correlation function $R(\cdot\,|\,\boldsymbol{\psi})$. Here $R(\cdot\,|\,\boldsymbol{\psi})$ denotes a parametric family of correlation functions. In a hierarchical model some (or all) of $\boldsymbol{\beta}$, σ_Z^2, and $\boldsymbol{\psi}$ are not specified but rather a 2^{nd} stage distribution that describes expert opinion about the relative likelihood of the parameter values.

To be specific, suppose it desired to place a 2^{nd} stage prior on all three parameters $\boldsymbol{\beta}$, σ_Z^2, and $\boldsymbol{\psi}$. Sometimes this task is facilitated because the prior $[\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\psi}]$ prior can be expressed in "pieces." Suppose that it is reasonable to assume that large scale location parameters $\boldsymbol{\beta}$ and the small scale variance, σ_Z^2, are independent of the correlation parameters, $\boldsymbol{\psi}$. This means that

$$[\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\psi}] = [\boldsymbol{\beta}, \sigma_Z^2] \times [\boldsymbol{\psi}] = [\boldsymbol{\beta}\,|\,\sigma_Z^2] \times [\sigma_Z^2] \times [\boldsymbol{\psi}] \ .$$

The second equality is true because $[\boldsymbol{\beta}, \sigma_Z^2] = [\boldsymbol{\beta}\,|\,\sigma_Z^2] \times [\sigma_Z^2]$ always holds. Thus the overall prior can be determined from these three pieces, which is often easier to do.

One complication with hierarchical models is that even when $[\boldsymbol{\beta}, \sigma_Z^2, \boldsymbol{\psi}]$ can be specified, it will usually be the case that the $Y(\boldsymbol{x})$ posterior cannot be expressed in closed form. Subsection 3.3.2 discusses the problem of computing the posterior mean in the context of various "empirical best linear unbiased predictors." See especially the discussion of "posterior mode empirical best linear unbiased predictors" beginning on page 66.

As an example, suppose that the input $\boldsymbol{x}$ is d-dimensional and that $R(\cdot\,|\,\boldsymbol{\psi})$ has the product Matérn correlation function

$$R(\boldsymbol{h}\,|\,\boldsymbol{\psi}) = \prod_{i=1}^{d} \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{2\sqrt{\nu}\,|h_i|}{\theta_i} \right)^{\nu} K_\nu \left(\frac{2\sqrt{\nu}\,|h_i|}{\theta_i} \right) \qquad (2.3.19)$$

with unknown common smoothness parameter and dimension-specific scale parameters; thus $\boldsymbol{\psi} = (\nu, \theta_1, \ldots, \theta_d)$. Consider specification of prior $[\boldsymbol{\psi} = (\nu, \theta_1, \ldots, \theta_d)]$. Suppose that any ν, $2 \leq \nu \leq 50$ is equally likely, which implies that the number of derivatives in each dimension is equally likely to range from 1 to 49. Given ν, 2^{nd} stage priors can be placed on each scale parameter by soliciting expert opinion about likelihood of correlation values between $Y(\boldsymbol{x}_1)$ and $Y(\boldsymbol{x}_2)$ where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ differ in exactly one coordinate direction. See Oakley (2002) for details and a case study. There are other examples of the construction of 2^{nd} stage prior distributions for parameters, mostly in the environmental literature. For example, Handcock and Wallis (1994) build a prior distribution for correlation parameters in their space-time model of the mean temperature of a region of the northern United States.

The references in the previous paragraph describe what might be thought of as "informative" 2^{nd} stage priors. Again returning to the Matérn correlation function (2.3.19), it may be difficult to choose even the means and variances of the smoothness parameter and the scale parameters for specific dimensions, much less the $[\boldsymbol{\psi}]$ joint distribution. in such cases it is tempting to develop and use so-called "non-informative" 2^{nd} stage priors, which give "equal" weight to all the legitimate parameter values. The reader should be warned that there is not always agreement in the statistical community about what constitutes a non-informative prior, even for parameters having finite ranges. Furthermore not every choice of a non-informative 2^{nd} stage prior dovetails with the 1^{st} stage model to produce a legitimate prior for $y(\cdot)$ (see the important paper by Berger, De Oliveira and Sansó (2001)). More will said about non-informative 2^{nd} stage priors in Subsection 3.3.2 on page 66, which discusses "posterior mode empirical best linear unbiased predictors." Such predictors assume that a hierarchical GRF model is specified having parametric correlation function $R(\cdot \mid \boldsymbol{\psi})$ with unknown $\boldsymbol{\psi}$.

A third possible choice for a 2^{nd} stage parameter prior is a "conjugate" prior. Conjugate priors lead to closed-form posterior calculations, and are sometimes reasonable. Section 4.1.2 discusses conjugate and non-informative 2^{nd} stage $[\boldsymbol{\beta}]$ distributions (with σ_Z^2 and $\boldsymbol{\psi}$ known). Section 4.1.3 gives the analogous conjugate and non-informative 2^{nd} stage $[\boldsymbol{\beta}, \sigma_Z^2]$ distributions (with $\boldsymbol{\psi}$ known). These two sections give closed-form expressions for the posterior of $Y(\boldsymbol{x})$ given the data.