TECHNICAL REPORT
ST-99-08

Bayesian inference in Gaussian model-based geostatistics

Paulo J. Ribeiro Jr.
Lancaster University / UFPR

and

Peter J. Diggle
Lancaster University

Department of Mathematics and Statistics
Lancaster University
LA1 4YF Lancaster-UK

Abstract

The term geostatistics refers to a collection of methods used in the analysis of a particular kind of spatial data, in which measured values Y_i at spatial locations u_i can be regarded as noisy observations from an underlying process in continuous space. In particular, in a geostatistical analysis spatial interpolation or smoothing of the observed values is often carried out by a procedure known as kriging. In its basic form, kriging involves the construction of a linear predictor for an unobserved value of the process, and the form of this linear predictor is chosen with reference to the covariance structure of the data as estimated by a data-analytic tool known as the variogram. Often, no explicit underlying stochastic model is declared.

In this text, we adopt a model-based approach to this class of problems, by which we mean that we start with an explict stochastic model and derive associated methods of parameter estimation, interpolation and smoothing by applying general statistical principles to the observed data under the assumed model. In particular, we use hierarchical spatial linear models whose components are Gaussian stochastic processes with specified parametric covariance structure, and Bayesian methods of inference with independent priors for the separate model parameters.

We present results using this model-based approach, and compare them with classical geostatistical solutions. We derive posterior distributions for model parameters, and predictive distributions for values of the underlying spatial process, taking into account different degrees of parameter uncertainty including uncertainty about some or all of the covariance parameters. We provide a catalogue of posterior and predictive distributions for particular combinations of prior choices and degrees of parameter uncertainty. We discuss computational aspects of the implementation, including non-iterative Monte Carlo inference. Finally, we give illustrative analyses of simulated data.

Keywords: Bayesian inference; geoestatistics; kriging; linear mixed models; spatial prediction.

Contents

1 Introduction

This text considers the analysis of data which can be considered as a partial realisation of a random function (stochastic process) over a region, i. e. a spatially continuous process, as characterised by Cressie (1993). Some examples of such kinds of data are: the concentration of a particular mineral in a soil volume; the water content (or other soil properties like porosity, permeability, fertility) in a soil layer; the concentration of pollutants within an area. Typically, samples are taken at a finite set of points in the region and used to estimate quantities of interest such as: the values of the property of interest at other locations; the mean, maximum or any other quantile at particular locations (or over a sub-area or the whole region); the probability of exceeding a particular threshold. Data of this kind are often called geostatistical data.

A methodological framework for dealing with problems of this kind was motivated by problems in the South African mining industry during the 1950's (Krige, 1951), and subsequently developed during the 1960's, principally by the French geostatistical school based at L'École des Mines, Fontainebleau. Parallel developments in statistics include the early work of Whittle (1954, 1962) and Matérn (1960). Geostatistical methods now find wide application, for example in soil science, meteorology, hydrology and ecology.

An important tool in geostatistics is the *kriging* predictor. The term kriging refers to a least squares linear predictor which, under certain stationarity assumptions, requires at least the knowledge of the covariance parameters and the functional form for the mean of the underlying random function. In most of the practice of geostatistics the parameters are not known. The kriging predictor does not take their uncertainty into account, but uses plug-in estimates as if they were the truth. Bayesian inference provides a way to incorporate parameter uncertainty in the prediction by treating the parameters as random variables and integrating over the parameter space to obtain the predictive distribution of any quantity of interest.

Relations between regression and kriging are discussed in Journel (1989) and Stein & Corsten (1991). In Omre (1987) the mean part of the model is assumed to be a random process with known parameters and an estimator is proposed for the variogram. Omre & Halvorsen (1989) and Omre, Halvorsen & Berteig (1989) assume a linear form for the mean part of the model. In their work the mean parameter is considered unknown and it is shown that the choice of prior distribution for the mean leads to a continuum of methods between simple and universal kriging. In their approach the covariance parameters are obtained using an estimator for the covariance function conditional on the mean parameters. Kitanidis (1978) includes also the covariance parameters as unknown quantities and derives results using conjugate prior distributions, together with a Gaussian assumption for the underlying random field. The reulting predictor is no longer linear. A similar approach is taken by Handcock & Stein (1993) using Jeffrey's prior for the mean and scale parameters and the Matérn class for the correlation function. Cressie (1993) summarises Bayesian approaches for geostatistics. Some non-Gaussian random fields are considered in De Oliveira, Kedem & Short (1997). They introduce a BTG (Bayesian transform Gaussian) model which considers a parametric family of monotonic transformations and accounts for parameter uncertainty in the predictions. A Bayesian variogram fitting procedure, using a finite mixture of Bessel functions, is suggested by Ecker & Gelfand (1997). Woodbury (1989) approaches the kriging problem from the Bayesian updating perspective which does not requires computation of variograms. Searching for methods suitable as a tool to interpolate over time Le & Zidek (1992) proposes a Bayesian alternative to kriging. Instead of a parametric correlation function they adopt an Inverse-

Wishart prior for the covariance matrix. Handcock & Wallis (1994) extends the methods in Handcock & Stein (1993) to spatial temporal modelling.

In this text, we take a model-based approach. By this, we mean that we declare at the outset an explicit stochastic model for the data, then use established statistical principles to derive associated methods of inference. In particular, we shall assume a hierarchical linear Gaussian model and use Bayesian methods of inference to allow for parameter uncertainty in the evaluation of predictive distributions for quantities of interest. Throughout, we emphasise connections between this model-based approach and classical geostatistical methods.

In Section 2 a general model specification is presented and discussed. Section 3 reviews some results from the Bayesian framework. In Section 4 a particular case of the model is considered and Bayesian results are derived for the different levels of parameter uncertainty. Extensions to more general models are considered in Section 4.5. Results for simulated data are presented in Section 5. Some final remarks are made in Section 6. The methods can be implemented using the S-PLUS library `geoS` (Ribeiro Jr & Diggle 1999). Appendix D provides details about this library.

2 A model for continuous spatial processes

This section describes a Gaussian spatial linear mixed model using two alternative specifications. A simulated data set illustrates the main features of the model. The Gaussian assumption and its consequences are discussed. In order to make predictions at unsampled locations, the model for the data is expanded including also the variable at these prediction locations. Based on the model, prediction results are derived assuming that all parameters are known, and are compared with conventional geostatistical methods.

2.1 Model specification

Consider a finite set of spatial sample locations $u_1, u_2, \ldots, u_n$ within a region $\mathcal{D}$ and denote $\boldsymbol{u} = (u_1, u_2, \ldots, u_n)$. Geostatistical data consist of measurements taken at the sample locations $\boldsymbol{u}$. For the model specification considered here, it will be assumed that only one measurement, of only one variable, is taken at each location. A more general model formulation is presented in Appendix A.

The data vector is denoted by $\boldsymbol{y}(\boldsymbol{u}) = (y(u_1), \ldots, y(u_n))$, measurements of a random vector $\boldsymbol{Y}$ which represents the variable under consideration. In geostatistics, the data are regarded as being a realisation of a spatial stochastic process (random function):

$$\{\boldsymbol{Y}(u); u \in \mathcal{D}\}, \tag{1}$$

at the sample locations (Cressie 1993). An arbitrary location is denoted by u and the region $\mathcal{D}$ under study is a fixed subset of $\mathbb{R}^d$ with positive d-dimensional volume. We assume that u varies continuously throughout the region $\mathcal{D}$.

2.1.1 A conditional model specification

Consider geostatistical data in the form $(u_i, y(u_i)) : i = 1, \ldots, n$, as measurements of the random variable $Y(u_i)$ taken at spatial location $u_i \in \mathbb{R}^d$. The model assumed here considers that the variable Y is a "noisy" version of a latent spatial process, the signal $S(u)$. The "noises" are assumed to be Gaussian and conditionally independent given $S(u)$. More general model assumptions are discussed in Section 4.5. The model is specified by:

1. covariates: the "mean part" of the model is given by the term $X(u_i)\boldsymbol{\beta}$. $X(u_i)'$ denotes a vector of spatially referenced non-random variables at location u_i and $\boldsymbol{\beta}$ is the mean parameter;

2. the underlying spatial process: $\{S(u) : u \in \mathbb{R}^d\}$ is a stationary Gaussian process with zero mean, variance σ^2 and correlation function $\rho(\boldsymbol{h}; \boldsymbol{\phi})$, where $\boldsymbol{\phi}$ is the correlation function parameter and $\boldsymbol{h}$ is the vector distance between two locations;

3. conditional independence: variables $Y(u_i), i = 1, \ldots, n$ are assumed to be Gaussian and conditionally independent given the signal:

$$Y(u_i)|S \overset{ind}{\sim} \mathcal{N}\left(X(u_i)'\boldsymbol{\beta} \; + \; S(u_i), \tau^2\right).$$

2.1.2 Refinement and an alternative formulation for Gaussian models

The model defined by 1. to 3. above corresponds to a spatial linear mixed model which can be specified in a hierarchical scheme. Furthermore, in some applications we may want to consider a decomposition of the signal $S(u)$ into a sum of latent processes $T_k(u)$ scaled by σ_k^2. Thus, the model can then be re-written as follows:

$$\begin{aligned}
\text{Level 1} \quad &: \quad \boldsymbol{Y}(\boldsymbol{u}) = \boldsymbol{X}(\boldsymbol{u})\boldsymbol{\beta} + S(\boldsymbol{u}) + \varepsilon(\boldsymbol{u}) \\
&\qquad = \boldsymbol{X}(\boldsymbol{u})\boldsymbol{\beta} + \sum_{k=1}^{K} \sigma_k T_k(\boldsymbol{u}) + \varepsilon(\boldsymbol{u})\,; && (2)\\
\text{Level 2} \quad &: \quad \boldsymbol{T}_k(\boldsymbol{u}) \sim \mathcal{N}\left(0, R_k(\boldsymbol{\phi}_k)\right), \; \boldsymbol{T}_1, \ldots, \boldsymbol{T}_K \;\text{ mutually independent and} \\
&\quad\;\; \boldsymbol{\varepsilon}(\boldsymbol{u}) \overset{ind}{\sim} \mathcal{N}\left(0, \tau^2 I\right)\,; && (3)\\
\text{Level 3} \quad &: \quad (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \tau^2) \sim pr(\cdot), \;\text{ a prior distribution.} && (4)
\end{aligned}$$

The model components are described by:

- $\boldsymbol{Y}(\boldsymbol{u})$ is a random vector with components $Y(u_1), \ldots, Y(u_n)$, related to the measurements at sample locations;

- $\boldsymbol{X}(\boldsymbol{u})\boldsymbol{\beta} = \boldsymbol{\mu}(\boldsymbol{u})$ is the expectation of $Y(\boldsymbol{u})$. $\boldsymbol{X}(\boldsymbol{u})$ (hereafter denoted only by $\boldsymbol{X}$) is a matrix of fixed covariates measured at locations $\boldsymbol{u}$. $\boldsymbol{\beta}$ is a vector parameter. If there are no covariates, $\boldsymbol{X} = \mathbb{1}$ and the mean reduces to a single constant value at all locations;

- $\boldsymbol{T}_k(\boldsymbol{u})$ denotes the random vector, at the sample locations, of a standardised latent stationary spatial process $\boldsymbol{T}_k$. It has zero mean, variance one and correlation matrix $R_k(\boldsymbol{\phi}_k)$. The elements of $R_k(\boldsymbol{\phi}_k)$ are given by a correlation function $\rho_k(\boldsymbol{h}; \boldsymbol{\phi}_k)$ with parameter $\boldsymbol{\phi}_k$. If the process is isotropic this parameter is denoted by ϕ_k and $\boldsymbol{h}$ is reduced to a scalar h, the Euclidean distance between two locations. The processes $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_K$ are mutually independent. The *signal* S is defined by the sum of scaled latent processes $S(u) = \sum_{k=1}^{K} \sigma_k T_k(u)$;

- σ_k is a scale parameter;

- $\boldsymbol{\varepsilon}(\boldsymbol{u})$ denotes the error (noise) vector at the sample locations $\boldsymbol{u}$. In other words, a spatially independent process (spatial white noise) with zero mean and variance τ^2, at the sample locations;

- in a Bayesian approach to inference, the third level specifies the prior for the model parameters.

The correlation function $\rho_k(\boldsymbol{h}; \boldsymbol{\phi}_k)$ should be a valid one and its choice will not be discussed here. A detailed review about valid correlation functions is given by Schlather (1999).

The choice of prior distribution for **Level 3** of the model is a delicate issue in Bayesian inference. Some priors commonly used in Bayesian linear models (Box & Tiao 1973) will be considered in Section 4.

This model-based specification can be related to conventional geostatistics terminology as follows:

- the term **trend** refers to the mean part of the model, $\boldsymbol{X\beta}$;
- a latent processes T_k correspond to a **structure in the variogram**;
- a value of σ_k^2 corresponds to a **partial sill**. The **sill** is the value of $\sum_1^K \sigma_k^2$;
- the **nugget effect** is quantified by τ^2. In the geostatistics literature this term refers to variation at small distances plus measurements errors. The exact interpretation of the nugget effect under the model perspective is discussed by Cressie (1993), p. 59-60 and p. 127-130.
- the **total sill** is given by the sum of the sill and the nugget effect.

2.2 The model components - an illustrative example

The model given by equation (2) consists of three basic components: the mean (or trend), the signal and the noise (nugget effect). This means that $Y(\boldsymbol{u}) - \mu(\boldsymbol{u})$ is assumed to be a zero mean "noisy" version of a latent spatially correlated random variable $S(\boldsymbol{u})$.

In order to illustrate the spatial linear model (2) consider the example:

$$\boldsymbol{Y}(\boldsymbol{u}) = 0.5 + 0.03\boldsymbol{u}^{(1)} + 0.07\boldsymbol{u}^{(2)} + \sqrt{4}T_1(\boldsymbol{u}) + \sqrt{5}T_2(\boldsymbol{u}) + \varepsilon(\boldsymbol{u}).$$

The model components are:

- a spatial linear *trend* on the coordinates $\boldsymbol{u} = (\boldsymbol{u}^{(1)}, \boldsymbol{u}^{(2)})$, with slopes $(0.03, 0.07)$ respectively. In other words, the coordinates are considered as covariates. $\boldsymbol{u}^{(1)}$ and $\boldsymbol{u}^{(2)}$ define a grid in a $100x100$ square area. In conformity with the notation in (2), $\boldsymbol{X}$ is a three column matrix with rows given by the concatenation of a vector of ones, $\mathbb{1}$, and the point coordinates. The mean parameter vector is $\boldsymbol{\beta}' = (0.5, 0.03, 0.07)$;
- a *signal* with two components;
  - a Gaussian *short range* process $T_1(\boldsymbol{u})$ with correlation function $\rho_1(h) = \exp(-h/6)$ and variance $\sigma_1^2 = 4$;
  - a Gaussian *long range* process $T_2(\boldsymbol{u})$ with correlation function $\rho_2(h) = \exp(-h/40)$ and variance $\sigma_2^2 = 5$;
- a *noise* or *nugget effect* with variance $\tau^2 = 1$.

Figure 1 shows a realization of this process (top right) and its components: the trend as a linear function of the coordinates (middle left), the noise or nugget effect process (bottom right) and the spatial processes with short range (middle right) and long range (bottom left). The plot in top left corner shows the theoretical variograms for the whole process (thick lines), for each of the latent processes (thin lines) and the empirical variogram for the obtained realisation (dashed line).

Let's imagine a possible real scenario which could be represented by such model. Consider that the variable Y represents a soil property, e. g.the soil porosity. Then, we can interpret each component of the model as follows:

- the mean part (trend) can be related to the inclination of the area, which affects soil porosity;

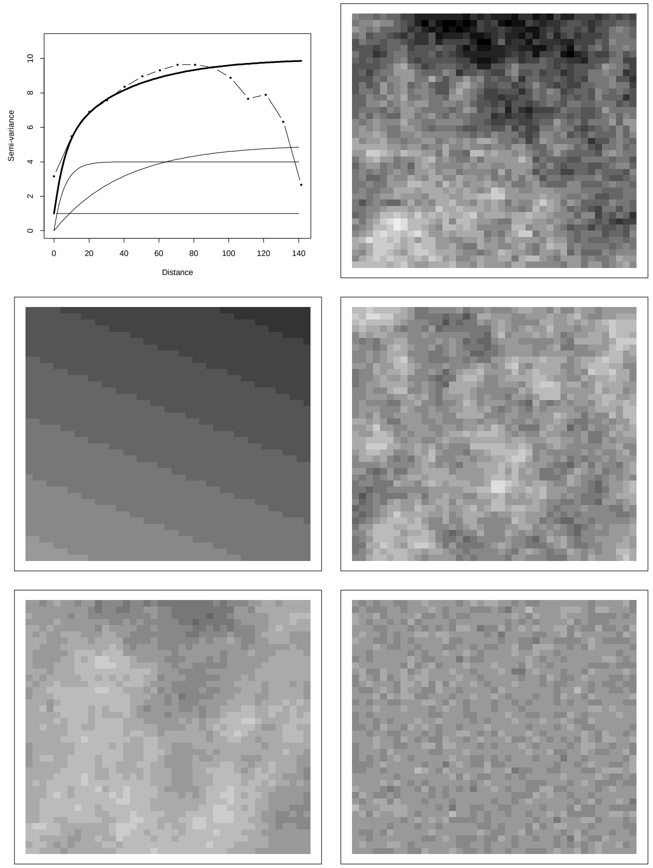

Figure 1: A simulated spatial process and its components

- the first structure (short range) can be induced by the soil management;
- the second structure (long range) can be the result of soil formation factors like rock types, etc;
- the noise (nugget) can be related to spatially uncorrelated events like action of insects and other soil fauna, laboratory errors, damages to the samples, etc.

The final result, i. e. the soil porosity, is assumed to be determined by the summation of components, which characterises a linear model. Note that, in real problems, only the Y variable is observable. The signal should be estimated as well as the trend and the noise i.e., the estimation tries to separate the individual components. When the signal is considered to be a sum of latent processes there is a potential problem related to the model size, namely the number of structures. Although statistical methods and criteria can be used to suggest the number of latent processes, these methods requires large amounts of data to be effective and external information related to knowledge of the physics of the problem should be used whenever possible to guide decisions about the model size. In other words, the structures in the model should, preferably, have a clear physical interpretation.

2.3 Spatial prediction

In geostatistical problems, often the main interest is not parameter estimation but prediction of the variable or functionals at a set of locations.

Denote by $Y(\boldsymbol{u}_0)$ (hereafter Y_0) the variable to be predicted at locations $\boldsymbol{u}_0 = (u_{0,1}, \ldots, u_{0,l},$ the prediction locations. The model (2) is expanded to include both, Y and Y_0. The prediction problem refers to statements about Y_0 after observing a sample $Y(\boldsymbol{u}) = y(\boldsymbol{u})$ (hereafter $Y = y$). In probability terms it means that the conditional distribution $pr\,(y_0|y)$ is required.

The *optimal* point predictor, defined here as the one which minimises the prediction *mean square error (MSE)*, is given by (DeGroot 1989):

$$\hat{\mathrm{Y}}_0 = E\left[Y_0|Y\right]. \tag{5}$$

This predictor is called the *least squares predictor* and its prediction variance is given by $\mathrm{Var}\left[Y_0|Y\right]$.

Note that the result (5) is valid not only for Y_0 a variable at the prediction locations but also for other quantities of interest. Some example of such quantities (functionals) are the overall mean or the mean within any sub-area, the maximum or minimum over the area, probabilities of being above a threshold, etc. If the quantity to be predicted is a linear functional of Y_0 then the prediction is obtained by applying the same functional to the predicted values for Y_0, but notice that this is not valid for non-linear functionals. In this text only prediction of Y_0 will be considered.

The values of the conditional expectation (5) can be calculated only if the model distributions are fully specified and the parameters are known. In practice the model parameters are unknown and an approximation to the conditional expectation may then be used. Finding the conditional expectation (5) or an approximation for it (an alternative estimator) is a central problem in geostatistics. Several methods have been suggested for this purpose. Some of these methods are now mentioned.

Linear predictor assuming known parameters: An alternative estimator can be obtained by restricting the class of predictors to the **linear** predictors. The linear predictor which minimises the MSE is called the *simple kriging* (SK) predictor. The SK predictor requires knowledge of the mean and covariance parameters, i. e. parameters of the the trend, signal and noise should be (or assumed to be) known. The SK predictor is of the form:

$$\hat{\mathrm{Y}}_{SK}(u_0) = \lambda_0 + \sum_i \lambda_i Y(u_i).$$

The weights $\lambda_i, i = 0, 1, 2, ..., n$ are such that the prediction MSE is minimum. Under the Gaussian model and if all the parameters are known, the SK predictor coincides with the conditional expectation (5). Therefore, under these assumptions the SK predictor is optimal.

Linear predictor filtering the mean and assuming known covariance parameters: Another estimator can be obtained from the narrower class of **unbiased linear** predictors, the *ordinary kriging* (OK) predictor. This predictor filters a constant mean requiring only the knowledge of the covariance parameters. The OK predictor is of the form:

$$\hat{\mathrm{Y}}_{OK}(u_0) = \sum_i \lambda_i Y(u_i).$$

The weights $\lambda_i, i = 1, 2, ..., n$ are such that the prediction MSE is minimum under the constraint $\sum \lambda_i = 1$. This constraint ensure the unbiasedness of the estimator. The results provided by OK coincide with the ones returned by SK with the scalar mean parameter β given by its generalised least squares estimator $\hat{\beta} = (\mathbb{1}'V_y^{-1}\mathbb{1})^{-1}\mathbb{1}'V_y^{-1}y$. The OK predictor is widely used in geostatistical applications and sometimes referred as *the* kriging predictor. Other kriging methods like *universal or trend kriging* and *kriging with external trend* are extensions of ordinary kriging allowing for covariates in the mean structure.

A non-linear predictor: A predictor of the form:

$$\hat{\mathrm{Y}}_{DK}(u_0) = \sum_i f_i\left(Y(u_i)\right),$$

where f_i is a measurable function, is called the *disjunctive kriging* (DK) predictor. For non-Gaussian processes, this predictor may provide a better approximation for (5) than the linear ones but requires the knowledge of the all the bivariate distributions of the process.

Non-parametric predictors : Methods like *indicator* and *probability kriging* uses indicator transforms of the data, but still require specification of covariance function parameters for the transformed variables.

A model-based approach: If complete parametric specification for the model components is assumed the conditional expectation (5) can, at least in theory, be assessed. For example, consider the Gaussian model specified in 2.1.2 extended to include both Y and Y_0. The joint distribution is given by:

$$\left(Y, Y_0|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \tau^2\right) \sim \mathcal{N}\left(\begin{bmatrix} X \\ X_0 \end{bmatrix}\boldsymbol{\beta}\,;\ \tau^2 I + \begin{bmatrix} V_y(\boldsymbol{\sigma}^2, \boldsymbol{\phi}) & v(\boldsymbol{\sigma}^2, \boldsymbol{\phi}) \\ v'(\boldsymbol{\sigma}^2, \boldsymbol{\phi}) & V_0(\boldsymbol{\sigma}^2, \boldsymbol{\phi}) \end{bmatrix}\right). \tag{6}$$

Under this model the conditional expectation (5) can be directly obtained if all the parameters are known. It coincides with the SK predictor. Predictions for this model when all of the parameters are known are given in Section 2.5. For the more realistic scenario of unknown parameters, both classical (likelihood based) and Bayesian paradigms can be adopted.

A model-based approach was proposed by Diggle, Tawn & Moyeed (1998). The authors consider a wider class of distributions for the variable under study conditional on a Gaussian underlying spatial process.

If the all the bivariate distributions of (Y, Y_0) are Gaussian, the disjunctive and simple kriging predictors coincide. Moreover, if the multivariate distribution of (Y, Y_0) is Gaussian, model-based, SK and DK predictors are coincident and equal to the conditional expectation for known parameters.

Some references for simple, ordinary kriging and its extensions are Matheron (1971), Journel & Huijbregts (1978), Wackernagel (1998), Goovaerts (1997) and Deutsch & Journel (1998). Disjunctive kriging is discussed, for example, in Kitanidis (1997), Rivoirard (1994), Journel & Huijbregts (1978). Non-parametric geostatistics is presented in Journel (1983), Journel (1984), Goovaerts (1997), Deutsch & Journel (1998), among others.

This text concentrates on Bayesian analysis for the model-based approach to geostatistical prediction. In Section 4 Bayesian prediction results are derived taking into account different levels of parameter uncertainty.

2.4 The Gaussian assumption

Assumptions only up to second order moments of the complete distribution of the process are widely used in the geostatistical literature (Journel & Huijbregts 1978, Goovaerts 1997). Based on them, several kriging techniques are available for data analysis claiming that they require only second order stationarity assumptions or the less restrictive *intrinsic hypothesis* (second order stationarity for increments). Without Gaussian assumptions, predictors given by kriging methods are not guaranteed to be optimal. As mentioned before, Gaussianity is a sufficient condition for the linear simple kriging predictor to be optimal, if all the parameters are considered known. If parameters are unknown the optimal predictor is, in general, non-linear. These results will be shown in Section 4.

On the other hand, a full probabilistic description of a spatial process of form (1) requires the complete specification of its multivariate distribution. A convenient assumption is that the process is Gaussian (maybe after some suitable data transformation). Considering the model (2), $Y(\boldsymbol{u})$ is Gaussian if all of the latent processes T_k and the error component $\varepsilon(\boldsymbol{u})$ are Gaussian.

However, Gaussianity is a strong assumption, not always realistic and cannot be tested in practice. Models based on this assumption can be better explored and many useful properties can be derived due to well known results for Gaussian distributions (see e.g. Goovaerts (1997), p. 266-267). For example, likelihood based methods can be used for parameter estimation, optimal properties of a spatial predictor can be reached, Bayesian inference and prediction is simpler. More comments on the Gaussian assumption can be found in Cressie (1993), pp. 110-111.

A more general model is proposed by Diggle et al. (1998), extending the Gaussian model in a similar way as generalised linear models (McCullagh & Nelder 1989) extend the classical linear Gaussian model for independent data. Considering the model specification in 2.1.1, the authors assume that, conditional on a Gaussian underlying process $S(\boldsymbol{u})$ the variables in $Y(\boldsymbol{u})$ are independent with a distribution in the exponential family. The expectation of this exponential family distribution is related to the covariates and the signal using a link function. The errors are no longer confined to the normal case, and the link is no longer confined to the identity. For this *generalised* model, even with known parameters, the predictor is, in general, non-linear. The authors rely on MCMC techniques (Gilks, Richardson & Spiegelhalter 1996) to make inferences about model parameters and prediction at unsampled locations.

2.5 Model-based prediction with known parameters

Under the model-based perspective, if we assume a Gaussian model and all parameters known, the prediction problem is straightforward. Hereafter, the index '$*$' indicates that the indexed parameter is assumed to be known. Denote $V(\boldsymbol{\sigma}^2_*, \boldsymbol{\phi}_*)$ and $v(\boldsymbol{\sigma}^2_*, \boldsymbol{\phi}_*)$ from (6) simply by V and v. Using properties of multivariate normal (Anderson 1984) marginal and conditional distributions can be directly derived from joint distribution (6) defined by the model. The predictive distribution is given by:

$$\begin{aligned}&\left(Y_0|Y, \boldsymbol{\beta}_*, \boldsymbol{\sigma}^2_*, \boldsymbol{\phi}_*, \tau^2_*\right) \sim \\ &\quad \mathcal{N}\left(X_0\boldsymbol{\beta}_* + v'(\tau^2_* I + V_y)^{-1}(y - X\boldsymbol{\beta}_*)\ ;\ \tau^2_* I + V_0 - v'(\tau^2_* I + V_y)^{-1} v\right). \end{aligned} \tag{7}$$

Therefore point predictors and associated uncertainty can be easily obtained. The mean of (7) coincides with the minimum MSE predictor, the conditional expectation (5).

2.6 Relations with conventional geostatistical methods

For $(\boldsymbol{\beta}, \boldsymbol{\theta})$ considered known the predictions given by (7) coincide with a method usually referred as *simple kriging* in the geostatistical literature. In simple kriging usage the mean is known and the covariance parameters are usually estimated by some method and 'plugged-in' for predictions, as if they were the truth.

Notice that if a constant mean over the region is assumed then $X = \mathbb{1}$. If the mean is a function of coordinates, the columns of X contain those coordinates and/or functions of them. If other covariates are available at prediction locations the columns of X contain these covariates and/or functions of them.

3 The Bayesian framework

In practice the parameters are often unknown. Bayesian inference treats unknown parameters as random variables. By this means, it allows for parameter uncertainty in the predictions. Therefore more realistic estimates of the prediction variance are obtained. Basic results of Bayesian theory are recalled in this section and will be used in Section 4 to perform Bayesian inference in the for the spatial model under investigation. More details and references about Bayesian methods can be found in Gelman, Carlin, Stern & Rubin (1995) and O'Hagan (1994).

3.1 Basic results from Bayesian inference

Consider a r.v. Y with probability distribution given by the function $pr(y|\boldsymbol{\vartheta})$, indexed by a unknown vector parameter $\boldsymbol{\vartheta}$. Considering that a sample $Y=y$ can be observed and writing $L(\boldsymbol{\vartheta}|y) \equiv pr(y|\boldsymbol{\vartheta})$, $L(\cdot)$ is a function of the parameter $\boldsymbol{\vartheta}$ and is called the *likelihood function*.

Consider the distribution of Y given by the model in 2.1.2:

$$\left(Y|\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \tau^2\right) \sim \mathcal{N}\left(X\boldsymbol{\beta}; \tau^2 I + V_y(\boldsymbol{\sigma}^2, \boldsymbol{\phi})\right). \tag{8}$$

The likelihood is a function of $\boldsymbol{\vartheta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \tau^2)'$:

$$L(\boldsymbol{\vartheta}|y) \propto |\tau^2 I + V_y(\boldsymbol{\sigma}^2, \boldsymbol{\phi})|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - X\boldsymbol{\beta})' \left(\tau^2 I + V_y(\boldsymbol{\sigma}^2, \boldsymbol{\phi})\right)^{-1}(y - X\boldsymbol{\beta})\right\}. \tag{9}$$

In the Bayesian approach, both the variable Y and parameters $\boldsymbol{\vartheta}$ are considered to be random quantities with joint distribution $pr(y, \boldsymbol{\vartheta}) = pr(y|\boldsymbol{\vartheta})pr(\boldsymbol{\vartheta})$. Information about model parameters external to the data is reflected in the *prior distribution* $pr(\boldsymbol{\vartheta})$. Bayes' Theorem combines prior and likelihood information in such way that the prior knowledge about the parameters is updated, after collecting data, using the relation:

$$pr(\boldsymbol{\vartheta}|Y) \propto pr(\boldsymbol{\vartheta})pr(Y|\boldsymbol{\vartheta}). \tag{10}$$

The distribution $pr(\boldsymbol{\vartheta}|y)$ is called *posterior distribution* and is the basis for Bayesian inference about model parameters. Decision theory methodology (Berger 1985) leads to optimal choices of point estimates. Loss functions define the quality of the estimators. For example, the mean square error corresponds to a quadratic loss function.

The posterior for the model (8) is:

$$\begin{aligned} pr\left(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \tau^2|y\right) \;\propto\;& pr(\boldsymbol{\beta}, \boldsymbol{\sigma}^2, \boldsymbol{\phi}, \tau^2)\;|\tau^2 I + V_y(\boldsymbol{\sigma}^2, \boldsymbol{\phi})|^{-\frac{1}{2}} \\ & \exp\left\{-\frac{1}{2}\,(y - X\boldsymbol{\beta})'\left(\tau^2 I + V_y(\boldsymbol{\sigma}^2, \boldsymbol{\phi})\right)^{-1}(y - X\boldsymbol{\beta})\right\}. \end{aligned} \tag{11}$$

Choice of priors is a delicate issue in Bayesian inference. Priors which leads to a posterior in the same family of distributions are called *conjugate priors*. Those priors can be computationally convenient although this alone should not justify their choice. Two extreme cases for prior choice are: 1) when parameters are perfectly known the priors can be regarded as degenerate distributions on the parameter values; 2) when the prior knowledge about

parameters is vague the so called *non-informative* and/or *flat* and/or *improper* priors can be adopted. For this case there are several coincidences of results provided by classical and Bayesian analysis.

The basis for Bayesian prediction is the so called *predictive distribution* $pr\,(y_0|y)$. The predictive distribution takes into account the parameter uncertainty by averaging over the parameter space the conditional distribution $pr(y_0|y, \boldsymbol{\vartheta})$, with weights given by the posterior distribution for the model parameters $pr(\boldsymbol{\vartheta}|y)$:

$$\begin{aligned} pr(y_0|y) &= \int pr(y_0, \boldsymbol{\vartheta}|y)\;\mathrm{d}\boldsymbol{\vartheta} \\ &= \int pr(y_0|y, \boldsymbol{\vartheta})\,pr(\boldsymbol{\vartheta}|y)\,\mathrm{d}\boldsymbol{\vartheta}. \end{aligned}$$

For the model in (2.1), the first probability distribution inside the integral is given by (7) which is easily derived from (6). The second one is the posterior for the model parameters given by (11).

The predictive distribution can also be written in terms of the distributions (4), (6) and (8), explicitly specified in the model.

$$pr(y_0|y) = \int \frac{pr(y, y_0|\boldsymbol{\vartheta})\;pr(\boldsymbol{\vartheta})}{\int pr(y|\boldsymbol{\vartheta})\,pr(\boldsymbol{\vartheta})\,\mathrm{d}\boldsymbol{\vartheta}}\,\mathrm{d}\boldsymbol{\vartheta}.$$

3.2 Relations with conventional geostatistical methods

The Bayesian approach acknowledges the parameter uncertainty treating parameters as random variables. The Bayesian prediction is based on the predictive distribution:

$$pr(y_0|y) = \int pr(y_0|y, \boldsymbol{\vartheta})\,pr(\boldsymbol{\vartheta}|y)\,\mathrm{d}\boldsymbol{\vartheta}. \tag{12}$$

Conventional geostatistical methods estimate the parameters and then plug-in their estimated values to perform predictions as if the estimates were the truth. The kriging predictor is based on the distribution:

$$(Y_0|Y) \sim pr(y_0|y, \hat{\boldsymbol{\vartheta}}). \tag{13}$$

Comparing (12) and (13) the Bayesian prediction can be interpreted as a weighted average of plug-in predictions. The weights are given by the posterior $pr(\boldsymbol{\vartheta}|y)$ which incorporates the data information. In comparison with likelihood-based methods, the Bayesian predictive distribution takes into account the complete likelihood surface rather than focusing on the maximum likelihood estimates of the covariance parameters (Handcock & Wallis 1994).

4 Bayesian inference for a geostatistical model

This section presents parameter estimation and prediction results for a Bayesian analysis of geostatistical data. Results are derived for different levels of uncertainty, according to which parameters are assumed to be unknown.

A simpler version of the spatial model presented in 2.1.2 will be considered throughout this section. Consider a model with only one latent spatial process and no measurement errors:

$$\begin{aligned} \text{Level 1} &: \ \boldsymbol{Y}(\boldsymbol{u}) = \boldsymbol{X}\beta + \sigma T(\boldsymbol{u})\,; \\ \text{Level 2} &: \ T(\boldsymbol{u}) \sim \mathcal{N}\left(0, R_y(\boldsymbol{\phi})\right)\,; \\ \text{Level 3} &: \ pr(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}). \end{aligned} \tag{14}$$

The likelihood function is given by:

$$L\left(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\phi}|Y\right) \propto (\sigma^2)^{-\frac{n}{2}} \left|R_y(\boldsymbol{\phi})\right|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y - X\boldsymbol{\beta})' \left(R_y(\boldsymbol{\phi})\right)^{-1}(y - X\boldsymbol{\beta})\right\}. \tag{15}$$

Results in the rest of this section will be derived through the steps:

- some choices of prior distribution for the model parameters are considered;
- joint and marginal distributions for the model parameters are obtained for some choices of prior distributions;
- the predictive distribution $pr(y_0|y)$ is derived for different choices of prior distributions.

4.1 Uncertainty in the mean parameter

In this section only the mean parameter $\boldsymbol{\beta}$ is considered unknown. The covariance parameters are known and the covariance matrix is written as $V(\sigma_*^2, \boldsymbol{\phi}_*) = \sigma_*^2 R(\boldsymbol{\phi}_*)$ (the product of the scale parameter and the correlation matrix) and denoted simply by $\sigma_*^2 R$. The results for this case still holds if a nugget effect (τ_*^2) is present and considered as a known parameter. The correlation parameter $\boldsymbol{\phi}$ include the correlation function parameters and, if the case, the anisotropy parameters. For instance, if the correlation function is given by the exponential model:

$$\rho_k(\boldsymbol{h}; \boldsymbol{\phi}_k) = \exp\left(-\frac{\boldsymbol{h}}{\boldsymbol{\phi}}\right),$$

and the process is isotropic, $\boldsymbol{\phi} = \phi$, a scalar parameter. For the powered exponential correlation function model:

$$\rho_k(\boldsymbol{h}; \boldsymbol{\phi}_k) = \exp\left(-\frac{\boldsymbol{h}}{\phi_1}\right)^{\phi_2},$$

the correlation function has two parameters, $\boldsymbol{\phi} = (\phi_1, \phi_2)$, if the process is isotropic. If anisotropy is present, at least two extra parameters are needed. For list of several correlation function models see Schlather (1999).

The results for unknown mean parameters are of particular interest because of their connections with conventional geostatistical methods. The model for prediction considered in this section corresponds to the common practice in geostatistics where the mean is 'filtered' and the covariance parameters are estimated by some method and just plugged-in for predictions. The relation between Bayesian and geostatistical results will be listed later in this section.

Considering the model (14), the joint probability distribution for (Y, Y_0) is a simpler version of (6), without the nugget effect term and with only one structure (latent process):

$$(Y, Y_0|\boldsymbol{\beta}, \sigma_*^2, \boldsymbol{\phi}_*) \sim \mathcal{N}\left(\begin{bmatrix} X \\ X_0 \end{bmatrix}\boldsymbol{\beta}\,;\, \sigma_*^2 \begin{bmatrix} R_y & r \\ r' & R_0 \end{bmatrix}\right),$$

the associated marginal and conditional distributions, simpler versions of (8) and (7), are:

$$\left(Y|\boldsymbol{\beta}, \sigma_*^2, \boldsymbol{\phi}_*\right) \sim \mathcal{N}\left(X\boldsymbol{\beta}\,;\, \sigma_*^2 R_y\right)$$

and

$$\left(Y_0|Y, \boldsymbol{\beta}, \sigma_*^2, \boldsymbol{\phi}_*\right) \sim \mathcal{N}\left(X_0\boldsymbol{\beta} + r'R_y^{-1}(y - X\boldsymbol{\beta})\,;\, \sigma_*^2(R_0 - r'R_y^{-1}r)\right). \tag{16}$$

An intuitive interpretation can be given for the two terms in the variance of the predictive distribution (16). The first one ($\sigma_*^2 R_0$) represents a *marginal* variance, i. e. the variance without taking account of the information provided by the sample. The second term ($\sigma_*^2 r' R_y^{-1} r$) is the reduction in the variance due to the information provided by the sample. The size of the second term depends on the data location configuration. If a monotone correlation function is assumed, the shorter the distances between the location to be predicted and neighbour data locations, the higher the correlation between the variable at a location to be predicted and the variables in the neighbourhood. It means that close samples provide more information about the variable to be predicted. The closer the neighbours, the bigger the reduction in the prediction variance i. e., the smaller the uncertainty about the predicted value.

4.1.1 Posterior for model parameters

Two different prior distributions for the mean parameter $\boldsymbol{\beta}$ will be considered:

- a conjugate prior,
- a flat prior.

Conjugate prior Assuming a Normal prior for the mean parameter,

$$(\boldsymbol{\beta}|\sigma_*^2, \boldsymbol{\phi}_*) \sim \mathcal{N}\left(m_\beta \,;\, \sigma_*^2 V_\beta\right), \tag{17}$$

the posterior is given by:

$$\begin{aligned}(\boldsymbol{\beta}|Y, \sigma_*^2, \boldsymbol{\phi}_*) &\sim \mathcal{N}\left((V_\beta^{-1} + X'R_y^{-1}X)^{-1}(V_\beta^{-1}m_\beta + X'R_y^{-1}y) \,;\, \sigma_*^2\,(V_\beta^{-1} + X'R_y^{-1}X)^{-1}\right) \\ &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_N \,;\, \sigma_*^2\, V_{\hat{\beta}_N}\right). \end{aligned} \tag{18}$$

Thus, the Normal distribution is a conjugate prior.

Flat prior For a flat prior $p(\boldsymbol{\theta}) \propto 1$, the posterior distribution is:

$$\begin{aligned}(\boldsymbol{\beta}|Y, \sigma_*^2, \boldsymbol{\phi}_*) &\sim \mathcal{N}\left((X'R_y^{-1}X)^{-1}X'R_y^{-1}y \,;\, \sigma_*^2(X'R_y^{-1}X)^{-1}\right) \\ &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}} \,;\, \sigma_*^2 V_{\hat{\beta}}\right). \end{aligned} \tag{19}$$

The results in (19) can be obtained from the conjugate case (18) taking $V_\beta \equiv \infty$ or $V_\beta^{-1} \equiv 0$.

4.1.2 Predictive distribution

Prediction of a random variable Y_0 is based on the posterior distribution $pr(y_0|y)$ which takes into account the uncertainty in $\boldsymbol{\beta}$. The predictive distribution can be obtained as follows:

$$\begin{aligned} pr\left(y_0|y, \sigma_*^2, \phi_*\right) &= \int pr\left(y_0, \boldsymbol{\beta}|y, \sigma_*^2, \phi_*\right)\, d\beta \\ &= \int pr\left(y_0|y, \boldsymbol{\beta}, \sigma_*^2, \phi_*\right)\; pr\left(\boldsymbol{\beta}|y, \sigma_*^2, \phi_*\right)\, d\boldsymbol{\beta}. \end{aligned}$$

The first probability distribution inside the last integral is the conditional distribution given by (16) and the second is the posterior distribution for $\boldsymbol{\beta}$. For both priors considered here, the term inside the integral is the expression of a bivariate Normal and therefore, the marginal $pr\,(y_0|y, \sigma_*^2, \phi_*)$ is also a Normal density of the form:

$$\left(Y_0|Y, \sigma_*^2, \phi_*\right) \sim \mathcal{N}\left(\mu_1, \sigma_*^2\,\Sigma_1\right). \tag{20}$$

The parameters of the posterior distributions depend on the prior. The results are given below for the two cases considered here.

Conjugate prior The posterior parameters in (20) are denoted by $(\mu_{1N}, \sigma_*^2\,\Sigma_{1N})$ and their values are given by the mean and variance of the predictive distribution, respectively:

$$\begin{aligned} E[Y_0|Y] &= (X_0 - r'R_y^{-1}X)(V_\beta^{-1} + X'R_y^{-1}X)^{-1}V_\beta^{-1}m_\beta \\ &\quad + \left[r'R_y^{-1} + (X_0 - r'R_y^{-1}X)(V_\beta^{-1} + X'R_y^{-1}X)^{-1}X'R_y^{-1}\right] y, \\ \mathrm{Var}\,[Y_0|Y] &= \sigma_*^2\left[R_0 - r'R_y^{-1}r + (X_0 - r'R_y^{-1}X)(V_\beta^{-1} + X'R_y^{-1}X)^{-1}(X_0 - r'R_y^{-1}X)'\right]. \end{aligned} \tag{21}$$

Flat prior For the flat prior the predictive distribution can be obtained by computing (21) with $V_\beta \equiv \infty$ ($V_\beta^{-1} \equiv 0$). The posterior parameters in (20) are now denoted by $(\mu_{1F}, \sigma_*^2\,\Sigma_{1F})$ and given by the mean and variance of the predictive distribution,respectively:

$$\begin{aligned} E[Y_0|Y] &= (X_0 - r'R_y^{-1}X)\hat{\boldsymbol{\beta}} + r'R_y^{-1}y \\ \mathrm{Var}\,[Y_0|Y] &= \sigma_*^2\left[R_0 - r'R_y^{-1}r + (X_0 - r'R_y^{-1}X)(X'R_y^{-1}X)^{-1}(X_0 - r'R_y^{-1}X)'\right]. \end{aligned} \tag{22}$$

Both predictive variances, (21) and (22), have three components. As for the case with all parameters known, the first and second components represent the marginal variance for Y_0 and the variance reduction after observing a sample $Y = y$, respectively. The third component is accounting for the uncertainty in the parameter $\boldsymbol{\beta}$.

Note: The posterior for known mean parameter $\boldsymbol{\beta}$ can be also obtained from (21) considering $V_\beta \equiv 0$ or $V_\beta^{-1} \gg X'R_y^{-1}X$. The posterior distribution is Normal with:

$$\begin{aligned} E[Y_0|Y] &= (X_0 - r'R_y^{-1}X)\boldsymbol{\beta} + r'R_y^{-1}y, \\ \mathrm{Var}\,[Y_0|Y] &= \sigma_*^2(R_0 - r'R_y^{-1}r). \end{aligned}$$

4.1.3 Relationships with conventional geostatistical methods

Some of the results shown before can be related to conventional geostatistical methods as the ones described in Journel & Huijbregts (1978); Isaaks & Srisvastava (1989) and Goovaerts (1997). Under the Bayesian perspective these geostatistical methods can be interpreted as prediction procedures which only take into account the uncertainty in the mean parameters.

- If $X \equiv \mathbb{1}$ and $X_0 \equiv \mathbb{1}$ (constant mean)
  - the mean and variance in (22) coincide with the *ordinary kriging (OK)* predictor and the *ordinary kriging* variance (σ_{OK}^2).
- If X and X_0 are trend matrices with rows given by data coordinates or a function of them:
  - the mean and variance in (22) coincide with the *universal or trend kriging (UK or KT)* predictor and the *universal or trend kriging* variance (σ_{KT}^2)
- If X and X_0 are trend matrices with covariates measured at data and prediction locations, respectively:
  - the mean and variance in (22) coincide with the *kriging with external trend (KTE)* predictor and the *kriging with external trend* variance (σ_{KTE}^2)

4.2 Uncertainty in the scale parameter

Consider now that the mean and correlation parameters are known and only the scale parameter σ^2 is unknown. This particular case might be less useful in practice but it will be discussed here as a useful step towards more general cases.

Three prior distribution choices are considered here:

- the conjugate prior, a Scaled-Inverse-χ^2,
- a flat prior,
- another improper prior.

4.2.1 Posterior for model parameters

The posterior distribution for σ^2 is obtained by:

$$pr(\sigma^2|y, \boldsymbol{\beta}_*, \phi_*) \propto pr(\sigma^2)\; pr(y|\boldsymbol{\beta}_*, \sigma^2, \phi_*),$$

where the first term in the right hand side is the prior distribution and the second is the likelihood given by (15).

For all the three prior distributions considered here, the posterior distribution is a Scaled-Inverse-χ^2 of the form:

$$\left(\sigma^2|Y, \boldsymbol{\beta}_*, \phi_*\right) \sim \chi^2_{ScI}\left(v, Q\right). \tag{23}$$

The parameters (v, Q) depends on the choice of prior distribution as follows.

Conjugate prior The conjugate prior for σ^2 is the Scaled-Inverse-χ^2, a particular case of the Inverse-Gamma distribution. This conjugate distribution is specified by two hyper-parameters:

$$(\sigma^2|\boldsymbol{\beta}_*, \boldsymbol{\phi}_*) \sim \chi^2_{ScI}(n_\sigma, S^2_\sigma),$$

which corresponds to saying that $\frac{n_\sigma S^2_\sigma}{\sigma^2} \sim \chi^2(n_\sigma)$

The posterior distribution is:

$$\left(\sigma^2|Y, \boldsymbol{\beta}_*, \phi_*\right) \sim \chi^2_{ScI}\left(n_\sigma + n, \frac{n_\sigma S^2_\sigma + n\hat{\sigma}^2}{n_\sigma + n}\right),$$

where

$$\hat{\sigma}^2 = \frac{1}{n}(y - X\boldsymbol{\beta})' R_y^{-1}(y - X\boldsymbol{\beta}), \tag{24}$$

is the maximum likelihood estimator for σ^2.

This prior distribution can be thought of as providing the information equivalent to n_σ observations with average variance S^2_σ (Gelman et al. 1995). This interpretation can be helpful for hyper-parameter specification.

A flat prior If the prior distribution is such that $pr(\sigma^2) \propto 1$, the posterior distribution is:

$$\left(\sigma^2|Y, \boldsymbol{\beta}_*, \phi_*\right) \sim \chi^2_{ScI}\left(n - 2, \frac{n}{n-2}\hat{\sigma}^2\right).$$

Improper prior In Bayesian inference for Gaussian linear models (Box & Tiao 1973), a prior distribution commonly adopted for σ^2 is:

$$pr(\sigma^2) \propto \frac{1}{\sigma^2}.$$

Like the flat prior, this prior is improper since it doesn't integrates to one. It also corresponds to the Jeffrey's prior (O'Hagan 1994). As with the flat prior it doesn't allows specification of hyper-parameters. Following the interpretation for the prior parameters given in the conjugate case, this improper prior corresponds to zero *prior observations* i. e., $n_\sigma = 0$ in the Scaled-Inverse-χ^2 prior distribution (Gelman et al. 1995).

The posterior is given by:

$$\left(\sigma^2|Y, \boldsymbol{\beta}_*, \phi_*\right) \sim \chi^2_{ScI}\left(n, \hat{\sigma}^2\right).$$

The equivalence with results for the conjugate case can be established taking $n_\sigma = 0$.

4.2.2 Predictive distribution

The predictive distribution is obtained by:

$$\begin{aligned} pr(y_0|y, \boldsymbol{\beta}_*, \phi_*) &= \int pr\left(y_0, \sigma^2|y, \boldsymbol{\beta}_*, \phi_*\right)\, d\sigma^2 \\ &= \int pr\left(y_0|y, \boldsymbol{\beta}_*, \sigma^2, \phi_*\right)\; pr\left(\sigma^2|y, \boldsymbol{\beta}_*, \phi_*\right)\, d\sigma^2, \end{aligned}$$

where the first term in the last integral is the conditional distribution (16) and the second is the posterior for σ^2 given by (23).

The analytical solution of the above integral is a multivariate-t distribution of the form:

$$(Y_0|Y, \boldsymbol{\beta}_*, \phi_*) \sim t_v\left(\mu_0, Q_0\, \Sigma_0\right)$$

where

$$\begin{aligned} \mu_0 &= X_0\boldsymbol{\beta} + r' R_y^{-1}(y - X\boldsymbol{\beta}) \\ \Sigma_0 &= R_0 - r' R_y^{-1} r. \end{aligned}$$

Therefore, the mean is the same as in (16) and the variance differ only by a multiplicative term, which replaces σ^2. Furthermore, using the results for the multivariate-t distribution given in Appendix B, the mean and the variance of the predictive distribution are given by:

$$\begin{aligned} E[Y_0|Y] &= \mu_0 = X_0\boldsymbol{\beta} + r' R_y^{-1}(y - X\boldsymbol{\beta}) \\ Var[Y_0|Y] &= \frac{v}{v-2}\, Q_0\, \Sigma_0 = \frac{v}{v-2}\, Q_0\, (R_0 - r' R_y^{-1} r). \end{aligned}$$

The results for the three priors considered here differs only for the term Q_0 as follows:

Conjugate prior The predictive distribution, the mean and variance are given by:

$$\begin{aligned}
(Y_0|Y,\boldsymbol{\beta}_*,\phi_*) &\sim t_{n_\sigma+n}\left(\mu_0, \frac{n_\sigma S_\sigma^2 + n\hat{\sigma}^2}{n_\sigma + n}\,\Sigma_0\right),\\
E[Y_0|Y] &= \mu_0,\\
Var[Y_0|Y] &= \frac{n_\sigma S_\sigma^2 + n\hat{\sigma}^2}{n_\sigma + n - 2}\,\Sigma_0.
\end{aligned}$$

Flat prior The predictive distribution, the mean and variance are given by:

$$\begin{aligned}
(Y_0|Y,\boldsymbol{\beta}_*,\phi_*) &\sim t_{n-2}\left(\mu_0, \frac{n}{n-2}\,\hat{\sigma}^2\,\Sigma_0\right),\\
E[Y_0|Y] &= \mu_0,\\
Var[Y_0|Y] &= \frac{n}{n-4}\,\hat{\sigma}^2\,\Sigma_0.
\end{aligned}$$

Improper prior The predictive distribution, the mean and variance are given by:

$$\begin{aligned}
(Y_0|Y,\boldsymbol{\beta}_*,\phi_*) &\sim t_n\left(\mu_0, \hat{\sigma}^2\,\Sigma_0\right),\\
E[Y_0|Y] &= \mu_0,\\
Var[Y_0|Y] &= \frac{n}{n-2}\,\hat{\sigma}^2\,\Sigma_0.
\end{aligned}$$

4.3 Uncertainty in the mean and scale parameters

Consider now that $(\boldsymbol{\beta}, \sigma^2)$ are unknown parameters. Posterior and predictive distributions are derived taking into account uncertainty in both. The correlation function parameter $\boldsymbol{\phi}_*$ is still considered fixed. Basically, the results in this section combines the ones obtained in Sections 4.1 and 4.2.

Two prior distributions for $(\boldsymbol{\beta}, \sigma^2)$ are considered:

- a conjugate prior,
- an improper prior.

4.3.1 Posterior for model parameters

The posterior distribution is obtained by:

$$pr\left(\boldsymbol{\beta}, \sigma^2|y, \phi_*\right) \;\propto\; pr\left(\boldsymbol{\beta}, \sigma^2|\phi_*\right)\; pr\left(y|\boldsymbol{\beta}, \sigma^2, \phi_*\right).$$

For both priors considered here the posterior distribution is a Normal-Scaled-Inverse-χ^2 , i. e. a product of Normal and Scaled-Inverse-χ^2 densities:

$$\left(\boldsymbol{\beta}, \sigma^2|Y, \phi_*\right) \sim \mathcal{N}\left(b, V\right)\; \chi^2_{ScI}\left(u, Q\right). \tag{25}$$

The parameters (b, V, u, Q) depend on the prior choice, as shown below.

Conjugate prior The conjugate prior family is the Normal-Scaled-Inverse-χ^2 :

$$\left(\boldsymbol{\beta}, \sigma^2|, \phi_*\right) \sim \mathcal{N}\left(m_\beta, \sigma^2 V_\beta\right)\; \chi^2_{ScI}\left(n_\sigma, S_\sigma^2\right),$$

which is equivalent to to the product of the distributions:

$$\left(\boldsymbol{\beta}|\sigma^2, \phi_*\right) \sim \mathcal{N}\left(m_\beta, \sigma^2 V_\beta\right) \quad \text{and} \quad \left(\sigma^2|\phi_*\right) \sim \chi^2_{ScI}\left(n_\sigma, S_\sigma^2\right).$$

The priors for $\boldsymbol{\beta}$ and σ^2 are the same as the ones in Sections 4.1 and 4.2; and the posterior distribution (25) is given by:

$$\left(\boldsymbol{\beta}, \sigma^2|Y, \phi_*\right) \sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_N, \sigma^2 V_{\hat{\beta}_N}\right)\; \chi^2_{ScI}\left(n_\sigma + n, S_1^2\right),$$

where

$$S_1^2 = \frac{n_\sigma S_\sigma^2 + n\hat{\sigma}^2 + \hat{\boldsymbol{\beta}}' V_{\hat{\beta}}{}^{-1}\hat{\boldsymbol{\beta}} + m_\beta' V_\beta{}^{-1} m_\beta - (V_{\hat{\beta}}{}^{-1}\hat{\boldsymbol{\beta}} + V_\beta{}^{-1} m_\beta)' V_{\hat{\beta}_N} (V_{\hat{\beta}}{}^{-1}\hat{\boldsymbol{\beta}} + V_\beta{}^{-1} m_\beta)}{n_\sigma + n} \tag{26}$$

and $\hat{\boldsymbol{\beta}}_N, V_{\hat{\beta}_N}, \hat{\boldsymbol{\beta}}, V_{\hat{\beta}}$ and $\hat{\sigma}^2$ are given in (18), (19) and (24).

Other posterior distributions can be obtained factorising this joint posterior distribution,

$$pr\left(\boldsymbol{\beta}, \sigma^2|y, \phi_*\right) = pr\left(\boldsymbol{\beta}|y, \sigma^2, \phi_*\right)\; pr\left(\sigma^2|y, \phi_*\right),$$

and then obtaining the conditional posterior for $\boldsymbol{\beta}$ and marginal posterior for σ^2 as follows:

$$\begin{aligned}(\boldsymbol{\beta}|Y,\sigma^2,\phi_*) &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}_N, \sigma^2 V_{\hat{\beta}_N}\right),\\ (\sigma^2|Y,\phi_*) &\sim \chi^2_{ScI}\left(n_\sigma + n, S_1^2\right).\end{aligned}$$

Finally the marginal posterior for $\boldsymbol{\beta}$ is obtained integrating the joint posterior with respect to σ^2:

$$(\boldsymbol{\beta}|Y,\phi_*) \sim t_{n_\sigma+n}\left(\hat{\boldsymbol{\beta}}_N, S_1^2 V_{\hat{\beta}_N}\right).$$

Improper prior If the prior is such that

$$pr(\boldsymbol{\beta},\sigma^2|\phi_*) = \frac{1}{\sigma^2},$$

the posterior distribution (25) is given by:

$$(\boldsymbol{\beta},\sigma^2|Y,\phi_*) \sim \mathcal{N}\text{–}\chi^2_{ScI}\left(\hat{\boldsymbol{\beta}}, V_{\hat{\beta}}, n-p, S^2\right),$$

where the values $(\hat{\boldsymbol{\beta}}, V_{\hat{\beta}})$ are given by (19) and

$$S^2 = \frac{1}{n-p}\ (y - X\hat{\boldsymbol{\beta}})' R_y^{-1} (y - X\hat{\boldsymbol{\beta}}), \tag{27}$$

where p the number of elements of $\boldsymbol{\beta}$.

The conditional posterior and marginal posterior distributions are:

$$\begin{aligned}(\boldsymbol{\beta}|Y,\sigma^2,\phi_*) &\sim \mathcal{N}\left(\hat{\boldsymbol{\beta}}\ ;\ \sigma^2 V_{\hat{\beta}}\right),\\ (\sigma^2|Y,\phi_*) &\sim \chi^2_{ScI}(n-p\ ;\ S^2),\\ (\boldsymbol{\beta}|Y,\phi_*) &\sim t_{n-p}\left(\hat{\boldsymbol{\beta}}, S^2 V_{\hat{\beta}}\right).\end{aligned}$$

4.3.2 Predictive distribution

The predictive distribution which takes into account the uncertainty in the mean and scale parameters is given by:

$$\begin{aligned}pr(y_0|y,\phi_*) &= \iint pr\left(y_0, \boldsymbol{\beta}, \sigma^2|y,\phi_*\right)\, d\boldsymbol{\beta}\, d\sigma^2\\ &= \iint pr\left(y_0, \boldsymbol{\beta}|y,\sigma^2,\phi_*\right)\ pr\left(\sigma^2|y,\phi_*\right)\, d\boldsymbol{\beta}\, d\sigma^2\\ &= \int pr\left(y_0|y,\sigma^2,\phi_*\right)\ pr\left(\sigma^2|y,\phi_*\right)\, d\sigma^2.\end{aligned}$$

The first term in the last integral is the predictive distribution (20) and the second is the marginal posterior distribution $pr(\sigma^2|y)$. Expressions of both depend on the choice of the prior distribution and consequently the posterior distribution also depends on choice of prior. For the prior distributions considered here analytical solutions can be obtained and the posterior is a multivariate-t density of the form:

$$(Y_0|Y,\phi_*) \sim t_v\left(\mu_1, Q_1\,\Sigma_1\right), \tag{28}$$

where $(\mu_1, Q_1\,\Sigma_1)$ depends on the choice of prior distribution as follows.

Conjugate prior The predictive distribution, the mean and variance are given by:

$$\begin{aligned}(Y_0|Y,\phi_*) &\sim t_{n_\sigma+n}\left(\mu_{1N}, S_1^2\,\Sigma_{1N}\right),\\ E[Y_0|Y] &= \mu_{1N},\\ \mathrm{Var}\,[Y_0|Y] &= \frac{S_1^2\,\Sigma_{1N}}{n_\sigma+n-2},\end{aligned}$$

where (μ_{1N}, Σ_{1N}) and S_1^2 are given by (21) and (26)

Improper prior The predictive distribution, the mean and variance are given by:

$$\begin{aligned}(Y_0|Y,\phi_*) &\sim t_{n-p}\left(\mu_{1F}, S^2\,\Sigma_{1F}\right),\\ E[Y_0|Y] &= \mu_{1F},\\ \mathrm{Var}\,[Y_0|Y] &= \frac{n-p}{n-p-2}\,S^2\,\Sigma_{1F},\end{aligned}$$

where (μ_{1F}, Σ_{1F}) and S^2 are given by (22) and (27).

4.4 Uncertainty in the mean, scale and correlation parameters

In this section a one parameter isotropic correlation function will be assumed. Therefore only a scalar correlation parameter $\boldsymbol{\phi} = \phi$ is considered. Extensions allowing for anisotropy and correlation functions with more than one parameters are outlined in the next section.

4.4.1 Posterior for model parameters

The prior distribution for the model parameters can be written as:

$$pr(\boldsymbol{\beta}, \sigma^2, \phi) = pr(\phi)\ pr(\boldsymbol{\beta}, \sigma^2|\phi). \tag{29}$$

No specific prior will be assumed for ϕ. For $(\boldsymbol{\beta}, \sigma^2|\phi)$ the flat prior $\frac{1}{\sigma^2}$ considered in 4.3 will be assumed. Results for other prior choices can be derived in a similar way.

The posterior distribution for the parameters is given by:

$$pr\left(\boldsymbol{\beta}, \sigma^2, \phi|y\right) = pr\left(\boldsymbol{\beta}, \sigma^2|y, \phi\right)\ pr(\phi|y)\,. \tag{30}$$

The first distribution in the right hand side is given by (25). The posterior $pr(\phi|y)$ can be obtained using the relation:

$$pr(\phi|y) \propto \frac{pr(\boldsymbol{\beta}, \sigma^2, \phi)\ \ pr(y|\boldsymbol{\beta}, \sigma^2, \phi)}{pr(\boldsymbol{\beta}|y, \sigma^2, \phi)\ \ pr(\sigma^2|y, \phi)}\,. \tag{31}$$

The distributions in the numerator are given by the prior (29) and the likelihood (15). The posterior distributions in the denominator are given in the Section 4.3.1.

For the prior $pr(\boldsymbol{\beta}, \sigma^2|\phi) \propto \frac{1}{\sigma^2}$ the posterior for the correlation parameter if of the form:

$$pr(\phi|y) \propto pr(\phi)\ |V_{\hat{\beta}}|^{\frac{1}{2}}\ |R_y|^{-\frac{1}{2}}\ (S^2)^{-\frac{n-p}{2}}. \tag{32}$$

This expression doesn't define a standard probability distribution. Tanner (1996) presents some methods to deal with such kind of distributions. Inference by simulation it the strategy adopted here. Samples are taken from the posterior and predictive distributions and used for inference and prediction, respectively.

In order to sample from the posterior distribution (30) an algorithm is given below.

Algorithm 1:

1. Discretise the distribution of $(\phi|y)$, i. e. choose a set of values for ϕ in a sensible interval considering the problem in hand and a discrete uniform prior for ϕ on the chosen support set.
2. Compute the posterior probabilities in this support set using (32). The results define the approximate discrete posterior distribution $\tilde{pr}(\phi|y)$.
3. Sample a value of ϕ from $\tilde{pr}(\phi|y)$.
4. Attach the sampled value of ϕ to $pr(\boldsymbol{\beta}, \sigma^2|y, \phi)$ given by (25) and sample from this distribution.
5. Iterate steps (3)-(4) as many times as the number of samples (triplets $(\boldsymbol{\beta}, \sigma^2, \phi)$) wanted from the posterior distribution of the parameters.

4.4.2 Predictive distribution

The predictive distribution is now derived taking also into account the uncertainty in the correlation parameters.

The predictive distribution is given by:

$$\begin{aligned} pr(y_0|y) &= \iiint pr\left(y_0, \boldsymbol{\beta}, \sigma^2, \phi|Y\right)\ d\boldsymbol{\beta}\, d\sigma^2\, d\phi \\ &= \iiint pr\left(y_0, \boldsymbol{\beta}, \sigma^2|y, \phi\right)\ d\boldsymbol{\beta}\, d\sigma^2\, pr(\phi|y)\ d\phi \\ &= \int pr(y_0|y, \phi)\ pr(\phi|y)\ d\phi. \end{aligned} \tag{33}$$

The first probability distribution in the integrand is the predictive distribution (28) and the second is given by (32). The result of the integral depends on the prior distribution adopted. Usually this predictive distribution is not a standard probability distribution and the integral can be solved by numerical methods. We use integration by simulation. The algorithm proposed for the predictive is similar to the one for the posterior mentioned above.

Algorithm 2:

1. Discretise the distribution of $(\phi|y)$, i. e. choose a set of values for ϕ in a sensible interval considering the problem in hand and a discrete uniform prior for ϕ on the chosen support set.
2. Compute the posterior probabilities in this support set using (32). The results define the approximate discrete posterior distribution $\tilde{pr}(\phi|y)$.
3. Sample a value of ϕ from $\tilde{pr}(\phi|y)$.
4. Attach the sampled value of ϕ to $pr(y_0|y, \phi)$ given by 28 and sample from it obtaining realisations of the predictive distribution.
5. Iterate from steps (3)-(4) as many times as the numbers of samples wanted from the predictive distribution.

4.4.3 Prior distribution for the correlation parameter

In the previous section generic results were derived without specifying any particular prior distribution for ϕ. Results of a data analysis can be sensitive to the prior adopted. This section discusses the choice of prior for the correlation parameter.

Consider the case of a scalar correlation parameter ϕ . For many of the correlation functions, including the widely used exponential and spherical models, ϕ is a 'range' parameter. It measures how quickly the correlation function decays when the separation distance between pairs of locations increases, i. e. the distance at which the correlation decays to a particular reference value.

In principle, the parameter ϕ varies in the interval $[0, \infty)$. Because of the direct interpretation of the parameter as a reference distance, the prior for ϕ represents a guess about this reference distance. Some possible types of prior, for $\phi \in [0, \infty)$, illustrated in Figure 2, are:

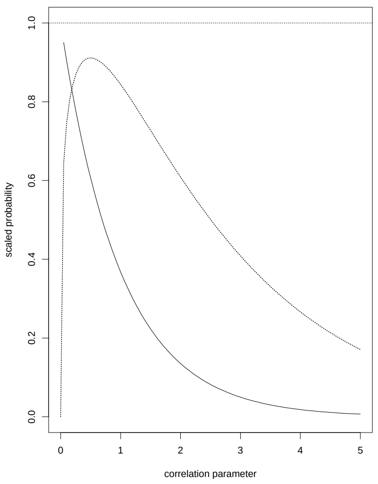

Figure 2: Shape of some prior distributions for the correlation parameter: a flat prior (thin dashed line), a decreasing prior (solid line) and a asymmetric prior (thick dashed line)

1. flat prior: $pr(\phi) \propto 1$,

2. a prior with flexibility to choose the shape, but not the scale hyperparameter, here a decreasing prior: $pr(\phi) \propto \frac{1}{\phi^{\delta}}$, $\delta > 0$ or $pr(\phi) \propto \exp(-\delta\,\phi)$, $\delta > 0$,

3. a prior with flexibility to choose both, shape and scale hyperparameters hera a asymmetric prior: e. g. a non-central χ^2, i. e. $pr(\phi) \sim \frac{a\chi^2}{b}$ where a and b are hyperparameters, or any other asymmetric distribution like the Gamma distribution, the Log-Normal distribution , etc.

The uniform prior (type 1) represents the belief that, a priori, all values in the interval $(0, \infty)$, or any other specified interval, are equally plausible. Priors of type 2 or 3 allow the user to express a prior belief that values in certain ranges are more likely than values in other ranges. All these priors can be easily implemented in algorithms **1** and **2**, where the continuous distribution for ϕ is replaced by a discrete approximation.

So far in this report, we have considered choices of prior distribution for the parameter of the correlation function, which can be associated with distances in the area. An alternative is specifying the prior distribution for correlation values instead. One of the advantages is the fact that the prior distribution will be restricted to the interval $[0, 1]$. For this specification is necessary to fix a 'reference distance'. In practice it corresponds to a distance for which a prior guess about the spatial correlation can be made.

To illustrate this prior specification consider the isotropic exponential model $\rho(h) = exp(-\frac{h}{\phi})$. Fixing the reference distance to 1, a prior is specified for values of the correlation at this distance. This corresponds to a reparametrisation of $\varphi = exp(-\frac{1}{\phi})$ for $\phi = 1$. The correlation function is written as $\rho(h) = \varphi^h$. The plots in Figure 3 illustrate this idea. Fixing the the

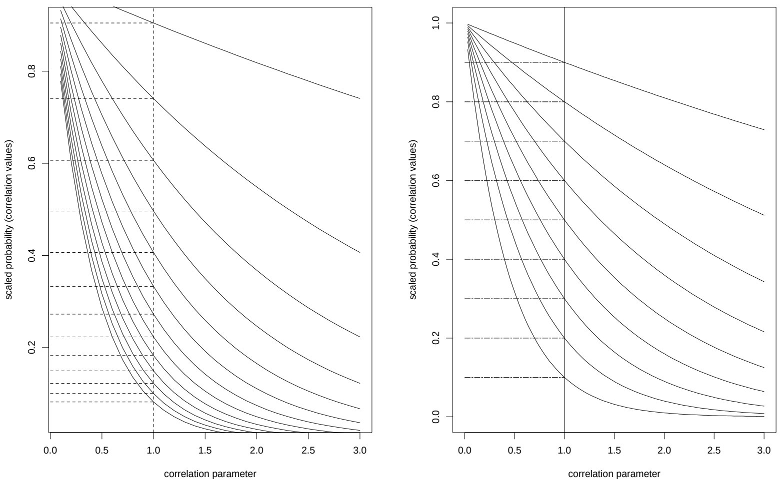

Figure 3: Illustration of the alternative prior specification: correlation values in a uneven set (left) and equally spaced set (right)

correlation parameter $\phi = 1$ a family of distributions can be chosen by taking values in a discrete set of correlation values. Notice that the scale probabilities in the $y - axis$ coincide with correlation values. For the left hand side plot an uneven set of correlation values was chosen and for the right hand side plot the correlation values are equally spaced.

4.5 Extensions for more general models

Some possible strategies to analyse extended versions of the simple model (14) considered so far are now presented.

Nugget effect: If measurement errors are included in the model the covariance matrix can still be written as a product of a scale parameter σ^2 and a matrix, now indexed by two parameters (ϕ, τ^2). The model can be written as:

$$\begin{aligned} \text{Level 1} &\;:\; \boldsymbol{Y}(\boldsymbol{u}) = \boldsymbol{X}\beta + \sigma T(\boldsymbol{u}) + \varepsilon(\boldsymbol{u})\,, \\ \text{Level 2} &\;:\; T(\boldsymbol{u}) \sim \mathcal{N}\left(0, R(\phi)\right) \text{ and } \varepsilon \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \tau^2\right), \end{aligned}$$

consequently,

$$\left(Y|\boldsymbol{\beta}, \sigma^2, \phi, \tau_R^2\right) \;\sim\; \mathcal{N}\left(X\boldsymbol{\beta}\;;\;\sigma^2\,R(\phi, \tau_R^2)\right),$$

where

$$R(\phi, \tau_R^2) = \sigma^2\,\left[R(\phi) + \tau_R^2\mathcal{I}\right] = \sigma^2\left[R(\phi) + \frac{\tau^2}{\sigma^2}\,\mathcal{I}\right].$$

Using this parametrisation, the parameter τ_R^2 can be regard as a *relative nugget effect*. The model without measurements errors corresponds to $\tau_R^2 = 0$ (or $\tau^2 = 0$).

For parameter estimation and to incorporate the uncertainty in τ_R^2 in the predictions, the algorithms **1** and **2** can be adapted changing the first step to:

1. Discretise the distribution of $(\phi, \tau_R^2|y)$, i. e. choose a set of values for (ϕ, τ_R^2) in an sensible grid considering the problem in hand and a discrete uniform prior for (ϕ, τ_R^2) on the chosen support set.

Anisotropy: The simplest way to incorporate anisotropy in the model is by adding two extra parameters in the vector correlation parameters $\boldsymbol{\phi}$. The two additional parameters are the *anisotropy ratio* and the *anisotropy angle*. In the geostatistical literature the term *geometric anisotropy* is used to describe models with this kind of anisotropic correlation function. More details about geometric anisotropy can be found in Isaaks & Srisvastava (1989).

For parameter estimation and prediction the step 1 of the algorithms **1** and **2** can be generalised defining a multi-dimensional grid for the components of $\boldsymbol{\phi}$. In practice, it becomes difficult to ensure adequate coverage of the parameter space as the dimensionality of $\boldsymbol{\phi}$ increases.

Correlation functions with more than one parameter: Some isotropic correlation functions have more than one parameter. See Schlather (1999) for an extensive list of correlations functions used in geostatistics. Again the general advice is, for the step 1 of the algorithms **1** and **2**, to define values for the correlation function parameters in a sensible grid.

A particular correlation function model with two parameters is given by the *Matérn class* (more details about this correlation function can be found e. g., in (Handcock & Wallis 1994)). The first parameter, ϕ_1, defines the decay in the correlation function. The second, ϕ_2, is associated with the differentiability of the underlying stochastic process. Under a certain parametrisation, $\phi_2 = \frac{1}{2}$ corresponds to the exponential model and integer values of ϕ_2 indicates the number of times the process is differentiable. For this correlation function we suggest building a fine grid for ϕ_1 and taking values of ϕ_2 only in the set $\{\frac{1}{2}, 1, 2\}$.

Several structures: The model considered in this section accommodates only one latent process. More general models with several latent process, as illustrated in 2.2, can be dealt with in a similar way. These models can be used as a tool to incorporate *zonal anisotropy* described in Journel & Huijbregts (1978) and Isaaks & Srisvastava (1989). But the caveat concerning high dimensional $\boldsymbol{\phi}$ still applies.

5 Results for simulated data

This section uses the results derived in Section 4 to analyse simulated. All the computations were performed using the `geoS` library (Ribeiro Jr & Diggle 1999) for S-PLUS (Mathsoft 1993). The commands used to generate and analyse the data set are listed in Appendix D.

REMARK: No exhaustive simulation exercise is reported here. The results are intended only to illustrate the methods and not to perform any comparative study or assess the performance of the methods. The following conclusions and comments are valid for these particular data set and should not be taken as general statements.

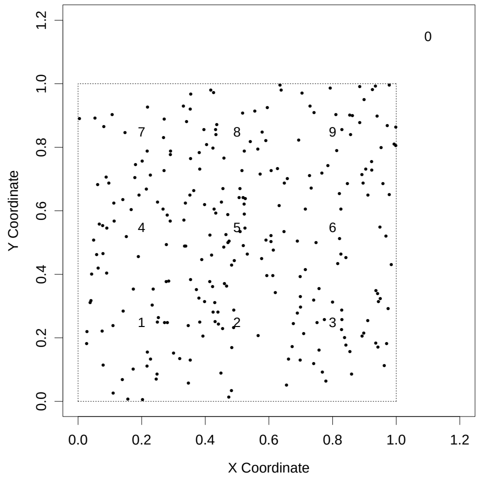

Figure 4: Data points locations (dots) and the 10 locations where prediction is aimed (numbers)

5.1 A first simulated data set

The first simulated data set was generated in a 1x1 square, with 256 data points randomly located within the area (irregular grid). The simulated values were generated for model (14), with no covariates, zero mean ($\beta = 0$) and covariance parameters $(\sigma^2, \phi) = (1, 0.3)$. Ten prediction locations were chosen with one point intentionally located outside of the 1x1 square area to illustrate the behaviour of the predictions in a extrapolation problem. The data points and locations to be predicted can be seen in Figure 4.

Parameters were estimated and predictions for each point were made considering different levels of uncertainty as discussed in Section 4. For the case with known parameters the true

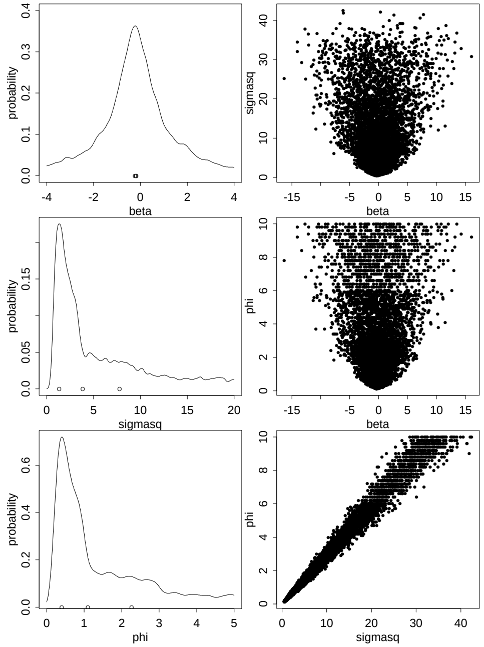

Figure 5: Posterior distributions for model parameters (β, σ^2, ϕ) on the left plots and correlations between parameters using the samples from the posterior distribution

parameters values were used. For the cases where the covariance parameters are unknown, but considered fixed in the prediction, they were estimated by weighted least squares. Differing from what is suggested in Cressie (1985), the weights are given only by the number of pairs in each bin, as suggested in Barry, Crowder & Diggle (1997). Other methods for covariance parameter estimation (Zimmerman & Zimmerman 1991) could have been adopted. Weighted least squares was chosen because it is an objective criterion and widely used in geostatistical analysis. The estimates were obtained from the empirical variogram shown in Figure 6.

The 'full Bayesian' model, where all the parameters were considered unknown, requires inference by simulation, as shown in Section 4.4. The prior $pr(\beta, \sigma^2, \phi) \propto \frac{1}{\sigma^2}$ was adopted. One thousand samples were taken from the posterior and predictive distributions following the Algorithms 1 and 2 in Section 4.4. These algorithms suggest the use of a discrete uniform prior for the *range* parameter ϕ, on the chosen support set. Here, an uneven support set was chosen with 151 values between 0 and 10, having higher concentration of points in regions with small values of ϕ. Therefore, a uniform prior in this support set was assumed. The commands in Appendix D provide more details about this prior specification.

Figure 5 shows the posterior distribution for model parameters (left hand-side plots) and their correlation using the samples from the posterior distribution (right hand-side plots). The posterior distributions were obtained by density estimation (Bowman & Azzalini 1997). The circles in the bottom of the posterior distributions plots, represents the posterior mode, median and mean, in this order (i. e. the mode is the smallest value and so on).

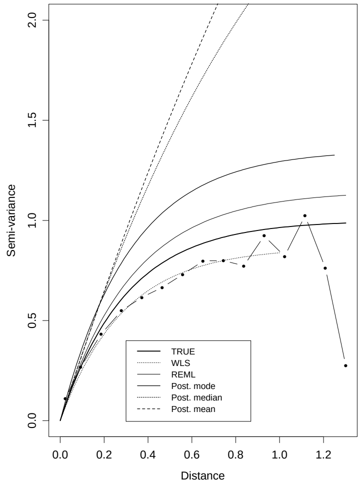

Figure 6: Empirical variogram and the variogram models fitted by different estimates

Some covariance parameters estimates can be compared in Figure 6. The empirical variogram is shown together with the fitted models for: 1) weighted least squares, 2) posterior modes, 3) posterior median and 4) posterior means. For this simulated data, all the Bayesian estimates fit a variogram model with higher variance than the weighted least squares fit. Although not shown here, the restricted maximum likelihood estimates are similar to the posterior modes.

For the predictions, four levels of uncertainty were considered: a) all parameters known, b) mean parameter unknown, c) mean and scale parameter unknown, d) all parameters unknown.

Figure 7 shows the predictive distributions at the 10 selected prediction locations for the different levels of uncertainty. The plots are in the same scale allowing direct comparison of the predictive distributions. As seen in Section 4, for uncertainty levels considered in (a)-(c), analytical solutions can be obtained. For (d) samples were drawn from the predictive distribution and a density estimation were performed. Notice that based on the variances of their predictive distribution the points can be grouped in three main groups:

- Group 1: locations with small variances (5,2,3), which correspond to locations with close neighbours.
- Group 2: locations with medium variances (1,4,6,7,8,9), which correspond to locations without close neighbours.
- Group 3: a location with large variance (0), which corresponds to a location outside of the area defined by data locations.

For all locations there are few differences between the results provided by simple kriging (known parameters), ordinary kriging (uncertainty in $\boldsymbol{\beta}$). A slightly higher variance was found when taking into account the uncertainty in the covariance parameters.

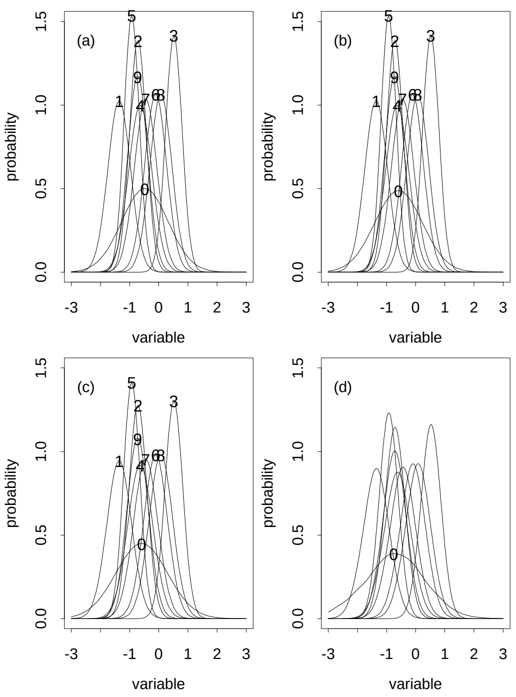

Figure 7: Predictive distributions at the 10 locations for (a) known parameters, (b) mean unknown, (c) mean and scale unknown and (d) all parameters unknown

The higher variance when considering $(\boldsymbol{\beta}, \sigma^2, \phi)$ unknown would be expected after inspecting the variogram in Figure 6. The reasoning is the fact that variogram model fitted by weighted least squares is below to the Bayesian ones, based on summaries of the posterior distribution. This does not happens always and the weighted least squares estimates can either over or under estimate the prediction variances.

Some locations were selected and for each of them the predictive distributions, for the different levels of uncertainty considered here, were plotted in Figure 8. Notice that in this graph the *y-axis* scales are not the same. For Location 5, the plot in the top left corner shows no differences between OK, SK and prediction with uncertainty only in the mean and variance parameters. The predictive distribution for the full Bayesian analysis has a higher variance. For Location 3 (right top corner), OK and SK have produces similar results, but now different from the other two Bayesian predictive distributions. The Bayesian predictive distributions have a slightly difference in their variances. For Location 8 (bottom left corner), again OK and SK produced similar results. The two Bayesian predictive distributions have produced similar results for this location. For the Location 0, which is located outside of the sampling window, the four methods have produced different predictive distributions. In general, the variance increases with the number of unknown parameters. In both, Figure 7 and Figure 8, it can be seen a much higher variance for the point located outside of the sampling window.

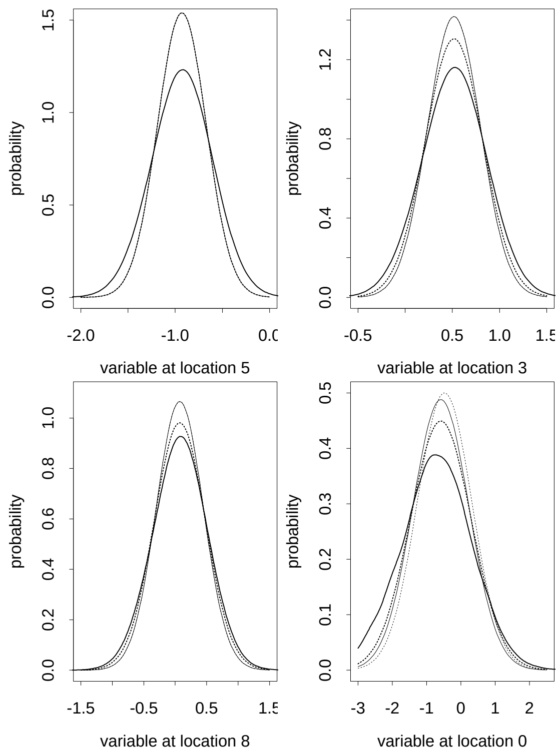

Figure 8: Predictive distributions for at some selected locations. SK (dashed line), OK (solid line), uncertainty in mean and variance (thick dashed line) and uncertainty in all parameters (thick solid line)

5.2 A second simulated data set

A second data set was simulated with 100 data locations randomly distributed in a irregular grid within a 1x1 square area. The parameters (β, σ^2, ϕ) are the same as for the first example. The correlation function is given by the exponential model. Four prediction locations were chosen. The plot in the top left corner of Figure 9 shows data and prediction locations. The prediction locations represent four basic neighbourhood configurations:

- Location 1 represents a location central in the area, with close neighbours,
- Location 2 represents a location central in the area, but without close neighbours,
- Location 3 represents a location in the border of the area
- Location represents a location outside the area

For Bayesian parameter estimation the **Algorithm 1** from Section 4 was applied. For this data set a different prior, $pr(\beta, \sigma^2, \phi) \propto \frac{1}{\phi^2}\frac{1}{\sigma^2}$ was adopted. The approximation for a discrete

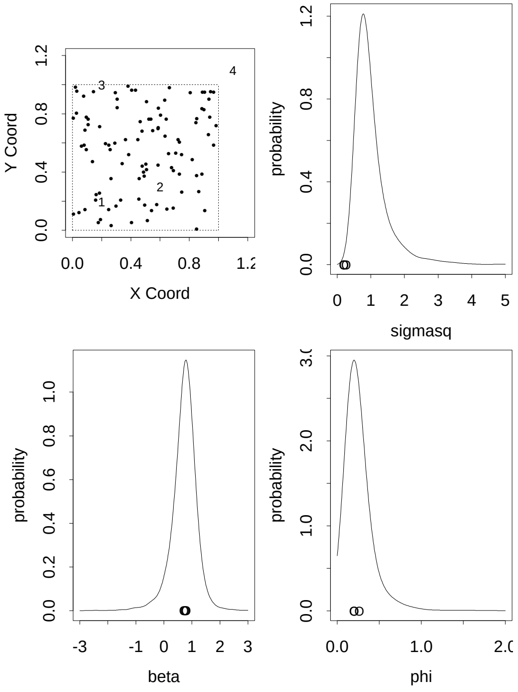

Figure 9: Data and prediction location for the second simulated data set with n=100 (top left) and the posterior distributions for model parameters

distribution uses the same support set as in the previous example. One thousand samples were drawn from the posterior distribution. The plots in the top right corner and bottom row of Figure 9 show density estimation of the marginal distributions for the model parameters. Comparing these posterior distributions with the ones in the previous examples, it can be seen that here the posterior distributions are more concentrated at small values. Furthermore the mode, median and mean (indicated in the plots by circles) are closer. This reflects the choice of prior for the parameter ϕ. The prior adopted here gives more weight for small values whilst the prior used in the previous example gives equal weight for all values of phi in the discrete prior distribution.

For prediction at the selected locations, two methods were used: ordinary kriging and Bayesian prediction considering all the parameters to be unknown.

For the ordinary kriging the covariance parameters were estimated by weighted least squares, as for the previous data set, using an empirical binned variogram up to distance 1. The weights are given by the number of pairs in each bin. Assuming Gaussianity, distributions based on the ordinary kriging results were obtained for each of the prediction locations. The mean is given by the predicted value and the variance is given by the ordinary kriging variance. The dashed lines in Figure 10 show the distributions at the for prediction locations.

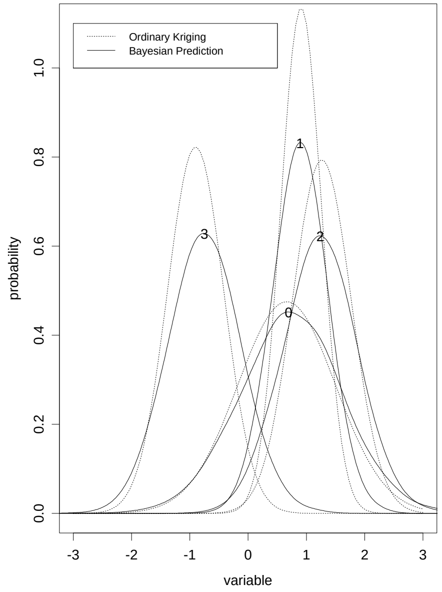

Figure 10: Predictive distributions at the 4 prediction locations for the second simulated data set

For the Bayesian prediction, the **Algorithm 2** from Section 4, was applied. One thousand samples were drawn from the predictive distribution, at each prediction location. The full lines in Figure 10 show non-parametric density density estimates of the predictive distributions.

The differences between the predictive distributions are higher than for the first example. Although not shown here, the predictive distributions using a uniform prior are similar to the ones obtained here (solid lines in Figure 10). The difference between the Bayesian and ordinary kriging predictive distributions is greater than in the first example because the data set is smaller and the effect of parameter uncertainty correspondingly greater.

6 Final remarks

A spatial linear model for analysing geostatistical data has been presented and some of its main features discussed and illustrated. Bayesian inference provides a elegant way to incorporate parameter uncertainty in the predictions. Results for different levels of uncertainty and prior choices have been derived. The methods were implemented using the `geoS` library (Ribeiro Jr & Diggle 1999). Analysis of a simulated data set illustrated the methods. Some possible extensions for more general models were indicated.

The results and differences for each level of uncertainty may depend on several factors: number of data points, the spatial arrangement of the locations, the size of the area, the value of each parameter, the presence and size of the nugget effect, the variogram fitting strategy, etc. Detailed simulation experiments should be performed in order to determine whether and/or for which cases the differences are relevant.

Within the Bayesian framework non-linear functionals, e. g. maximum in a sub-area, can be easily predicted using the samples of the posterior distribution. There is no obvious analogue plug-in methods apart from some circumstances where indicator kriging or sequential simulation methods (Deutsch & Journel 1998) can be used.

Although one could be expected to have a bigger variance for the case where all the parameters are unknown, compared with the cases where the covariance parameters are estimated from the empirical variogram, that's not always true. For a particular data set (simulated or not) the covariance parameter estimates based on the empirical variogram can be substantially different from the Bayesian estimates, in either direction. Covariance parameter estimates from the posterior mode and restricted maximum likelihood estimates (Patterson & Thompson 1971) should be similar, if the prior $pr(\boldsymbol{\beta}, \sigma^2, \phi) = \frac{1}{\sigma^2}$ is assumed. This is due the fact that the expression of the predictive distribution for the covariance parameters coincides with their profile likelihood expression. This suggests a 'sub-optimal' alternative to the Bayesian approach where the covariance parameters are estimated by restricted maximum likelihood and plugged-in for predictions.

When all the parameters are considered unknown we assume a discrete prior for the parameter ϕ, leading to a discrete posterior distribution. The prior can be defined in a fine enough grid to ensure a good approximation. Furthermore, the grid can be more concentrated in the regions with higher prior density. The correlation parameter usually relates to a distance at which the correlation assumes a particular value. Any knowledge about the variable and the region under study should be useful to guide the prior elicitation.

Acknowledgement

The authors thank Anders Brix, Martin Schlather and Stefano Tonellato for their helpful comments on earlier drafts of this report. This report was written during the first author's PhD program at the Department of Mathematics and Statistics – Lancaster University, supported by CAPES/BRASIL, grant BEX 1676/96-2.

References

Anderson, T. (1984). *An introduction to multivariate statistical analysis*, 2nd edn, Wiley.

Barry, J., Crowder, M. & Diggle, P. (1997). Parametric estimation of the variogram, *Technical Report ST-97-06*, Dept. Maths and Stats, Lancaster University, Lancaster, UK.

Berger, J. (1985). *Statistical decision theory and Bayesian analysis*, 2 edn, Springer.

Bowman, A. & Azzalini, A. (1997). *Applied smoothing techniques for data analysis*, Oxford University Press.

Box, G. & Tiao, G. (1973). *Bayesian inference in statistical analysis*, Wiley.

Cressie, N. (1985). Fitting variogram models by weighted least squares, *Mathematical Geology* **47**: 563–586.

Cressie, N. (1993). *Statistics for Spatial Data - revised edition*, Wiley, New York.

De Oliveira, V., Kedem, B. & Short, D. (1997). Bayesian prediction of transformed gaussian random fields, *JASA* **92**(440): 1422–1433.

DeGroot, M. (1989). *Probability and statistics*, 2nd edn, Addison-Wesley.

Deutsch, C. & Journel, A. (1998). *Geostatistics software library and user's guide*, second edn, Oxford University Press, New York.

Diggle, P., Tawn, J. & Moyeed, R. (1998). Model based geostatistics (with discussion), *Journal of the Royal Statistical Society, Series C* **47**(3): 299–350.

Ecker, M. & Gelfand, A. (1997). Bayesian variogram modeling for an isotropic spatial process, *JABES* **2**(4): 347–369.

Gelman, A., Carlin, J., Stern, H. & Rubin, D. (1995). *Bayesian data analysis*, Chapman and Hall, London.

Gilks, W., Richardson, S. & Spiegelhalter, D. (eds) (1996). *Markov Chain Monte Carlo in Practice*, Chapman and Hall, London.

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*, Oxford University Press, New York.

Handcock, M. & Stein, M. (1993). A Bayesian analysis of kriging, *Technometrics* **35**(4): 403–410.

Handcock, M. & Wallis, J. (1994). An approach to statistical spatial temporal modeling of meteorological fields (with discussion), *JASA* **89**(426): 368–390.

Isaaks, E. & Srisvastava, R. (1989). *An Introduction to Applied Geostatistics*, Oxford University Press, New York.

Journel, A. (1983). Non-parametric estimation of spatial distribution, *Mathematical Geology* **24**: 269–286.

Journel, A. (1984). The place of non-parametric geostatistics, *in* G. e. a. Verly (ed.), *Geostatistics for natural resources characterization*, Vol. 1, D. Reidel Pub. Co., pp. 307–335.

Journel, A. (1989). *Fundamentals of geostatistics in five lessons*, American Geophysical Union.

Journel, A. & Huijbregts, C. (1978). *Mining Geostatistics*, Academic Press, London.

Kitanidis, P. (1978). Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *WRR* **22**(4): 499–507.

Kitanidis, P. (1997). *Introduction to geostatistics: applications in hydrogeology*, Cambridge University Press, New York.

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

Le, N. & Zidek, J. (1992). Interpolation with uncertain spatial covariances: a Bayesian alternative to kriging, *Journal of multivariate analysis* **43**(375): 351–374.

Matérn, B. (1960). Spatial variation, *Technical report*, Statens Skogsforsningsinstitut, Stockholm.

Matheron, G. (1971). The theory of regionalized variables and its applications, *Technical Report 5*, Cah. Centr. Morph. Math. Fontainebleau.

Mathsoft (1993). *S-PLUS User's Manual. Version 3.4.*, Mathsoft Inc.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*, second edn, Chapman and Hall, London.

O'Hagan, A. (1994). *Bayesian Inference*, Vol. 2b of *Kendall's advanced theory of statistics*, Edward Arnold.

Omre, H. (1987). Bayesian kriging - merging observations and qualified guesses in kriging, *Mathematical Geology* **19**(1): 25–38.

Omre, H., Halvorsen, B. & Berteig, V. (1989). A Bayesian approach to kriging, *in* M. Armstrong (ed.), *Geostatistics*, Vol. I, pp. 109–126.

Omre, H. & Halvorsen, K. (1989). The Bayesian bridge between simple and universal kriging, *Mathematical Geology* **21**(7): 767–786.

Patterson, H. & Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal, *Biometrika* **58**: 545–554.

Ribeiro Jr, P. & Diggle, P. (1999). `geoR/geoS`: A geostatistical library for `r/s-plus`, *Technical Report ST-99-09*, Dept. of Maths and Stats, Lancaster University, Lancaster, UK.

Rivoirard, J. (1994). *Introduction to disjunctive Kriging and non-Linear Geostatistics*, Oxford.

Schlather, M. (1999). Introduction to positive definite functions and to unconditional simulation of random fields, *Technical Report ST-99-10*, Dept. Maths and Stats, Lancaster University, Lancaster, UK.

Stein, A. & Corsten, L. (1991). Universal kriging and cokriging as a regression procedure, *Biometrics* **47**(2): 575–587.

Tanner, M. (1996). *Tools for Statistical Inference*, Springer, New York.

Wackernagel, H. (1998). *Multivariate geostatistics*, Springer, Berlin.

Whittle, P. (1954). On stationary processes in the plane, *Biometrika* **41**: 434–49.

Whittle, P. (1962). Topographic correlation, power-law covariance functions, and diffusion, *Biometrika* **49**: 305–314.

Woodbury, A. (1989). Bayesian updating revisited, *Mathematical Geology* **21**(3): 285–308.

Zimmerman, D. & Zimmerman, M. (1991). A comparison of spatial semivariogram estimators and corresponding kriging predictors, *Technometrics* **33**(1): 77–91.

A A more general model specification

The model described in Section 2.1 allows for only one measurement at each location and for just one variable under study. Although this corresponds to a common situation in geostatistical analysis, sometimes it is possible to have more than one measurement at each location, of one or more variables (attributes) under study.

For this more general case, $\underset{\sim}{Y}_j(u)$ is a random vector at locations $u \in \mathcal{D}$. This denotes the j^{th} measurement , at sample locations, of attributes under study $\underset{\sim}{Y}(u) = (Y^{(1)}(u) \ldots, Y^{(l)}(u))$. If the number of measurements for each variable is not the same, j^{th} is replaced by j_l^{th}.

Consider the following example: in order to study the soil porosity and moisture in a region, several soil samples were taken and their locations were recorded. The samples were analysed at a laboratory. Three measurements of porosity and moisture were taken for each sample. Therefore, three soil porosity and moisture data are available for each location. For the notation adopted here, $j = 1, 2, 3$ and $\underset{\sim}{y}_j(u_i) = \left(y_j^{(1)}(u_i), y_j^{(2)}(u_i)\right)$ denotes the j^{th} measurements, at location u_i, of the soil porosity ($y^{(1)}$) and soil moisture ($y^{(2)}$). Although the mean value of the triplets could be used for each variable, by considering data from all of the measurements within the model it's possible to assess the measurement errors.

To contemplate multiple variables and measurements at each location, the model specifications in Section 2.1 will be rephrased.

A conditional model specification

Consider now geostatistical data of the form $(u_i, \underset{\sim}{y}_j(u_i)) : i = 1, \ldots, n\ ;\ j = 1, \ldots, m$, where $\underset{\sim}{y}_j(u_i)$ denotes the j^{th} measurement of the random variables $\underset{\sim}{Y} = (Y^{(1)} \ldots, Y^{(l)})$ taken at spatial location $u_i \in \mathbb{R}^d$. The model is specified by:

1. covariates: the "mean part" of the model is given by the term $X(u_i)\boldsymbol{\beta}$. $X(u_i)$ denotes spatially referenced non-random variables at location u_i and $\boldsymbol{\beta}$ is the mean parameter;

2. the underlying spatial process: $\{S(u) : u \in \mathbb{R}^d\}$ is a stationary Gaussian process with mean zero, variance σ^2 and correlation function $\rho(\boldsymbol{h}; \boldsymbol{\phi})$ where $\boldsymbol{\phi}$ is the correlation function parameter and $\boldsymbol{h}$ is the vector distance between two locations;

3. conditional independence: variables $\underset{\sim}{Y}_j(u_i)$ are assumed to be Gaussian and conditionally independent given the signal. For $i = 1, \ldots, n$ and $j = 1, \ldots, m$:

$$\underset{\sim}{Y}_j(u_i)|S \overset{ind}{\sim} \mathcal{N}\left(X(u_i)\boldsymbol{\beta} \ + S(u_i), \tau^2 I\right).$$

An alternative formulation for Gaussian models

For a j^{th} measurement of the variables in $\underset{\sim}{Y}$ at a finite set of sample locations $u_i : i = 1, ..., n$, and denoting $\boldsymbol{u} = (u_1, u_2, ..., u_n)$, the model can be written as a spatial linear mixed model:

$$\begin{aligned}
\text{Level 1} \quad : \quad \underset{\sim}{\boldsymbol{Y}}_j(\boldsymbol{u}) &= \boldsymbol{X}(\boldsymbol{u})\boldsymbol{\beta} + S(\boldsymbol{u}) + \varepsilon_j(\boldsymbol{u}) \\
&= \boldsymbol{X}(\boldsymbol{u})\boldsymbol{\beta} + \sum_{k=1}^{K} \sigma_k T_k(\boldsymbol{u}) + \varepsilon_j(\boldsymbol{u}); \\
\text{Level 2} \quad : \quad \boldsymbol{T}_k(\boldsymbol{u}) &\sim \mathcal{N}\left(0, R_k(\boldsymbol{\phi}_k)\right) \quad \text{and} \quad \boldsymbol{\varepsilon}_j(\boldsymbol{u}) \overset{ind}{\sim} \mathcal{N}\left(0, \tau^2 I\right).
\end{aligned}$$

B Some probability distributions

Gaussian (m-dimentional multivariate normal)

$$\begin{aligned}
\text{Notation} &: \; z \sim \mathcal{N}(\mu, \Sigma) \\
\text{Density} &: \; pr(z) = (2\pi)^{-\frac{m}{2}} |\Sigma|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(z-\mu)'\Sigma^{-1}(z-\mu)\} \\
\text{Expectation} &: \; E(z) = \mu \\
\text{Variance} &: \; \mathrm{Var}(z) = \Sigma \\
\text{Mode} &: \; mode(z) = \mu
\end{aligned}$$

Multivariate-t (m-dimentional)

$$\begin{aligned}
\text{Notation} &: \; z \sim t_\nu(\mu, \Sigma) \\
\text{Density} &: \; pr(z) = \frac{\Gamma(\frac{\nu+m}{2})}{\Gamma(\frac{\nu}{2})(\pi\nu)^{\frac{m}{2}}} |\Sigma|^{-\frac{1}{2}} \{1 + \frac{1}{\nu}(z-\mu)'\Sigma^{-1}(z-\mu)\}^{-\frac{\nu+m}{2}} \\
\text{Expectation} &: \; E(z) = \mu \; , \; \nu > 1 \\
\text{Variance} &: \; \mathrm{Var}(z) = \frac{\nu}{\nu-2}\Sigma \; , \; \nu > 2 \\
\text{Mode} &: \; mode(z) = \mu
\end{aligned}$$

Scale-Inverse χ^2

$$\begin{aligned}
\text{Notation} &: \; z \sim \chi^2_{ScI}(\nu, S^2) \\
\text{Density} &: \; pr(z) = \frac{\frac{\nu}{2}^{\frac{\nu}{2}}}{\Gamma(\frac{\nu}{2})} (S^2)^{\frac{\nu}{2}} \; z^{-(\frac{\nu}{2}+1)} \; \exp\{-\frac{\nu S^2}{2z}\} \\
\text{Expectation} &: \; E(z) = \frac{\nu}{\nu-2} S^2 \; , \; \nu > 2 \\
\text{Variance} &: \; \mathrm{Var}(z) = \frac{2\nu^2}{(\nu-2)^2(\nu-4)} S^2 \; , \; \nu > 4 \\
\text{Mode} &: \; mode(z) = \frac{\nu}{\nu+2} S^2
\end{aligned}$$

Inverse-Gamma

$$\begin{aligned}
\text{Notation} &: \; z \sim IG(\alpha, \beta) \\
\text{Density} &: \; pr(z) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-(\alpha+1)} \; \exp\{-\frac{\beta}{z}\} \\
\text{Expectation} &: \; E(z) = \frac{\beta}{\alpha-1} \; , \; \alpha > 1 \\
\text{Variance} &: \; \mathrm{Var}(z) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)} \; , \; \alpha > 2 \\
\text{Mode} &: \; mode(z) = \frac{\beta}{\alpha+1}
\end{aligned}$$

C Geostatistics terminology

This Appendix compiles the relations between some geostatistical terminology and terms of the model considered here.

The process and assumptions

Regionalised variable: refers to an attribute observed in a region. No probabilistic interpretations is attached to the term.

Random function: the same as stochastic process. The term is often used to describe spatial stochastic processes.

Structures and variogram structures: each structure corresponds to a latent process. Each structure in the variogram describes the spatial dependence of the associated latent process.

Second order (or weak) stationarity: the spatial stochastic process is (weakly) stationary if:

$$\begin{aligned}
E[Y(u)] &= \mu \\
\mathrm{Cov}[Y(u+\boldsymbol{h}), Y(u)] &= 2\,C(\boldsymbol{h})
\end{aligned}$$

where $C(\boldsymbol{h})$ is the **covariogram** and $\boldsymbol{h}$ is the distance between two locations.

Here second order stationarity is implicit in the Gaussian assumption.

Intrinsic hypothesis: corresponds to a second order stationarity for increments. It is defined by:

$$\begin{aligned}
E[Y(u+\boldsymbol{h}) - Y(u)] &= 0 \\
\mathrm{Var}[Y(u+\boldsymbol{h}) - Y(u)] &= 2\,\gamma(\boldsymbol{h})
\end{aligned}$$

where $\gamma(\boldsymbol{h})$ is the **variogram** (or **semivariogram**).

This assumption is less restrictive than the second order stationarity, i. e. the class of processes which are intrinsically stationary is wider than the class of weakly stationary process. Under the Gaussian model adopted here, the intrinsic hypothesis coincides with second order stationarity.

Process parameters

Sill: the summation of the variance of the latent processes.

Total sill: the summation of the latent process variances and the measurement error variance.

Partial sill: the variance of each latent process.

Range: the parameter of the correlation (or variogram) function which measures the strength of the spatial dependence.

Nugget effect: variance of the measurement error plus (usually non-identifiable) short range process.

Drift: refers to the change in the level of an intrinsic stationary process.

Trend: the mean part of the model.

Nested variograms: variograms with more than one structure. It refers to models with more than one latent process.

Some kriging types

Simple kriging: prediction considering that the mean and correlation parameters are known or using their estimates as if they were the truth.

Ordinary kriging: prediction without covariates and assuming a constant mean. From the Bayesian perspective, corresponds to assume a flat prior for the (scalar) mean parameter. The covariance parameters are considered known or their estimates are used as if they were the truth.

External trend kriging: prediction including covariates in the mean part of the model. From the Bayesian perspective, corresponds to assume a flat prior for the (vector) mean parameter. The covariance parameters are considered known or their estimates are used as if they were the truth.

Universal kriging: prediction including the coordinates (or a function of them) as covariates in the mean part of the model. From the Bayesian perspective, corresponds to assume a flat prior for the (vector) mean parameter. The covariance parameters are considered known or their estimates are used as if they were the truth.

Kriging with trend model: the same as universal kriging.

D geoS library commands

The numerical results and graphics showed in this text were obtained using the `geoS` library (Ribeiro Jr & Diggle 1999) for S-PLUS. They can be reproduced using the commands given in this appendix, provided that S-PLUS and the `geoS` library are available.

Simulation of a Gaussian random field and its components

The realisation of a Gaussian random field and its components in Section 2.2 can be reproduced used the commands indicated below.

Generation of the model components:
The nugget effect (noise) component:

```
.Random.seed <- c(21, 20, 30, 54, 23, 1, 49, 25, 40, 21, 62, 3)
yt0 <- grf(41*41,xgrid=c(0,100),ygrid=c(0,100),grid='reg',
           nugget=1,cov.pars=c(0,0))
```

The short range process:

```
.Random.seed <- c(37, 1, 40, 29, 28, 1, 32, 53, 9, 15, 60, 3)
yt1 <- grf(41*41,xgrid=c(0,100),ygrid=c(0,100),grid='reg',cov.pars=c(4,6))
```

The long range process:

```
.Random.seed <- c(5, 61, 12, 17, 8, 0, 63, 5, 22, 9, 45, 2)
yt2 <- grf(41*41,xgrid=c(0,100),ygrid=c(0,100),grid='reg',cov.pars=c(5,40))
```

The trend:

```
ytrend <- list(coords=yt0$coords, data=0.5+0.03*yt0$coords[,1] +
               0.07*yt0$coords[,2])
class(ytrend) <- 'grfsim'
```

The realisation without trend:

```
yres <- yt0
yres$data <- yt0$data + yt1$data + yt2$data
yres$cov.pars <- rbind(yt1$cov.pars,yt2$cov.pars)
yres$call <- grf(n = 41*41, grid ="reg", xgrid = c(0, 100), ygrid =c(0, 100),
            cov.pars = matrix(c(4, 6, 5, 40), 2, 2, byrow = T), nugget = 1)
yres$message <- messa.grf(2, nugget=1, sigmasq=c(4,5), phi=c(6,40),kappa=0.5,
                          cov.model = 'exponential')
```

Remark: If the components `yt0, yt1, yt2` are not required i. e., only `yres` is intended to be generated, it can be done direct using the command in `yres$call` above. Notice that the realisation obtained from that can be different because of seed for the random number generator.

The realisation with trend:

```
y <- yres
y$data <- yres$data + ytrend$data
```

Plotting the variograms, the realisation and its components

```
par(mfrow=c(3,2))
```

```
plot(yres)
lines.variomodel(yt0,dmax=140)
lines.variomodel(yt1,dmax=140)
lines.variomodel(yt2,dmax=140)
par(mar=c(1,1,1,1))
image(y,zlim=range(y$data))
image(ytrend,zlim=range(y$data))
image(yt1,zlim=range(y$data))
image(yt2,zlim=range(y$data))
image(yt0,zlim=range(y$data)
```

Analysing of the first simulated data set

The main commands to reproduce the results in Section 5 are given below.

Generating the simulated data set:

```
.Random.seed <- c(49,25,59,3,23,2,3,63,63,8,48,2)
s256i <- grf(256,cov.pars=c(1,0.3))
```

Selecting the prediction locations:

```
s256i$loci <- matrix(c(0.2,0.25,0.5,0.25,0.8,0.25,0.2,0.55,
       0.5,0.55,0.8,0.55,0.2,0.85,0.5,0.85,0.8,0.85,1.1,1.15)
       ,ncol=2,byrow=T)
```

Weighted least squares estimates:

```
s256i$bin <- variog(s256i, max.dist = 1)
s256i$wls <- wlsfit(s256i$bin, ini = c(0.5,0.5), fix.nugget=T)
```

Maximum likelihood and restricted maximum likelihood parameter estimation:

```
s100$ml <- likfit(s256i, ini=c(.5,.5), fix.nugget=T, res=F)
s100$reml <- likfit(s256i, ini=c(.5,.5), fix.nugget=T, meth='REML', res=F)
```

Simple Kriging (no parameter uncertainty):

```
s256i$sk.wls <- krige.conv(s256i, loc=s256i$loci, krige =
   krige.control(type.krige='sk', beta=0, cov.pars=s256i$wls$cov.pars))
```

Ordinary Kriging (uncertainty (β)):

```
s256i$ok.wls <- krige.conv(s256i, loc=s256i$loci, krige =
    krige.control(cov.pars=s256i$wls$cov.pars))
```

Uncertainty in (β, σ^2):

```
s256i$bs.wls <- krige.bayes(s256i, loc=s256i$loci, prior =
       prior.control(range.prior="fixed", range=s256i$wls$cov.pars[2]))
```

The full Bayesian model, i. e. uncertainty in (β, σ^2, ϕ):

```
s256i$bsp <- krige.bayes(s256i, locations=s256i$loci, prior=
   prior.control(range.discrete=c(seq(0,10,l=201))), output =
   output.control(n.posterior = 10000))
```


The second simulated data set

Generating the data set:

```
.Random.seed <- c(37,22,26,1,45,3,60,52,36,6,8,3)
s100 <- grf(100,cov.pars=c(1,.3))
```

Selecting the prediction locations:

```
s100$loci <- matrix(c(0.2,0.6,0.2,1.1,0.2,0.3,1.0,1.1),ncol=2)
```