3
Gaussian models for geostatistical data

Gaussian stochastic processes are widely used in practice as models for geostatistical data. These models rarely have any physical justification. Rather, they are used as convenient empirical models which can capture a wide range of spatial behaviour according to the specification of their correlation structure. Historically, one very good reason for concentrating on Gaussian models was that they are uniquely tractable as models for dependent data. With the increasing use of computationally intensive methods, and in particular of simulation-based methods of inference, the analytic tractability of Gaussian models is becoming a less compelling reason to use them. Nevertheless, it is still convenient to work within a standard model class in routine applications. The scope of the Gaussian model class can be extended by using a transformation of the original response variable, and with this extra flexibility the model often provides a good empirical fit to data. Also, within the specific context of geostatistics, the Gaussian assumption is the model-based counterpart of some widely used geostatistical prediction methods, including simple, ordinary and universal kriging (Journel and Huijbregts, 1978; Chilès and Delfiner, 1999). We shall use the Gaussian model initially as a model in its own right for geostatistical data with a continuously varying response, and later as an important component of a hierarchically specified generalised linear model for geostatistical data with a discrete response variable, as previously discussed in Section 1.4.

3.1 Covariance functions and the variogram

A *Gaussian spatial process*, $\{S(x) : x \in \mathbb{R}^2\}$, is a stochastic process with the property that for any collection of locations $x_1, \ldots, x_n$ with each $x_i \in \mathbb{R}^2$,

the joint distribution of $S = \{S(x_1), \ldots, S(x_n)\}$ is multivariate Gaussian. Any process of this kind is completely specified by its *mean function*, $\mu(x) = \mathrm{E}[S(x)]$, and its *covariance function*, $\gamma(x, x') = \mathrm{Cov}\{S(x), S(x')\}$.

In any such process, consider an arbitrary set of locations $x_1, \ldots, x_n$, define $S = \{S(x_1), \ldots, S(x_n)\}$, write μ_S for the n-element vector with elements $\mu(x_i)$ and G for the $n \times n$ matrix with elements $G_{ij} = \gamma(x_i, x_j)$. Then, S follows a multivariate Gaussian distribution with mean vector μ_S and covariance matrix G. We write this as $S \sim \mathrm{MVN}(\mu_S, G)$.

Now, let $T = \sum_{i=1}^{n} a_i S(x_i)$. Then T is univariate Gaussian with mean $\mu_T = \sum_{i=1}^{n} a_i \mu(x_i)$ and variance

$$\sigma_T^2 = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j G_{ij} = a'Ga,$$

where $a = (a_1, \ldots, a_n)$. It must therefore be the case that $a'Ga \geq 0$. This condition, which must hold for all choices of n, $(x_1, \ldots, x_n)$ and $(a_1, \ldots, a_n)$ constrains G to be a *positive definite matrix*, and the corresponding $\gamma(\cdot)$ to be a *positive definite function*. Conversely, any positive definite function $\gamma(\cdot)$ is a legitimate covariance function for a spatial Gaussian process.

A spatial Gaussian process is *stationary* if $\mu(x) = \mu$, a constant for all x, and $\gamma(x, x') = \gamma(u)$, where $u = x - x'$ i.e., the covariance depends only on the vector difference between x and x'. Additionally, a stationary process is *isotropic* if $\gamma(u) = \gamma(||u||)$, where $||\cdot||$ denotes Euclidean distance i.e., the covariance between values of $S(x)$ at any two locations depends only on the distance between them. Note that the variance of a stationary process is a constant, $\sigma^2 = \gamma(0)$. We then define the *correlation function* to be $\rho(u) = \gamma(u)/\sigma^2$. The correlation function is symmetric in u i.e., $\rho(-u) = \rho(u)$. This follows from the fact that for any u, $\mathrm{Corr}\{S(x), S(x-u)\} = \mathrm{Corr}\{S(x-u), S(u)\} = \mathrm{Corr}\{S(x), S(x+u)\}$, the second equality following from the stationarity of $S(x)$. Hence, $\rho(u) = \rho(-u)$. From now on, we will use u to mean either the vector $x - x'$ or the scalar $||x - x'||$ according to context. We will also use the term *stationary* as a shorthand for stationary and isotropic. A process for which $S(x) - \mu(x)$ is stationary is called *covariance stationary*. Processes of this kind are very widely used in practice as models for geostatistical data.

In Chapter 2, we introduced the empirical variogram as a tool for exploratory data analysis. We now consider the theoretical variogram as an alternative characterisation of the second-order dependence in a spatial stochastic process.

The *variogram* of a spatial stochastic process $S(x)$ is the function

$$V(x, x') = \frac{1}{2}\mathrm{Var}\{S(x) - S(x')\}. \tag{3.1}$$

Note that $V(x, x') = \frac{1}{2}[\mathrm{Var}\{S(x)\} + \mathrm{Var}\{S(x')\} - 2\mathrm{Cov}\{S(x), S(x')\}]$. In the stationary case, this simplifies to $V(u) = \sigma^2\{1 - \rho(u)\}$ which, incidentally, explains why the factor of one-half is conventionally included in the definition of the variogram. The variogram is also well defined as a function of u for a limited class of non-stationary processes; a one-dimensional example is a simple random walk, for which $V(u) = \alpha u$. Processes which are non-stationary but for which

$V(u)$ is well-defined are called *intrinsic random functions* (Matheron, 1973). We discuss these in more detail in Section 3.9.

In the stationary case the variogram is theoretically equivalent to the covariance function, but it has a number of advantages as a tool for data analysis, especially when the data locations form an irregular design. We discuss the data analytic role of the variogram in Chapter 5. Conditions for the theoretical validity of a specified class of variograms are usually discussed in terms of the corresponding family of covariance functions. Gneiting, Sasvári and Schlather (2001) present analogous results in terms of variograms.

3.2 Regularisation

In Section 1.2.1 we discussed briefly how the support of a geostatistical measurement could affect our choice of a model for the data. When the support for each measured value extends over an area, rather than being confined to a single point, the modelled signal $S(x)$ should strictly be represented as

$$S(x) = \int w(r)S^*(x-r)dr, \tag{3.2}$$

where $S^*(\cdot)$ is an underlying, unobserved signal process and $w(\cdot)$ is a weighting function. In this case, the form of $w(\cdot)$ constrains the allowable form for the covariance function of $S(\cdot)$. Specifically, if $\gamma(\cdot)$ and $\gamma^*(\cdot)$ are the covariance functions of $S(\cdot)$ and $S^*(\cdot)$, respectively, it follows from (3.2) that

$$\gamma(u) = \int\int w(r)w(s)\gamma^*(u+r-s)drds. \tag{3.3}$$

Now make a change of variable in (3.3) from s to $t = r - s$, and define

$$W(t) = \int w(r)w(t-r)dr.$$

Then (3.3) becomes

$$\gamma(u) = \int W(t)\gamma^*(u+t)dt. \tag{3.4}$$

Typical weighting functions $w(r)$ would be radially symmetric, non-negative valued and non-increasing functions of $||r||$; this holds for the effect of the gamma camera integration in Example 1.3, where $w(r)$ is not known explicitly but is smoothly decreasing in $||r||$, and for the soil core data of Example 1.4, where $w(\cdot)$ is the indicator corresponding to the circular cross section of each core. In general, the effect of weighting functions of this kind is to make $S(x)$ vary more smoothly than $S^*(x)$, with a similar effect on $\gamma(u)$ by comparison with $\gamma^*(u)$.

An analogous result holds for the relationship between the variograms of $S(\cdot)$ and $S^*(\cdot)$. Using the relationship that $V(u) = \gamma(0) - \gamma(u)$ it follows from (3.4) that

$$V(u) = \int W(t)\{V^*(t+u) - V^*(t)\}dt. \tag{3.5}$$

If the form of the weighting function $w(\cdot)$ is known, it would be possible to incorporate it into our model for the data. This would mean specifying a model for the covariance function of $S^\star(\cdot)$ and evaluating (3.4) to derive the corresponding covariance function of $S(\cdot)$. Note that this would enable data with different supports to be combined naturally, for example soil core data using different sizes of core. A more pragmatic strategy, and the only available one if $w(\cdot)$ is unknown, is to specify directly an appropriately smooth model for the covariance function of $S(\cdot)$.

The question of regularisation can also arise in connection with prediction, rather than model formulation. The simplest geostatistical prediction problem is to map the spatial signal $S(x)$, but in some applications a more relevant target for prediction might be a map of a regularised signal,

$$T(x) = \int S(u)du,$$

where the integral is over a disc with centre x i.e., $T(x)$ is a spatial average over the disc. We return to questions of this kind in Chapter 6.

3.3 Continuity and differentiability of stochastic processes

The specification of the covariance structure of a spatial process $S(x)$ directly affects the smoothness of the surfaces which the process generates. Accepted mathematical descriptors of the smoothness of a surface are its continuity and differentiability. However, for stochastically generated surfaces $S(x)$ we need to distinguish two kinds of continuity or differentiability. In what follows, we shall consider a one-dimensional space x, essentially for notational convenience.

We first consider *mean-square* properties, defined as follows. A stochastic process $S(x)$ is *mean-square continuous* if $\mathrm{E}[\{S(x+h)-S(x)\}^2] \to 0$ as $h \to 0$. Also, $S(x)$ is *mean-square differentiable*, with mean-square derivative $S'(x)$, if

$$\mathrm{E}\left[\left\{\frac{S(x+h)-S(x)}{h} - S'(x)\right\}^2\right] \to 0$$

as $h \to 0$. Higher-order mean-square differentiability is then defined sequentially in the obvious way; $S(x)$ is twice mean-square differentiable if $S'(x)$ is mean-square differentiable, and so on.

An important result, described for example in Bartlett (1955), is the following.

Theorem 3.1. A stationary stochastic process with correlation function $\rho(u)$ is k times mean-square differentiable if and only if $\rho(u)$ is $2k$ times differentiable at $u = 0$.

To examine differentiability at the origin of any particular correlation function $\rho(u)$, we need to consider the extended form of $\rho(u)$ in which u can take positive

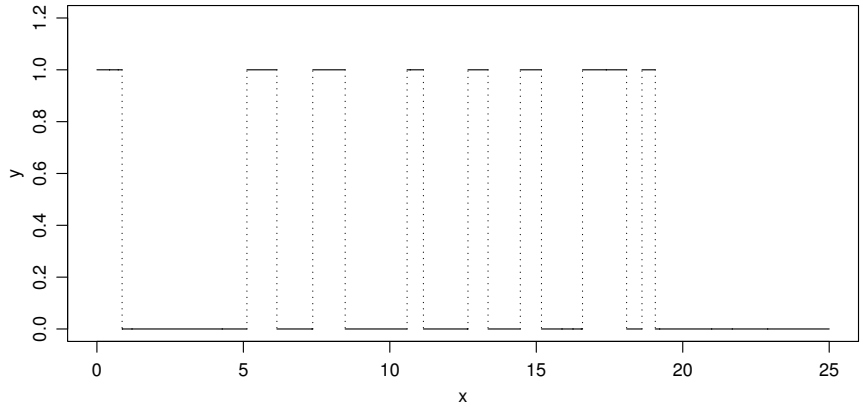

Figure 3.1. Realisation of a binary-values, mean-square continuous stochastic process (see text for details).

or negative arguments with $\rho(-u) = \rho(u)$. Hence, for example, the exponential correlation function $\rho(u) = \exp(-u/\phi)$ is continuous but not differentiable at the origin. In contrast, the Gaussian correlation function, defined by $\rho(u) = \exp\{-(u/\phi)^2\}$, is infinitely differentiable.

A second version of continuity and differentiability properties concerns *path continuity and differentiability.* A process $S(x)$ is *path-continuous*, or more generally *k times path-differentiable* if its realisations are continuous or k times differentiable functions, respectively.

In general, there need be no link between mean-square and path properties of stochastic processes. As a simple example, we can consider a binary-valued process $S(x)$ in which the real line is partitioned into a sequence of random intervals, whose lengths are independent realisations from a unit-mean exponential distribution, the value of $S(x)$ within each interval is zero with probability p, one otherwise, and the values of $S(x)$ on successive intervals are determined independently. Figure 3.1 shows a realisation with $p = 0.5$. Clearly, this process is not path-continuous. However, its correlation function is the exponential, $\rho(u) = \exp(-u)$, which is continuous at $u = 0$, hence $S(x)$ is mean-square continuous.

Kent (1989) gives a rigorous theoretical discussion of path-continuity for stationary, not necessarily Gaussian processes. Write $\rho(u) = p_m(u) + r_m(u)$, where $p_m(u)$ is the polynomial of degree m given by the Taylor series expansion of $\rho(u)$ about $u = 0$. Then, a sufficient condition for the existence of a path-continuous two-dimensional stationary process with correlation function $\rho(\cdot)$ is that $\rho(\cdot)$ is twice continuously differentiable and $|r_2(u)| = O(u^2/|\log u|^{3+\gamma})$ as $u \to 0$, for some $\gamma > 0$. A slightly stronger condition which is easier to check in practice is that $|r_2(u)| = O(u^{2+\epsilon})$ for some $\epsilon > 0$. For stationary Gaussian processes in two dimensions, a sufficient condition for path-continuity is that $\rho(0) - \rho(u) = O(1/|\log u|^{1+\epsilon})$, which is only slightly stronger than the requirement for mean-square continuity, namely that $\rho(\cdot)$ is continuous at the origin.

This justifies using mean-square differentiability as a convenient measure of the smoothness of stationary Gaussian processes when considering their suitability as empirical models for natural phenomena.

3.4 Families of covariance functions and their properties

Positive definiteness is the necessary and sufficient condition for a parametric family of functions to define a legitimate class of covariance functions, but this is not an easy condition to check directly. For this reason, it is useful to have available a range of standard families which are known to be positive definite but in other respects are sufficiently flexible to meet the needs of applications to geostatistical data. In this section, we give the details of several such families and outline their properties. Our concern here is with models for processes in two spatial dimensions. All of the covariance families which we describe are also valid in one or three dimensions. In general, a valid covariance family in $\mathbb{R}^d$ does not necessarily remain valid in more than d spatial dimensions, but is automatically valid in dimensions less than d.

3.4.1 The Matérn family

The most common form of empirical behaviour for stationary covariance structure is that the correlation between $S(x)$ and $S(x')$ decreases as the distance $u = ||x - x'||$ increases. It is therefore natural to look for models whose theoretical correlation structure behaves in this way. In addition, we can expect that different applications may exhibit different degrees of smoothness in the underlying spatial process $S(x)$.

The Matérn family of correlation functions, named after Matérn (1960), meets both of these requirements. It is a two-parameter family,

$$\rho(u) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(u/\phi)^{\kappa}K_{\kappa}(u/\phi), \tag{3.6}$$

in which $K_{\kappa}(\cdot)$ denotes a modified Bessel function of order κ, $\phi > 0$ is a scale parameter with the dimensions of distance, and $\kappa > 0$, called the *order*, is a shape parameter which determines the analytic smoothness of the underlying process $S(x)$. Specifically, $S(x)$ is $\lceil \kappa - 1$ times mean-square differentiable, where $\lceil \kappa$ denotes the smallest integer greater than or equal to κ

Figure 3.2 shows the Matérn correlation function for each of $\kappa = 0.5$, 1.5 and 2.5, corresponding to processes $S(x)$ which are mean-square continuous, once differentiable and twice differentiable, respectively. In the diagram, the values of ϕ have been adjusted so as to give all three functions the same *practical range*, which we define here as the distance u at which the correlation is 0.05. For Figure 3.2 we used $u = 0.75$ as the value of the practical range. For $\kappa = 0.5$, the Matérn correlation function reduces to the exponential, $\rho(u) = \exp(-u/\phi)$, whilst as $\kappa \to \infty$, $\rho(u) \to \exp\{-(u/\phi)^2\}$ which is also called the Gaussian correlation function or, somewhat confusingly in the present context, the Gaus-

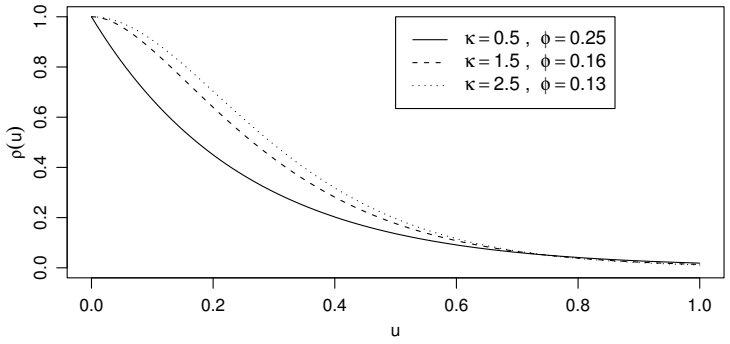

Figure 3.2. Matérn correlation functions with, $\kappa = 0.5$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2.5$ (dotted line), and adjusted values of ϕ for equivalent practical ranges.

sian model. Whittle (1954) proposed the special case of the Matérn correlation function with $\kappa = 1$.

Note that the parameters ϕ and κ in (3.6) are non-orthogonal, in the following sense. If the true correlation structure is Matérn with parameters ϕ and κ, then the best-fitting approximation with order $\kappa^* \neq \kappa$ will also have $\phi^* \neq \phi$. In other words, scale parameters corresponding to different orders of Matérn correlation are not directly comparable. The relationship between the practical range and the scale parameter ϕ therefore depends on the value of κ. For instance, the practical range as defined above is approximately 3ϕ, 4.75ϕ and 5.92ϕ for the Matérn functions with $\kappa = 0.5$, 1.5 and 2.5, respectively, and $\sqrt{3}\phi$ for the Gaussian correlation function. For this reason, Handcock and Wallis (1994) suggest a re-parametrisation of (3.6) from κ and ϕ to a more nearly orthogonal pair κ and $\alpha = 2\phi\sqrt{\kappa}$. The re-parametrisation does not, of course, change the model but is relevant to our discussion of parameter estimation in Chapters 5 and 7.

Figure 3.3 shows a one-dimensional trace through a simulated realisation of a spatial Gaussian process with each of the Matérn correlation functions above, using the same random seed for all three realisations. The increasing analytic smoothness of the process as κ increases is reflected in the visual appearance of the three realisations, but the more noticeable difference is between the non-differentiable and the differentiable case i.e., between $\kappa = 0.5$ on the one hand and $\kappa = 1.5$ or $\kappa = 2.5$ on the other.

Figure 3.4 shows simulated two-dimensional realisations of Gaussian processes whose correlation functions are Matérn with $\kappa = 0.5$ and $\kappa = 2.5$, again using the same random number seed to make the realisations directly comparable. The difference in smoothness between the non-differentiable and differentiable cases is again visually striking.

3.4.2 The powered exponential family

This family is defined by the correlation function

$$\rho(u) = \exp\{-(u/\phi)^\kappa\}. \tag{3.7}$$

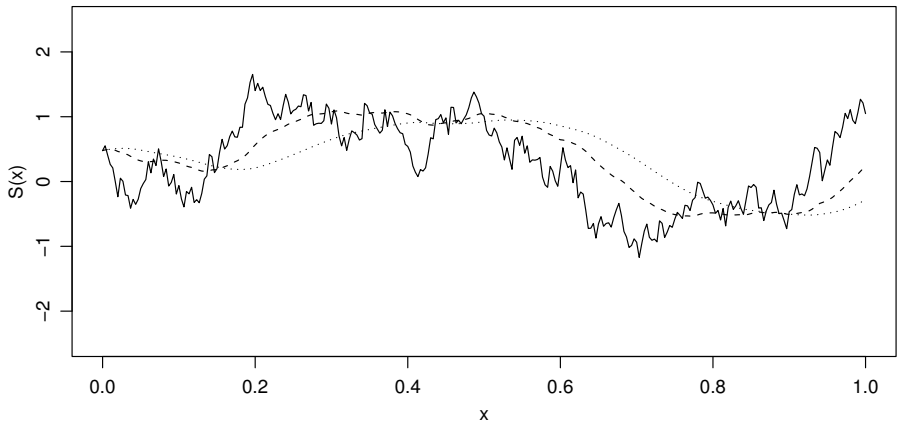

Figure 3.3. One-dimensional realisations of spatial Gaussian processes whose correlation functions are Matérn with $\kappa = 0.5$ (solid line), $\kappa = 1.5$ (dashed line) and $\kappa = 2.5$ (dotted line).

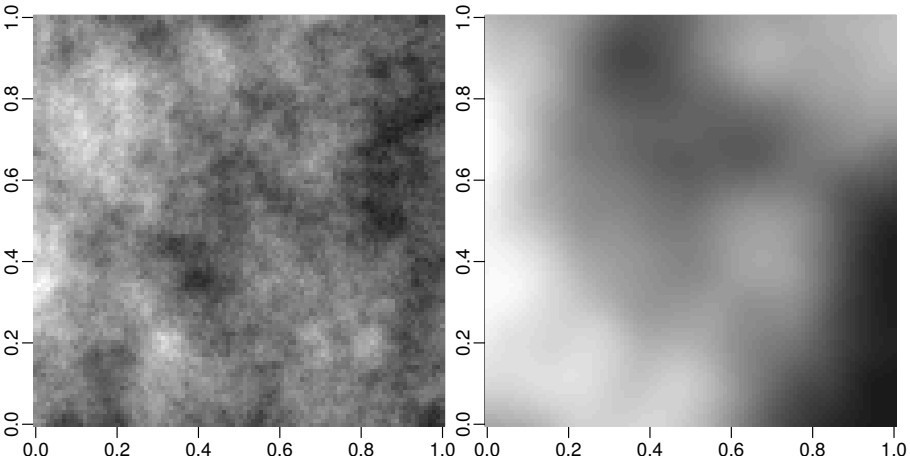

Figure 3.4. Simulations of Gaussian processes with Matérn correlation functions with $\kappa = 0.5$ and $\phi = 0.25$ (left) and $\kappa = 2.5$ and $\phi = 0.13$ (right).

Like the Matérn family, it has a scale parameter $\phi > 0$, a shape parameter κ, in this case bounded by $0 < \kappa \leq 2$, and generates correlation functions which are monotone decreasing in u. Also like the Matérn family the relation between the practical range and the parameter ϕ will depend on the value of κ. However, the family is less flexible than the Matérn, in the sense that the underlying Gaussian process $S(x)$ is mean-square continuous and not mean-square differentiable for all $0 < \kappa < 2$ but infinitely mean square differentiable when $\kappa = 2$, the maximum legitimate value. Figure 3.5 shows the powered exponential correlation function for each of $\kappa = 0.7$, 1 and 2, and with values of ϕ adjusted to provide the same practical range of 0.75. Figure 3.6 shows one-

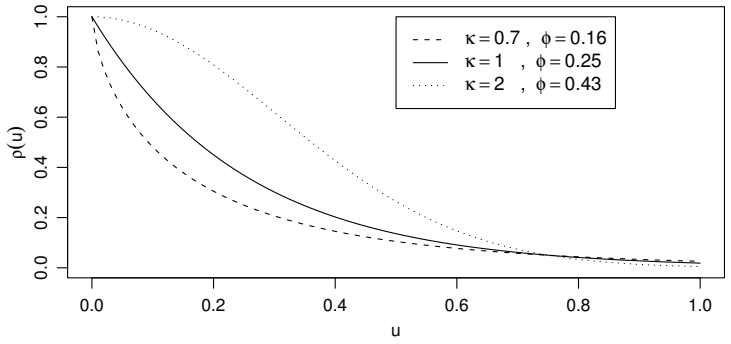

Figure 3.5. Powered exponential correlation functions with $\kappa = 0.7$ (dashed line), $\kappa = 1$ (solid line) and $\kappa = 2$ (dotted line) and values of *phi* adjusted such that the practical range is 0.75.

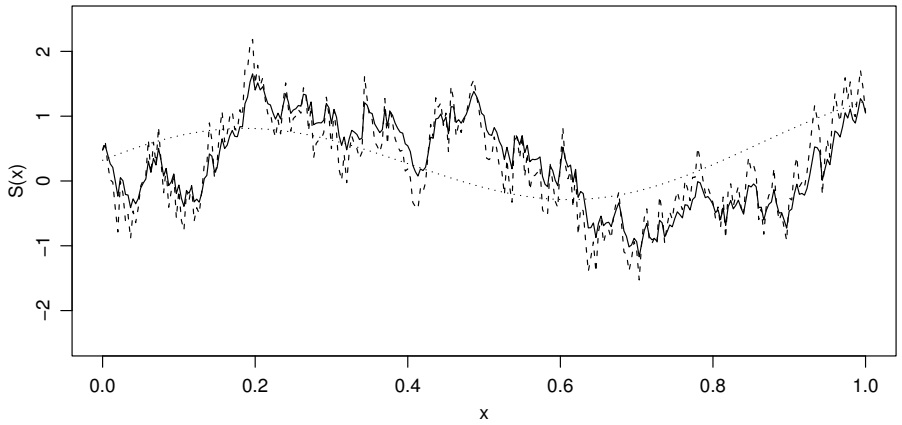

Figure 3.6. One-dimensional realisations of spatial Gaussian processes whose correlation functions are powered exponential, $\kappa = 0.7$ (dashed line), $\kappa = 1$ (solid line) and $\kappa = 2$ (dotted line).

dimensional realisations of the corresponding Gaussian processes $S(x)$. We used the same seed as for the earlier simulations of the Matérn model. The realisation for the powered exponential model with $\kappa = 1$ is therefore the same as for the Matérn model with $\kappa = 0.5$. Notice that the realisations for $\kappa = 0.7$ and $\kappa = 1$, both of which correspond to mean-square continuous but non-differentiable processes, look rather similar in character.

The extreme case $\kappa = 2$, which is equivalent to the limiting case of a Matérn correlation function as $\kappa \to \infty$, can generate very ill-conditioned covariance structure. A process $S(x)$ with this correlation function has the theoretical property that its realisation on an arbitrarily small, continuous interval determines the realisation on the whole real line. For most applications, this would be considered unrealistic.

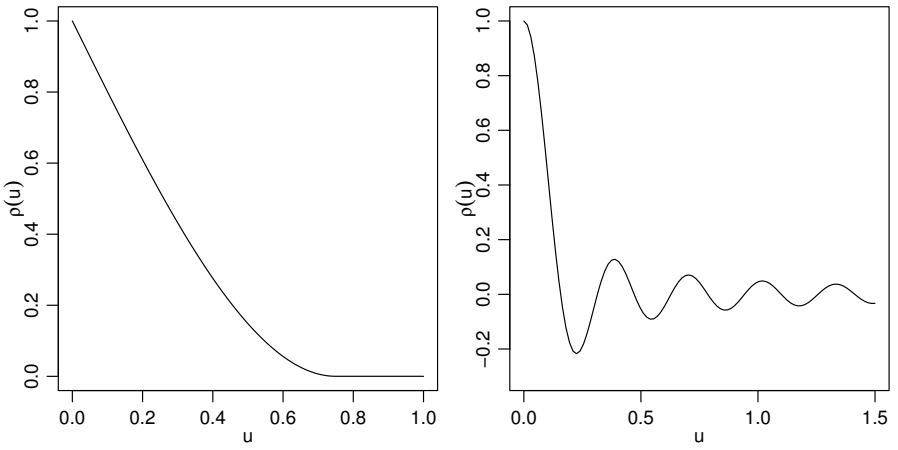

Figure 3.7. Correlation functions, the spherical (left) with $\phi = 0.75$ and wave (right) with $\phi = 0.05$.

3.4.3 Other families

In classical geostatistics, the *spherical family* is widely used. This has correlation function

$$\rho(u) = \begin{cases} 1 - \frac{3}{2}(u/\phi) + \frac{1}{2}(u/\phi)^3 & : \quad 0 \leq u \leq \phi \\ 0 & : \quad u > \phi \end{cases} \tag{3.8}$$

where $\phi > 0$ is a single parameter with the dimensions of distance. One qualitative difference between this and the families described earlier is that it has a finite range i.e., $\rho(u) = 0$ for sufficiently large u, namely $u > \phi$. The spherical family lacks flexibility by comparison with the two-parameter Matérn class. Also, $\rho(u)$ is only once differentiable at $u = \phi$, which causes technical difficulties with maximum likelihood estimation (Warnes and Ripley, 1987; Mardia and Watkins, 1989). The left-hand panel in Figure 3.7 shows the spherical correlation function when $\phi = 0.75$. The corresponding Gaussian process $S(x)$ is mean-square continuous but non-differentiable. The name and algebraic form of the spherical family derives from the geometry of intersecting spheres; see Exercise 3.3.

Non-monotone correlation functions are rare in practice. One example of a valid non-monotone family is

$$\rho(u) = (u/\phi)^{-1} \sin(u/\phi) \tag{3.9}$$

where $\phi > 0$ is a single parameter, again with the dimension of distance. The right-hand panel of Figure 3.7 illustrates the characteristic damped oscillatory behaviour of this correlation function, whilst Figure 3.8 shows a realisation of the corresponding process $S(x)$. Notice how the oscillatory nature of the correlation function is reflected in the oscillatory behaviour of the simulated realisation.

Other classes of correlation function, and criteria to check the validity of candidate functions, are described in Schlather (1999), who in turn draws on

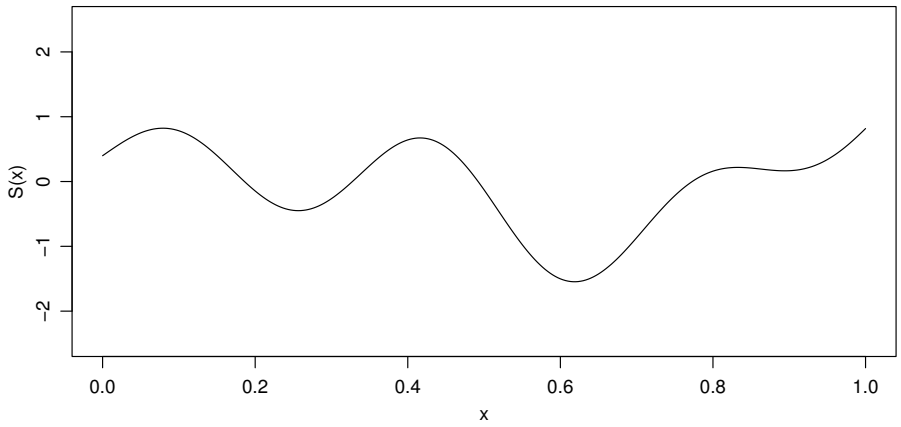

Figure 3.8. One-dimensional realisation of a spatial Gaussian process whose correlation function is $\rho(u) = (u/0.05)^{-1}\sin(u/0.05)$.

material in Gneiting (1997). However, for most geostatistical applications the families described here should be sufficient, if only because more elaborate models are hard to identify unless the available data are abundant. In general, we favour the Matérn family because of its flexibility, coupled with the tangible interpretation of the shape parameter κ as a measure of the differentiability of the underlying process $S(x)$. Also, because of the difficulty of identifying all the parameters of this model empirically, we would usually either fix the value of κ according to the context of the application, or choose amongst a limited set of values of κ, for example $\kappa = 0.5$, 1.5, 2.5 as illustrated in Figure 3.2.

3.5 The nugget effect

In geostatistical practice, the term "nugget effect" refers to a discontinuity at the origin in the variogram. Within our model-based framework, its literal interpretation is as the measurement error variance, τ^2, or equivalently the conditional variance of each measured value Y_i given the underlying signal value $S(x_i)$. Formally, this amounts to modelling the measurement process, $Y(x)$ say, as a Gaussian process whose correlation function is discontinuous at the origin, hence

$$\mathrm{Corr}\{Y(x), Y(x')\} = \begin{cases} 1 & : \quad x = x' \\ \sigma^2\rho(||x - x'||)/(\sigma^2 + \tau^2) & : \quad x \neq x' \end{cases}$$

where $\rho(\cdot)$ is the (continuous) correlation function of $S(x)$ and $||\cdot||$ denotes distance.

In practice, when the sampling design specifies a single measurement at each of n distinct locations, the nugget effect has a dual interpretation as either measurement error or spatial variation on a scale smaller than the smallest distance between any two points in the sample design, or any combination

of these two effects. These two components of the nugget effect can only be separately identified if the measurement error variance is either known, or can be estimated directly using repeated measurements taken at coincident locations.

3.6 Spatial trends

The simplest form of departure from stationarity is to allow the mean response, $\mu(x)$, to depend on location. We call any such varying mean a *spatial trend*. In applications, we may choose to model $\mu(x)$ directly as a function of x. In practice, this is most often done through a polynomial regression model, using powers and cross products of the Cartesian coordinates of x as explanatory variables. Models of this kind are called *trend surface* models. They rarely have any scientific foundation. Our view is that linear or quadratic trend surfaces can provide useful empirical descriptions of simple, unexplained spatial trends, but that higher-degree surfaces should be avoided because complicated trends are better described through the stochastic component of the model. See, for example, our illustrative analysis of the surface elevation data reported in Chapter 2.

A more interesting kind of spatial trend arises when the mean function can be modelled using spatially referenced covariates, hence for example $\mu(x) = \alpha + d(x)\beta$ where $d(x)$ is a scientifically relevant property of the location x. In our opinion, models of this kind are more interesting than trend surface models because they seek to explain, rather than merely to describe, the spatial variation in the response variable. For example, in the Gambia malaria data of Example 1.3 modelling the spatial variation in prevalence as a function of greenness has a natural scientific interpretation because the greenness index is a surrogate measure of the suitability of each location for mosquitos to breed. If, hypothetically, greenness showed a smooth east-west trend, then modelling malaria prevalence as a function of greenness or as a function of longitude might give equally good empirical fits to the data, but modelling prevalence as a function of greenness would offer the more satisfying explanation and would be the more likely to translate to other study regions.

As discussed in Section 1.2.2, when values of a potential explanatory variable $d(x)$ are only recorded at the same locations as give rise to the basic geostatistical data (x_i, y_i), we need to consider whether we should treat $d(x)$ as a second, stochastic variable to be analysed jointly with the primary signal process, $S(x)$, rather than as a deterministic quantity.

3.7 Directional effects

Another form of non-stationarity is non-stationarity in the covariance structure. One specific way to relax the stationarity assumption is to allow directional effects so that, for example, the rate at which the correlation decays with increasing distance is allowed to depend also on the relative orientation between pairs of locations.

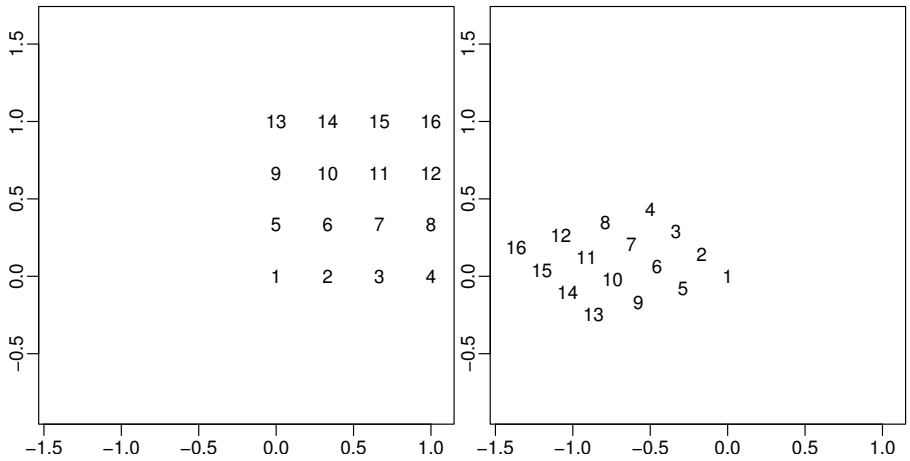

Figure 3.9. Rotation of the data configuration by the anisotropy parameters. The left-hand panel shows the original locations, the right-hand panel the transformed locations in isotropic space when $\psi_A = 2\pi/3$ and $\psi_R = 2$.

The simplest form of directional effect on the covariance structure is called *geometrical anisotropy.* This arises when a stationary covariance structure is transformed by a differential stretching and rotation of the coordinate axes. Hence, geometrical anisotropy is defined by two additional parameters. Algebraically, a model with geometrical anisotropy in spatial coordinates $x = (x_1, x_2)$ can be converted to a stationary model in coordinates $x' = (x'_1, x'_2)$ by the transformation

$$(x'_1, x'_2) = (x_1, x_2) \begin{bmatrix} \cos(\psi_A) & -\sin(\psi_A) \\ \sin(\psi_A) & \cos(\psi_A) \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & \psi_R^{-1} \end{bmatrix} \tag{3.10}$$

where ψ_A is called the *anisotropy angle* and $\psi_R > 1$ is called the *anisotropy ratio.* The direction along which the correlation decays most slowly with increasing distance is called the *principal axis.*

These operations are illustrated in Figure 3.9. The original locations are shown in the left-hand panel. Suppose that the anisotropy angle is $\psi_A = 2\pi/3$, and the anisotropy ratio is $\psi_R = 2$. Then, applying the coordinate transformation (3.10) we obtain the locations in the right-hand panel, which are now in an isotropic space, and proceed to fit an isotropic model in this transformed space. In practice, ψ_A and ψ_R are unknown, and the model fit would be optimised by treating ψ_A and ψ_R as additional parameters to be estimated.

Figure 3.10 shows realisations of two Gaussian spatial process with geometrical anisotropy. The directional effects are visually clear, with the principal axis in each case running diagonally over the square region shown. For the left panel the anisotropy angle is $\pi/3$ radians and the anisotropy ratio is 4. For the right panel the anisotropy angle is $3\pi/4$ radians and the anisotropy ratio is 2. The two processes have common parameter values $\mu = 0$, $\sigma^2 = 1$ and exponential correlation function with $\phi = 0.25$, and the two realisations were generated using the same random seed.

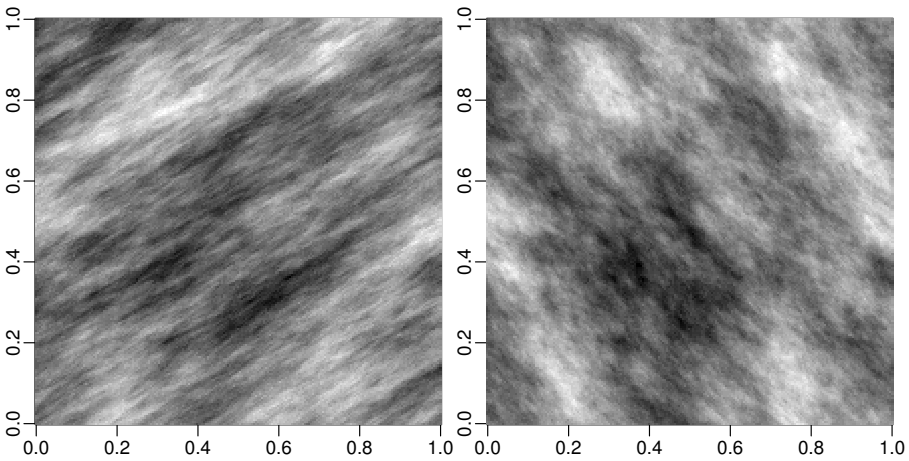

Figure 3.10. A realisation of a geometrically anisotropic Gaussian spatial process whose principal axis runs diagonally across the square region with anisotropy parameters $(\pi/3, 4)$ for the left-hand panel and $(3\pi/4, 2)$ for the right-hand panel.

Note that geometric anisotropy cannot describe local directional features of a spatial surface, only global ones. On the other hand, the presence of local directional features in a realisation of a spatial process need not imply that the underlying process is anisotropic. Consider, for example, a surface constructed as the superposition of profiles $f(\cdot)$ translated by the points of a homogeneous Poisson point process. Thus,

$$S(x) = \sum_{i=1}^{\infty} f(x - X_i) \tag{3.11}$$

where the X_i are the points of the Poisson process. Figure 3.11 compares realisations of two such processes in which the intensity of the Poisson process is 16 points per unit area and the profile function is the probability density of a bivariate Gaussian distribution with zero mean, standard deviation 0.1 in each coordinate direction and correlation 0.75. In the left-hand panel, the global directional feature along the diagonal direction is clear. In the right-hand panel, each profile has been randomly rotated so that, whilst local directional effects can still be seen, the resulting model is isotropic with no global directional effects. Higdon (1998, 2002) has proposed constructions similar to, but more general than, (3.11) to define a general class of non-stationary, non-Gaussian models.

Geometrical anisotropy deals with a particular form of non-stationarity by transforming the space x using stretching and rotation, so that the underlying process is stationary in the transformed space. Sampson and Guttorp (1992), Guttorp, Meiring and Sampson (1994) and Guttorp and Sampson (1994) develop a more general version of this approach. Their method seeks a smooth deformation of the x-space, equivalent to a transformation from x to x^* say, so that the covariance function depends only on distance in the deformed space, hence for any two locations x and y in the original space,

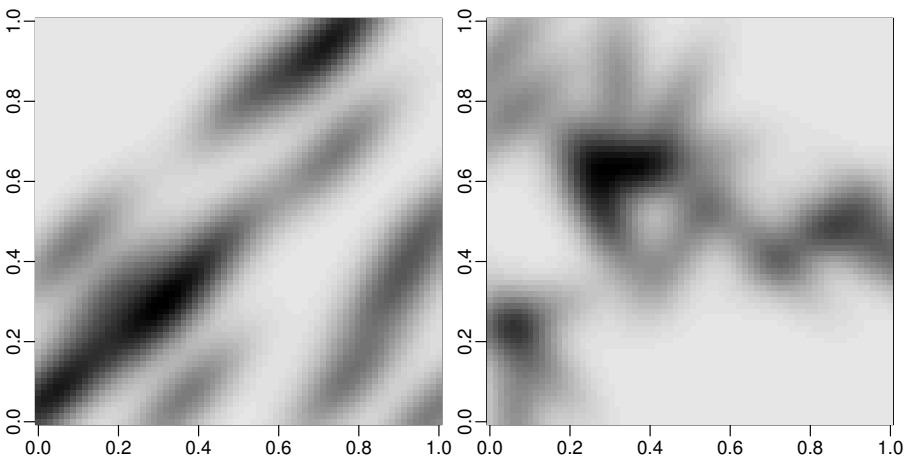

Figure 3.11. Realisations of two spatial processes with global (left-hand panel) and local (right-hand panel) directional effects. See text for detailed explanation.

$\text{Cov}\{S(x), S(y)\} = \gamma(||x^* - y^*||)$. Perrin and Meiring (1999) discuss identifiability issues for this class of models, whilst Schmidt and O'Hagan (2003) develop a Bayesian version. Replicated observations are needed at each sampling location in order to identify the required transformation. In practice, the approach is feasible when a time series is collected at each location as this gives the necessary, albeit dependent, replication.

Non-stationarity can also arise because Euclidean distance is not an appropriate measure of spatial separation. For example, Rathbun (1998) considers non-Euclidean distances in modelling spatial variation in an estuary where, amongst other considerations, the line segment joining two locations within the estuary may cross a stretch of land.

3.8 Transformed Gaussian models

We now expand the discussion of Section 2.2, where we mentioned briefly that the range of applicability of the Gaussian model can be extended by assuming that the model holds after a marginal transformation of the response variable.

As in other areas of statistics, there are at least three different reasons for using a transformation of the data. Firstly, a particular transformation might be suggested by qualitative arguments, or even by convention. For example, if effects are thought to be operating multiplicatively, then a log-transformation converts the problem to a scale on which effects are, more conveniently, additive. Secondly, a transformation may be used as a variance-stabilising device for a known, non-Gaussian sampling distribution. For example, square root and arc-sine transformations approximately stabilise the sampling variance under Poisson and binomial sampling, respectively. Note, however, that there is no reason why a transformation which stabilises the variability in the measurements conditional on the signal should also stabilise the variability in the signal,

or *vice versa*. The transformation approach to variance instability used to be widespread in regression modelling of non-Gaussian data, but has largely been replaced by the use of generalized linear models (McCullagh and Nelder, 1989). Section 1.4 and, in more detail, Chapter 4 describe an extension of classical generalized linear models to accommodate non-Gaussian geostatistical data. Finally, we can introduce a parametric family of transformations simply as an empirical generalisation of the Gaussian model, in which case the choice of a particular transformation corresponds to the estimation of an additional parameter. The most widely used example of this approach is the Box-Cox family of transformations (Box and Cox, 1964),

$$Y^* = \begin{cases} (Y^\lambda - 1)/\lambda & : \quad \lambda \neq 0 \\ \log Y & : \quad \lambda = 0. \end{cases} \tag{3.12}$$

The log-transformation is perhaps the most widely used in practice, and explicit expressions can be derived for its mean and covariance structure. Suppose that $T(x) = \exp\{S(x)\}$, where $S(x)$ is a stationary Gaussian process with mean μ, variance σ^2 and correlation function $\rho(u)$. The moment generating function of $S(x)$ is

$$M(a) = \mathrm{E}[\exp\{aS(x)\}] = \exp\{a\mu + \frac{1}{2}a^2\sigma^2\}. \tag{3.13}$$

It follows from (3.13), setting $a = 1$, that $T(x)$ has expectation

$$\mu_T = \exp\left(\mu + \frac{1}{2}\sigma^2\right). \tag{3.14}$$

Similarly, setting $a = 2$ in (3.13) gives $\mathrm{E}[T(x)^2]$, and hence the variance of $T(x)$ as

$$\sigma_T^2 = \exp(2\mu + \sigma^2)\{\exp(\sigma^2) - 1\}. \tag{3.15}$$

Finally, for any two locations x and x', $T(x)T(x') = \exp\{S(x) + S(x')\}$, and $S(x) + S(x')$ is Gaussian with mean $m = 2\mu$ and variance $v = 2\sigma^2\{1 + \rho(||x - x'||)\}$. It follows that $\mathrm{E}[T(x)T(x')] = \exp(m+v/2)$, and straightforward algebra gives the correlation function of $T(x)$ as

$$\rho_T(u) = [\exp\{\sigma^2\rho(u)\} - 1]/[\exp\{\sigma^2\} - 1]. \tag{3.16}$$

Note that the mean and variance of $T(x)$ depend on both μ and σ^2, whereas the correlation function of $T(x)$ does not depend on μ.

Log-Gaussian processes exhibit, to a greater or lesser extent depending on the values of the model parameters, asymmetric behaviour with local patches of values close to zero, interspersed with relatively sharp peaks. In particular, we can write any Gaussian process $S(x)$ as $\mu + \sigma Z(x)$, and the corresponding log-Gaussian process as $T(x) = \alpha T_0(x)^\sigma$, where $\alpha = \exp(\mu)$ and $T_0(x) = \exp\{Z(x)\}$. Hence, for any given $Z(x)$, the value of μ affects the scale of the surface $T(x)$, whilst σ affects its shape, with larger values of σ producing sharper peaks and flatter troughs

The two panels of Figure 3.12 illustrate this affect. They show realisations of two log-Gaussian processes of the form $T(x) = \exp\{\sigma Z(x)\}$, where $Z(x)$ is

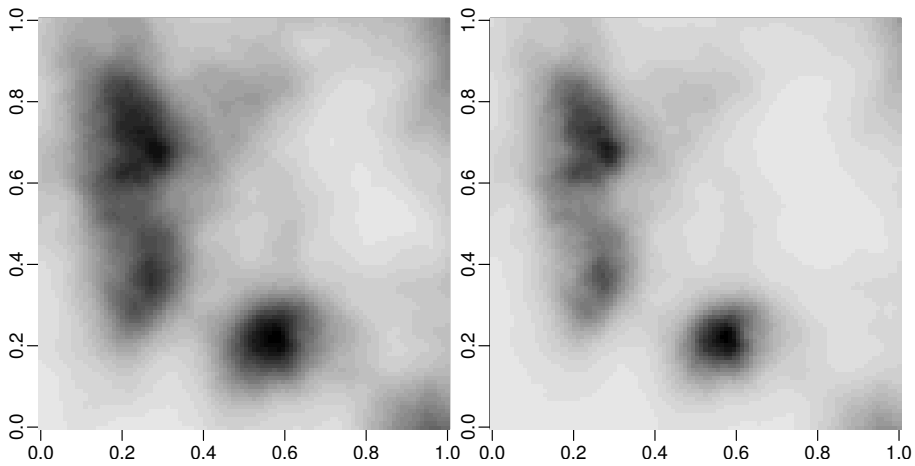

Figure 3.12. Realisations of two log-Gaussian processes. See text for parameter specifications.

a Gaussian process with zero mean, unit variance and Matérn correlation of order $\kappa = 1.5$ and with range parameter $\phi = 0.2$. Both panels use the same realisation of $Z(x)$ and differ only in that the left-hand panel has $\sigma = 0.1$ and the right-hand panel $\sigma = 0.7$.

The two panels of Figure 3.13 compare a realisation of a log-Gaussian process and a Gaussian process with the same mean and variance, and closely matched correlation structure. The log-Gaussian process used for the left-hand panel of Figure 3.13 has its correlation structure $\rho_T(u)$ induced by an underlying Matérn correlation function $\rho_0(u)$ with parameters $\kappa = 1.5$ and $\phi = 0.2$, and variance $\sigma^2 = 1$. We then used a simple least squares criterion to obtain a Matérn correlation function, $\rho_A(u)$ say, which approximated $\rho_T(u)$ as closely as possible, resulting in the parameter values $\phi_a = 0.18$ and $\kappa_a = 1.32$. To obtain the right-hand panel of Figure 3.13 we then simulated a Gaussian process using the correlation function $\rho_A(u)$ in conjunction with a mean and variance chosen so as to match those of the log-Gaussian process. As usual, we used the same random number seed for the two realisations being compared. Figure 3.14 compares the correlation functions $\rho_T(u)$, $\rho_A(u)$ and $\rho_0(u)$. We see that the correlation functions of the processes used to generate the two realisations shown in Figure 3.13 are almost identical, yet the realisations themselves are very different in character because of their different distributional properties.

3.9 Intrinsic models

In Section 3.6 we discussed a simple form of non-stationary model, namely the sum of a deterministic spatial trend and a stochastic, spatially correlated residual. Similarly, in Section 3.7 we discussed a deterministic strategy for dealing with non-stationarity, in this case a transformation of the spatial coordinate system to deal with a global directional effect in the underlying process. An

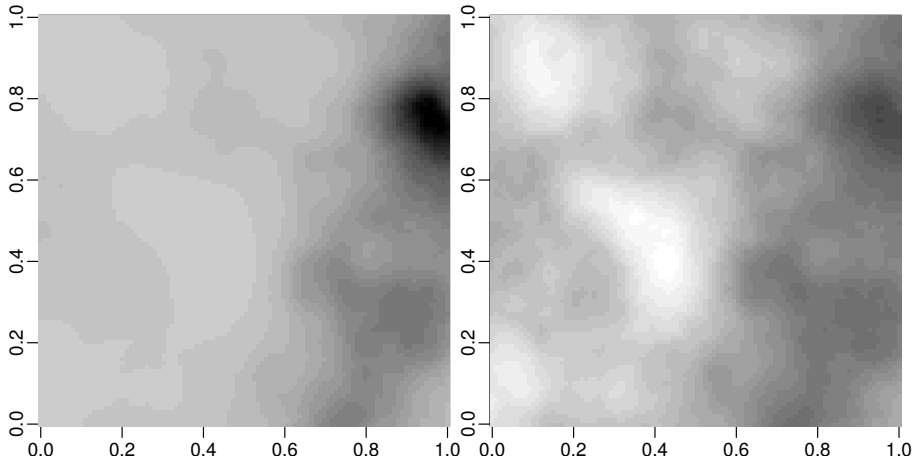

Figure 3.13. Realisations of a log-Gaussian process (left-hand panel) and a Gaussian process with closely matched correlation structure (right-hand panel). See text for parametric specifications.

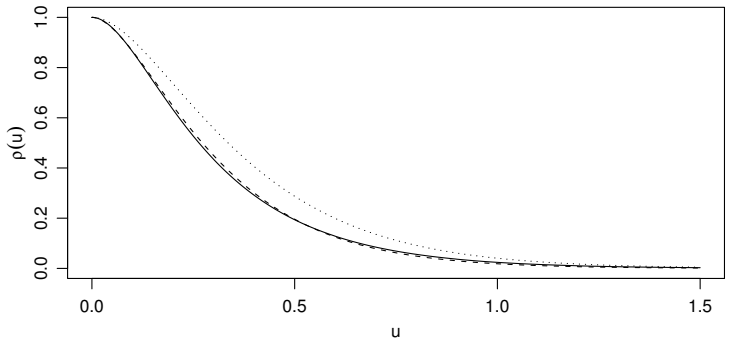

Figure 3.14. Correlation functions $\rho_T(u)$ (solid line) and $\rho_A(u)$ (dashed line) for the log-Gaussian and Gaussian processes whose realisations are compared in Figure 3.13. The dotted line shows the Matérn correlation function $\rho_0(u)$. See text for parametric specifications.

alternative strategy is to treat non-stationarity as an inherently stochastic phenomenon.

As a simple, spatially discrete one-dimensional example of an intrinsic model we consider a random walk, $S(x)$, defined recursively by

$$S(x) = S(x-1) + Z(x) : x = 0, 1, \ldots \tag{3.17}$$

where the $Z(x)$ are mutually independent, normally distributed with mean 0 and variance 1. Conventionally, we add the initial condition that $S(0) = 0$, in which case $\mathrm{E}[S(x)] = 0$ for all x and $\mathrm{Var}\{S(x)\} = x$. However, an alternative interpretation, which is perhaps more natural in the spatial setting where any ordering along the coordinate axes is arbitrary, is that $S(x)$ fluctuates

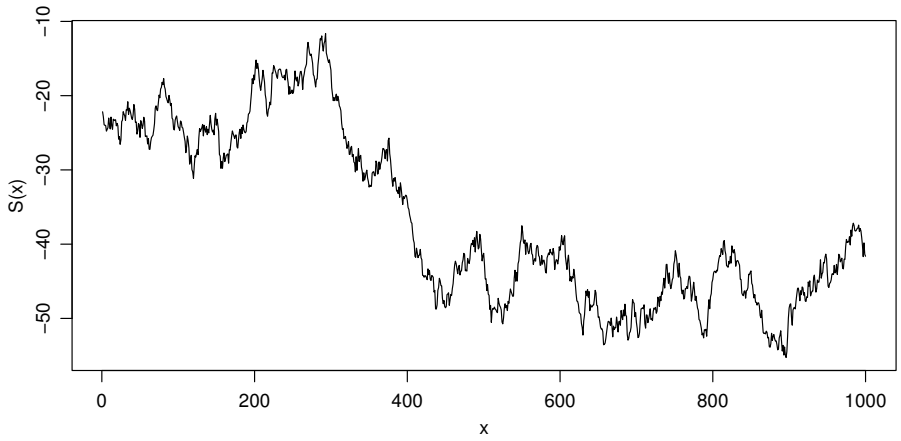

Figure 3.15. Realisation of a one-dimensional random walk. See text for detailed explanation

randomly about an arbitrary level i.e., the average is indeterminate and the variation observed within a finite segment of space increases with the length of the segment.

Figure 3.15 shows a simulated realisation of such a process. The process was initialised at zero, allowed to run for 5000 steps, then observed and plotted for an additional 1000 steps. By chance, the plotted values of $S(x)$ vary over the approximate range -60 to -10 although their theoretical expectation over repeated realisations is zero. More interestingly, the initial and final portions of Figure 3.15 appear on casual inspection to be approximately stationary whereas, the portion between $x = 300$ and $x = 450$ suggests a decreasing, approximately linear trend. One lesson which we take from this example is that when our data consist of a single realisation of a correlated stochastic process, it is often the case that qualitatively wrong models can give a reasonable empirical fit to the data.

The random walk model (3.17) is an example of a general class of non-stationary stochastic processes known as *intrinsic random functions* (Matheron, 1973). An intrinsic random function is a stochastic process $S(x)$ with stationary increments. This means that for any $u \in \mathbb{R}^2$, the process $D_u(x)$ defined by

$$D_u(x) = S(x) - S(x-u)$$

is stationary. Suppose that $\mathrm{Var}(D_u) = \sigma_u^2$. Then, $\frac{1}{2}\sigma_u^2$, regarded as a function of u, is also the variogram of $S(x)$. Hence, intrinsic random functions can be thought of as processes for which the variogram, but not necessarily the covariance function, depends only on u. For the random walk process (3.17), the variogram is $V(u) = \frac{1}{2}\mathrm{Var}\{S(x) - S(x-u)\} = \frac{1}{2}u$, for $u \geq 0$, whereas the covariance function is $\gamma(x,u) = \mathrm{Cov}\{S(x), S(x-u)\} = |x-u|$, which depends on both u and x.

Examples of legitimate intrinsic variogram models include power law and logarithmic forms. The power law model, $V(u) = (u/\phi)^\kappa$ is valid for $0 < \kappa < 2$.

The most widely used special case is the linear variogram, $V(u) = u/\phi$. The logarithmic model,

$$V(u) = \log(u/\phi), \tag{3.18}$$

occupies a special place in classical geostatistics because of its connection to an empirical law discovered by De Wijs (1951, 1953). De Wijs observed that when a sample of ore was broken into smaller pieces, the variability between the grades of the pieces in relation to the average grade of the original sample appeared to depend only on the ratio of the volume of the pieces to the volume of the original, and not on the absolute volume of the original. Viewed as a model for a variogram, (3.18) has the unattractive property that $V(u) \to -\infty$ as $u \to 0$ which is incompatible with the definition of the variogram as a variance. However, suppose that (3.18) holds for an unobserved process $S^*(x)$, and that we observe

$$S(x) = \int w(r)S^*(x-r)dr, \tag{3.19}$$

where $w(u)$ is a non-negative valued weighting function. As discussed in Section 1.2.1 this corresponds to each observed measurement having a finite support deriving from a finite spatial neighbourhood centred on the point x. Now, as in the derivation of (3.5), write

$$W(t) = \int w(r)w(t-r)dr.$$

Combining (3.18) and (3.5) then gives the variogram of the regularised process as

$$\begin{aligned} V(u) &= \int W(t)[\log\{(t+u)/\phi\} - \log(t/\phi)]dt \\ &= \int W(t)\{\log(t+u) - \log(t)\}dt, \end{aligned} \tag{3.20}$$

which is non-negative valued for all $u \geq 0$ and does not depend on ϕ. This rather surprising result is the theoretical analogue of De Wijs's empirical law. Besag and Mondal (2005) establish a close theoretical link between the De Wijs process and intrinsic autoregressive processes on a two-dimensional lattice and show that, by making the lattice spacing sufficiently fine, the spatially discrete autoregressive process can give an excellent approximation to the spatially continuous De Wijs process. The lattice formulation also brings substantial computational benefits for large data-sets.

Intrinsic random functions embrace a wider class of models than do stationary random functions. With regard to spatial prediction, the main difference between predictions obtained from intrinsic and from stationary models is that if intrinsic models are used, the prediction at a point x is influenced by the local behaviour of the data i.e., by the observed measurements at locations relatively close to x, whereas predictions from stationary models are also affected by global behaviour. One way to understand this is to remember that the mean of an intrinsic process is indeterminate. As a consequence, predictions derived

from an assumed intrinsic model tend to fluctuate around a local average. In contrast, predictions derived from an assumed stationary model tend to revert to the global mean of the assumed model in areas where the data are sparse. Which of these two types of behaviour is the more natural depends on the scientific context in which the models are being used.

3.10 Unconditional and conditional simulation

Simulation plays an important role in geostatistical practice, both in conducting Monte Carlo experiments to gain insight into the properties of particular models and associated statistical methods, and as a fundamental tool in conducting geostatistical inference when the required analytical results are intractable.

The most basic simulation problem is to simulate a realisation, say $Y = (Y_1, \ldots, Y_n)$, of a Gaussian model at a set of n locations $x_i \in \mathbb{R}^2$. Note firstly that if the model for Y includes a nugget effect, with nugget variance τ^2, we can represent Y as $Y = \mu + S + \tau T$ where $\mu = \mathrm{E}[Y]$, $T = (T_1, .., T_n)$ is a set of mutually independent $\mathrm{N}(0, 1)$ random variables, and the spatial signal $S = (S_1, \ldots, S_n)$ follows a zero-mean multivariate Gaussian distribution, namely $S \sim \mathrm{MVN}(0, \Sigma)$.

The standard method for simulating a realisation of S is to simulate an independent random sample $Z = (Z_1, \ldots, Z_n)$ from the standard Gaussian distribution, $\mathrm{N}(0, 1)$, and apply a linear transformation,

$$S = AZ, \tag{3.21}$$

where A is any matrix such that $AA' = \Sigma$. Two ways to construct A are through Cholesky factorisation and singular value decomposition.

The Cholesky factorisation of Σ is $\Sigma = LL'$, where L is a lower-triangular matrix. Hence in (3.21) we take $A = L$. Because A is lower triangular, this method of simulating S can be interpreted as first simulating S_1 from its marginal, univariate Gaussian distribution, then successively simulating $S_2, \ldots, S_n$ from the conditional distributions of each S_i given $S_1, \ldots, S_{i-1}$, each of which is again univariate Gaussian.

The singular value decomposition of Σ is $\Sigma = U\Lambda U'$, where Λ is a diagonal matrix whose diagonal elements $\lambda = (\lambda_1, \ldots, \lambda_n)$ are the eigenvalues of Σ, ordered from largest to smallest, whilst the columns of U contain the corresponding eigenvectors, hence $U'U = I$. Because Σ is positive definite, all of the λ_i are positive. Hence, a second possible choice for A in (3.21) is $A = U\Lambda^{\frac{1}{2}}$, where $\Lambda^{\frac{1}{2}}$ is the diagonal matrix with diagonal elements $\sqrt{\lambda_i}$.

Simulating realisations of the stationary Gaussian model by either of these methods becomes difficult in practice when n is very large, because of the computational burden associated with the necessary matrix operations. Typically, to simulate a realisation of a process $S(\cdot)$ over a spatial region, A say, we would approximate the spatially continuous surface $S(x)$ by its values on a fine grid to cover the region of interest. For this situation, Wood and Chan (1994) provide an ingenious algorithm which uses circulant embedding in conjunction with fast

Fourier transform methods to achieve very substantial reductions in both computing time and storage requirements when the number of grid points is large; for example, simulation on a grid of size 256 by 256 becomes computationally straightforward.

A completely different approach is to use a Markov chain Monte Carlo method known as Gibbs sampling (Gilks et al., 1996). Define the *full conditional distributions* of $S = (S_1, \ldots, S_n)$ as the n univariate Gaussian distributions of each S_i given all other S_j. Choose any initial set of values for S, say $S_0 = (S_{01}, \ldots, S_{0n})$. Now, simulate a new set of values, $S_1 = (S_{11}, \ldots, S_{1n})$ successively from the full conditionals of each S_i given the *new* values $S_{1j} : j = 1, \ldots, i-1$ and the *old* values $S_{0j} : j = i+1, \ldots, n$, with the obvious interpretations for $i = 1$ and $i = n$. This defines a single *sweep* of the Gibbs sampler. Re-set S_0 to be the newly simulated S_1 and repeat. If we iterate this process over many sweeps, the distribution of the resulting sequence of simulations S_1 converges to the required multivariate Gaussian.

For the models considered in this chapter, the Gibbs sampler is generally not a sensible option because the evaluation of each full conditional distribution requires the inversion of an $(n-1) \times (n-1)$ covariance matrix. However, the method becomes very attractive if we *define* our models by the form of their full conditionals, especially so if the full conditionals are sparse i.e., the full conditional of each S_i depends only on a small number of S_j, called the *neighbours* of S_i. Models of this kind are known as Gaussian Markov random fields and are discussed in Rue and Held (2005). For general geostatistical applications, Markov random field models have the unattractive feature that they are tied to a specified set of locations rather than being defined in a spatially continuous way. Hence, they cannot be used directly to make spatially continuous predictions. However, Rue and Tjelmeland (2002) have shown how a spatially continuous Gaussian process can be approximated by a Gaussian Markov random field on a fine grid. Hence, a feasible strategy is to define a spatially continuous model but use its approximating Markov random field for computation.

In the geostatistical literature, simulating a realisation of a spatial process $S(x)$ on a set of locations $x_i : i = 1, \ldots, n$ is called *unconditional simulation*, to distinguish it from *conditional simulation*. The latter refers to simulation of a spatial process $S(x)$ at locations $x_i^* : i = 1, \ldots, N$, conditional on observed values $S(x_i)$ at locations $x_i : i = 1, \ldots, n$ or, more generally, conditional on data $Y = (Y_1, \ldots, Y_n)$ which are stochastically related to $S(\cdot)$. In the present context, the underlying model for Y is that $Y_i = S(x_i) + Z_i$, where the Z_i are mutually independent and normally distributed, $Z_i \sim \mathrm{N}(0, \tau^2)$. Conditional simulation is used informally to investigate to what extent the observed data do or do not identify the essential features of the underlying spatially continuous surface $S(x)$. It is also an essential tool in formal geostatistical inference, and as such will arise naturally in later chapters. Here, we note only that for the Gaussian model, the conditional distribution of the values of the process $S(x)$ at any set of locations, say $S^* = \{S(x_i^*), .., S(x_N^*)\}$, given the data Y, is multivariate Gaussian with a variance matrix which does not depend on Y. Hence, both unconditional and conditional simulation require computationally feasible ways

of simulating from high-dimensional multivariate Gaussian distributions with particular kinds of structured covariance matrices.

3.11 Low-rank models

A low-rank model (Hastie, 1996) for a random vector S is one whose distributional dimension is less than the dimension of S itself. To motivate this idea in the context of geostatistical modelling, we briefly re-visit the singular value decomposition method for simulating realisations of S when the underlying model is a Gaussian process.

Recall that the singular value decomposition method simulates S as $S = AZ$ where Z is a vector of mutually independent $\mathrm{N}(0,1)$ random variables and $A = U\Lambda^{\frac{1}{2}}$. Here, the diagonal matrix Λ contains the eigenvalues of the required covariance matrix of S, whilst U contains the corresponding eigenvectors. If the eigenvalues are ordered from largest to smallest, then we could obtain an approximate simulation of S by using only the first $m < n$ columns of A to give

$$S = A_m Z \tag{3.22}$$

where now Z consists of only m independent $\mathrm{N}(0,1)$ variates (see Exercise 3.4). The resulting S has a singular multivariate Gaussian distribution, which can be regarded as a low-rank approximation to the target, non-singular distribution. Because A is derived from the covariance matrix of $S = \{S(x_1, \ldots, S(x_n)\}$ its elements are, implicitly, functions of the sampling locations x_i and we could therefore think of (3.22) as a specification of the form

$$S(x_i) = \sum_{j=1}^{m} Z_j f_j(x_i) : i = 1, \ldots, n. \tag{3.23}$$

This suggests that, rather than considering the low-rank approximation only as a computational short-cut, we could also use it as a way of defining a model for $S(\cdot)$. The general idea is to represent a spatially continuous stochastic process $S(x)$ as a linear combination of functions $f_j(x)$ and random coefficients A_j, so that for any $x \in \mathbb{R}^2$,

$$S(x) = \sum_{j=1}^{m} A_j f_j(x). \tag{3.24}$$

If the A_j follow a zero-mean multivariate Gaussian distribution with $\mathrm{Cov}(A_j, A_k) = \gamma_{jk}$, then $S(\cdot)$ is a zero-mean Gaussian process with covariance structure given by

$$\mathrm{Cov}\{S(x), S(x')\} = \sum_{j=1}^{m} \sum_{k=1}^{m} \gamma_{jk} f_j(x) f_k(x'). \tag{3.25}$$

In general, the covariance structure (3.25) is non-stationary. Whether or not it has an intuitively appealing form depends on the choices made for the functions

$f_j(\cdot)$ and for the covariances amongst the A_j. The $f_k(\cdot)$ would usually be chosen to form an orthonormal basis, meaning that

$$\int f_j(x) f_k(x) dx = 1$$

if $k = j$ and is zero otherwise. Typically, the coefficients A_j would then be specified as mutually independent.

A familiar example of (3.24) in one dimension is the spectral representation of a time-series as a superposition of sine and cosine waves with mutually independent random coefficients. For an exact representation of a time-series $S(x) : x = 1, .., n$ we define n functions $f_k(x)$ which correspond to $n/2$ sine-cosine pairs at frequencies $2\pi jx/n : j = 0, 1, \ldots, [n/2]$. The associated coefficients are then assigned large or small variances corresponding to frequencies which account for large or small proportions, respectively, of the overall variation in the series. A low-rank approximation is obtained by setting some of the coefficients to zero. Spectral representations can also be used in two spatial dimensions and are discussed for example in Stein (1999).

Low-rank models for spatial processes can also be constructed using splines. Splines (Wahba, 1990) are piece-wise polynomial functions. By choosing the pieces to be cubics, constrained to be continuously differentiable at the joins, or "knots" connecting successive pieces, we obtain a very flexible method for approximating any smooth function. In two spatial dimensions, the same idea can be used to construct a flexible class of smooth surfaces by joining together locally polynomial pieces, known as *thin-plate splines* (Duchon, 1977). Thin-plate spline models are discussed in Wood (2003). Kammann and Wand (2003) emphasise the connection between splines and linear random effect models which is hinted at in (3.24) above. Laslett (1994) compares predictions obtained from spline models and from more conventional geostatistical models of the kind discussed earlier in this chapter. Ruppert, Wand and Carroll (2003) discuss the use of low-rank splines in semiparametric regression modelling.

3.12 Multivariate models

Multivariate geostatistical models are relevant when two or more different response variables are measured at spatial locations within a continuous spatial region. As discussed in Section 1.2.2 this situation can arise either because the variables are all of equal scientific interest and we wish to describe their joint spatial distribution, or because we wish to describe the conditional distribution of a response variable of primary interest given one or more spatially referenced covariates. When a covariate is only available at a finite set of sample locations we may choose to treat it as a set of sampled values from an underlying stochastic process. A third situation in which multivariate methods are useful is when the variable of primary interest, Y say, is difficult or expensive to measure, but it is easy to measure a second variable, Z, which is known to be correlated with Y. In this situation, for efficient prediction of Y the most cost-effective design

may be one in which a small number of measurements of Y are combined with a large number of cheaper measurements of Z.

In the remainder of this section we describe some possible multivariate extensions to the univariate Gaussian models considered so far in this chapter. All of the general ideas discussed for univariate processes carry over, but with additional aspects introduced by the multivariate setting. We focus on the specification of valid models for stationary variation about a trend, including the distinction between the observation process $Y(x)$ and an unobserved signal process $S(x)$.

3.12.1 Cross-covariance, cross-correlation and cross-variogram

The covariance and correlation functions of a multivariate spatial process are easily defined as follows. A d-dimensional spatial process is a collection of random variables $Y(x) = \{Y_1(x), \ldots, Y_d(x)\}$, where $x \in \mathbb{R}^2$. Then, the covariance function of $Y(x)$ is a $d \times d$ matrix-valued function $\Gamma(x, x')$, whose $(j,k)^{th}$ element is

$$\gamma_{jk}(x, x') = \mathrm{Cov}\{Y_j(x), Y_k(x')\}. \tag{3.26}$$

For each pair of locations (x, x'), the matrix $\Gamma(x, x')$ is symmetric i.e., $\gamma_{jk}(x, x') = \gamma_{kj}(x, x')$.

When $Y(x)$ is stationary, $\gamma_{jj}(x, x) = \mathrm{Var}\{Y_j(x)\} = \sigma_j^2$ does not depend on x, and for $j \neq k$, $\gamma_{jk}(x, x')$ depends only on $u = ||x - x'||$. We then define the correlation function of $Y(x)$ as the matrix-valued function $R(u)$ whose $(j,k)^{th}$ element is $\rho_{jk}(u) = \gamma_{jk}(u)/(\sigma_j \sigma_k)$. When $k = j$, the functions $\rho_{jj}(u)$ are the correlation functions of the univariate processes $Y_j(x)$ and are symmetric in u i.e., $\rho_{jj}(-u) = \rho_{jj}(u)$. When $k \neq j$, the functions $\rho_{jk}(u)$, called the cross-correlation functions of $Y(x)$, are not necessarily symmetric but must satisfy the condition that $\rho_{jk}(u) = \rho_{kj}(-u)$.

To define a cross-variogram for $Y(x)$, there are at least two possibilities. The first, and the more traditional, is

$$V_{jk}^*(u) = \frac{1}{2}\mathrm{Cov}[\{Y_j(x) - Y_j(x-u)\}\{Y_k(x) - Y_k(x-u)\}]. \tag{3.27}$$

See, for example, Journel and Huijbregts (1978) or Chilès and Delfiner (1999). Expanding the right-hand side of (3.27) we find that

$$\begin{aligned} V_{jk}^*(u) &= \gamma_{jk}(0) - \frac{1}{2}\{\gamma_{jk}(u) + \gamma_{jk}(-u)\} \\ &= \sigma_j \sigma_k [1 - \frac{1}{2}\{\rho_{jk}(u) + \rho_{jk}(-u)\}]. \end{aligned} \tag{3.28}$$

The similarity between (3.28) and the corresponding relationship between univariate covariance, correlation and variogram functions, as discussed in Section 3.4, is clear.

The second possibility, introduced by Cressie and Wikle (1998) and called by them the *variance-based cross-variogram*, is

$$V_{jk}(u) = \frac{1}{2}\mathrm{Var}\{Y_j(x) - Y_k(x-u)\}. \tag{3.29}$$

Expanding the right-hand side of (3.29) gives

$$V_{jk}(u) = \frac{1}{2}(\sigma_j^2 + \sigma_k^2) - \sigma_j \sigma_k \rho_{jk}(u). \tag{3.30}$$

The expansion (3.30) highlights an apparent objection to (3.29), namely that it mixes incompatible physical dimensions. However, we can overcome this by working with standardised, and therefore dimensionless, variables. An advantage of (3.29) over (3.27) is that it suggests a way of estimating the variogram empirically which does not require the different variables to be measured at a common set of sampling locations.

Using standardised variables reduces the two definitions of the cross-variogram in (3.30) and (3.28) to

$$V_{jk}^*(u) = 1 - \frac{1}{2}\{\rho_{jk}(u) + \rho_{jk}(-u)\}$$

and

$$V_{jk}(u) = 1 - \rho_{jk}(u),$$

respectively, hence

$$V_{jk}^*(u) = \frac{1}{2}\{V_{jk}(u) + V_{jk}(-u)\}.$$

In particular, provided that we use standardised variables, we see that $V_{jk}^*(u) = V_{jk}(u)$ whenever the cross-correlation function $\rho_{jk}(u)$ is symmetric in u.

3.12.2 Bivariate signal and noise

To construct a stationary Gaussian model for bivariate data $(Y_{ij} : i = 1, ..., n_j, j = 1, 2)$ measured at locations x_{ij} we first specify a model for an unobserved bivariate stationary Gaussian process $\{S(x) = (S_1(x), S_2(x)) : x \in \mathbb{R}^2\}$, with bivariate mean zero, variances $\sigma_j^2 = \mathrm{Var}\{S_j(x)\}$ and correlation structure determined by three functions $\rho_{11}(u) = \mathrm{Corr}\{S_1(x), S_1(x-u)\}$, $\rho_{22}(u) = \mathrm{Corr}\{S_2(x), S_2(x-u)\}$ and $\rho_{12}(u) = \mathrm{Corr}\{S_1(x), S_2(x-u)\}$.

The simplest assumption we can make about the data Y_{ij} is that $Y_{ij} = S_j(x_{ij})$ i.e., the signal at any location x can be observed without error. When the data are subject to measurement error, the simplest assumption is that the Y_{ij} are mutually independent given $S(\cdot)$ and normally distributed,

$$Y_{ij} \sim \mathrm{N}\{\mu_j(x_{ij}) + S_j(x_{ij}), \tau_j^2\} : i = 1, \ldots, n_j; j = 1, 2. \tag{3.31}$$

Under this model, each dimension of the response separately follows a univariate Gaussian model, whilst dependence between the two response dimensions is modelled indirectly through the structure of the unobserved process $S(\cdot)$. The conditional independence assumption in (3.31) invites the interpretation that the parameters τ_j^2 represent the measurement error variances in each of the two response dimensions. A less restrictive assumption than (3.31) would be to allow the measurement errors associated with $Y(x) = \{Y_1(x), Y_2(x)\}$ to be correlated. This would only affect the model at locations where both of $Y_1(x)$ and $Y_2(x)$ are measured; where only one of the $Y_j(x)$ is measured, (3.31) would still hold.

Correlated measurement errors might be particularly appropriate if, as already discussed in the univariate setting, we want the nugget effect to include spatial variation on scales smaller than the smallest inter-point distance in the sampling design.

In the case of spatially independent error terms, the mean and covariance structure of the data, Y_{ij}, are given by

$$\mathrm{E}[Y_{ij}] = \mu_j(x_{ij}),$$

$$\mathrm{Var}\{Y_{ij}\} = \tau_j^2 + \sigma_j^2$$

and, for $(i,j) \neq (i',j')$,

$$\mathrm{Cov}\{Y_{ij}, Y_{i'j'}\} = \sigma_j \sigma_{j'} \rho_{jj'}(||x_{ij} - x_{i'j'}||).$$

Note in particular that non-zero error variances τ_j^2 induce discontinuities at the origin in the covariance structure of the measurement process.

3.12.3 Some simple constructions

In order to construct particular bivariate models, we need to specify explicit forms for the two mean functions $\mu_j(x)$ and for the covariance structure of $S(\cdot)$. With regard to the means, in practice the easiest models to handle are those in which the means are linear functions of spatial explanatory variables, as was also true in the univariate case. With regard to the covariance structure, the univariate models discussed earlier are a natural starting point. However, in extending these to the bivariate case, we need to be sure that the required positive definiteness conditions are not violated. Note that these require that arbitrary linear combinations of either or both of the response dimensions should have non-negative variances. A simple way to ensure that this is the case is to build a bivariate model explicitly from univariate components. The same holds, with the obvious modifications, for multivariate processes of dimension $d > 2$.

A common-component model

One example of an explicit bivariate construction is the following. Suppose that $S_0^*(\cdot)$, $S_1^*(\cdot)$ and $S_2^*(\cdot)$ are independent univariate stationary Gaussian processes with respective covariance functions $\gamma_j(u) : j = 0, 1, 2$. Define a bivariate process $S(\cdot) = \{S_1(\cdot), S_2(\cdot)\}$ to have components

$$S_j(x) = S_0^*(x) + S_j^*(x) : j = 1, 2.$$

Then, by construction, $S(\cdot)$ is a valid bivariate process with covariance structure

$$\mathrm{Cov}\{S_j(x), S_{j'}(x-u)\} = \gamma_0(u) + I(j = j')\gamma_j(u)$$

where $I(\cdot)$ is the indicator function, equal to one if its logical argument is true, zero otherwise. Note that if, as is typically the case, the covariance functions $\gamma_j(u)$ are non-negative valued, then this construction can only generate non-negative cross-covariances between $S_1(\cdot)$ and $S_2(\cdot)$. In practice this is often the case or, if the two variables are inversely related, can be made so by reversing the sign of one of the components. The common-component construction

extends to processes of dimension $d > 2$ in which all of the components $S_j(x)$ share an underlying common component $S_0^*(x)$. Note, however, that the simple device of applying a change of sign to $S_0(x)$ obviously cannot induce an arbitrary mix of positive and negative cross-covariances. Also, as written the construction implicitly assumes a common measurement scale for all of the component processes. When this is not the case, the model requires an additional $d-1$ scaling parameters so that the common component $S_0^*(x)$ is replaced by $S_{0j}^*(x) = \sigma_{0j} R(x)$ where $R(x)$ has unit variance.

Linear combinations of independent components

Another simple construction is to begin with two, or more generally d, independent univariate processes $U_k(x)$ and define $S_j(x)$ as a linear combination,

$$S_j(x) = \sum_{j=1}^{d} a_{kj} U_j(x),$$

or in vector-matrix notation,

$$S(x) = AU(x). \tag{3.32}$$

Without loss of generality, we can assume that each process $U_k(x)$ has unit variance. If $U_k(x)$ has correlation function $\rho_k(\cdot)$, it follows that the matrix-valued covariance function of $S(x)$ is

$$\Gamma(x, x') = ARA', \tag{3.33}$$

where R is the diagonal matrix with diagonal entries $R_{kk} = \rho_k(x - x')$. In the special case where $\rho_k(u) = \rho(u)$, (3.33) reduces to $\Gamma(x, x') = B\rho(x - x')$. This is sometimes called the *proportional covariance model* (Chilès and Delfiner, 1999). The assumption that all of the $U_k(x)$ share a common correlation function reduces the number of parameters in the model to manageable proportions, but otherwise does not seem particularly natural.

Schmidt and Gelfand (2003) use a variant of (3.32) in which there is a natural ordering of the components of $S(x)$ so that $S_1(x)$ depends on $U_1(x)$ only, $S_2(x)$ depends on $U_1(x)$ and $U_2(x)$, and so on. Gelfand, Schmidt, Banerjee and Sirmans (2004) extend this model to allow the non-zero elements of the A_i to depend on location, x.

The linear model of co-regionalisation

By construction, we can also obtain valid models by adding linear combinations of $p \geq 2$ models with independent components. Hence, we can define a model for a d-dimensional process $S(x)$ as

$$S(x) = \sum_{i=1}^{p} A_i U^i(x), \tag{3.34}$$

where now each $U^i(x) = \{U_1^i(x), \ldots, U_d^i(x)\}$ is a set of d independent univariate processes and A_i is a $d \times d$ matrix. In practice, models of this kind would be

very poorly identified without some restrictions being placed beforehand on the processes $U_k^i(x)$. In the *linear model of co-regionalisation*, these restrictions are that each term on the right-hand side of (3.34) is a proportional covariance model. This again raised the question of whether the resulting savings in the number of unknown parameters has a natural scientific interpretation or is merely a pragmatic device.

How useful are standard classes of multivariate model?

The question is worth asking because, as the examples above illustrate, even very simple multivariate constructions quickly lead to models with either large numbers of parameters and consequent problems of poor identifiability, or potentially severe restrictions on the allowable form of cross-correlation structure. A better modelling strategy than an empirical search through a richly parameterised standard model class may be to build multivariate models by incorporating structural assumptions suggested by the context of each specific application; see, for example, Knorr-Held and Best (2001), who use the common component model in an epidemiological setting where it has a natural interpretation.

3.13 Computation

We first show how to use **geoR** to compute and plot standard correlation functions. The function `cov.spatial()` has an argument `cov.model` which allows the user to choose from a set of correlation families. Options include the Matérn, powered exponential, spherical and wave families discussed earlier in this chapter; a complete list can be obtained by typing *`help(cov.spatial)`*. Below, we show the commands used to produce Figure 3.2. Similar commands were used for Figure 3.5 and Figure 3.7.

```
> x <- seq(0, 1, l = 101)
> plot(x, cov.spatial(x, cov.model = "mat", kappa = 0.5,
+     cov.pars = c(1, 0.25)), type = "l", xlab = "u", ylab = expression(rho(u)),
+     ylim = c(0, 1))
> lines(x, cov.spatial(x, cov.model = "mat", kappa = 1.5,
+     cov.pars = c(1, 0.16)), lty = 2)
> lines(x, cov.spatial(x, cov.model = "mat", kappa = 2.5,
+     cov.pars = c(1, 0.13)), lty = 3)
```

We now illustrate the use of the **geoR** function `grf()` for generating simulations of two-dimensional Gaussian processes. We encourage the reader to experiment with different input parameters so as to obtain an intuitive understanding of the different ways in which the model parameters affect the appearance of the simulated realisations. The arguments to `grf()` specify the model and the locations for which simulated values are required. The locations can be specified to form a regular lattice, a completely random pattern, or a

configuration supplied explicitly as a set of (x, y) coordinates. For example, to produce Figure 3.4 we used the following commands.

```
> set.seed(159)
> image(grf(100^2, grid = "reg", cov.pars = c(1, 0.25)),
+     col = gray(seq(1, 0.1, l = 51)), xlab = "", ylab = "")
> set.seed(159)
> image(grf(100^2, grid = "reg", cov.pars = c(1, 0.13),
+     cov.model = "mat", kappa = 2.5), col = gray(seq(1,
+     0.1, l = 51)), xlab = "", ylab = "")
```

Using the R function `set.seed()` ensures that simulations are generated with the same random number seed, hence differences between the simulated realisations are due only to the different values of the model parameters. In the example above, the realisation covers $n = 100^2 = 10,000$ locations, whilst the argument `grid="reg"` instructs the function to generate the locations in a 100 by 100 regular square lattice.

For the simulations of the anisotropic model in Figure 3.10 we used the argument `aniso.pars` to specify the anisotropy angle and ratio, as follows.

```
> set.seed(421)
> image(grf(201^2, grid = "reg", cov.pars = c(1, 0.25),
+     aniso.pars = c(pi/3, 4)), col = gray(seq(1, 0, l = 51)),
+     xlab = "", ylab = "")
> set.seed(421)
> image(grf(201^2, grid = "reg", cov.pars = c(1, 0.25),
+     aniso.pars = c(3 * pi/4, 2)), col = gray(seq(1, 0,
+     l = 51)), xlab = "", ylab = "")
```

The function `grf()` allows the user to select from several algorithms for generating the simulated realisations, including an automatic link to the function `GaussRF()` within the R package **RandomFields** written by Martin Schlather. To invoke this link, the user specifies the optional argument `method="RF"` in the call to the function `grf()`. At the time of writing, the default in the latest version of the **geoR** package is to use the Choleski factorisation for $n \leq 500$, and the link to `GaussRF()` for $n > 500$. The **RandomFields** package is also available at the CRAN website, *http://cran.r-project.org*.

Note also that Håvard Rue has written very efficient code, available for download at *http://www.math.ntnu.no/∼hrue/GMRFLib*, for simulation of Gaussian processes on very large numbers of locations using an approximating Markov random field, as described in Section 3.10. Rue and Held (2005) provide details on the methods and on the use of the software.

Figure 3.16 shows two further examples of simulations generated by `grf()`, using the commands below. The first call to the function produces the simulation shown in the left-hand panel, a realisation of a stationary Gaussian model with mean $\mu = 0$, variance $\sigma^2 = 1$ and Matérn correlation function with $\kappa = 1.5$ and $\phi = 0.15$. The simulation generates 100 values at locations distributed completely at random over the unit square. The right panel shows simulated values at the 52 locations of the elevation data from Example 1.1. In this case,

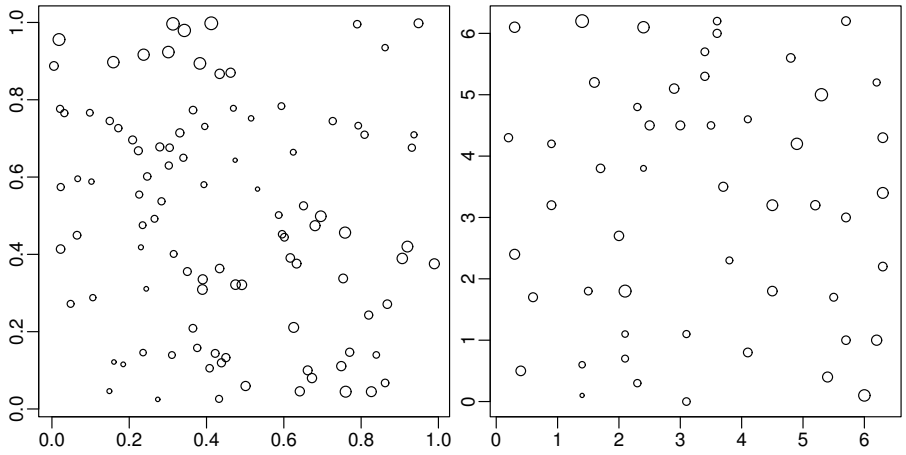

Figure 3.16. Realisations of two stationary Gaussian processes on irregularly distributed sets of locations. See text for detailed specifications.

we have used a stationary Gaussian model with mean $\mu = 850$, nugget variance $\tau^2 = 100$, signal variance $\sigma^2 = 3500$ and Matérn correlation function with $\kappa = 2.5$ and $\phi = 0.8$.

```
> sim1 <- grf(100, cov.pars = c(1, 0.15), cov.model = "matern",
+     kappa = 1.5)
> points(sim1)
> data(elevation)
> sim2 <- grf(grid = elevation$coords, cov.pars = c(3500,
+     0.8), nugget = 100)
> sim2$data <- sim2$data + 850
> points(sim2)
```

3.14 Exercises

3.1. Consider a one-dimensional spatial process $S(x) : x \in \mathbb{R}$ with mean μ, variance σ^2 and correlation function $\rho(u) = \exp(-u/\phi)$. Define a new process $R(x) : x \in \mathbb{R}$ by the equation

$$R(x) = (2\theta)^{-1} \int_{x-\theta}^{x+\theta} S(u)du.$$

Derive the mean, variance and correlation function of $R(\cdot)$. Comment briefly.

3.2. Is the following a legitimate correlation function for a one-dimensional spatial process $S(x) : x \in \mathbb{R}$?

$$\rho(u) = \begin{cases} 1-u & : \quad 0 \leq u \leq 1 \\ 0 & : \quad u > 1 \end{cases}$$

Give either a proof or a counter-example.

3.3. Derive a formula for the volume of the intersection of two spheres of equal radius, ϕ, whose centres are a distance u apart. Compare the result with the formula (3.8) for the spherical variogram and comment.

3.4. Consider the following method of simulating a realisation of a one-dimensional spatial process on $S(x) : x \in \mathbb{R}$, with mean zero, variance 1 and correlation function $\rho(u)$. Choose a set of points $x_i \in \mathbb{R} : i = 1, \ldots, n$. Let R denote the correlation matrix of $S = \{S(x_1), \ldots, S(x_n)\}$. Obtain the singular value decomposition of R as $R = D\Lambda D'$ where λ is a diagonal matrix whose non-zero entries are the eigenvalues of R, in order from largest to smallest. Let $Y = \{Y_1, \ldots, Y_n\}$ be an independent random sample from the standard Gaussian distribution, $\mathrm{N}(0, 1)$. Then the simulated realisation is

$$S = D\Lambda^{\frac{1}{2}}Y. \tag{3.35}$$

Write an R function to simulate realisations using the above method for any specified set of points x_i and a range of correlation functions of your choice. Use your function to simulate a realisation of S on (a discrete approximation to) the unit interval $(0, 1)$.
Now investigate how the appearance of your realisation S changes if in (3.35) you replace the diagonal matrix Λ by a truncated form in which you replace the last k eigenvalues by zeros.

3.5. Consider a spatial process $S(\cdot)$ defined by

$$S(x) = \int w(u)S^*(x-u)du$$

where $w(u) = (2\pi)^{-1}\exp(-||u||^2/2)$ and $S^*(\cdot)$ is another stationary Gaussian process. Derive an expression for the correlation function, $\rho(u)$ say, of $S(\cdot)$ in terms of $w(\cdot)$ and the correlation function, $\rho^*(u)$ say, of $S^*(\cdot)$. Give explicit expressions for $\rho(u)$ when $\rho^*(u)$ is of the form:

(a) pure nugget, $\rho^*(u) = 1$ if $u = 0$, zero otherwise;
(b) spherical;
(c) Gaussian.
(d) In each case, comment on the mean square continuity and differentiability properties of the process $S(\cdot)$ in relation to its corresponding $S^*(\cdot)$.