3

Classical Geostatistical Methods

Dale L. Zimmerman and Michael Stein

CONTENTS

3.1 Overview

Suppose that a spatially distributed variable is of interest, which in theory is defined at every point over a bounded study region of interest, $D \subset R^d$, where $d = 2$ or 3. We suppose further that this variable has been observed (possibly with error) at each of n distinct points in D, and that from these observations we wish to make inferences about the process that governs how this variable is distributed spatially and about values of the variable at locations where it was not observed. The geostatistical approach for achieving these objectives is to assume that the observed data are a sample (at the n data locations) of one realization of a continuously indexed spatial stochastic process (random field) $Y(\cdot) \equiv \{Y(\mathbf{s}) : \mathbf{s} \in D\}$. Chapter 2 reviewed some probabilistic theory for such processes. In this chapter, we are concerned with how to use the sampled realization to make statistical inferences about the process. In particular, we discuss a body of spatial statistical methodology that has come to be known as "classical geostatistics." Classical geostatistical methods focus on estimating the first-order (large-scale or global trend) structure and especially the second-order (small-scale or local) structure of $Y(\cdot)$, and on predicting or interpolating (kriging) values of $Y(\cdot)$ at unsampled locations using linear combinations of the observations and evaluating the performance of these predictions by their (unconditional) mean squared errors. However, if the process Y is sufficiently non-Gaussian, methods based on considering just the first two moments of Y may be misleading. Furthermore, some common practices in classical geostatistics are problematic even for Gaussian processes, as we shall note herein.

Because good prediction of $Y(\cdot)$ at unsampled locations requires that we have at our disposal estimates of the structure of the process, the estimation components of a geostatistical analysis necessarily precede the prediction component. It is not clear, however, which structure, first-order or second-order, should be estimated first. In fact, an inherent circularity exists—to properly estimate either structure, it appears we must know the other. We note that likelihood-based methods (see Chapter 4) quite neatly avoid this circularity problem, although they generally require a fully specified joint distribution and a parametric model

for the covariance structure (however, see Im, Stein, and Zhu, 2007). The classical solution to this problem is to provisionally estimate the first-order structure by a method that ignores the second-order structure. Next, use the residuals from the provisional first-order fit to estimate the second-order structure, and then finally reestimate the first-order structure by a method that accounts for the second-order structure. This chapter considers each of these stages of a classical geostatistical analysis in turn, plus the kriging stage. We begin, however, with a description of the geostatistical model upon which all of these analyses are based.

3.2 Geostatistical Model

Because only one realization of $Y(\cdot)$ is available, and the observed data are merely an incomplete sample from that single realization, considerable structure must be imposed upon the process for inference to be possible. The classical geostatistical model imposes structure by specifying that

$$Y(\mathbf{s}) = \mu(\mathbf{s}) + e(\mathbf{s}), \tag{3.1}$$

where $\mu(\mathbf{s}) \equiv E[Y(\mathbf{s})]$, the mean function, is assumed to be deterministic and continuous, and $e(\cdot) \equiv \{e(\mathbf{s}) : \mathbf{s} \in D\}$ is a zero-mean random "error" process satisfying a stationarity assumption. One common stationarity assumption is that of second-order stationarity, which specifies that

$$\text{Cov}[e(\mathbf{s}), e(\mathbf{t})] = C(\mathbf{s} - \mathbf{t}), \text{ for all } \mathbf{s}, \mathbf{t} \in D. \tag{3.2}$$

In other words, this asserts that the covariance between values of $Y(\cdot)$ at any two locations depends on only their *relative* locations or, equivalently, on their spatial lag vector. The function $C(\cdot)$ defined in (3.2) is called the covariance function. Observe that nothing is assumed about higher-order moments of $e(\cdot)$ or about its joint distribution. Intrinsic stationarity, another popular stationary assumption, specifies that

$$\frac{1}{2}\text{var}[e(\mathbf{s}) - e(\mathbf{t})] = \gamma(\mathbf{s} - \mathbf{t}), \text{ for all } \mathbf{s}, \mathbf{t} \in D. \tag{3.3}$$

The function $\gamma(\cdot)$ defined by (3.3) is called the semivariogram (and the quantity $2\gamma(\cdot)$ is known as the variogram). A second-order stationary random process with covariance function $C(\cdot)$ is intrinsically stationary, with semivariogram given by

$$\gamma(\mathbf{h}) = C(\mathbf{0}) - C(\mathbf{h}), \tag{3.4}$$

but the converse is not true in general. In fact, intrinsically stationary processes exist for which $\text{var}[Y(\mathbf{s})]$ is not even finite at any $\mathbf{s} \in D$. An even weaker stationarity assumption is that satisfied by an intrinsic random field of order k (IRF-k), which postulates that certain linear combinations of the observations known as kth-order generalized increments have mean zero and a (generalized) covariance function that depends only on the spatial lag vector. IRF-ks were introduced in Chapter 2, to which we refer the reader for more details.

Model (3.1) purports to account for large-scale spatial variation (trend) through the mean function $\mu(\cdot)$, and for small-scale spatial variation (spatial dependence) through the process $e(\cdot)$. In practice, however, it is usually not possible to unambiguously identify and separate these two components using the available data. Quoting from Cressie (1991, p. 114), "One person's deterministic mean structure may be another person's correlated error structure."

Consequently, the analyst will have to settle for a plausible, but admittedly nonunique, decomposition of spatial variation into large-scale and small-scale components.

In addition to capturing the small-scale spatial variation, the error process $e(\cdot)$ in (3.1) accounts for measurement error that may occur in the data collection process. This measurement error component typically has no spatial structure; hence, for some purposes it may be desirable to explicitly separate it from the spatially dependent component. That is, we may write

$$e(\mathbf{s}) = \eta(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{3.5}$$

where $\eta(\cdot)$ is the spatially dependent component and $\epsilon(\cdot)$ is the measurement error. Such a decomposition is discussed in more detail in Section 3.5.

The stationarity assumptions introduced above specify that the covariance or semivariogram depends on locations $\mathbf{s}$ and $\mathbf{t}$ only through their lag vector $\mathbf{h} = \mathbf{s} - \mathbf{t}$. A stronger property, not needed for making inference from a single sampled realization but important nonetheless, is that of isotropy. Here we describe just intrinsic isotropy (and anisotropy); second-order isotropy differs only by imposing an analogous condition on the covariance function rather than the semivariogram. An intrinsically stationary random process with semivariogram $\gamma(\cdot)$ is said to be (intrinsically) isotropic if $\gamma(\mathbf{h}) = \gamma(h)$, where $h = (\mathbf{h}'\mathbf{h})^{1/2}$; that is, the semivariogram is a function of the locations only through the (Euclidean) distance between them. If the process is not isotropic, it is said to be anisotropic. Perhaps the most tractable form of anisotropy is geometric anisotropy, for which $\gamma(\mathbf{h}) = \gamma((\mathbf{h}'\mathbf{A}\mathbf{h})^{1/2})$ where $\mathbf{A}$ is a positive definite matrix. Isotropy can be regarded as a special case of geometric anisotropy in which $\mathbf{A}$ is an identity matrix. Contours along which the semivariogram is constant (so-called isocorrelation contours when $Y(\cdot)$ is second-order stationary) are d-dimensional spheres in the case of isotropy and d-dimensional ellipsoids in the more general case of geometric anisotropy.

The objectives of a geostatistical analysis, which were noted in general terms in Section 3.1, can now be expressed more specifically in terms of model (3.1). Characterization of the spatial structure is tantamount to the estimation of $\mu(\cdot)$ and either $C(\cdot)$ or $\gamma(\cdot)$. The prediction objective can be reexpressed as seeking to predict the value of $Y(\mathbf{s}_0) = \mu(\mathbf{s}_0) + e(\mathbf{s}_0)$ at an arbitrary site $\mathbf{s}_0$.

3.3 Provisional Estimation of the Mean Function

The first stage of a classical geostatistical analysis is to specify a parametric model, $\mu(\mathbf{s}; \boldsymbol{\beta})$, for the mean function of the spatial process, and then provisionally estimate this model by a method that requires no knowledge of the second-order dependence structure of $Y(\cdot)$. The most commonly used parametric mean model is a *linear* function, given by

$$\mu(\mathbf{s}; \boldsymbol{\beta}) = \mathbf{X}(\mathbf{s})^T\boldsymbol{\beta}, \tag{3.6}$$

where $\mathbf{X}(\mathbf{s})$ is a vector of covariates (explanatory variables) observed at $\mathbf{s}$, and $\boldsymbol{\beta}$ is an unrestricted parameter vector. Alternative choices include nonlinear mean functions, such as sines/cosines (with unknown phase, amplitude, and period) or even semiparametric or nonparametric (locally smooth) mean functions, but these appear to be used very rarely.

One possible approach to spatial interpolation is to place all of the continuous variation of the process into the mean function, i.e., assume that the observations equal a true but unknown continuous mean function plus independent and identically distributed errors, and use nonparametric regression methods, such as kernel smoothers, local polynomials, or splines. Although nonparametric regression methods provide a viable approach to spatial

interpolation, we prefer for the following reasons the geostatistical approach when $\mathbf{s}$ refers to a location in physical space. First, the geostatistical approach allows us to take advantage of properties, such as stationarity and isotropy, that do not usually arise in nonparametric regression. Second, the geostatistical approach naturally generates uncertainty estimates for interpolated values even when the underlying process is continuous and is observed with little or no measurement error. Uncertainty estimation is problematic with nonparametric regression methods, especially if the standard deviation of the error term is not large compared to the changes in the underlying function between neighboring observations. It should be pointed out that smoothing splines, which can be used for nonparametric regression, yield spatial interpolants that can be interpreted as kriging predictors (Wahba, 1990). The main difference, then, between smoothing splines and kriging is in how one goes about estimating the degree of smoothing and in how one provides uncertainty estimates for the interpolants.

The covariates associated with a point $\mathbf{s}$ invariably include an overall intercept term, equal to one for all data locations. Note that if this is the only covariate and the error process $e(\cdot)$ in (3.1) is second-order (or intrinsically) stationary, then $Y(\cdot)$ itself is second-order (or intrinsically) stationary. The covariates may also include the geographic coordinates (e.g., latitude and longitude) of $\mathbf{s}$, mathematical functions (such as polynomials) of those coordinates, and attribute variables. For example, in modeling the mean structure of April 1 snow water equivalent (a measure of how much water is contained in the snowpack) over the western United States in a given year, one might consider, in addition to an overall intercept, latitude and longitude, such covariates as elevation, slope, aspect, average wind speed, etc., to the extent that data on these attribute variables are available. If data on potentially useful attribute variables are not readily available, the mean function often is taken to be a polynomial function of the geographic coordinates only. Such models are called *trend surface models*. For example, the first-order (planar) and second-order (quadratic) polynomial trend surface models for the mean of a two-dimensional process are respectively as follows, where $\mathbf{s} = (s_1, s_2)$:

$$\mu(\mathbf{s}; \beta) = \beta_0 + \beta_1 s_1 + \beta_2 s_2,$$
$$\mu(\mathbf{s}; \beta) = \beta_0 + \beta_1 s_1 + \beta_2 s_2 + \beta_{11} s_1^2 + \beta_{12} s_1 s_2 + \beta_{22} s_2^2.$$

Using a "full" qth-order polynomial, i.e., a polynomial that includes all pure and mixed monomials of degree $\leq q$, is recommended because this will ensure that the fitted surface is invariant to the choice of origin and orientation of the (Euclidean) coordinate system.

It is worth noting that realizations of a process with constant mean, but strong spatial correlation, frequently appear to have trends; therefore, it is generally recommended that one refrain from using trend surfaces that cannot be justified apart from examining the data.

The standard method for fitting a provisional linear mean function to geostatistical data is ordinary least squares (OLS). This method yields the OLS estimator $\hat{\beta}_{OLS}$ of β, given by

$$\hat{\beta}_{OLS} = \text{argmin} \sum_{i=1}^{n} [Y(\mathbf{s}_i) - \mathbf{X}(\mathbf{s}_i)^T \beta]^2.$$

Equivalently, $\hat{\beta}_{OLS} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ where $\mathbf{X} = [\mathbf{X}(\mathbf{s}_1), \mathbf{X}(\mathbf{s}_2), \ldots, \mathbf{X}(\mathbf{s}_n)]^T$ and $\mathbf{Y} = [Y(\mathbf{s}_1), Y(\mathbf{s}_2), \ldots, Y(\mathbf{s}_n)]^T$, it being assumed without loss of generality that $\mathbf{X}$ has full column rank. Fitted values and fitted residuals at data locations are given by $\hat{\mathbf{Y}} = \mathbf{X}^T\hat{\beta}_{OLS}$ and $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}}$, respectively. The latter are passed to the second stage of the geostatistical analysis, to be described in the next section. While still at this first stage, however, the results of the OLS fit should be evaluated and used to suggest possible alternative

mean functions. For this purpose, the standard arsenal of multiple regression methodology, such as transformations of the response, model selection, and outlier identification, may be used, but in an exploratory rather than confirmatory fashion since the independent errors assumption upon which this OLS methodology is based is likely not satisfied by the data.

As a result of the wide availability of software for fitting linear regression models, OLS fitting of a linear mean function to geostatistical data is straightforward. However, there are some practical limitations worth noting, as well as some techniques/guidelines for overcoming these limitations. First, and in particular for polynomial trend surface models, the covariates can be highly multicollinear, which causes the OLS estimators to have large variances. This is mainly a numerical problem, not a statistical one, unless the actual value of the regression coefficients is of interest and it can be solved by centering the covariates (i.e., subtracting their mean values) or, if needed, by orthogonalizing the terms in some manner prior to fitting. Second, the fitted surface in portions of the spatial domain of interest where no observations are taken may be distorted so as to better fit the observed data. This problem is avoided, however, if the sampling design has good spatial coverage. Finally, as with least squares estimation in any context, the OLS estimators are sensitive to outliers and thus one may instead wish to fit the mean function using one of many available general procedures for robust and resistant regression. If the data locations form a (possibly partially incomplete) rectangular grid, one robust alternative to OLS estimation is median polish (Cressie, 1986), which iteratively sweeps out row and column medians from the observed data (and thus is implicitly based on an assumed row–column effects model for the first-order structure). However, the notion of what constitutes an outlier can be tricky with spatially dependent data, so robust methods should be used with care.

3.4 Nonparametric Estimation of the Semivariogram

The second stage of a geostatistical analysis is to estimate the second-order dependence structure of the random process $Y(\cdot)$ from the residuals of the fitted provisional mean function. To describe this in more detail, we assume that $e(\cdot)$ is intrinsically stationary, in which case the semivariogram is the appropriate mode of description of the second-order dependence. We also assume that $d = 2$, though extensions to $d = 3$ are straightforward.

Consider first a situation in which the data locations form a regular rectangular grid. Let $\mathbf{h}_1 = \begin{pmatrix} h_{11} \\ h_{12} \end{pmatrix}, \ldots, \mathbf{h}_k = \begin{pmatrix} h_{k1} \\ h_{k2} \end{pmatrix}$ represent the distinct lags between data locations (in units of the grid spacings), with displacement angles $\phi_u = \tan^{-1}(h_{u2}/h_{u1}) \in [0, \pi)$ $(u = 1, \ldots, k)$. Attention may be restricted to only those lags with displacement angles in $[0, \pi)$ without any loss of information because $\gamma(\mathbf{h})$ is an even function. For $u = 1, \ldots, k$, let $N(\mathbf{h}_u)$ represent the number of times that lag $\mathbf{h}_u$ occurs among the data locations. Then the *empirical semivariogram* is defined as follows:

$$\hat{\gamma}(\mathbf{h}_u) = \frac{1}{2N(\mathbf{h}_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}_u} \{\hat{e}(\mathbf{s}_i) - \hat{e}(\mathbf{s}_j)\}^2 \qquad (u = 1, \ldots, k),$$

where $\hat{e}(\mathbf{s}_i)$ is the residual from the fitted provisional mean function at the ith data location and is thus the ith element of the vector $\hat{\mathbf{e}}$ defined in the previous section. We call $\hat{\gamma}(\mathbf{h}_u)$ the uth ordinate of the empirical semivariogram. Observe that $\hat{\gamma}(\mathbf{h}_u)$ is a method-of-moments type of estimator of $\gamma(\mathbf{h}_u)$. Under model (3.1) with constant mean, this estimator is unbiased; if the mean is not constant in model (3.1), the estimator is biased as a consequence of

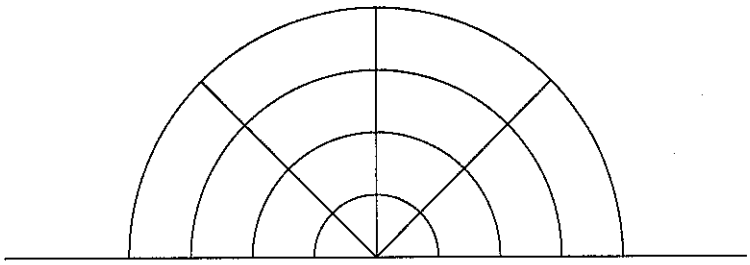

FIGURE 3.1
A polar partition of the lag space.

estimating the mean structure, but the bias is not large in practice (provided that the mean structure that is estimated is correctly specified).

When data locations are irregularly spaced, there is generally little to no replication of lags among the data locations. To obtain quasireplication of lags, we first partition the lag space $H = \{\mathbf{s} - \mathbf{t}: \mathbf{s}, \mathbf{t} \in D\}$ into lag classes or "bins" $H_1, \ldots, H_k$, say, and assign each lag with displacement angle in $[0, \pi)$ that occurs among the data locations to one of the bins. Then, we use a similar estimator:

$$\hat{\gamma}(\mathbf{h}_u) = \frac{1}{2N(H_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j \in H_u} \{\hat{e}(\mathbf{s}_i) - \hat{e}(\mathbf{s}_j)\}^2 \qquad (u = 1, \ldots, k). \tag{3.7}$$

Here $\mathbf{h}_u$ is a representative lag for the entire bin H_u, and $N(H_u)$ is the number of lags that fall into H_u. The bin representative, $\mathbf{h}_u$, is sometimes taken to be the centroid of H_u, but a much better choice is the average of all the lags that fall into H_u. The most common partition of the lag space is a "polar" partition, i.e., a partitioning into angle and distance classes, as depicted in Figure 3.1. A polar partition naturally allows for the construction and plotting of a directional empirical semivariogram, i.e., a set of empirical semivariogram ordinates corresponding to the same angle class, but different distance classes, in each of several directions. It also allows for lags to be combined over all angle classes to yield the ordinates of an omnidirectional empirical semivariogram. The polar partition of the lag space is not the only possible partition; however, some popular software for estimating semivariograms use a rectangular partition instead.

Each empirical semivariogram ordinate in the case of irregularly spaced data locations is approximately unbiased for its corresponding true semivariogram ordinate, as it is when the data locations form a regular grid, but there is an additional level of approximation or blurring in the irregularly spaced case due to the grouping of unequal lags into bins.

How many bins should be used to obtain the empirical semivariogram, and how large should they be? Clearly, there is a trade-off involved: The more bins that are used, the smaller they are and the better the lags in H_u are approximated by $\mathbf{h}_u$, but the fewer the number of observed lags belonging to H_u (with the consequence that the sampling variation of the empirical semivariogram ordinate corresponding to that lag is larger). One popular rule of thumb is to require $N(\mathbf{h}_u)$ to be at least 30 and to require the length of $\mathbf{h}_u$ to be less than half the maximum lag length among data locations. But, there may be many partitions that meet these criteria, and so the empirical semivariogram is not actually uniquely defined when data locations are irregularly spaced. Furthermore, as we shall see in the simulation below, at lags that are a substantial fraction of the dimensions of the observation domain, $\hat{\gamma}(\mathbf{h}_u)$ may be highly variable even when $N(\mathbf{h}_u)$ is much larger than 30. The problem is that the various terms making up the sum in (3.7) are not independent and the dependence can be particularly strong at larger lags.

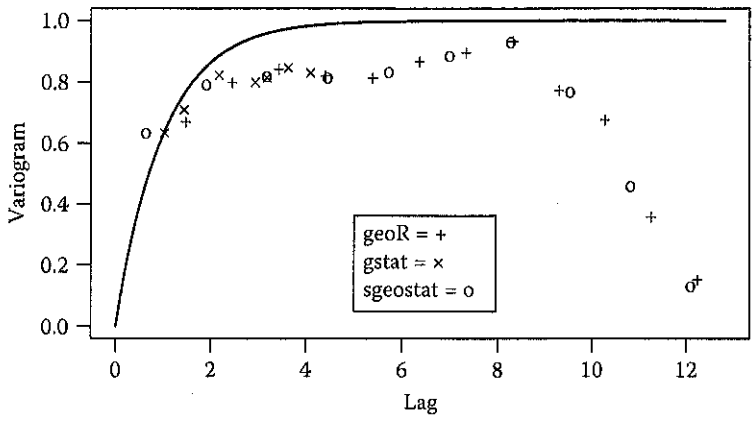

FIGURE 3.2
Empirical semivariograms of simulated data obtained via three R programs.

One undesirable feature of the empirical semivariogram is its sensitivity to outliers, a consequence of each of its ordinates being a scaled sum of squares. An alternative and more robust estimator, due to Cressie and Hawkins (1980), is

$$\bar{\gamma}(\mathbf{h}_u) = \frac{\{\frac{1}{N(H_u)} \sum_{s_i - s_j \in H_u} |\hat{e}(\mathbf{s}_i) - \hat{e}(\mathbf{s}_j)|^{1/2}\}^4}{.914 + [.988/N(H_u)]} \qquad (u = 1, \ldots, k).$$

As an example, let us consider empirical semivariograms obtained from three programs available in R with all arguments left at their default values. Specifically, we simulate an isotropic Gaussian process Y with constant mean and exponential semivariogram with sill and range parameters equal to 1 on a 10×10 square grid with distance 1 between neighboring observations. (See Section 3.5 for definitions of the exponential semivariogram and its sill and range parameters.) Figure 3.2 shows the resulting empirical semivariograms using the command `variog` from geoR, the command `est.variogram` from sgeostat, and the command `variogram` from gstat. The first two programs do not automatically impose an upper bound on the distance lags and we can see that the estimates of γ at the longer lags are very poor in this instance, even though, for example, for `est.variogram` from sgeostat, the estimate for the second longest lag (around 10.8) is based on 80 pairs of observations and the estimate for the third longest lag (aound 9.5) is based on 326 pairs. For `variogram` in gstat, the default gives a largest lag of around 4.08. Another important difference between the gstat program and the other two is that gstat, as we recommend, uses the mean distance within the bin rather than the center of the bin as the ordinate on the horizontal axis. For haphazardly sited data, the differences between the two may often be small, but here we find that for regular data, the differences can be dramatic. In particular, gstat and sgeostat give the same value for $\hat{\gamma}$ at the shortest lag (0.6361), but gstat gives the corresponding distance as 1, whereas sgeostat gives this distance as 0.6364. In fact, with either program, every pair of points used in the estimator is exactly distance 1 apart, so the sgeostat result is quite misleading. It would appear that, in this particular setting, the default empirical variogram in gstat is superior to those in geoR and sgeostat. However, even with the best of programs, one should be very careful about using default parameter values for empirical semivariograms. Furthermore, even with well-chosen bins, it is important to recognize that empirical semivariograms do not necessarily contain all of the information in the data about the true semivariogram, especially, as noted by Stein (1999, Sec. 6.2), for differentiable processes.

3.5 Modeling the Semivariogram

Next, it is standard practice to smooth the empirical semivariogram by fitting a parametric model to it. Why smooth the empirical semivariogram? There are several reasons. First, it is often quite bumpy; a smoothed version may be more reliable (have smaller variance) and therefore may increase our understanding of the nature of the spatial dependence. Second, the empirical semivariogram will often fail to be conditionally nonpositive definite, a property which must be satisfied to ensure that at the prediction stage to come, the prediction error variance is nonnegative at every point in D. Finally, prediction at arbitrary locations requires estimates of the semivariogram at lags not included among the bin representatives $\mathbf{h}_1, \ldots, \mathbf{h}_k$ nor existing among the lags between data locations, and smoothing can provide these needed estimates.

To smooth the empirical semivariogram, a valid parametric model for the semivariogram and a method for fitting that model must be chosen. The choice of model among the collection of valid semivariogram models is informed by an examination of the empirical semivariogram, of course, but other considerations (prior knowledge, computational simplicity, sufficient flexibility) may be involved as well. The following three conditions are necessary and sufficient for a semivariogram model to be valid (provided that they hold for all $\theta \in \Theta$, where Θ is the parameter space for θ):

1. Vanishing at 0, i.e., $\gamma(\mathbf{0}; \theta) = 0$
2. Evenness, i.e., $\gamma(-\mathbf{h}; \theta) = \gamma(\mathbf{h}; \theta)$ for all $\mathbf{h}$
3. Conditional negative definiteness, i.e., $\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \gamma(\mathbf{s}_i - \mathbf{s}_j; \theta) \leq 0$ for all n, all $\mathbf{s}_1, \ldots, \mathbf{s}_n$, and all $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 0$

Often, the empirical semivariogram tends to increase roughly with distance in any given direction, up to some point at least, indicating that the spatial dependence decays with distance. In other words, values of $Y(\cdot)$ at distant locations tend to be less alike than values at locations in close proximity. This leads us to consider primarily those semivariogram models that are monotone increasing functions of the intersite distance (in any given direction). Note that this is not a requirement for validity, however. Moreover, the modeling of the semivariogram is made easier if isotropy can be assumed. The degree to which this assumption is tenable has sometimes been assessed informally via "rose diagrams" (Isaaks and Srivastava, 1989) or by comparing directional empirical semivariograms. It is necessary to make comparisons in at least three, and preferably more, directions so that geometric anisotropy can be distinguished from isotropy. Moreover, without some effort to attach uncertainty estimates to semivariogram ordinates, we consider it dangerous to assess isotropy based on visual comparisons of directional empirical semivariograms. Specifically, directional empirical semivariograms for data simulated from an isotropic model can appear to show clear anisotropies (e.g., the semivariogram in one direction being consistently higher than in another direction) that are due merely to random variation and the strong correlations that occur between estimated semivariogram ordinates at different lags. More formal tests for isotropy have recently been developed; see Guan, Sherman, and Calvin (2004).

A large variety of models satisfy the three aforementioned validity requirements (in R^2 and R^3), plus monotonicity and isotropy, but the following five appear to be the most commonly used:

- *Spherical*

$$\gamma(h; \theta) = \begin{cases} \theta_1 \left(\frac{3h}{2\theta_2} - \frac{h^3}{2\theta_2^3} \right) & \text{for } 0 \leq h \leq \theta_2 \\ \theta_1 & \text{for } h > \theta_2 \end{cases}$$

- *Exponential*

$$\gamma(h;\boldsymbol{\theta}) = \theta_1\{1 - \exp(-h/\theta_2)\}$$

- *Gaussian*

$$\gamma(h;\boldsymbol{\theta}) = \theta_1\left\{1 - \exp\left(-h^2/\theta_2^2\right)\right\}$$

- *Matérn*

$$\gamma(h;\boldsymbol{\theta}) = \theta_1\left(1 - \frac{(h/\theta_2)^\nu \mathcal{K}_\nu(h/\theta_2)}{2^{\nu-1}\Gamma(\nu)}\right)$$

  where $\mathcal{K}_\nu(\cdot)$ is the modified Bessel function of the second kind of order ν

- *Power*

$$\gamma(h;\boldsymbol{\theta}) = \theta_1 h^{\theta_2}$$

These models are displayed in Figure 3.3. For each model, θ_1 is positive; similarly, θ_2 is positive in each model except the power model, for which it must satisfy $0 \le \theta_2 < 2$. In the Matérn model, $\nu > 0$. It can be shown that the Matérn models with $\nu = 0.5$ and $\nu \to \infty$ coincide with the exponential and Gaussian models, respectively.

Several attributes of an isotropic semivariogram model are sufficiently important to single out. The *sill* of $\gamma(h;\boldsymbol{\theta})$ is defined as $\lim_{h\to\infty}\gamma(h;\boldsymbol{\theta})$ provided that the limit exists. If this limit exists, then the process is not only intrinsically stationary, but also second-order stationary, and $C(\mathbf{0};\boldsymbol{\theta})$ coincides with the sill. Note that the spherical, exponential, Gaussian, and Matérn models have sills (equal to θ_1 in each of the parameterizations given above), but the power model does not. Furthermore, if the sill exists, then the *range* of $\gamma(h;\boldsymbol{\theta})$ is the smallest value of h for which $\gamma(h;\boldsymbol{\theta})$ equals its sill, if such a value exists. If the range does not exist, there is a related notion of an *effective range*, defined as the smallest value of h for which $\gamma(h;\boldsymbol{\theta})$ is equal to 95% of its sill; in this case, the effective range is often a function of a single parameter called the *range parameter*. Of those models listed above that have a sill, only the spherical has a range (equal to θ_2); however, the exponential and Gaussian models have effective ranges of approximately $3\theta_2$ and $\sqrt{3}\theta_2$, respectively, with θ_2 then being the range parameter. Range parameters can be difficult to estimate even with quite large datasets, in particular when, as is often the case, the range is not much smaller than the dimensions of the observation region (see Chapter 6). This difficulty is perhaps an argument for using the power class of variograms, which is essentially the Matérn class for $\nu < 1$ with the range set to infinity, thus, avoiding the need to estimate a range.

The Matérn model has an additional parameter ν known as the *smoothness parameter*, as the process $Y(\cdot)$ is m times mean square differentiable if and only if $\nu > m$. The smoothness of the semivariogram near the origin (i.e., at small lags) is a key attribute for efficient spatial prediction (Stein, 1988; Stein and Handcock, 1989). Finally, the *nugget effect* of $\gamma(h;\boldsymbol{\theta})$ is defined as $\lim_{h\to 0}\gamma(h;\boldsymbol{\theta})$. The nugget effect is zero for all the models listed above, but a nonzero nugget effect can be added to any of them. For example, the exponential model with nugget effect θ_3 is given by

$$\gamma(h;\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } h = 0 \\ \theta_3 + \theta_1\{1 - \exp(-h/\theta_2)\} & \text{if } h > 0. \end{cases} \tag{3.8}$$

One rationale for the nugget effect can be given in terms of the measurement error model (3.5). If $\eta(\cdot)$ in that model is intrinsically stationary and mean square continuous with a nuggetless exponential semivariogram, if $\epsilon(\cdot)$ is an iid (white noise) measurement error process with variance θ_3, and if $\eta(\cdot)$ and $\epsilon(\cdot)$ are independent, then the semivariogram of $e(\cdot)$ will coincide with (3.8).

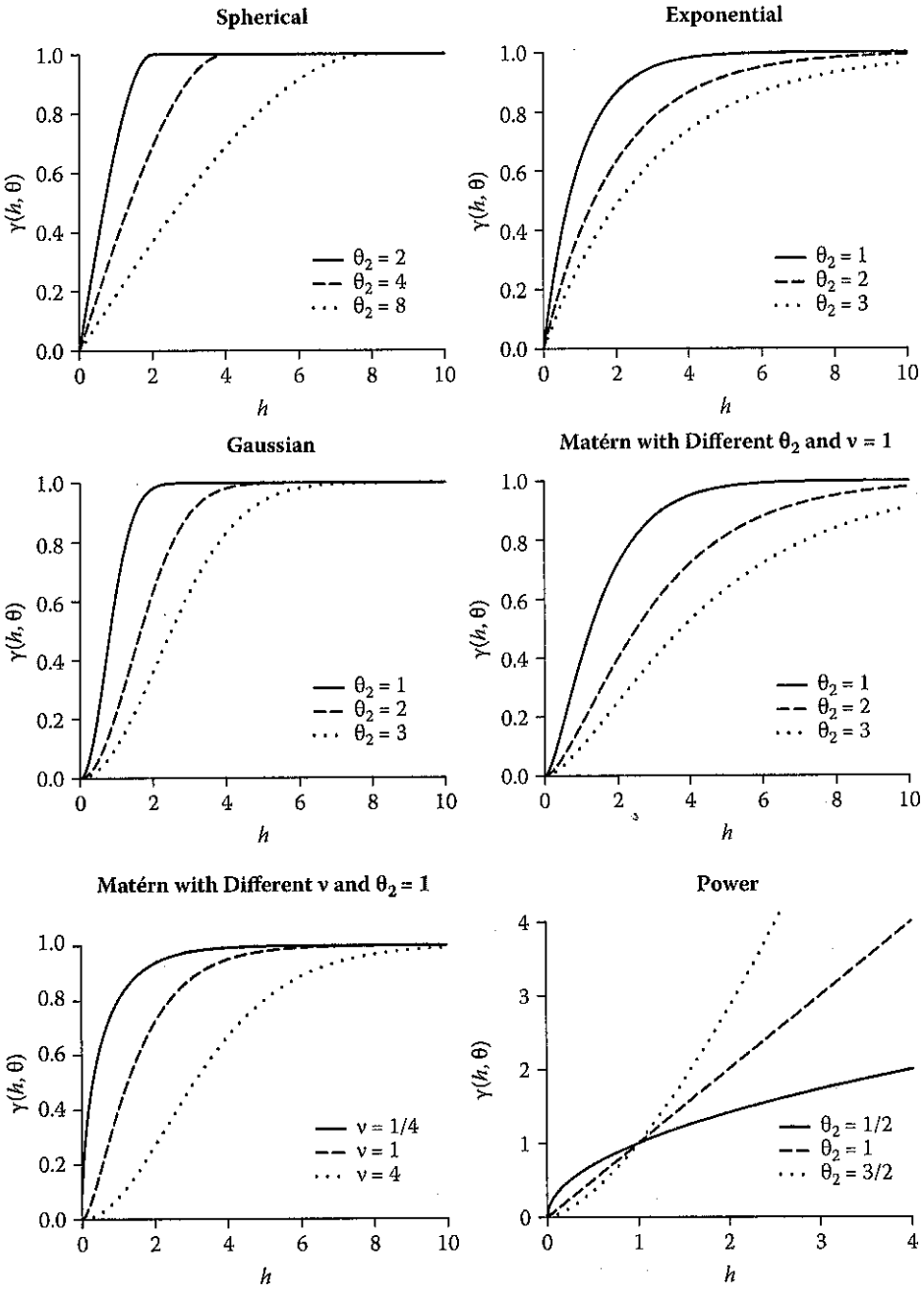

FIGURE 3.3
Semivariogram models.

Gaussian semivariograms correspond to processes that are extremely smooth—too much so to generally serve as good models for natural processes. For differentiable spatial processes, a Matérn model with $\nu > 1$, but not very large, is generally preferable. However, if one has an underlying smooth process with a sufficiently large nugget effect, it may sometimes not matter much whether one uses a Gaussian or Matérn model. Spherical semivariograms are very popular in the geostatistical community, but less so among statisticians, in part because the semivariogram is only once differentiable in θ_2 at $\theta_2 = h$,

which leads to rather odd looking likelihood functions for the unknown parameters. There can be computational advantages to using semivariograms with a finite range if this range is substantially smaller than the dimensions of the observation domain, but even if one wants to use a semivariogram with finite range for computational reasons, there may be better alternatives than the spherical semivariogram (Furrer, Genton, and Nychka, 2006).

Any valid isotropic semivariogram model can be generalized to make it geometrically anisotropic, simply by replacing the argument h with $(\mathbf{h}'\mathbf{A}\mathbf{h})^{1/2}$, where $\mathbf{A}$ is a $d \times d$ positive definite matrix of parameters. For example, a geometrically anisotropic exponential semivariogram in R^2 is given by

$$\gamma(\mathbf{h};\boldsymbol{\theta}) = \theta_1 \left\{1 - \exp\left[-\left(h_1^2 + 2\theta_3 h_1 h_2 + \theta_4 h_2^2\right)^{1/2} / \theta_2^2\right]\right\}.$$

Thus, for example, if $\theta_3 = 0$ and $\theta_4 = 4$, the effective range of the spatial correlation is twice as large in the E–W direction as in the N–S direction, and the effective range in all other directions is intermediate between these two. The isotropic exponential semivariogram corresponds to the special case in which $\theta_3 = 0$, $\theta_4 = 1$. Anisotropic models that are not geometrically anisotropic—so-called zonally anisotropic models—have sometimes been used, but they are problematic, both theoretically and practically (see Zimmerman (1993)).

Two main procedures for estimating the parameters of a chosen semivariogram model have emerged: weighted least squares (WLS) and maximum likelihood (ML) or its variant, restricted (or residual) maximum likelihood (REML). The WLS approach is very popular among practitioners due to its relative simplicity, but, because it is not based on an underlying probabilistic model for the spatial process, it is suboptimal and does not rest on as firm a theoretical footing as the likelihood-based approaches (though it is known to yield consistent and asymptotically normal estimators under certain regularity conditions and certain asymptotic frameworks) (see Lahiri, Lee, and Cressie (2002)). Nevertheless, at least for nondifferentiable processes, its performance is not greatly inferior to those that are likelihood-based (Zimmerman and Zimmerman, 1991; Lark, 2000). The remainder of this section describes the WLS approach only; likelihood-based approaches are the topic of the next chapter.

The WLS estimator of $\boldsymbol{\theta}$ in the parametric model $\gamma(\mathbf{h};\boldsymbol{\theta})$ is given by

$$\hat{\boldsymbol{\theta}} = \operatorname{argmin} \sum_{u \in \mathcal{U}} \frac{N(\mathbf{h}_u)}{[\gamma(\mathbf{h}_u;\boldsymbol{\theta})]^2} [\hat{\gamma}(\mathbf{h}_u) - \gamma(\mathbf{h}_u;\boldsymbol{\theta})]^2 \tag{3.9}$$

where all quantities are defined as in the previous section. Observe that the weights, $N(\mathbf{h}_u)/[\gamma(\mathbf{h}_u;\boldsymbol{\theta})]^2$, are small if either $N(\mathbf{h}_u)$ is small or $\gamma(\mathbf{h}_u;\boldsymbol{\theta})$ is large. This has the effect, for the most commonly used semivariogram models (which are monotone increasing) and for typical spatial configurations of observations, of assigning relatively less weight to ordinates of the empirical semivariogram corresponding to large lags. For further details on the rationale for these weights, see Cressie (1985), although the argument is based on an assumption of independence between the terms in the sum (3.9), so it may tend to give too much weight to larger lags. Since the weights depend on $\boldsymbol{\theta}$, the WLS estimator must be obtained iteratively, updating the weights on each iteration until convergence is deemed to have occurred.

Comparisons of two or more fitted semivariogram models are usually made rather informally. If the models are non-nested and have the same number of parameters (e.g., the spherical and exponential models, with nuggets), the minimized weighted residual

sum of squares (the quantity minimized in (3.9)) might be used to choose from among the competing models. However, we are unaware of any good statistical arguments for such a procedure and, indeed, Stein (1999) argues that an overly strong emphasis on making parametric estimates of semivariograms match the empirical semivariogram represents a serious flaw in classical geostatistics.

3.6 Reestimation of the Mean Function

Having estimated the second-order dependence structure of the random process, there are two tacks the geostatistical analysis may take next. If the analyst has no particular interest in estimating the effects of covariates on $Y(\cdot)$, then he/she may proceed directly to kriging, as described in the next section. If the analyst has such an interest, however, the next stage is to estimate the mean function again, but this time accounting for the second-order dependence structure. The estimation approach of choice in classical geostatistics is estimated generalized least squares (EGLS), which is essentially the same as generalized least squares (GLS) except that the variances and covariances of the elements of $\mathbf{Y}$, which are assumed known for GLS, are replaced by estimates. Note that second-order stationarity, not merely intrinsic stationarity, of $e(\cdot)$ must be assumed here to ensure that these variances and covariances exist and are functions of lag only.

A sensible method for estimating the variances and covariances, and one which yields a positive definite estimated covariance matrix, is as follows. First, estimate the common variance of the $Y(\mathbf{s}_i)$s by the sill of the fitted semivariogram model, $\gamma(\mathbf{h}; \hat{\theta})$, obtained at the previous stage; denote this estimated variance by $\hat{C}(\mathbf{0})$. Then, motivated by (3.4), estimate the covariance between $Y(\mathbf{s}_i)$ and $Y(\mathbf{s}_j)$ for $i \neq j$ as $\hat{C}(\mathbf{s}_i - \mathbf{s}_j) = \hat{C}(\mathbf{0}) - \gamma(\mathbf{s}_i - \mathbf{s}_j; \hat{\theta})$. These estimated variances and covariances may then be arranged appropriately to form an estimated variance–covariance matrix

$$\hat{\Sigma} = (\hat{C}(\mathbf{s}_i - \mathbf{s}_j)).$$

The EGLS estimator of β, $\hat{\beta}_{EGLS}$ is then given by

$$\hat{\beta}_{EGLS} = (\mathbf{X}^T \hat{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^T \hat{\Sigma}^{-1} \mathbf{Y}.$$

The sampling distribution of $\hat{\beta}_{EGLS}$ is much more complicated than that of the OLS or GLS estimator. It is known, however, that $\hat{\beta}_{EGLS}$ is unbiased under very mild conditions, and that, if the process is Gaussian, the variance of $\hat{\beta}_{EGLS}$ is larger than that of the GLS estimator were θ to be known, i.e., larger than $(\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}$ (Harville, 1985). (Here, by "larger," we mean that the difference, $\text{var}(\hat{\beta}_{EGLS}) - (\mathbf{X}^T \Sigma^{-1} \mathbf{X})^{-1}$, is nonnegative definite.) Nevertheless, for lack of a simple satisfactory alternative, the variance of $\hat{\beta}_{EGLS}$ is usually estimated by the plug-in estimator, $(\mathbf{X}^T \hat{\Sigma}^{-1} \mathbf{X})^{-1}$.

If desired, the EGLS residuals, $\mathbf{Y} - \mathbf{X}\hat{\beta}_{EGLS}$, may be computed and the semivariogram reestimated from them. One may even iterate between mean estimation and semivariogram estimation several times, but, in practice, this procedure usually stops with the first EGLS fit. REML, described in Chapter 4, avoids this problem by estimating θ using only linear combinations of the observations whose distributions do not depend on β.

3.7 Kriging

The final stage of a classical geostatistical analysis is to predict the values of $Y(\cdot)$ at desired locations, perhaps even at all points, in D. Methods dedicated to this purpose are called *kriging*, after the South African mining engineer D. G. Krige, who was the first to develop and apply them. Krige's original method, now called ordinary kriging, was based on the special case of model (3.1) in which the mean is assumed to be constant. Here, we describe the more general method of universal kriging, which is identical to best linear unbiased prediction under model (3.1) with mean function assumed to be of the linear form (3.6).

Let $\mathbf{s}_0$ denote an arbitrary location in D; usually this will be an unsampled location, but it need not be. Consider the prediction of $Y(\mathbf{s}_0)$ by a predictor, $\hat{Y}(\mathbf{s}_0)$, that minimizes the prediction error variance, $\text{var}[\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0)]$, among all predictors satisfying the following two properties:

1. Linearity, i.e., $\hat{Y}(\mathbf{s}_0) = \boldsymbol{\lambda}^T\mathbf{Y}$, where $\boldsymbol{\lambda}$ is a vector of fixed constants
2. Unbiasedness, i.e., $E[\hat{Y}(\mathbf{s}_0)] = E[Y(\mathbf{s}_0)]$, or equivalently $\boldsymbol{\lambda}^T\mathbf{X} = \mathbf{x}(\mathbf{s}_0)$

Suppose for the moment that the semivariogram of $Y(\cdot)$ is known. Then the solution to this constrained minimization problem, known as the universal kriging predictor of $Y(\mathbf{s}_0)$, is given by

$$\hat{Y}(\mathbf{s}_0) = [\boldsymbol{\gamma} + \mathbf{X}(\mathbf{X}^T\boldsymbol{\Gamma}^{-1}\mathbf{X})^{-1}(\mathbf{x}_0 - \mathbf{X}^T\boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma})]^T\boldsymbol{\Gamma}^{-1}\mathbf{Y}, \tag{3.10}$$

where $\boldsymbol{\gamma} = [\gamma(\mathbf{s}_1 - \mathbf{s}_0), \ldots, \gamma(\mathbf{s}_n - \mathbf{s}_0)]^T$, $\boldsymbol{\Gamma}$ is the $n \times n$ symmetric matrix with ijth element $\gamma(\mathbf{s}_i - \mathbf{s}_j)$ and $\mathbf{x}_0 = \mathbf{x}(\mathbf{s}_0)$. This result may be obtained using differential calculus and the method of Lagrange multipliers. However, a geometric proof is more instructive and following is an example.

Let us assume that the first component of $\mathbf{x}(\mathbf{s})$ is identically 1, which guarantees that the error of any linear predictor of $Y(\mathbf{s}_0)$ that satisfies the unbiasedness constraint is a contrast, so that its variance can be obtained from the semivariogram of $Y(\cdot)$. Let us also assume that there exists a linear predictor satisfying the unbiasedness constraint. Suppose $\boldsymbol{\lambda}^T\mathbf{Y}$ is such a predictor. Consider any other such predictor $\boldsymbol{\nu}^T\mathbf{Y}$ and set $\boldsymbol{\mu} = \boldsymbol{\nu} - \boldsymbol{\lambda}$. Since $E(\boldsymbol{\lambda}^T\mathbf{Y}) = E(\boldsymbol{\nu}^T\mathbf{Y})$ for all $\boldsymbol{\beta}$, we must have $\mathbf{X}^T\boldsymbol{\mu} = \mathbf{0}$. And,

$$\begin{aligned}
\text{var}\{\boldsymbol{\nu}^T\mathbf{Y} - Y(\mathbf{s}_0)\} &= \text{var}[\boldsymbol{\mu}^T\mathbf{Y} + \{\boldsymbol{\lambda}^T\mathbf{Y} - Y(\mathbf{s}_0)\}] \\
&= \text{var}(\boldsymbol{\mu}^T\mathbf{Y}) + \text{var}\{\boldsymbol{\lambda}^T\mathbf{Y} - Y(\mathbf{s}_0)\} + 2\,\text{Cov}\{\boldsymbol{\mu}^T\mathbf{Y}, \boldsymbol{\lambda}^T\mathbf{Y} - Y(\mathbf{s}_0)\} \\
&\geq \text{var}\{\boldsymbol{\lambda}^T\mathbf{Y} - Y(\mathbf{s}_0)\} + 2\,\text{Cov}\{\boldsymbol{\mu}^T\mathbf{Y}, \boldsymbol{\lambda}^T\mathbf{Y} - Y(\mathbf{s}_0)\} \\
&= \text{var}\{\boldsymbol{\lambda}^T\mathbf{Y} - Y(\mathbf{s}_0)\} + 2\boldsymbol{\mu}^T(-\boldsymbol{\Gamma}\boldsymbol{\lambda} + \boldsymbol{\gamma}).
\end{aligned}$$

If we can choose $\boldsymbol{\lambda}$ such that $\boldsymbol{\mu}^T(-\boldsymbol{\Gamma}\boldsymbol{\lambda} + \boldsymbol{\gamma}) = 0$ for all $\boldsymbol{\mu}$ satisfying $\mathbf{X}^T\boldsymbol{\mu} = \mathbf{0}$, then $\boldsymbol{\lambda}$ is the solution we seek, since we then have $\text{var}\{\boldsymbol{\nu}^T\mathbf{Y} - Y(\mathbf{s}_0)\} \geq \text{var}\{\boldsymbol{\lambda}^T\mathbf{Y} - Y(\mathbf{s}_0)\}$ for any predictor $\boldsymbol{\nu}^T\mathbf{Y}$ satisfying the unbiasedness constraint. But, since the column space of $\mathbf{X}$ is the orthogonal complement of its left null space, this condition holds if and only if $-\boldsymbol{\Gamma}\boldsymbol{\lambda} + \boldsymbol{\gamma}$ is in the column space of $\mathbf{X}$, which is equivalent to the existence of a vector $\boldsymbol{\alpha}$ satisfying $-\boldsymbol{\Gamma}\boldsymbol{\lambda} + \mathbf{X}\boldsymbol{\alpha} = -\boldsymbol{\gamma}$. Putting this condition together with the unbiasedness constraint yields the system of linear equations for $\boldsymbol{\lambda}$ and $\boldsymbol{\alpha}$

$$\begin{pmatrix} -\boldsymbol{\Gamma} & \mathbf{X} \\ \mathbf{X}^T & \mathbf{O} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\alpha} \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\gamma} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ and $\mathbf{O}$ indicate a vector and a matrix of zeroes, respectively. If Γ is invertible and $\mathbf{X}$ is of full rank, then simple row reductions yields λ as in (3.10).

The minimized value of the prediction error variance is called the (universal) kriging variance and is given by

$$\sigma^2(\mathbf{s}_0) = \gamma^T \Gamma^{-1}\gamma - (\mathbf{X}^T\Gamma^{-1}\gamma - \mathbf{x}_0)^T(\mathbf{X}^T\Gamma^{-1}\mathbf{X})^{-1}(\mathbf{X}^T\Gamma^{-1}\gamma - \mathbf{x}_0). \tag{3.11}$$

The universal kriging predictor is an example of the best linear unbiased predictor, or BLUP, as it is generally abbreviated. If $Y(\cdot)$ is Gaussian, the kriging variance can be used to construct a nominal $100(1-\alpha)\%$ prediction interval for $Y(\mathbf{s}_0)$, which is given by

$$\hat{Y}(\mathbf{s}_0) \pm z_{\alpha/2}\sigma(\mathbf{s}_0),$$

where $0 < \alpha < 1$ and $z_{\alpha/2}$ is the upper $\alpha/2$ percentage point of a standard normal distribution. If $Y(\cdot)$ is Gaussian and $\gamma(\cdot)$ is known, then $\hat{Y}(\mathbf{s}_0) - Y(\mathbf{s}_0)$ is normally distributed and the coverage probability of this interval is exactly $1-\alpha$.

If the covariance function for Y exists and $\sigma = [C(\mathbf{s}_1 - \mathbf{s}_0), \ldots, C(\mathbf{s}_n - \mathbf{s}_0)]^T$, then the formula for the universal kriging predictor (3.10) holds with γ replaced by σ and Γ by Σ. It is worthwhile to compare this formula to that for the best (minimum mean squared error) linear predictor when β is known: $\mathbf{x}_0^T\beta + \sigma^T\Sigma^{-1}(\mathbf{Y} - \mathbf{X}\beta)$. A straightforward calculation shows that the universal kriging predictor is of this form with β replaced by $\hat{\beta}_{GLS}$. Furthermore, the expression (3.11) for the kriging variance is replaced by

$$C(0) - \sigma^T\Sigma^{-1}\sigma + (\mathbf{x}_0 - \mathbf{X}^T\Sigma^{-1}\sigma)^T(\mathbf{X}^T\Sigma^{-1}\mathbf{X})^{-1}(\mathbf{x}_0 - \mathbf{X}^T\Sigma^{-1}\sigma).$$

The first two terms, $C(\mathbf{0}) - \sigma^T\Sigma^{-1}\sigma$, correspond to the mean squared error of the best linear predictor, so that the last term, which is always nonnegative, is the penalty for having to estimate β.

In practice, two modifications are usually made to the universal kriging procedure just described. First, to reduce the amount of computation required, the prediction of $Y(\mathbf{s}_0)$ may be based not on the entire data vector $\mathbf{Y}$, but on only those observations that lie in a specified neighborhood around $\mathbf{s}_0$. The range, the nugget-to-sill ratio, and the spatial configuration of data locations are important factors in choosing this neighborhood (for further details, see Cressie (1991, Sec. 3.2.1)). Generally speaking, larger nuggets require larger neighborhoods to obtain nearly optimal predictors. However, there is no simple relationship between the range and the neighborhood size. For example, Brownian motion is a process with no finite range for which the kriging predictor is based on just the two nearest neighbors. Conversely, there are processes with finite ranges for which observations beyond the range play a nontrivial role in the kriging predictor (Stein, 1999, p. 67). When a spatial neighborhood is used, the formulas for the universal kriging predictor and its associated kriging variance are of the same form as (3.10) and (3.11), but with γ and $\mathbf{Y}$ replaced by the subvectors, and Γ and $\mathbf{X}$ replaced by the submatrices, corresponding to the neighborhood.

The second modification reckons with the fact that the semivariogram that appears in (3.10) and (3.11) is in reality unknown. It is common practice to substitute $\hat{\gamma} = \gamma(\hat{\theta})$ and $\hat{\Gamma} = \Gamma(\hat{\theta})$ for γ and Γ in (3.10) and (3.11), where $\hat{\theta}$ is an estimate of θ obtained by, say, WLS. The resulting *empirical* universal kriging predictor is no longer a linear function of the data, but remarkably it remains unbiased under quite mild conditions (Kackar and Harville, 1981). The empirical kriging variance tends to underestimate the actual prediction error

variance of the empirical universal kriging predictor because it does not account for the additional error incurred by estimating θ. Zimmerman and Cressie (1992) give a modified estimator of the prediction error variance of the empirical universal kriging predictor, which performs well when the spatial dependence is not too strong. However, Bayesian methods are arguably a more satisfactory approach for dealing with the uncertainty of spatial dependence parameters in prediction (see Handcock and Stein (1993)). Another possibility is to estimate the prediction error variance via a parametric bootstrap (Sjöstedt-de Luna and Young, 2003).

Universal kriging yields a predictor that is a "location estimator" of the conditional distribution of $Y(\mathbf{s}_0)$ given $\mathbf{Y}$; indeed, if the error process $e(\cdot)$ is Gaussian, the universal kriging predictor coincides with the conditional mean, $E(Y(\mathbf{s}_0)|\mathbf{Y})$ (assuming $\gamma(\cdot)$ is known and putting a flat improper prior on any mean parameters). If the error process is non-Gaussian, then generally the optimal predictor, the conditional mean, is a nonlinear function of the observed data. Variants, such as disjunctive kriging and indicator kriging, have been developed for spatial prediction of conditional means or conditional probabilities for non-Gaussian processes (see Cressie, 1991, pp. 278–283), but we are not keen about them, as the first is based upon strong, difficult to verify assumptions and the second tends to yield unstable estimates of conditional probabilities. In our view, if the process appears to be badly non-Gaussian and a transformation doesn't make it sufficiently Gaussian, then the analyst should "bite the bullet" and develop a decent non-Gaussian model for the data.

The foregoing has considered *point kriging*, i.e., prediction at a single point. Sometimes a block kriging predictor, i.e., a predictor of the average value $Y(B) \equiv \int_B Y(\mathbf{s})d\mathbf{s}/|B|$ over a region (block) $B \subset D$ of positive d-dimensional volume $|B|$ is desired, rather than predictors of $Y(\cdot)$ at individual points. Historically, for example, mining engineers were interested in this because the economics of mining required the extraction of material in relatively large blocks. Expressions for the universal block kriging predictor of $Y(B)$ and its associated kriging variance are identical to (3.10) and (3.11), respectively, but with $\gamma = [\gamma(B, \mathbf{s}_1), \ldots, \gamma(B, \mathbf{s}_n)]^T$, $\mathbf{x}_0 = [X_1(B), \ldots, X_p(B)]^T$ (where p is the number of covariates in the linear mean function), $\gamma(B, \mathbf{s}_i) = |B|^{-1} \int_B \gamma(\mathbf{u} - \mathbf{s}_i)\, d\mathbf{u}$ and $X_j(B) = |B|^{-1} \int_B X_j(\mathbf{u})\, d\mathbf{u}$.

Throughout this chapter, it was assumed that a single spatially distributed variable, namely $Y(\cdot)$, was of interest. In some situations, however, there may be two or more variables of interest, and the analyst may wish to study how these variables co-vary across the spatial domain and/or predict their values at unsampled locations. These problems can be handled by a multivariate generalization of the univariate geostatistical approach we have described. In this multivariate approach, $\{\mathbf{Y}(\mathbf{s}) \equiv [Y_1(\mathbf{s}), \ldots, Y_m(\mathbf{s})]^T : \mathbf{s} \in D\}$ represents the m-variate spatial process of interest and a model $\mathbf{Y}(\mathbf{s}) = \boldsymbol{\mu}(\mathbf{s}) + \mathbf{e}(\mathbf{s})$ analogous to (3.1) is adopted in which the second-order variation is characterized by either a set of m semivariograms and $m(m-1)/2$ cross-semivariograms $\gamma_{ij}(\mathbf{h}) = \frac{1}{2}\text{var}[Y_i(\mathbf{s}) - Y_j(\mathbf{s}+\mathbf{h})]$, or a set of m covariance functions and $m(m-1)/2$ cross-covariance functions $C_{ij}(\mathbf{h}) = \text{Cov}[Y_i(\mathbf{s}), Y_j(\mathbf{s}+\mathbf{h})]$, depending on whether intrinsic or second-order stationarity is assumed. These functions can be estimated and fitted in a manner analogous to what we described for univariate geostatistics; likewise, the best (in a certain sense) linear unbiased predictor of $\mathbf{Y}(\mathbf{s}_0)$ at an arbitrary location $\mathbf{s}_0 \in D$, based on observed values $\mathbf{Y}(\mathbf{s}_1), \ldots, \mathbf{Y}(\mathbf{s}_n)$, can be obtained by an extension of kriging known as cokriging. Good sources for further details are Ver Hoef and Cressie (1993) and Chapter 27 in this book.

While we are strong supporters of the general geostatistical framework to analyzing spatial data, we have, as we have indicated, a number of concerns about common geostatistical practices. For a presentation of geostatistics from the perspective of "geostatisticians" (that is, researchers who can trace their lineage to Georges Matheron and the French School of Geostatistics), we recommend the book by Chilès and Delfiner (1999).

References

Chilès, J.-P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty.* New York: John Wiley & Sons.

Cressie, N. (1985). Fitting variogram models by weighted least squares. *Journal of the International Association for Mathematical Geology*, **17**, 563–586.

Cressie, N. (1986). Kriging nonstationary data. *Journal of the American Statistical Association*, 81, 625–634.

Cressie, N. (1991). *Statistics for Spatial Data.* New York: John Wiley & Sons.

Cressie, N. and Hawkins, D.M. (1980). Robust estimation of the variogram, I. *Journal of the International Association for Mathematical Geology*, **12**, 115–125.

Furrer, R., Genton, M.G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523.

Guan, Y., Sherman, M., and Calvin, J.A. (2004). A nonparametric test for spatial isotropy using subsampling. *Journal of the American Statistical Association*, **99**, 810–821.

Handcock, M.S. and Stein, M.L. (1993). A Bayesian analysis of kriging. *Technometrics*, **35**, 403–410.

Harville, D.A. (1985). Decomposition of prediction error. *Journal of the American Statistical Association*, **80**, 132–138.

Im, H.K., Stein, M.L., and Zhu, Z. (2007). Semiparametric estimation of spectral density with irregular observations. *Journal of the American Statistical Association*, **102**, 726–735.

Isaaks, E.H. and Srivastava, R.M. (1989). *An Introduction to Applied Geostatistics.* London: Academic Press.

Kackar, R.N. and Harville, D.A. (1981). Unbiasedness of two-stage estimation and prediction procedures for mixed linear models. *Communications in Statistics—Theory and Methods*, **10**, 1249–1261.

Lahiri, S.N., Lee, Y., and Cressie, N. (2002). On asymptotic distribution and asymptotic efficiency of least squares estimators of spatial variogram parameters. *Journal of Statistical Planning and Inference*, **103**, 65–85.

Lark, R.M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. *European Journal of Soil Science*, **51**, 717–728.

Sjöstedt-de Luna, S. and Young, A. (2003). The bootstrap and kriging prediction intervals. *Scandinavian Journal of Statistics*, **30**, 175–192.

Stein, M.L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Annals of Statistics*, **16**, 55–63.

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging.* New York: Springer.

Stein, M.L. and Handcock, M.S. (1989). Some asymptotic properties of kriging when the covariance function is misspecified. *Mathematical Geology*, **21**, 171–190.

Ver Hoef, J.M. and Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology*, **25**, 219–240.

Wahba, G. (1990). *Spline Models for Observational Data.* Philadelphia: SIAM.

Zimmerman, D.L. (1993). Another look at anisotropy in geostatistics. *Mathematical Geology*, **25**, 453–470.

Zimmerman, D.L. and Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Annals of the Institute of Statistical Mathematics*, **44**, 27–43.

Zimmerman, D.L. and Zimmerman, M.B. (1991). A comparison of spatial semivariogram estimators and corresponding ordinary kriging predictors. *Technometrics*, **33**, 77–91.