Contents

Chapter 1

Basics on Random Fields

In this chapter we start by considering the definition of a Gaussian random field and review its main geometric properties. We briefly consider the spectral representation of the correlation functions. We discuss the problem of making inference for Gaussian processes from a Bayesian viewpoint and discuss the main computational issues.

1.1 Basic definitions

A *random field, random function* or *stochastic process* $X(s)$, defined on $S = \mathbb{R}^n$, is a function whose values are random variables, for any value of s. In this notes we will use all those names indistinguishably. When explicit mention is to be made of the fact that $X(s)$ is a random variable, we will use the notation $X(s, \omega)$. A more formal definition can be found, for example, in Billingsley (1986). In this notes S will often be $\mathbb{R}^2$. We will often refer to the elements of S as the *sites* or *locations*. A random field is described by the finite dimensional distributions corresponding to collections $s_1, \ldots, s_n$ of points in S. A *Gaussian* random field corresponds to the case where all such finite dimensional distributions are normal. More specifically

Definition 1.1 *A* **Gaussian random field** *is a random field where all the finite-dimensional distributions, say, $F(s_1, \ldots, s_n)$ are multivariate normal distributions, for any choice of n and $s_1, \ldots, s_n$.*

A multivariate normal distributions is fully specified by its mean and its covariance matrix. So all we need in order to specify a Gaussian process are functions $m(s)$ and $C(s, s')$ that correspond, respectively, to the mean at location s and the covariance between the process at location s and location s', i.e. $m(s) = E(X(s))$ and $C(s, s') = \text{cov}(X(s), X(s')$. This is a consequence of the so called Kolmogorov existence theorem. The reader is referred to Billingsley (1986) for more details. $C(s, s')$ is a valid covariance function if it satisfies the following property:

Definition 1.2 *$C(s, s')$ is* **positive definite** *if for any positive integer n, $s_j \in S$ and $c_j \in \mathbb{R}$ for $j = 1, \ldots, n$*

$$\sum_{i,j} c_i c_j C(s_i, s_j) > 0$$

and the expression above is equal to 0 *if and only if* $c_i = 0$ *for all* i.

Determining if a function is positive definite is in general a difficult task. One of the most commonly used tools to obtain classes of valid covariance functions is spectral analysis. Very often the *correlation function* $\rho(s, s') = C(s, s')/\sqrt{C(s, s)C(s', s')}$ is used. Notice that $C(s, s')$ defines a variance function $\sigma^2(s) = C(s, s)$.

1.1.1 Stationarity

Making inference for stochastic processes is usually simplifies by the assumption of some sort of symmetry in its distribution. The most common one is the notion that arbitrary translations don not change the distribution of the process.

Definition 1.3 *A random field is* **strictly stationary** *if for any finite collection of sites* $s_1, \ldots, s_n$ *and any* $u \in S$*, the joint distribution of* $(X(s_1), \ldots, X(s_n))$ *is the same as that of* $(X(s_1 + u), \ldots, X(s_n + u))$.

A less restrictive definition of stationarity requires that the mean and covariance functions be invariant under translations, i.e. $m(s) = m, \forall s \in S$ and $C(s, s + u) = C(u)$. This type of stationarity if usually referred to as *weak stationarity* (Cressie, 1993). Stationarity in the strict sense clearly implies weak stationarity. The opposite is not true in general. For Gaussian processes the two conditions are equivalent. Notice that any stationary process must have constant variance, so $C(u) = \sigma^2\rho(u)$.

Following the tradition in the geostatistical literature (Matheron, 1963), another type of stationarity can be defined using the *variogram*.

Definition 1.4 *Assume that* $E(X(s + u) - X(s)) = 0$ *, then the* **variogram** *is defined as*

$$E(X(s + u) - X(s))^2 = var(X(s + u) - X(s)).$$

The process is **intrinsically stationary** *if the variogram depends only on* u*. So we can write* $var(X(s + u) - X(s)) = 2\gamma(u)$ *where* $\gamma(u)$ *is called the* **semi-variogram**.

Notice that the former definition is based on the second moment of the difference $X(s+u)-X(s)$. If the covariance of the process exists, then $\gamma(u) = C(0)-C(u)$. So, we can recover the semi-variogram from the covariance function. Also, if the process is weakly stationary, then it is intrinsically stationary. If the semi-variogram is given, we need an additional condition on γ to obtain the covariance function. In fact we have that

$$C(u) = C(0) - \gamma(u) = \lim_{||h||\to\infty} \gamma(h) - \gamma(u).$$

The limit is valid only if the association between two locations vanishes as the locations becomes infinitely distant. Clearly the limit may not exist, so strict stationarity does not imply weak stationarity.

A much stronger form of symmetry is provided by the condition of *isotropy*. This corresponds to a radial symmetry where the dependence between sites is determined solely by their distance.

Model	$C(\tau)$
Spherical	$C(\tau) = \begin{cases} \sigma^2(1 - 3/2\phi\tau + 1/2(\phi\tau)^3) & \text{if } 0 < \tau \leq 1/\phi \\ 0 & \text{if } \tau \geq 1/\phi \end{cases}$
Powered exponential	$C(\tau) = \sigma^2 \exp(-\lvert\phi\tau\rvert^\nu) \;\; \tau > 0 \;\; 0 < \nu \leq 2$
Rational Quadratic	$C(\tau) = \sigma^2 \left(1 + \frac{\tau^2}{(1+\phi\tau^2)}\right) \;\; \tau > 0$
Wave	$C(\tau) = \sigma^2 \frac{\sin(\phi\tau)}{\phi\tau} \;\; \tau > 0$
Matérn	$C(\tau) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)}(2\sqrt{\nu}\tau\phi)^\nu K_\nu(2\sqrt{\nu}\tau\phi) \;\; \tau > 0 \;\; \nu > 0$

Table 1.1: Examples of commonly used covariance functions for isotropic random fields. In all cases $C(\tau) = 0$ if $\tau < 0$. A 'nugget' ψ^2 is added to the model by assuming that $C(0) = \psi^2$.

Model	$\gamma(\tau)$
Spherical	$\gamma(\tau) = \begin{cases} \psi^2 + \sigma^2(3/2\phi\tau - 1/2(\phi\tau)^3) & \text{if } 0 < \tau \leq 1/\phi \\ \psi^2 + \sigma^2 & \text{if } \tau \geq 1/\phi \end{cases}$
Powered exponential	$\gamma(\tau) = \psi^2 + \sigma^2(1 - \exp(-\lvert\phi\tau\rvert^\nu)) \;\; \tau > 0 \;\; 0 < \nu \leq 2$
Rational Quadratic	$\gamma(\tau) = \psi^2 + \frac{\sigma^2\tau^2}{(1+\phi\tau^2)} \;\; \tau > 0$
Wave	$\gamma(\tau) = \psi^2 + \sigma^2 \left(1 - \frac{\sin(\phi\tau)}{\phi\tau}\right) \;\; \tau > 0$
Matérn	$\gamma(\tau) = \psi^2 + \sigma^2 \left(1 - \frac{(2\sqrt{\nu}\tau\phi)^\nu}{2^{\nu-1}\Gamma(\nu)} K_\nu(2\sqrt{\nu}\tau\phi)\right) \;\; \tau > 0 \;\; \nu > 0$

Table 1.2: Examples of commonly used semi-variograms for isotropic random fields. For all these cases $\gamma(\tau) = 0$ for $\tau \leq 0$.

Table 1.3: Semivariograms

Definition 1.5 *A stationary random field is* **isotropic** *if the covariance function depends on distance alone, i.e.* $C(s, s') = C(\tau)$ *where* $\tau = ||s - s'||$.

Isotropic random processes may seem very restrictive. Nevertheless they are extremely popular as model fitting tools. This is due to the fact that their properties are clearly understood and that it is possible to specify flexible classes of isotropic covariance functions that capture wide ranges of natural phenomena. Isotropic covariances are often used as building blocks for more complicated models. In Table (1.1) and Table (1.3) we report some of the most commonly used covariances and semivariograms, respectively. For details see Banerjee et al. (2004).

According to the traditional geostatistical terminology, $\psi^2 = \lim_{\tau \to 0^+} \gamma(\tau)$ is called the *nugget*. $\psi^2 + \sigma^2 = \lim_{\tau \to \infty} \gamma(\tau)$ is called the *sill*. When $\gamma(\tau)$ reaches the sill for a finite value of τ, the inverse of such value, which is a function of ϕ is called the range. As can be seen from Table (1.3) the range is often infinite. Nevertheless $1/\phi$ is usually referred to as the range. More precisely, one can define an *effective range* as the value of τ for which the correlation drops below a small value (say, 0.05).

1.1.2 Other forms of symmetry

The definition of isotropy requires the use of a norm for the domain S. Since S is an Euclidean space we have implicitly assumed that the Euclidean norm is used. A more general norm $||s||_K = \sqrt{s'Ks}$ for a positive definite matrix K can be used. If ρ is a valid correlation function for an isotropic random field, we can define $\rho_K(s, s') = \rho(||s - s'||_K)$. This provides a valid stationary correlation function which is *geometric anisotropic*. Geometric anisotropy can be a useful way of relaxing the assumption of isotropy without adding too many parameters to the description of the correlation function.

A simple way of creating valid correlation functions for high dimensional spaces is by assuming separability. Let $u = (u_1, \ldots, u_k),\ k \leq n,\ u_i \in \mathbb{R}^{n_i}$, then

$$\rho(h) = \rho_1(h_1) \cdots \rho_k(h_k)$$

is a valid correlation function if and only if each ρ_i is a valid correlation function and $\sum_i n_i = n$. ρ is said to be a *separable* correlation function. A very common example of the use of separable correlation functions is that of random processes in space and time, where $\rho(x, y, z, t) = \rho(x, y, z)\rho(t)$.

1.2 Geometric properties

When considering random functions we are interested in characterizing their smoothness and differentiability. These are key properties that have to be taken into account when choosing the family of models that is most suited for a given problem. Covariance functions are responsible for the smoothness properties of Gaussian random fields. Depending on the choice of covariance function, a Gaussian process can have sample paths that range from discontinuous to analytic. Abrahamsen (1997) and Paciorek (2003) give very clear accounts of the theory of smoothness properties for random fields. Stein (1999) discusses asymptotic results for several popular models.

To study the smoothness of a random field one needs to consider converge of sequences $X(s_n)$ of random variables. This can be done in several ways. Four types of convergence are usually considered, one based on *mean squares* distance, one based on *almost surely* proximity a third based on proximity of *sample paths* and a fourth based on *proximity in probability*. For the continuity of a random field we have the following definitions. Analogous definitions can be obtained for differentiability.

Definition 1.6

1. *A random field X has* **continuous sample paths with probability one** *in B if, for every sequence s_n such that $||s_n - s|| \to 0$ as $n \to \infty$, then*

$$Pr(\omega : |X(s_n, \omega) - X(s, \omega)| \to 0, \quad as \ \ n \to \infty,\ \forall s \in B) = 1$$

2. *A random field X is* **almost surely continuous** *in B if for every sequence s_n such that $||s_n - s|| \to 0$ as $n \to \infty$, then*

$$Pr(\omega : |X(s_n, \omega) - X(s, \omega)| \to 0, \quad as \ \ n \to \infty) = 1 \quad \forall s \in B$$

3. *A random field X is* **mean square continuous** *in B if for every sequence s_n such that $||s_n - s|| \to 0$ as $n \to \infty$, then*

$$E(|X(s_n) - X(s)|^2) \to 0, \quad as \;\; n \to \infty \quad \forall s \in B$$

provided the expectation exists.

4. *A random field X is* **continuous in probability** *in B if for every sequence s_n such that $||s_n - s|| \to 0$ as $n \to \infty$, then*

$$\lim_{n\to\infty} Pr(\omega : |X(s_n, \omega) - X(s, \omega)| > \delta) = 0, \quad for \; any \;\; \delta > 0 \quad \forall s \in B$$

Sample path continuity means that the probability of a discontinuous sample path is zero. Geometrically speaking, this is the type of continuity that is most interesting. Almost sure continuity allows discontinuities in B with probability zero. Thus, sample path continuity is a stronger property that almost sure continuity. For Gaussian random fields, mean square continuity is a necessary and almost sufficient condition for sample path continuity. See Paciorek (2003) for details. Continuity in probability is implied by mean square continuity, as can be seen by direct application of Chebyshev's inequality. Mean square properties are the most tractable ones. Thus, especially for Gaussian processes, are the ones that have received the most attention. The following theorem provides the link between the behavior of the covariance function and the mean square continuity of a random process.

Theorem 1.1 *Assume that $E(X(t))$ is continuous. Then, a random field $X(t)$ is mean square continuous at t if and only if its covariance function $C(s, s')$ is continuous at $s = s' = t$.*

A proof of this theorem can be found in Abrahamsen (1997). A direct consequence of this theorem is the following corollary.

Corollary 1.1 *A stationary random field $X(s)$ is mean square continuous at $s \in S$ if and only if its correlation function $\rho(h)$ is continuous at 0.*

Notice that the above result implies that when a nugget is added to an isotropic correlation function like the ones presented in Table (1.1), the resulting random field is not mean square continuous.

Theorem 1.1 can be extended for to the differentiability of a process.

Theorem 1.2 *Consider a random field $X(s)$ with covariance function C and expectation function sufficiently smooth. If the derivative*

$$\frac{\partial^{2\nu} C(s,t)}{\partial s_1^{\nu_1} \cdots \partial s_n^{\nu_n} \partial t_1^{\nu_1} \cdots \partial t_n^{\nu_n}} \tag{1.1}$$

where $\nu = \sum_i \nu_i$, exists and is finite for all $i = 1, \ldots n$ at (s, s), then $X(s)$ is $|\nu|$ times differentiable at s. Moreover, the covariance function of

$$\frac{\partial^{2\nu} X(s)}{\partial s_1^{\nu_1} \cdots \partial s_n^{\nu_n}}$$

is given by (1.1).

A proof of this theorem can be found in Cramér and Leadbetter (1967). Clearly, for a stationary random field, mean square smoothness is a consequence of the smoothness of the covariance function at zero. So, when choosing a particular covariance family for modeling purposes, it is important to pay attention to the smoothness that it will induce in the corresponding random field. Families that are either too smooth or too rough may not realistically capture the variability of the phenomenon under study.

Stein (1999) advocates the use of correlation families that provide different smoothness properties. Of the covariances listed in (1.1) the powered exponential and the Matérn are the ones that have a parameter that determines the smoothness. The powered exponential is a popular family since it encompasses the exponential ($\nu = 1$) and the Gaussian ($\nu = 2$), two of the most commonly used correlation functions. As can be easily seen by differentiating the correlation function, the family has a first derivative equal to $-\infty$ for $0 < \nu < 1$, corresponding to a very erratic behavior of the sample paths. For $1 \leq \nu \leq 2$ the derivative is a constant. But the second derivative exists only for $\nu = 2$. In this case the correlation is not only differentiable, it is analytic. The fact that the random processes in this family jump from being erratic to being extremely smooth when $\nu = 2$ makes the family very unappealing for modeling purposes. In contrast, the Matérn family depends on a parameters ν that produces a gradual transition from erratic to increasingly smooth sample paths as ν increases in value from 0 to ∞. These results are presented in Yaglom (1986). We notice that the exponential correlation corresponds to $\nu = 1/2$. Whittle (1954) advocates the use of $\nu = 1$ as the default choice of a correlation function for processes on the plane. Paciorek (2003) proves that for $\nu > M$ processes in the Matérn family are at least M times mean square differentiable and for $\nu > 2M$ they are at least M times sample path differentiable. Thus $\nu = 1$ is the threshold value for processes to show differentiability.

1.3 Spectral representation

Consider a correlation function $\rho(\tau)$, then it is possible to use Fourier transforms to obtain a spectral representation of ρ. Finding a spectral representation is useful to establish positive definiteness. The following theorem is known as the Wiener-Khintchine's Theorem and it is a consequence of Bochner's Theorem (Bochner, 1959). According to Bochner's Theorem a real function on $\mathbb{R}^n$ is positive definite if and only if it can be represented as the Fourier transform of a non-negative bounded measure.

Theorem 1.3 *A real function $\rho(\tau)$ on $\mathbb{R}^n$ is a correlation function if and only if it can be represented in the form*

$$\rho(\tau) = \int_{\mathbb{R}^n} e^{i\tau' k} dF(k)$$

where the function $F(k)$ on $\mathbb{R}^n$ is an n-dimensional distribution function.

Wiener-Khintchine's Theorem can be re-phrased by saying that ρ is a correlation function if and only if it can be expressed as the characteristic function of some n-dimensional random variable. Since ρ is real valued, the Fourier integral simplifies to

$$\rho(\tau) = \int \cos(\tau' k) dF(k).$$

When F is continuous, a spectral density f exists and

$$\rho(\tau) = \int e^{i\tau' k} f(k) dk = \int \cos(\tau' k) f(k) dk \ .$$

The spectral density can be obtained from the correlation using the inverse Fourier transform, thus

$$f(k) = (2\pi)^{-n} \int_{\mathbb{R}^n} e^{-i\tau' k} \rho(\tau) d\tau \ .$$

A general strategy for determining if a given function is a valid correlation is to evaluate its spectral density and check if it is non-negative for any $k \in \mathbb{R}^n$. On the other hand, a general strategy for creating valid correlation functions is to consider a non-negative function as a spectral density and find its Fourier transform. It is also useful to notice that the class of all valid correlation function is closed under addition, multiplication, limits and integration.

For isotropic correlation function the Wiener-Khintchine's Theorem takes a simpler form. This is because the n-dimensional Fourier integral can be replaced by a one dimensional integral. For details on the following theorem see Yaglom (1986).

Theorem 1.4 *A real function $\rho(\tau)$, $\tau \in \mathbb{R}$ is correlation function if and only if*

$$\rho(\tau) = 2^{(n-2)/2} \Gamma(n/2) \int_0^\infty \frac{J_{(n-2)/2}(k\tau)}{(k\tau)^{(n-2)/2}} d\Phi(k),$$

where Φ is a distribution function on $\mathbb{R}$ and J are Bessel function of the first kind.

Notice that the representation formula for an isotropic correlation depends on n, the dimension of the location space. Indeed, for higher dimensions, the conditions for an isotropic correlation are more restrictive than for lower dimensions. A correlation that is valid in $\mathbb{R}^n$ must be valid in $\mathbb{R}^{n-1}$, but not the opposite. In general, if we denote by $\mathcal{D}_n$ the class of valid isotropic correlations in $\mathbb{R}^n$ then we have that

$$\mathcal{D}_1 \supset \mathcal{D}_2 \supset \cdots \supset \mathcal{D}_\infty \ .$$

When a spectral isotropic density exists it is related to Φ by the formula

$$\Phi(k) = \frac{2\pi^{n/2}}{\Gamma(n/2)} \int_0^k w^{n-1} f(w) dw \ .$$

Also, it can be seen that the representation theorem takes the following particular forms

$$\rho(\tau) = \tfrac{1}{2} \int_0^\infty \cos(k\tau) f(k) dk \quad \text{for} \rho \in \mathcal{D}_1$$

$$\rho(\tau) = \int_0^\infty J_0(k\tau) f(k) dk \quad \text{for} \rho \in \mathcal{D}_2$$

$$\rho(\tau) = \int_0^\infty \tfrac{\sin(k\tau)}{k\tau} k^2 f(k) dk \quad \text{for} \rho \in \mathcal{D}_3 \ .$$

1.4 White noise

It is of great importance to consider the case where the correlation between $X(s)$ and $X(s+u)$ falls off very rapidly with u. We can approximate the correlation of such a process with the exponential correlation function $\rho(\tau) = \exp(-\phi|\tau|)$, for a very large ϕ. The spectral density for the exponential correlation can be calculated as

$$f(k) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-\phi|\tau|-ik\tau}d\tau = \frac{1}{2\pi}\left(\int_{-\infty}^{0} e^{(\phi-ik)\tau}d\tau + \int_{0}^{\infty} e^{(\phi-ik)\tau}d\tau\right) = \frac{\phi}{\pi(\phi^2+k^2)}.$$

So, if ϕ is very large compared to k, the value of $f(k)$ is almost constant. In other words, we can approximate the spectrum of X by a constant. We can then define the white noise as a Gaussian process with constant spectrum. This corresponds to a correlation whose mass is all concentrated at zero. Such a process is discontinuous and as such is hardly a realistic process for most natural phenomena. Nevertheless, Matérn (1986) proves the following theorem where the covariance of a stationary correlation function is decomposed into a linear combination of white noise correlation and a continuous correlation function.

Theorem 1.5 *If ρ is a stationary correlation function that is continuous everywhere except possibly at zero, then*

$$\rho(\tau) = a\rho_w(\tau) + b\rho_c(\tau), \quad a, b \geq 0$$

where $\rho_w(0) = 1$ and $\rho_w(\tau) = 0$, if $\tau \neq 0$. ρ_c is a stationary correlation function that is continuous everywhere.

So, a random field X can be decomposed into a completely chaotic part, say X_w, and a continuous part, say X_c. Furthermore, the two components are independent. Such decomposition provides a justification for the use of a nugget effect, as is customary in the geostatistical literature.

1.5 Inference for Gaussian processes

The classical approach to spatial prediction is usually termed *kriging*, in honor of a South African mining engineer named D.G. Krige (Krige, 1951). For a mothern presentation of kriging methods and extensions see Cressie (1993). The Bayesian version of kriging is based on considering the likelihood that is induced by a Gaussian process. Assuming a prior on the parameters of such likelihood and obtain the corresponding posterior distribution. Spatial predictions are obtained from the predictive posterior distribution. When the Gaussian process has a correlation function that depends on unknown parameters, the exploration of the posterior usually requires the use of Markov chain Monte Carlo methods (MCMC). In such case, spatial predictions is usually done by obtaining samples from the predictive posterior distribution.

Let $\boldsymbol{X} = (X(s_1), \ldots, X(s_n))'$ be a vector of data corresponding to realizations of a Gaussian process X at n locations. Suppose that the mean function of X is linear on some, possibly location dependent, covariates. So we can assume that $E(\boldsymbol{X}) = K\beta$, for a $n \times p$ matrix of covariates, K. We assume that the correlation can be expressed as a function of

some parameters Ψ. So that $\text{var}(\boldsymbol{X}) = V(\Psi)$. When the variance of X is constant, we have that $\text{var}(\boldsymbol{X}) = \sigma^2 W(\Psi)$. Letting $\theta = (\beta, \sigma^2, \Psi)$, the likelihood is then given by

$$\boldsymbol{X}|\theta \sim N(K\beta, \sigma^2 W(\Psi)) \ .$$

We complete the model by assuming a prior for θ. Notice that we have specified a completely general setting that can be used for the analysis of non-stationary Gaussian processes. The challenging problem of building suitable covariance structures that are not isotropic will be left for Chapter **??**. For the moment we will focus on the isotropic case. If we assume that the correlation of X is given by $\rho(\tau)$ for $\tau \in \mathbb{R}$, then $\text{var}(X(s_i), X(s_j)) = \rho(||s_i - s_j||)$. As an example consider the powered exponential correlation from Table (1.1), then $\Psi = (\phi, \nu)$ and $W(\Psi)_{i,j} = \exp(-\phi||s_i - s_j||^\nu)$.

Selecting a prior for θ can be tricky. Consider the simplest possible model where $\Psi = \phi$ is a scalar that corresponds to the range parameter of an isotropic correlation. Berger et al. (2001) observe that the likelihood is not integrable as a function of θ. So it can be expected that the prior will have a substantial influence on the posterior, especially on the marginal for ϕ. Limiting the domain of ϕ to a bounded set given, for example, by the effective range, would not solve the problem. In fact, theoretically, there will still be an infinite amount of posteriori mass above the upper limit for ϕ. This may imply lack of robustness with respect to the choice of such limit. It is interesting to notice that the likelihood does not seem to carry much information regarding the range parameter. Mardia and Watkins (1989) wrote a controversial paper on the multi-modality of the likelihood for the spherical correlation. They propose the use of a profile likelihood as a solution. Berger et al. (2001) use an integrated likelihood to obtain the reference prior. This is somehow pointing out that integrating or fixing the other parameters is required in order for the likelihood to provide information about the range. In fact Zhang (2004) finds that, in the Matérn family, σ^2 and ϕ can not be consistently estimated.

The reference prior proposed in in Berger et al. (2001), which will be referred to as BDS prior, has the form $\pi(\beta, \sigma^2, \phi) \propto 1/\sigma^2 \pi^R(\phi)$, where

$$\pi^R(\phi) \propto \left\{ tr[H_\phi^2] - \frac{1}{(n-p)}(tr[H_\phi])^2 \right\}^{1/2} ,$$

$$H_\phi = \left(\frac{\partial}{\partial \phi} W(\phi) \right) W(\phi)^{-1} P_\phi^W, \quad P_\phi^W = I - K(K'W(\phi)^{-1}K)^{-1}K'W(\phi)^{-1} \ .$$

Unfortunately when evaluating this prior it is easy to run into numerical problems. Theorem 3 in Berger et al. (2001) provides a second order expansion that increases the numerical accuracy. Paulo (2005) provides a generalization of this prior for separable correlation functions. The posterior distribution for ϕ that results from the use of the reference prior is

$$\pi(\phi|\boldsymbol{X}) \propto |W(\phi)|^{-1/2}|K'W(\phi)K|^{1/2}(S^2(\phi))^{(n-p)/2}\pi^R(\phi) \tag{1.2}$$

where $S^2(\phi) = (\boldsymbol{X} - K\hat{\beta}(\phi))'W^{-1}(\phi)(\boldsymbol{X} - K\hat{\beta}(\phi))$ and $\hat{\beta}(\phi) = (K'W^{-1}(\phi)K)^{-1}K'W^{-1}(\phi)\boldsymbol{X}$. This posterior is proper, moreover, the marginal $\pi^R(\phi)$ is an integrable function. As discussed in Berger et al. (2001), posterior intervals based on the BDS prior are likely to provide better

coverage than prior other non-informative priors as well as methods based on maximum likelihood. These results are confirmed in Schmidt et al. (2005).

The BDS prior provides a default choice for a model with unknown mean, variance and range parameters. A class like Matérn's requires the specification of a smoothness parameter. The formula defining the BDS prior can be applied to a bivariate parameter. Whether this produces a proper posterior for both is an open question. Alternatively, it is possible to estimate the smoothness parameter using a model comparison approach. Fix ν and write the reference prior as

$$\pi(\beta, \sigma^2, \phi|\nu) = \frac{C(\nu)\pi^R(\phi|\nu)}{\sigma^2}.$$

Here $c(\nu) = (\int_0^\infty \pi^R(\phi|\nu)d\phi)^{-1}$. Then the marginal density for $\boldsymbol{X}$ under the model that considers a given value of ν is

$$m_\nu(\boldsymbol{X}) = C(\nu)\int_0^\infty |W(\phi)|^{-1/2}|K'W(\phi)K|^{1/2}(S^2(\phi))^{(n-p)/2}\pi^R(\phi|\nu)d\phi \ .$$

As a consequence of the results in Berger et al. (1998) the Bayes factor that is obtained by the ratio of $m_{\nu_1}(\boldsymbol{X})/m_{\nu_2}(\boldsymbol{X})$ is well calibrated for the comparison of the model that assumes that the true value of ν is ν_1 versus the one that assumes that the true value is ν_2. In other words, we can estimate the value of ν by considering a grid of values $\nu_1, \ldots, \nu_k$ and maximizing $m_\nu(\boldsymbol{X})$ over such grid. This corresponds to choosing the most likely model assuming that they are a priori all equally likely and using a properly calibrated procedure. Clearly, it is also possible to perform model averaging.

The posterior predictive distribution of a vector, say, $\boldsymbol{Z}$ is given by $\pi(\boldsymbol{Z}|\boldsymbol{X})$. Writing the joint distribution of $\boldsymbol{Z}$ and $\boldsymbol{X}$ as

$$\begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{X} \end{pmatrix} \sim N\left(\begin{pmatrix} K_Z \\ K_X \end{pmatrix}\beta, \sigma^2 \begin{pmatrix} W_Z(\phi) & W_{ZX}(\phi) \\ W_{XZ}(\phi) & W_X(\phi) \end{pmatrix}\right)$$

we have that $\pi(\boldsymbol{Z}|\boldsymbol{X}, \beta, \sigma^2, \phi) = N(m_{Z|X}, W_{Z|X}(\phi))$ where $m_{Z|X} = K_Z\beta + W_{ZX}(\phi)W_X^{-1}(\phi)(\boldsymbol{X} - K_X\beta)$ and $W_{Z|X} = \sigma^2(W_Z(\phi) - W_{ZX}(\phi)W_X^{-1}(\phi)W_{XZ}(\phi))$. Then

$$\pi(\boldsymbol{Z}|\boldsymbol{X}) = \int\int\int \pi(\boldsymbol{Z}|\boldsymbol{X}, \beta, \sigma^2, \phi)\pi(\beta, \sigma^2, \phi|\boldsymbol{X})d\beta d\sigma^2 d\phi \ .$$

This provides a joint distribution for spatial estimation at any collection of points in S. Point estimation can be done by calculating the posterior predictive expectation $E(\boldsymbol{Z}|\boldsymbol{X})$. This is given by

$$\int\int\int (K_Z\beta + W_{ZX}(\phi)W_X^{-1}(\phi)(\boldsymbol{X} - K_X\beta))\pi(\beta, \sigma^2, \phi|\boldsymbol{X})d\beta d\sigma^2 d\phi \ .$$

Notice that $E(\boldsymbol{X}|\boldsymbol{X}) = \boldsymbol{X}$, and so the spatial estimation is actually interpolating the observations.

1.5.1 Computational methods

When a BDS prior is used to perform Bayesian inference for Gaussian processes the computational issues reduce to one dimensional integrations. This is clear from Equation (1.2) as well as the formulæ for the predictive posterior distribution. Similar results are obtained when conjugate priors are used for β and σ^2. When performing the calculations for spatial interpolation, careful attention has to be given to the numerical accuracy of the methods used to decomposed the matrices. Matrix computations will have to be performed repeatedly and have to be fast and accurate. There are three golden rules to be considered: never invert a matrix explicitly, prefer methods that are based on orthogonal transformations like QR or SVD (see, for example, Golub and Van Loan, 1983) and use methods that account for the particular structure of the matrix. Matrix inversion is a very time consuming operation which is usually unnecessary. Methods based on orthogonal transformation are usually very stable numerically. Some problems produce matrices that have a particular structure and this can be used to speed calculations. For example, sparse matrices can be handled with iterative methods. Toeplitz or circulant matrices, that arise when the data correspond to regular grids, can be decomposed with ad hoc methods like Fast Fourier Transform that are much faster than general decompositions.

Simulation methods are complicated by the fact that the parameters that define the likelihood tend to show very strong correlations. To illustrate this fact, consider the plots in Figure 1.5.1. We used simulated data to obtain the likelihood of a model that assumes an isotropic spatial correlation in the Matérn class. As is observed from the plots the joint posterior has an elongated shape which is difficult to sample from.

The full conditional distributions of β and σ^2 are easily obtained. These correspond, respectively, to a multivariate normal distribution with mean $\hat{\beta}(\phi)$ and variance $\sigma^2(K'W(\phi)^{-1}K)^{-1}$ and an inverse gamma with parameters $n/2$ and $1/2[(\beta-\hat{\beta}(\phi))'K'W(\phi)^{-1}K(\beta-\hat{\beta}(\phi))+S^2(\phi)]$. Samples of ϕ can be obtained using a Metropolis-Hastings (MH) step. We have observed that in many applications this procedure produces a chain that converges very slowly to the equilibrium distribution. In fact, consecutive realizations of ϕ tend to be highly correlated. Although fine tuning of the proposal distribution in the MH is likely to improve things, this will depend on the specific structure of the problem.

An alternative strategy is given by sampling the three parameters as a block. The posterior distribution can be factorized as $\pi(\beta,\sigma^2,\phi|\boldsymbol{X})=\pi(\beta|\sigma^2,\phi,\boldsymbol{X})\pi(\sigma^2|\phi,\boldsymbol{X})\pi(\phi|\boldsymbol{X})$ where $\pi(\beta|\sigma^2,\phi,\boldsymbol{X})$ has been described above, $\pi(\sigma^2|\phi,\boldsymbol{X})$ is an inverse gamma distribution with parameters $(n-p)/2$ and $1/2S(\phi)^2$ and $\pi(\phi|\boldsymbol{X})$ is given Equation (1.2). Since the efficiency of a MH algorithm improves as the jumping distribution is closer to the target distribution we use the proposal distribution

$$g(\beta,\sigma^2,\phi)=\pi(\beta|\sigma^2,\phi,\boldsymbol{X})\pi(\sigma^2|\phi,\boldsymbol{X})g(\phi)$$

for some $g(\phi)$. Then, the acceptance probability of a new sample $(\beta_p,\sigma_p^2,\phi_p)$ when the chain has reached the state $(\beta_c,\sigma_c^2,\phi_c)$, is given by $\min\{1,\alpha\}$, with

$$\alpha=\frac{\pi(\beta_p|\sigma_p^2,\phi_p,\boldsymbol{X})\pi(\sigma_p^2|\phi_p,\boldsymbol{X})\pi(\phi_p|\boldsymbol{X})}{\pi(\beta_c|\sigma_c^2,\phi_c,\boldsymbol{X})\pi(\sigma_c^2|\phi_c,\boldsymbol{X})\pi(\phi_c|\boldsymbol{X})}\frac{\pi(\beta_c|\sigma_c^2,\phi_c,\boldsymbol{X})\pi(\sigma_c^2|\phi_c,\boldsymbol{X})g(\phi_c)}{\pi(\beta_p|\sigma_p^2,\phi_p,\boldsymbol{X})\pi(\sigma_p^2|\phi_p,\boldsymbol{X})g(\phi_p)}=\frac{\pi(\phi_p|\boldsymbol{X})g(\phi_c)}{\pi(\phi_c|\boldsymbol{X})g(\phi_p)}.$$

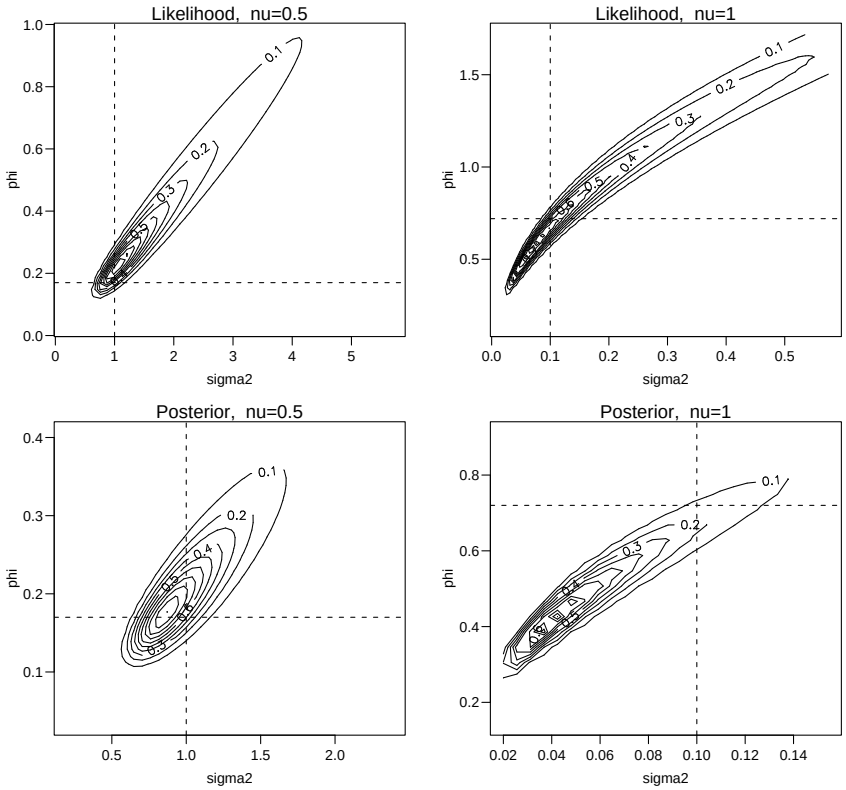

Figure 1.1: Top panels: likelihood function for the range and scale parameters of a Gaussian process model using simulated data. Bottom panels: posterior distribution obtained using the likelihood in the top panels and the BDS prior for the range, location and scale parameters. The true values are denoted by dashed lines. The correlation belongs to the Matèrn class. Two different values of the smoothness parameter are considered: $\nu = .5$ (exponential) and $\nu = 1$ (Whittle).

Table 1.4: Raftery and Lewis' Dependence Factors (Raftery and Lewis, 1992) for simulated data sets with different values of σ^2, ϕ and ν. BLK denotes the blocked Metropolis, RPM denotes Metropolis after re-parameterization and IND denotes independent Metropolis sampling.

ϕ	ν		$\sigma^2 = 0.1$			$\sigma^2 = 1.0$		
			BLK	RPM	IND	BLK	RPM	IND
.17	.5	range	5.0	8.3	83.0	5.0	5.6	265.7
		scale	1.0	7.6	2.3	1.0	4.5	3.6
.17	1	range	6.3	7.3	149.8	4.8	5.7	347.9
		scale	1.0	7.8	4.4	1.0	3.5	2.2
.72	.5	range	4.8	4.9	150.0	6.3	5.7	66.6
		scale	1.1	4.7	5.9	1.1	3.7	4.6
.72	1	range	4.7	5.0	110.8	6.0	7.3	146.2
		scale	1.1	4.7	5.9	1.0	3.7	8.7

Thus, at each step, we only need to generate new values of β and σ^2 if the value of ϕ is acceptable. This means a gain in simulation time additional to the likely mixing improvement due to the blocking. A possible choice for $g(\phi)$ is to consider a random walk on the log scale. In such a case we can further simplify α to

$$\alpha = \frac{\pi(\phi_p|\boldsymbol{X})\phi_c}{\pi(\phi_c|\boldsymbol{X})\phi_p} \quad .$$

We performed a simulation study using a model where the scale, location and range are assumed unknown and estimated via Monte Carlo sampling. We used a sample of one hundred points from a Gaussian random field on the unit square. We produced two sets of simulations using correlations corresponding to the values $\nu = 0.5$ and $\nu = 1$ of the smoothness parameters of a Matèrn class. We considered two different values of the range parameter, $\phi = .17$ and $\phi = .72$ and two values of the scale, $\sigma^2 = .1$ and $\sigma^2 = 1$. We fixed the value of the location β at 0. We used the BDS prior.

In Table 1.5.1 we report the comparison between three sampling schemes using the convergence criteria proposed in Raftery and Lewis (1992). The independence sampler is based of the full conditional distributions of the three parameters sampled independently. The re-parameterization consists in sampling the triple $(\beta, \sigma^2, \phi/\sigma^2)$ and the blocking consisted on using the Monte Carlo scheme proposed in the previous section. We observe that for the scale parameter the dependence factors are uniformly smaller when the blocking is used. For the range parameter the blocking produces results as good or better than the re-parameterization. What is very evident is that that sampling the parameters independently produces very bad results, especially for the range parameter.

In Appendix A.6.2 Banerjee et al. (2004) a method based on slice sampling is discussed. The method is suitable for the case when prior information on β using a normal prior. A clever decomposition of the covariance matrices reduces the amount of matrix computations needed to perform the sampling. Unfortunately it is likely that the method will depend strongly on the choice of prior for ϕ. This issue is not discussed by the authors.

1.6 Incorporating a nugget effect

The traditional geostatistical approach incorporates a nugget in the modeling of spatial variability. A theoretical justification for this fact is provided by Theorem 1.5. Abrahamsen (1997) vehemently opposes the idea that realistic natural phenomena be modeled with discontinuous processes. The only justification for the use of a nugget being the presence of measurement error. Cressie (1993) discusses the nugget effect in detail and proposes to decompose the nugget into two sources of variability, one due to measurement error and a second one due to *micro-scale* effects that correspond to distances smaller than those between locations. the very rapid decay associated with the micro-scale variability should be approximated by a white noise. In practice the data usually carry very little information about such micro-scale variability, if at all present. It is often impossible to separate measurement error from micro-scale variations.

An alternative justification for the nugget is purely computational. Covariance matrices that correspond to isotropic processes observed at many location are usually very ill conditioned. The problem is worse when the random field is very smooth. In particular, for Gaussian correlations and for correlation in the Matérn family with smoothness parameter above 3, most covariance matrices are singular. A numerical solution to this problem is to add a small perturbation or *jitter*, as it is termed in the machine learning literature.

A model that incorporates measurement error can be written in a hierarchical fashion as

$$\begin{aligned} \boldsymbol{X} &= \boldsymbol{\mu} + \varepsilon, \qquad \varepsilon \sim N(0, \psi^2 \boldsymbol{I}) \\ \boldsymbol{\mu} &= K\beta + v, \qquad v \sim N(0, \sigma^2 W(\phi)) \; . \end{aligned}$$

Integrating $\boldsymbol{\mu}$ out of the model produces $\boldsymbol{X} \sim N(K\beta, \psi^2 \boldsymbol{I} + \sigma^2 W(\phi))$. Specifying a default prior for the parameters $(\beta, \psi^2, \sigma^2, \phi)$ is an open problem. In our experience the prior for β will have little influence on the posterior, unless it is very informative. The situation for the other three parameters is quite different. The relationship between ψ^2 and σ^2 is quite important and the prior for ϕ can have a strong influence on the posterior. If ϕ is a range parameter, then, recalling that $\pi^R(\phi)$ is an integrable function, we can use it as a default choice. Following Sansó and Guenni (2004), we find that the specification of a prior for ψ^2 and σ^2 is easier if we re-parameterize using $\chi^2 = \psi^2/\sigma^2$, so that $\boldsymbol{X} \sim N(K\beta, \psi^2(\boldsymbol{I} + 1/\chi^2 W(\phi)))$. This transformation also facilitates some of the calculations for the full conditionals. Recall that the sill corresponds to $\sigma^2 + \psi^2$, so the nugget to sill ratio is equal to $\psi^2/(\psi^2 + \sigma^2) = (1 + 1/\chi^2)^{-1}$. We can specify a priori the amount of variance that we expect to be due to nugget effect with respect to the total variance and use that information to obtain a prior for χ^2. In our experience, the prior for χ^2 has a strong influence on the posterior and, most importantly, the spatial predictions. In fact, it controls how closely the predictive surface follows the observed data. The prior for ψ^2, on the contrary, has little influence on the posterior. We conjecture that the prior $\pi(\beta, \chi^2, \psi^2, \phi) \propto \pi(\chi^2)\pi^R(\phi)/\psi^2$ produces a proper posterior, and could be used as a default, where only one density has to be carefully elicited.

Bibliography

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions. Technical Report 917, Norwegian Computing Center.

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis of Spatial Data*. Chapman and Hall, New York.

Berger, J., De Oliveira, V., and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96:1361–1374.

Berger, J., Pericchi, L., and Varshavsky, J. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya, Ser. A*, 86:79–92.

Billingsley, P. (1986). *Probability and Measure*. John Wiley & Sons, New York, USA, second edition.

Bochner, S. (1959). *Lectures on Fourier Integrals*. Princeton University Press, Princeton, N.J.

Cramér, H. and Leadbetter, M. (1967). *Stationary and Related Stochastic Processes*. John Wiley & Sons, New York.

Cressie, N. A. C. (1993). *Statistics for Spatial Data, Revised Edition*. John Wiley and Sons, New York.

Golub, G. H. and Van Loan, C. F. (1983). *Matrix Computations*. The John Hopkins University Press, Baltimore.

Krige, D. (1951). A statistical approach to some basic mine evaluation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52:119–139.

Mardia, K. and Watkins, A. (1989). On multimodality of the likelihood in the psatial linearl model. *Biometrika*, 76:289–295.

Matérn, B. (1986). *Spatial Variation*. Spinger Verlag, Berlin, second edition.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

Paciorek, C. (2003). *Nonstationary Gaussian Processes fir Regression and Spatial Modelling*. PhD thesis, Department of Statistics, Carnegie Mellon University, Pittsburg, Pennsylvania, USA.

Paulo, R. (2005). Default priors for Gaussian processes. *Annals of Statistics*, 33:556–582.

Raftery, A. E. and Lewis, S. M. (1992). How many iterations in the Gibbs sampler? In Bernardo, J. M., Berger, J. O., Dawid, P., and Smith, A. F. M., editors, *Bayesian Statistics 4*, pages 765–776. Oxford University Press.

Sansó, B. and Guenni, L. (2004). A Bayesian approach to compare observed rainfall data to deterministic simulations. *Environmetrics*, 15:597–612.

Schmidt, A., Conceiçao, M., and Moreira, G. (2005). Investigating the sensitivity of Gaussian processes to the choice of their correlation functions and prior specifications. Technical report, Universidade Federal do Rio de Janeiro.

Stein, M. (1999). *Interpolation of Spatial Data*. Springer-Verlag, New York, USA.

Whittle, P. (1954). On stationary processes in the plane. *Biometrika*, 41:434–449.

Yaglom, A. (1986). *Correlation Theory of Stationary and Related Random Functions I: Basic Results*. Springer Series in Statistics. Springer-Verlag, New York.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpoalations in model-based geostatistics. *Journal of the American Statistical Association*, 76:250–61.