Non-Gaussian Bayesian Geostatistical Modelling

M.B. Palacios and M.F.J. Steel*

Abstract

Sampling models for geostatistical data are usually based on Gaussian processes. However, real data often display non-Gaussian features, such as heavy tails.

In this paper we propose a more flexible class of sampling models. We start from the spatial linear model which has a spatial trend plus a stationary Gaussian error process. We extend the sampling model to non-Gaussianity by including a scale parameter at each location. We make sure that we obtain a valid stochastic process. The scale parameters are spatially correlated to ensure the process is mean square continuous. We derive expressions for the moments and the kurtosis of the process. This more general stochastic process allows us to accommodate and identify observations that would be outliers under a Gaussian sampling process. For the spatial correlation structure we adopt the flexible Matèrn class with unknown smoothness parameter. Furthermore, a nugget effect is included in the model.

Bayesian inference (posterior and predictive) is performed using a Markov chain Monte Carlo algorithm. The choice of the prior distribution is discussed and its importance is assessed in a sensitivity analysis. We also examine identifiability of the parameters. Our methods are illustrated with two data sets.

Keywords: Identifiability; Nugget effect; Outliers; Prior sensitivity; Scale mixing; Smoothness; Spatial heteroskedasticity.

1 Introduction

Geostatistical data are considered a partial realization of an underlying random field indexed by locations that vary continuously through a fixed region in space. Sampling models for such data are usually based

*Blanca Palacios is lecturer, Department of Econometrics and Statistics, Universidad del Pais Vasco, 48015 Bilbao, Spain (Email: etppanab@bs.ehu.es) and Mark Steel is Professor, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (Email: M.F.Steel@stats.warwick.ac.uk). Mark Steel was affiliated with the Institute of Mathematics and Statistics (IMS), University of Kent at Canterbury at the start of this research, and Blanca Palacios acknowledges the hospitality of the IMS. We are deeply indebted to Carmen Fernández for many insightful discussions and for pointing out an error in a previous version. We also thank Paul Blackwell, Tilmann Gneiting and Doug Nychka for very useful comments and are grateful to the Basque Meteorological Service for providing us with the temperature data.

on Gaussian processes. This assumption facilitates prediction and provides some justification for the use of the spatial prediction method called kriging. However, distributions of real data often deviate from the Gaussian distribution, presenting *e.g.* heavy tails or skewness. To cope with these problems, various approaches have been suggested. De Oliveira *et al.* (1997) developed the Bayesian Transformed Gaussian model based on the Box-Cox family of power transformations. Diggle *et al.* (1998) proposed generalized linear spatial models to widen the class of distributions.

In this paper we propose an alternative flexible class of sampling models for geostatistical data, specifically aimed at accommodating non-Gaussian tail behaviour. We start from the basic model with a stationary Gaussian error process around a spatial trend. This model is extended to non-Gaussianity by scale mixing. Two different types of mixing variables are discussed and we focus mainly on the case where the mixing variables are location-specific. In particular, we propose the so-called Gaussian-log-Gaussian (GLG) model, which is based on a log-Gaussian mixing process. We make sure that we obtain a stochastic process satisfying Kolmogorov's conditions. In addition, the mixing variables are spatially correlated to induce mean square continuity of the resulting process on the observables. We derive some properties of this new sampling model, and show that its smoothness is not affected by the particular scale mixing proposed. The expressions for the moments highlight the relation between the extra hyper-parameter introduced in the mixing distribution and the marginal kurtosis of the observables. This more general stochastic process allows us to accommodate and identify observations that would be "outliers" under a Gaussian sampling process, by means of the posterior distribution of the mixing variables. In the context of our model, outliers are defined in terms of spatial heteroskedasticity, as observations belonging to a subregion with larger observational variance. Prediction is fairly straightforward as we can exploit the Gaussian structure of the sampling process, conditional on the mixing variables.

For the spatial correlation structure of the underlying Gaussian process, we adopt the Matèrn class where the smoothness parameter is also treated as unknown. Furthermore, a nugget effect is included in the model to capture measurement errors and/or microscale variations.

A fully Bayesian approach to inference is adopted, which leads to a coherent treatment of all parameter uncertainty given the model as well as model uncertainty (which can be dealt with by Bayesian model averaging). We choose to elicit a proper prior, based on inducing "reasonable" behaviour of the process. We consider different prior structures and a wide range of prior hyperparameters in a sensitivity analysis. We also develop and implement a Markov chain Monte Carlo (MCMC) sampling strategy for inference, which is found to work quite well. Fortran code implementing the samplers discussed in this paper is freely available on the website http://www.warwick.ac.uk/go/msteel/steel_homepage/software/.

As applications, we use the topographic data set of Davis (1973), which was also used by Handcock and Stein (1993) and Berger *et al.* (2001), as well as temperature data collected in the Basque country. In the context of the first example, we conduct a prior sensitivity analysis and conclude that the prior (in certain directions) is critical enough to warrant serious attention to prior elicitation in practically relevant situations. Using simulated data, we also assess the identifiability of the parameters and the ability of the GLG model to correctly identify outlying observations, which were introduced in a sample simulated from a Gaussian model. Besides posterior inference, we also conduct predictive inference at specified points and over a regular grid. In addition, we assess which observations would be outliers

under the Gaussian model. We compare our GLG model with the Gaussian model through Bayes factors, and show that in both applications (and particularly in the second application) the data favour the GLG model. For the temperature data, the GLG model also provides a much closer fit to a robust empirical semivariogram. Throughout, Bayes factors between models are computed using the modified harmonic mean estimator $\hat{p}_4$ of Newton and Raftery (1994).

Besides its use in a model of considerable practical interest here, the basic Gaussian-log-Gaussian process (defined as a ratio of a Gaussian and a log-Gaussian process and denoted by GLG process) can be applied to a host of more complicated settings, simply as a more flexible substitute to the Gaussian process which often underlies spatial models. We can, for example, think of models accommodating non-stationary as in Fuentes (2002), but now based on a convolution of local stationary GLG rather than Gaussian processes. Likewise, extensions to spatio-temporal models are conceptually straightforward.

Section 2 introduces our new class of non-Gaussian sampling models, based on scale mixing a Gaussian process and establishes a condition that ensures a well-defined stochastic process. The use of this model for outlier detection and for prediction is explained. Section 3 describes the elicitation of a proper prior structure and the following section outlines the Markov chain Monte Carlo algorithm, used for Bayesian inference. Section 5 examines the important issues of prior sensitivity and identifiability of the parameters. The applications are discussed in Section 6 and the final section concludes.

2 The Sampling Model

Let $Z(\mathbf{x})$ be a random process defined for locations $\mathbf{x}$ in some spatial region $D \subset \Re^m$. Our starting point will be the model

$$Z(\mathbf{x}) = \mathbf{f}'(\mathbf{x})\boldsymbol{\beta} + \sigma\epsilon(\mathbf{x}) + \tau\rho(\mathbf{x}) \tag{1}$$

where the mean surface is assumed to be a linear function of $\mathbf{f}'(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_k(\mathbf{x}))$, a vector of k known functions of the spatial coordinates, with unknown coefficient vector $\boldsymbol{\beta} \in \Re^k$. Further, $\epsilon(\mathbf{x})$ is a second-order stationary error process with zero mean and unit variance and with a correlation function depending only on the distance between points (isotropy)

$$\text{corr}\left[\epsilon(\mathbf{x}_i), \epsilon(\mathbf{x}_j)\right] = C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|), \tag{2}$$

where $C_{\boldsymbol{\theta}}(d)$ is a valid correlation function of distance d, parameterized by a vector $\boldsymbol{\theta}$. Finally, $\rho(\mathbf{x})$ denotes an uncorrelated Gaussian process with mean zero and unitary variance, modelling the so-called "nugget effect", which allows for measurement error and small scale variation. This appears to be important in many applications (see *e.g.* Stein, 1999, Ch.3.7, Ecker and Gelfand, 1997, and De Oliveira and Ecker, 2002). The scale parameters σ and τ are both defined in $\Re_+$ and the ratio $\omega^2 = \tau^2/\sigma^2$ indicates the relative importance of the nugget effect.

We assume that we have observed a single realization from this random process at n different locations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and we denote the vector observation by $\mathbf{z} = (z_1, \ldots, z_n)'$, where we use the notation $z_i = Z(\mathbf{x}_i)$. In the literature, by far the most commonly made stochastic assumption is that $\epsilon(\mathbf{x})$ (and thus $Z(\mathbf{x})$) is a Gaussian process, which implies that $\mathbf{z}$ follows an n-variate Normal distribution with

$E[\mathbf{z}] = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X} = (\mathbf{f}(\mathbf{x}_1), \ldots, \mathbf{f}(\mathbf{x}_n))'$, and $\text{var}[\mathbf{z}] = \sigma^2\mathbf{C}_{\boldsymbol{\theta}} + \tau^2\mathbf{I}_n$, where $\mathbf{C}_{\boldsymbol{\theta}}$ is the $n \times n$ correlation matrix with $C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|)$ as its (i,j)th element.

In this paper we propose an alternative stochastic specification based on scale mixing the Gaussian process $\epsilon(\mathbf{x})$. In particular, a mixing variable $\lambda_i \in \Re_+$ is assigned to each observation $i = 1, \ldots, n$, and the joint distribution of $\mathbf{z}$ is changed to the following conditional Normal distribution:

$$p(\mathbf{z}|\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\theta}, \boldsymbol{\Lambda}) = f_N^n\left(\mathbf{z}|\mathbf{X}\boldsymbol{\beta}, \sigma^2(\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{C}_{\boldsymbol{\theta}}\boldsymbol{\Lambda}^{-\frac{1}{2}}) + \tau^2\mathbf{I}_n\right), \tag{3}$$

where we have defined the matrix $\boldsymbol{\Lambda} = \text{Diag}(\lambda_1, \ldots, \lambda_n)$ and $f_N^k(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is the pdf of a k-variate Normal distribution on $\mathbf{y}$ with mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$ (denoted as $\mathbf{y} \sim \text{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma})$). The sampling model for $\mathbf{z}$ is obtained by integrating out $\boldsymbol{\Lambda}$. Alternatively, if we use the notation $\epsilon_i = \epsilon(\mathbf{x}_i)$ and $\rho_i = \rho(\mathbf{x}_i)$, we can write the model for the ith location, $i = 1, \ldots, n$, as

$$z_i = \mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta} + \sigma\frac{\epsilon_i}{\sqrt{\lambda_i}} + \tau\rho_i, \tag{4}$$

where $\rho_i \sim \text{N}_1(0,1)$, i.i.d. and independent of $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)' \sim \text{N}_n(\mathbf{0}, \mathbf{C}_{\boldsymbol{\theta}})$. The mixing variables λ_i are independent of ρ_i and $\boldsymbol{\epsilon}$. Note that the scale mixing only affects the spatially dependent process and the nugget effect remains Gaussian. We feel this is the most natural specification, but the case where we mix the nugget effect process could also be considered, leading to a somewhat different interpretation. See Subsection 2.2.2 for more discussion on this issue.

First, we verify that the sampling model assumed is consistent with some well-defined stochastic process, in the sense that the Kolmogorov consistency conditions are verified. The latter hold for the process defined through the finite-dimensional distributions in (3), integrated with respect to the mixing variables $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$, provided the distribution of $\boldsymbol{\lambda}$, say $P_{\boldsymbol{\lambda}}$, satisfies a weak symmetry condition under permutation:

Proposition 1. The finite-dimensional distributions in (3) or (4) support a stochastic process $Z(\mathbf{x})$, if $P_{\lambda_1,\ldots,\lambda_n}(A_1, \ldots, A_n) = P_{\lambda_{\pi_1},\ldots,\lambda_{\pi_n}}(A_{\pi_1}, \ldots, A_{\pi_n})$, where $\pi_1, \ldots, \pi_n$ is any permutation of $1, \ldots, n$.
Proof: See Appendix A.

In the context of conditionally independent observations, scale mixing has been used in hierarchical models to accommodate and detect outlying observations in *e.g.* Wakefield *et al.* (1994). In this spatial model, however, the scale mixing introduces a potential problem with the continuity of the resulting random field Z. Let us, therefore, consider a stationary process $\lambda(\mathbf{x})$ for the mixing variables, so that $\lambda_i = \lambda(\mathbf{x}_i)$. The representation in (4) makes clear that we are now effectively replacing the Gaussian stochastic process $\epsilon(\mathbf{x})$ by a ratio of independent stochastic processes $\epsilon(\mathbf{x})/\sqrt{\lambda(\mathbf{x})}$. Mean square continuity of the spatial process $\epsilon(\mathbf{x})/\sqrt{\lambda(\mathbf{x})}$ is defined by $\text{E}[\{\epsilon(\mathbf{x}_i)/\sqrt{\lambda(\mathbf{x}_i)} - \epsilon(\mathbf{x}_j)/\sqrt{\lambda(\mathbf{x}_j)}\}^2]$ tending to zero as $\mathbf{x}_i \to \mathbf{x}_j$. We obtain

$$\text{E}\left[\left\{\frac{\epsilon_i}{\sqrt{\lambda_i}} - \frac{\epsilon_j}{\sqrt{\lambda_j}}\right\}^2\right] = 2\left\{\text{E}[\lambda_i^{-1}] - C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|)\text{E}[\lambda_i^{-1/2}\lambda_j^{-1/2}]\right\},$$

which in the limit as $\|\mathbf{x}_i - \mathbf{x}_j\| \to 0$ tends to $2\left\{\text{E}[\lambda_i^{-1}] - \lim_{\|\mathbf{x}_i - \mathbf{x}_j\| \to 0}\text{E}[\lambda_i^{-1/2}\lambda_j^{-1/2}]\right\}$. If λ_i and λ_j are independent, then $\lim_{\|\mathbf{x}_i - \mathbf{x}_j\| \to 0}\text{E}[\lambda_i^{-1/2}\lambda_j^{-1/2}] = \{\text{E}[\lambda_i^{-1/2}]\}^2 \leq \text{E}[\lambda^{-1}]$ from Jensen's inequality

and, thus, $\epsilon(\mathbf{x})/\sqrt{\lambda(\mathbf{x})}$ is not mean square continuous. This discontinuity essentially derives from the fact that two separate locations, no matter how close, are assigned independent mixing variables. Thus, in order to induce mean square continuity, we need to correlate the mixing variables in $\boldsymbol{\lambda}$, so that locations that are close will have very similar values of λ_i. In particular, if $\lambda^{-1/2}(\mathbf{x})$ is itself mean square continuous, then, by definition, $\lim_{\|\mathbf{x}_i-\mathbf{x}_j\|\to 0} \mathrm{E}[\lambda_i^{-1/2}\lambda_j^{-1/2}] = \mathrm{E}[\lambda^{-1}]$ and $\epsilon(\mathbf{x})/\sqrt{\lambda(\mathbf{x})}$ is a mean square continuous process. We now consider two situations for which this continuity holds.

2.1 Common mixing variables

One extreme implementation of this would be to assume that $\lambda_i = \lambda \sim P_\lambda, i = 1, \ldots, n$ so that all locations share a common mixing variable, irrespective of where they are. Let us briefly examine the implications of that for the models with and without a nugget effect.

2.1.1 Common mixing without nugget effect

In the absence of a nugget effect ($\tau^2 = 0$) the process is mean square continuous for any correlation function $C_{\boldsymbol{\theta}}(\cdot)$ which is continuous at zero (see *e.g.* Stein, 1999 for this general equivalence). The correlation structure will not be affected by the mixing, in the sense that $\mathrm{corr}\,[z_i, z_j] = C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|)$. This mixing generates multivariate scale mixtures of Normals. For example, if $P_\lambda = \mathrm{Ga}(\nu/2, \nu/2)$, the distribution in (3) marginalised with respect to λ will be n-variate Student-t with ν degrees of freedom. In addition, from the results in Fernández *et al.* (1995, 1997), we know that under the commonly used improper prior on the scale parameter, *i.e.* for the prior structure

$$p(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma) \propto \sigma^{-1} p(\boldsymbol{\beta}, \boldsymbol{\theta}), \tag{5}$$

posterior inference on $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and prediction will be totally unaffected by the choice of P_λ. The prior structure in (5) was used in Kitanidis (1986), Handcock and Stein (1993), Handcock and Wallis (1994) and Berger *et al.* (2001). Thus, this tells us that under a popular class of improper priors, the common scale mixing does not affect inference, and the latter is exactly the same as in the Gaussian model (where P_λ is Dirac at 1). This result also appears in Kim and Mallick (2003).

Within the class of jointly spherical (or elliptical) distributions as used in Kim and Mallick (2003), we know from Kelker (1970, Th.10) or Fang *et al.* (1990, Th.2.21) that the only finite-dimensional distributions able to support a stochastic process are scale mixtures of multivariate Normals, as obtained in this subsection.

2.1.2 Common mixing with nugget effect

Once we introduce a nugget effect, the correlation structure is affected by the scale mixing, as we obtain

$$\mathrm{corr}\,[z_i, z_j] = C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|) \frac{\mathrm{E}[\lambda^{-1}]}{\mathrm{E}[\lambda^{-1}] + \omega^2},$$

where $\omega^2 = \tau^2/\sigma^2$, and the process is no longer mean square continous. The exact inference robustness results of Fernández *et al.* (1995, 1997) under a prior compatible with (5) no longer apply. We know

from Proposition 1 that the finite-dimensional distributions in (3) combined with common mixing are consistent with a stochastic process. Of course, these finite-dimensional distributions are no longer scale mixtures of Normals (except trivially in the Gaussian case with Dirac $R\lambda$) but correspond to a superposition of a Gaussian and a mixed Gaussian process, and are outside the spherical (elliptical) class, as a consequence of the result stated at the end of Subsection 2.1.1.

2.2 Individual mixing variables

Even though this is exactly a case where one observation for all of the sites is the standard, and the robustness results mentioned in Subsection 2.1.1 (which do not hold for repeated sampling) have a real practical meaning, the common mixing idea above is perhaps not the most interesting way to extend the class of Gaussian processes. In particular, we would really like a situation that can account for individual "outliers", which we will implement by using a separate mixing variable for every location.

In particular, we will take the following mixing distribution:

$$\ln(\boldsymbol{\lambda}) = (\ln(\lambda_1), \ldots, \ln(\lambda_n))' \sim \mathrm{N}_n \left(-\frac{\nu}{2}\mathbf{1}, \nu \mathbf{C}_{\boldsymbol{\theta}} \right), \tag{6}$$

where $\mathbf{1}$ is a vector of ones, and we correlate the elements of $\ln(\boldsymbol{\lambda})$ through the same correlation matrix as $\boldsymbol{\epsilon}$ in (2). Equivalently, we assume a Gaussian process for $\ln(\lambda(\mathbf{x}))$ with constant mean surface at $-\nu/2$ and covariance function $\nu \mathbf{C}_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|)$. One scalar parameter $\nu \in \Re_+$ is introduced in (6), and we can easily see that the latter implies a log-Normal distribution for λ_i with $\mathrm{E}[\lambda_i] = 1$ and $\mathrm{var}[\lambda_i] = \exp(\nu) - 1$. Thus, the marginal distribution of λ_i is tight around one for very small ν (of the order $\nu = 0.01$) and as ν increases, the distribution becomes more spread out and more right skewed, while the mode shifts towards zero. For example, for $\nu = 3$, the variance is 19.1 and there is a lot of mass close to zero. It is exactly values of λ_i close to zero that will lead to an inflation of the scale in (4) and will allow us to accommodate outliers.

In principle, the correlation structure used in (6) need not coincide with that of $\boldsymbol{\epsilon}$. However, if we would use a different correlation matrix, separate inference on its parameters would be extremely difficult with practically relevant sample sizes.

We feel the process described by the model in (4) with the mixing distribution in (6) has many interesting properties for our purposes and we will focus on this model in the sequel of the paper. In the sequel, we shall denote the latter model by Gaussian-log-Gaussian (GLG) model, and the process $\epsilon(\mathbf{x})/\sqrt{\lambda(\mathbf{x})}$ by GLG process.

2.2.1 Properties of the Gaussian-log-Gaussian sampling model

From Proposition 1 and the permutation symmetry of (6), we know that $\mathbf{z}|\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\theta}$ is consistent with a well-defined stochastic process. The correlation structure is given by

$$\mathrm{corr}\,[z_i, z_j] = C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|) \frac{\exp\left(\nu\left\{1 + \frac{1}{4}[C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|) - 1]\right\}\right)}{\exp(\nu) + \omega^2}. \tag{7}$$

Thus, in the case without nugget effect ($\omega^2 = 0$) we see that if the distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ tends to zero, the correlation between z_i and z_j tends to one, so that the mixing does not induce a discontinuity at zero. From the continuity of the correlation (and thus covariance) function at zero, we know that the process $Z(\mathbf{x})$ is mean square continuous (which can also be immediately verified by the continuity of $\lambda^{-1/2}(\mathbf{x})$ from the discussion just preceding Subsection 2.1). More information on the smoothness of the surface is given by the degree of mean square differentiability (see Stein, 1999, and Banerjee and Gelfand, 2003, for an extension to higher dimensions m). We know that a process is q times mean square differentiable if and only if the $2q$th derivative of the covariance function at zero exists and is finite.

The covariance function of the model in (4) with individual mixing as in (6) is the following function of distance $d = \|\mathbf{x}_i - \mathbf{x}_j\| > 0$:

$$\mathrm{cov}_Z(d) = \sigma^2 C_{\boldsymbol{\theta}}(d) \exp\left(\nu\left\{1 + \frac{1}{4}[C_{\boldsymbol{\theta}}(d) - 1]\right\}\right), \tag{8}$$

which also hold for $d = 0$ in the absence of a nugget effect. Thus, we can derive:

Proposition 2. The process $Z(\mathbf{x})$ defined by (4) with the scale mixing distribution in (6) and no nugget effect ($\tau = 0$) is q times mean square differentiable ($q \geq 1$) if and only if the underlying process $\epsilon(\mathbf{x})$ is q times mean square differentiable.

Proof: See Appendix A.

In other words, the smoothness properties of $\epsilon(\mathbf{x})$ carry over to the process $Z(\mathbf{x})$ and the mixing does not affect the degree of mean square differentiability.

In order to evaluate how the tail behaviour of the finite-dimensional distributions corresponding to $Z(\mathbf{x})$ are affected by the log-Normal scale mixing in (6), we can compute the marginal moments of z_i around the mean $\mu_i = \mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta}$, which in the absence of a nugget effect are given as

$$\begin{aligned} \mathrm{E}[z_i - \mu_i]^r &= 0 && \text{for } r \text{ odd} \\ \mathrm{E}[z_i - \mu_i]^r &= \sigma^r \textstyle\prod_{i=1}^{r/2}(r - 2i + 1) \exp\left[\frac{\nu r}{4}(1 + \frac{r}{2})\right] && \text{for } r \text{ even}, \end{aligned} \tag{9}$$

where $r \geq 0$ takes integer values. An interesting property of log-Normal mixing is that all positive moments will exist (as opposed to *e.g.* Gamma mixing). From (9) we immediately find an expression for the kurtosis, which is simply given by $\mathrm{kurt}[z_i] = 3\exp(\nu)$, clearly indicating that ν governs tail behaviour. In particular, as $\nu \to 0$ we retrieve Normal tails, while larger values of ν will induce thicker tails. In order to relate the behaviour to that of Student-t tails with a degrees of freedom, Table 1 displays the kurtosis and the value of $a > 4$ that would lead to the same kurtosis for various values of ν.

ν	0.01	0.1	0.5	1	2	3	4
$\mathrm{kurt}[z_i]$	3.03	3.32	4.95	8.15	22.2	60.3	163.8
a	203.0	23.0	7.08	5.16	4.31	4.10	4.04

Table 1: Kurtosis values as a function of ν and comparison with Student-t.

2.2.2 Spatial heteroskedasticity and detection

As explained in the beginning of Section 2, our chosen specification of mixing the spatially dependent process as in (4) requires a smooth $\lambda(\mathbf{x})$ process and this means that observations with particularly small values of λ_i will tend to cluster together. Thus, what we are identifying through small values of λ_i are regions of the space where the observations tend to be relatively far away from the estimated mean surface. Therefore, we can interpret the presence of small values of λ_i in terms of spatial heteroskedasticity, rather than the usual concept of outlying observations. The regions characterized by observations with small λ_i's require a relatively large variance around the underlying trend surface. However, for want of a better term we will continue to call observations with small λ_i "outliers", even though we really mean that they belong to a region with larger observational variance, relative to the rest of the space.

As already mentioned after (4), we could alternatively apply the mixing to the independent nugget error process. Then independent mixing variables would be a natural choice, and the interpretation of the mixing variables would be somewhat different. In fact, the occurrence of particularly small λ_i values would then simply indicate an unusually large nugget effect (most likely measurement error) and would really correspond to the traditional notion of "outliers", which would have no tendency to cluster.

Even though our chosen model with mixing through (6) (the GLG model) can account for outliers and does not need to identify them for inference, it may be useful to have an indication of which observations are particularly out of line with the rest (*i.e.* which areas of the space require an inflated variance). Indicating regions of the space where the Gaussian model fails to fit the data well might, in some cases, suggest extensions to the underlying trend surface that would make a Gaussian model a more feasible option. For example, it could indicate missing covariates in the chosen form for the mean surface. Outliers are not defined here as observations that can not be modelled properly, but rather those observations that necessitate the use of the GLG model. Clearly, the distribution of λ_i is informative about the outlying nature of observation i. In the extreme case that $\lambda_i = 1$, the marginal sampling distribution of z_i is Gaussian. Thus, we propose to compute the ratio between the posterior and the prior density functions for λ_i evaluated at $\lambda_i = 1$, *i.e.*

$$R_i = \frac{p(\lambda_i|\mathbf{z})}{p(\lambda_i)}|_{\lambda_i=1}\,. \tag{10}$$

In fact, the ratio R_i is the Savage-Dickey density ratio and this would be the Bayes factor in favour of the model with $\lambda_i = 1$ (and all other elements of $\boldsymbol{\lambda}$ free) versus the model with free λ_i (*i.e.* the full mixture model proposed here) if $C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|) = 0$ for all $j \neq i$. Thus, R_i would be very close to the Bayes factor against z_i being an outlier if $\mathbf{x}_i$ was sufficiently far from all other locations. In general, the Savage-Dickey density ratio is no longer the exact Bayes factor, but has to be adjusted as discussed in Verdinelli and Wasserman (1995). Here, the ratio R_i needs to be multiplied by a factor

$$\mathrm{E}\left[\frac{p(\ln \boldsymbol{\lambda}_{-i})}{p(\ln \boldsymbol{\lambda}_{-i}|\lambda_i = 1)}\right] = \mathrm{E}\left[\frac{f_N^{n-1}\left(\ln \boldsymbol{\lambda}_{-i}| - \frac{\nu}{2}\mathbf{1}, \nu\mathbf{C}_{\boldsymbol{\theta}}^{(-i)}\right)}{f_N^{n-1}\left(\ln \boldsymbol{\lambda}_{-i}| - \frac{\nu}{2}(\mathbf{1} - c_i), \nu(\mathbf{C}_{\boldsymbol{\theta}}^{(-i)} - c_i c_i')\right)}\right]\,, \tag{11}$$

where $\boldsymbol{\lambda}_{-i}$ is $\boldsymbol{\lambda}$ with the ith element deleted, c_i is the vector of elements $C_{\boldsymbol{\theta}}(\|\mathbf{x}_i - \mathbf{x}_j\|), j \neq i$ and $\mathbf{C}_{\boldsymbol{\theta}}^{(-i)}$ is the matrix $\mathbf{C}_{\boldsymbol{\theta}}$ without the ith row and column. The expectation in (11) is with respect to $p(\boldsymbol{\lambda}_{-i}|\lambda_i = 1, \mathbf{z})$.

2.2.3 Prediction

An important reason for geostatistical modelling is to be able to predict the response variable at unsampled sites. The standard approach to this problem is kriging (see *e.g.* Cressie, 1993), which is essentially linear prediction using optimal least squares interpolation of the random field. For Gaussian processes, standard kriging can be given a Bayesian interpretation (with an improper uniform prior on $\boldsymbol{\beta}$) if we condition on the covariance structure.

Here we are, of course, dealing with a non-Gaussian process and wish to fully incorporate all uncertainty in the problem, including the covariance function. Obtaining the full predictive distribution will also immediately allow us to conduct inference on other aspects of interests, such as the probability of exceedance of threshold values (see *e.g.* De Oliveira and Ecker, 2002).

Let $\mathbf{z} = (\mathbf{z}_o', \mathbf{z}_p')'$ where $\mathbf{z}_o$ correspond to the $n - f$ observed locations and $\mathbf{z}_p$ is a vector of values to predict at f given sites. We are interested in the posterior predictive distribution of $\mathbf{z}_p$, *i.e.*

$$p(\mathbf{z}_p|\mathbf{z}_o) = \int p(\mathbf{z}_p|\mathbf{z}_o, \boldsymbol{\lambda}, \boldsymbol{\zeta}) p(\boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \boldsymbol{\zeta}, \mathbf{z}_o) p(\boldsymbol{\lambda}_o, \boldsymbol{\zeta}|\mathbf{z}_o) \mathrm{d}\boldsymbol{\lambda} \mathrm{d}\boldsymbol{\zeta}, \tag{12}$$

where we have partitioned $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_o', \boldsymbol{\lambda}_p')'$ conformably with $\mathbf{z}$ and $\boldsymbol{\zeta} = (\boldsymbol{\beta}, \sigma, \tau, \boldsymbol{\theta}, \nu)$. The integral in (12) will be approximated to any desired level of accuracy by Monte Carlo simulation, where the drawings for $(\boldsymbol{\lambda}_o, \boldsymbol{\zeta})$ are obtained directly from the MCMC inference chain (which is described below), and, since $p(\boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \boldsymbol{\zeta}, \mathbf{z}_o) = p(\boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \nu)$ we can evaluate (12) by using drawings for $\boldsymbol{\lambda}_p$ from

$$p(\ln \boldsymbol{\lambda}_p|\boldsymbol{\lambda}_o, \nu) = f_N^f \left(\ln \boldsymbol{\lambda}_p \,|\, \frac{\nu}{2}[\mathbf{C}_{po}\mathbf{C}_{oo}^{-1}\mathbf{1}_n - \mathbf{1}_f] + \mathbf{C}_{po}\mathbf{C}_{oo}^{-1} \ln \boldsymbol{\lambda}_o,\, \nu[\mathbf{C}_{pp} - \mathbf{C}_{po}\mathbf{C}_{oo}^{-1}\mathbf{C}_{op}] \right), \tag{13}$$

where we have partitioned

$$\mathbf{C}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathbf{C}_{oo} & \mathbf{C}_{op} \\ \mathbf{C}_{po} & \mathbf{C}_{pp} \end{pmatrix}$$

conformably with $\mathbf{z}$. Thus, for each posterior drawing of $(\boldsymbol{\lambda}_o, \boldsymbol{\zeta})$, we will generate a drawing from (13) and evaluate

$$p(\mathbf{z}_p|\mathbf{z}_o, \boldsymbol{\lambda}, \boldsymbol{\zeta}) = f_N^f \left((\mathbf{X}_p - \mathbf{A}\mathbf{X}_o)\boldsymbol{\beta} + \mathbf{A}\mathbf{z}_o, \sigma^2 \left(\boldsymbol{\Lambda}_p^{-\frac{1}{2}} \mathbf{C}_{pp} \boldsymbol{\Lambda}_p^{-\frac{1}{2}} + \omega^2 \mathbf{I}_f - \mathbf{A}\boldsymbol{\Lambda}_o^{-\frac{1}{2}} \mathbf{C}_{op} \boldsymbol{\Lambda}_p^{-\frac{1}{2}} \right) \right), \tag{14}$$

where $\mathbf{A} = \boldsymbol{\Lambda}_p^{-\frac{1}{2}} \mathbf{C}_{po} \boldsymbol{\Lambda}_o^{-\frac{1}{2}} \left[\boldsymbol{\Lambda}_o^{-\frac{1}{2}} \mathbf{C}_{oo} \boldsymbol{\Lambda}_o^{-\frac{1}{2}} + \omega^2 \mathbf{I}_n \right]^{-1}$ and $\mathbf{X}$ and $\boldsymbol{\Lambda}$ are partitioned conformably to $\mathbf{z}$. The average of the densities in (14) will give us the required posterior predictive density function.

Due to the fact that we can think of our GLG model as Gaussian given $\boldsymbol{\lambda}$, we can still base prediction on a mixture of simple conditional Normal densities, just like for the Gaussian model (where $\boldsymbol{\Lambda} = \mathbf{I}$).

2.3 The Correlation Function

For the correlation function $C_{\boldsymbol{\theta}}(d)$, where d is the Euclidean distance, we use the flexible Matèrn class:

$$C_{\boldsymbol{\theta}}(d) = \frac{1}{2^{\theta_2 - 1}\Gamma(\theta_2)} \left(\frac{d}{\theta_1} \right)^{\theta_2} \mathcal{K}_{\theta_2} \left(\frac{d}{\theta_1} \right), \tag{15}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)'$ with $\theta_1 > 0$ the range parameter and $\theta_2 > 0$ the smoothness parameter and where $\mathcal{K}_{\theta_2}(\cdot)$ is the modified Bessel function of the third kind of order θ_2. Whereas the range parameter θ_1 indicates how fast the correlation decreases with d, θ_2 controls the smoothness of the random field. In particular $\epsilon(\mathbf{x})$ and thus, by Proposition 2, $Z(\mathbf{x})$ are q times mean square differentiable if and only if $\theta_2 > q$ (see Stein, 1999, p.31).

3 The Prior Distribution

Berger *et al.* (2001) derive a reference prior for the simpler Gaussian model without nugget effect and with fixed smoothness parameter θ_2. This benchmark is based on formal arguments and is very important for obtaining insight into the properties of these models. However, the reference prior algorithm has a number of drawbacks in this particular context:

- The reference prior idea was originally developed for independent replications from an experiment (see Berger and Bernardo, 1992, Bernardo and Smith, 1994, p.299). In the context of dependent observations, these priors can lead to a number of problems, as has already been documented for time series models. In particular, the prior depends on sample size n, which means that we do not get the same result if we process part of the sample first and the rest later. This is a rather unsatisfactory property of the reference (and Jeffreys') prior in this context.
- There is a variety of essentially arbitrary decisions in applying the reference prior algorithm that could affect the result. Whereas no limiting sequences of sets are mentioned in the derivation of the reference prior in Berger *et al.* (2001), there still remain the issues of how to group the parameters and how to order them in inferential importance.
- Given the reported difficulties in Berger *et al.* (2001) to extend the reference prior to include unknown smoothness parameter θ_2, it is unlikely that a full reference prior for our model, where we add θ_2 as well as the nugget effect and the parameter ν of the mixing distribution, is both explicitly available and practical to use. Paulo (2005) explores objective priors for multiparameter correlation structures. However, his results do not cover, *e.g.* smoothness parameters, but focus on separable correlation functions with Cartesian product structures for the locations.

Thus, we will take a different approach and choose a proper prior instead, where we attempt to induce reasonable properties by a careful elicitation process. The propriety of the prior will also mean that we do not need to verify posterior existence.

The prior will be continuous with a density function of the form:

$$p(\boldsymbol{\beta}, \sigma^{-2}, \omega^2, \nu, \boldsymbol{\theta}) = p(\boldsymbol{\beta})p(\sigma^{-2})p(\omega^2)p(\nu)p(\boldsymbol{\theta}), \tag{16}$$

where we have parameterized in terms of $\omega^2 = \tau^2/\sigma^2$. In the sequel, we describe our choice for the most "reasonable" prior, used as a benchmark, and list some alternatives. In Subsection 5.1 we will investigate sensitivity to the prior assumptions.

Prior on $\boldsymbol{\beta}$: For the purposes of posterior and predictive inference we can simply take

$$\boldsymbol{\beta} \sim \mathrm{N}_k(\mathbf{0}, c_1 \mathbf{I}_k), \tag{17}$$

where we can allow c_1 to vary. As a benchmark value for c_1 we choose 10^4. For posterior and predictive inference, this is not a critical prior and any suitably large value of c_1 will lead to very similar results. However, if we wish to compare models with different trend specifications, the prior on $\boldsymbol{\beta}$ becomes much more critical, as it will have a direct impact on the relevant Bayes factors (see *e.g.* Kass and Raftery, 1995). Then it becomes critical to tune the prior on these model-specific parameters $\boldsymbol{\beta}$ very carefully. One simple way of implementing this is to standardize the data, so that the interpretations of all coefficients are comparable and not affected by merely rescaling or relocating the data. This is equivalent to adopting the following prior for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)'$ with the original data:

$$p(\boldsymbol{\beta}) = f_N^1(\beta_1 | m_z, \tilde{c}_1 s_z^2) \prod_{j=2}^{k} f_N^1(\beta_j | 0, \tilde{c}_1 \frac{s_z^2}{s_{x_j}^2}), \tag{18}$$

where m_z and s_z^2 are the sample mean and variance of $\mathbf{z}$, $s_{x_j}^2$ is the sample variance of the jth column of $\mathbf{X}$ and we have assumed that $f_1(\mathbf{x}) = 1$ throughout, accommodating an intercept. Values of $\tilde{c}_1$ in the order of 10 or so seem reasonable in the context of (18). Thus, this latter prior allows us a formal way of choosing between competing trend specifications at the cost of making the prior somewhat data-dependent. In practice, the priors (17) and (18) lead to virtually identical results for posterior and predictive inference, and it is only for the purpose of assessing the evidence in favour of different trend specifications that we will use (18).

Prior on σ^{-2}: Ideally, we would like a prior that is invariant to rescaling the observations. Whereas that can not be achieved exactly with a proper prior, we approximate this by adopting

$$\sigma^{-2} \sim \mathrm{Ga}(c_2, c_3)$$

with very small values of c_2 and c_3 (as benchmark values we take 10^{-6}).

Prior on ω^2: For ω^2 we often expect values that are smaller than unity, and most prior mass could be around small values (no nugget effect), so that any strong nugget effect comes from the data rather than the prior. However, we do not want the prior to exclude a large nugget effect either and propose to use the flexible Generalized Inverse Gaussian prior as defined in Appendix B:

$$\omega^2 \sim \mathrm{GIG}(0, c_4, c_5),$$

where we fix the first hyperparameter to zero, as this leads to reasonable shapes of the prior density function without any real loss of flexibility. Choosing the value 0.2 for the prior mode throughout, we take as benchmark values $c_4 = 0.66$ and $c_5 = 1$, which implies a prior mean of 1.07 with a standard deviation of 1.20. In the prior sensitivity analysis in Subsection 5.1 we will adopt two rather extreme alternative sets of values. In particular, we take $(c_4, c_5) = (0.87, 3)$ leading to a mean of 0.34 and a standard deviation of 0.21 (very concentrated case), and $(c_4, c_5) = (0.63, 0.25)$ with mean 7.82 and standard deviation 13.98 (very dispersed case). To assess the impact of allowing for more prior mass very close to zero, the prior sensitivity analysis also considers an exponential prior with mean 0.5.

Prior on ν: The parameter ν is also expected to have a rather restricted range in practice. From Table 1 we know that very small values (around 0.01) correspond to near Normality and large values (of the order of say 3) indicate very thick tails. Again we adopt the Generalized Inverse Gaussian prior class

$$\nu \sim \text{GIG}(0, c_6, c_7).$$

If we fix the mode at 0.1, reasonable benchmark values are $c_6 = 0.5$ and $c_7 = 2$, corresponding to a prior mean of 0.36 with a standard deviation of 0.33. We shall also use the very different alternative pairs of values $(0.75, 6)$ and $(0.45, 0.5)$ for (c_6, c_7). The mode stays at 0.1, while the implied means and standard deviations are 0.14 and 0.06 for the first pair (very concentrated case) and 2.3 and 3.73 for the second pair (very dispersed case). If we want to put a fair amount of prior mass close to Normality without excluding really thick tails, a reasonable prior might also be an exponential. The latter prior with mean 0.2 is also used in the prior sensitivity analysis.

Prior on $\boldsymbol{\theta}$: We will distinguish two separate prior structures here corresponding to independence in two different parameterisations. The correlation function in (15) is parameterised in terms of (θ_1, θ_2). However, Stein (1999, p. 51) shows that the alternative range parameter $\rho = 2\theta_1\sqrt{\theta_2}$ is linked with the autocovariance function in a way that is less dependent on the smoothness parameter θ_2. This suggests the alternative parameterisation (ρ, θ_2).

Prior independence between θ_1 and θ_2: The prior on θ_1 should take into account that the value of this range parameter is critically dependent on the scaling of the distance d. In fact, the correlation structure only depends on θ_1 through θ_1/d. We propose an exponential with mean $\text{med}(d)/c_8$, denoted as

$$\theta_1 \sim \text{Exp}\left(\frac{c_8}{\text{med}(d)}\right), \tag{19}$$

where $\text{med}(d)$ is the median value of all distances in the data, so that if there is a change in measurement system (scaling of $\mathbf{x}$) this is accurately reflected in the prior on θ_1. The exponential form ensures that there is a lot of mass close to $\theta_1 = 0$, which is the limiting case of no correlation; so if we find evidence of substantial correlation, it would mostly come from the data. We try a range of values of c_8 to assess prior sensitivity. A rough idea of useful values for c_8 could be obtained by considering the implied correlation in the simple exponential correlation case (which is the special case of (15) for $\theta_2 = 1/2$) at the median distance when replacing θ_1 by its prior mean. This correlation would be $\exp(-c_8)$ and thus we can obtain guidelines on c_8 by considering reasonable values for this correlation. $c_8 = 2.3, 0.92, 0.22$ corresponds to correlation values of $0.1, 0.4, 0.8$, respectively.

The smoothness parameter θ_2 is linked to the degree of mean square differentiability of the process, and will be given a simple exponential prior

$$\theta_2 \sim \text{Exp}(c_9).$$

A reasonable value for c_9 as a benchmark is 0.5. Then, the mean will be 2 and the variance 4. In the prior sensitivity analysis we will also use the values 0.25 and 2 for c_9.

Prior independence between ρ and θ_2: Now, instead of a product structure prior on (θ_1, θ_2), we use a similar structure in terms of ρ and θ_2, in line with the fact that the interpretation of ρ as a range parameter

is less dependent on θ_2. In particular, we take $p(\rho, \theta_2) = p(\rho)p(\theta_2)$ with the same marginal prior $p(\theta_2)$ as before and

$$\rho \sim \text{Exp}\left(\frac{c_8}{\sqrt{2}\,\text{med}(d)}\right),$$

which corresponds to the same conditional prior on θ_1 as above for $\theta_2 = 1/2$. This is equivalent to $p(\theta_1, \theta_2) = p(\theta_1|\theta_2)p(\theta_2)$, where $p(\theta_2)$ is as before and now

$$\theta_1|\theta_2 \sim \text{Exp}\left(\frac{c_8\sqrt{2\theta_2}}{\text{med}(d)}\right). \tag{20}$$

4 The MCMC Sampler

The parameters introduced into our model (3), (6) and (15) are $(\boldsymbol{\beta}, \sigma^{-2}, \boldsymbol{\theta}, \omega^2, \nu)$, and in order to facilitate the sampling, we will augment with the mixing variables $\boldsymbol{\lambda}$ and the vector $\boldsymbol{\epsilon}$ defined in (3). Thus, we run an MCMC algorithm on $\boldsymbol{\beta}, \sigma^{-2}, \omega^2, \nu, \boldsymbol{\theta}, \boldsymbol{\lambda}, \boldsymbol{\epsilon}|\mathbf{z}$. If we want to compute the Bayes factor in favour of the model with $\lambda_i = 1$ (see Subsection 2.2.2), we run a separate sampler on $\boldsymbol{\beta}, \sigma^{-2}, \omega^2, \nu, \boldsymbol{\theta}, \boldsymbol{\lambda}_{-i}, \boldsymbol{\epsilon}|\lambda_i = 1, \mathbf{z}$ to evaluate (11).

The conditional posteriors for $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ are Normal distributions which are treated through Gibbs sampling. The full conditional for $\boldsymbol{\epsilon}$ is given by

$$p(\boldsymbol{\epsilon}|\boldsymbol{\beta}, \sigma^{-2}, \boldsymbol{\theta}, \omega^2, \boldsymbol{\lambda}, \nu, \mathbf{z}) \propto f_N^n\left(\boldsymbol{\epsilon} \mid \mathbf{W}^{-1}\mathbf{d}, \mathbf{W}^{-1}\right),$$

where $\mathbf{W} = \mathbf{C}_{\boldsymbol{\theta}}^{-1} + \omega^{-2}\boldsymbol{\Lambda}^{-1}$ and $\mathbf{d} = (d_1, \ldots, d_n)'$ with $d_i = \sigma(z_i - \mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta})/(\tau^2\sqrt{\lambda_i})$.

The parameters $\sigma^{-2}, \omega^2, \nu, \theta_1$ and θ_2 are all drawn separately using random walk Metropolis-Hastings algorithms, tuned to give a reasonable acceptance rate. The only parameter we discuss in some detail here is $\boldsymbol{\lambda}$ as drawings from its conditional are not that trivial to conduct.

4.1 Conditional posterior for $\boldsymbol{\lambda}$

The main problem is that because of the correlation induced by (6), the elements of $\boldsymbol{\lambda}$ are not conditionally independent given the other parameters and the data. This complicates matters in view of the large dimension of $\boldsymbol{\lambda}$.

We will partition the elements of $\boldsymbol{\lambda}$ in blocks, where each block corresponds to a cluster of observations that are relatively close together. Thus, most of the dependence between the λ_i's will be confined to the same cluster, and by drawing the entire block at once, mixing will be improved. For each cluster we use a Metropolis-Hastings step, so we need to come up with a proposal that has a reasonable acceptance probability, *i.e.* we need to find a good approximation to the relevant conditional. The latter consideration will, in practice, limit the size of the clusters.

Let $\boldsymbol{\lambda}_{(i)}$ denote the n_i elements of $\boldsymbol{\lambda}$ corresponding to cluster i, and indicate by $\boldsymbol{\lambda}_{-(i)}$ the remaining elements, so that we partition

$$\boldsymbol{\lambda} = \begin{pmatrix} \boldsymbol{\lambda}_{-(i)} \\ \boldsymbol{\lambda}_{(i)} \end{pmatrix} \text{ and conformably } \mathbf{C}_{\boldsymbol{\theta}} = \begin{pmatrix} \mathbf{C}_{11} & \mathbf{C}_{12} \\ \mathbf{C}_{21} & \mathbf{C}_{22} \end{pmatrix}. \tag{21}$$

The conditional posterior distribution for $\boldsymbol{\lambda}_{(i)}$ given $\boldsymbol{\lambda}_{-(i)}$ is given by

$$p(\boldsymbol{\lambda}_{(i)}|\boldsymbol{\lambda}_{-(i)},\boldsymbol{\beta},\sigma^{-2},\boldsymbol{\theta},\omega^2,\nu,\boldsymbol{\epsilon},\mathbf{z}) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i\in(i)}\left(z_i-\mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta}-\frac{\sigma\epsilon_i}{\sqrt{\lambda_i}}\right)^2\right\} p(\boldsymbol{\lambda}_{(i)}|\boldsymbol{\lambda}_{-(i)},\boldsymbol{\theta},\nu), \tag{22}$$

where $i \in (i)$ indicates all observations corresponding to cluster i and $p(\boldsymbol{\lambda}_{(i)}|\boldsymbol{\lambda}_{-(i)},\boldsymbol{\theta},\nu)$ is the prior conditional derived from the mixing distribution in (6), characterized by

$$\ln(\boldsymbol{\lambda}_{(i)})|\boldsymbol{\lambda}_{-(i)},\boldsymbol{\theta},\nu \sim N\left(\frac{\nu}{2}(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{1}-\mathbf{1})+\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\ln(\boldsymbol{\lambda}_{-(i)}),\nu(\mathbf{C}_{22}-\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12})\right). \tag{23}$$

The first factor in (22), *i.e.* the likelihood contribution, is proportional to a product of pdf's of truncated Normal distributions on $\lambda_i^{-1/2}, i \in (i)$. In order to construct a suitable candidate generator, we will approximate these truncated Normal distributions by log-Normal distributions on λ_i. By matching the first two moments of $\lambda_i^{-1/2}$ we obtain as approximating distribution of the likelihood contribution to λ_i

$$\ln(\lambda_i)|\boldsymbol{\beta},\sigma^{-2},\omega^2,\epsilon_i,z_i \sim N(m_i,s_i^2), \tag{24}$$

with

$$m_i = \ln\frac{\epsilon_i^2[\eta_i+\eta_i\delta(\eta_i)+1]}{\omega^2[\eta_i+\delta(\eta_i)]^4}$$

$$\text{and } s_i^2 = 4\ln\frac{\eta_i^2+\eta_i\delta(\eta_i)+1}{[\eta_i+\delta(\eta_i)]^2},$$

where we have defined $\eta_i = \tau^{-1}[z_i-\mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta}]\text{sign}(\epsilon_i)$ and $\delta(\cdot) = \phi(\cdot)/\Phi(\cdot)$ denotes the Mill's ratio (the ratio of the pdf and the cdf of the standard Normal). Combining (23) and (24) we will propose a candidate value for $\boldsymbol{\lambda}_{(i)}$ from

$$p(\ln(\boldsymbol{\lambda}_{(i)})|\boldsymbol{\lambda}_{-(i)},\boldsymbol{\beta},\sigma^{-2},\boldsymbol{\theta},\omega^2,\nu,\boldsymbol{\epsilon},\mathbf{z}) = f_N^{n_i}\left(\boldsymbol{\mu}_{(i)},\boldsymbol{\Sigma}_{(i)}\right), \tag{25}$$

where

$$\boldsymbol{\Sigma}_{(i)}^{-1} = \nu^{-1}\left(\mathbf{C}_{22}-\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)^{-1} + \text{Diag}_{i\in(i)}(s_i^{-2}) \text{ and}$$

$$\boldsymbol{\mu}_{(i)} = \boldsymbol{\Sigma}_{(i)}\left\{\nu^{-1}\left(\mathbf{C}_{22}-\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}\right)^{-1}\left[\frac{\nu}{2}\left(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{1}-\mathbf{1}\right)+\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\ln(\boldsymbol{\lambda}_{-(i)})\right]+(s_i^{-2}m_i)_{i\in(i)}\right\}.$$

Due to the construction of this generator, the acceptance probability of the Metropolis-Hastings step will only depend on the ratio of the likelihood contribution to $\boldsymbol{\lambda}_{(i)}$ and its approximation (with the same first two moments) in (24). Experience with a number of examples suggests that this works very well with cluster sizes of the order of 10 to 15 observations.

In the Gaussian case, we run the sampler without the steps for $\boldsymbol{\epsilon}$, $\boldsymbol{\lambda}$ (which is equal to $\mathbf{1}$) and ν.

5 Prior Sensitivity and Identifiability

5.1 Prior Sensitivity Analysis

Section 3 described a choice of prior structures on $\boldsymbol{\theta}$ and various marginal prior distributions on $\boldsymbol{\beta}$, ω^2 and ν. In addition, it indicated some benchmark values and some extreme values for the various

hyperparameters used in these priors. We now investigate the sensitivity of posterior results to changes in the prior. We conduct this sensitivity analysis separately for both prior structures on $\boldsymbol{\theta}$, described by either (19) or (20). It is worth stressing that the range of priors used here is very wide, and the sample used is rather small. This was done specifically in order to assess where the data is little informative and in which directions we need to concentrate the effort in prior elicitation. We do not expect the posterior results to be stable in this sensitivity analysis. Of course, our interest is often in prediction rather than the parameters per se, and we expect predictive inference to be much more robust with respect to prior changes than posterior inference.

As data we use the spatial data set of 52 topographic elevations (in feet) within a 310 yard square introduced in Davis (1973) and used later in *e.g.* Handcock and Stein (1993) and Berger *et al.* (2001). As in the latter paper, we consider a model with a quadratic trend in the spatial coordinates $\mathbf{x} = (x_1, x_2)'$:

$$\mathbf{f}(\mathbf{x})'\boldsymbol{\beta} = \beta_1 + \beta_2 x_1 + \beta_3 x_1^2 + \beta_4 x_2 + \beta_5 x_1 x_2 + \beta_6 x_2^2. \tag{26}$$

Results are based on every 5th draw from an MCMC chain of length 2,000,000 with a burn-in of 500,000. This proved more than enough for convergence and much shorter runs led to virtually identical results.

In this sensitivity analysis, we change the prior of one parameter at a time, say of parameter $\boldsymbol{\zeta}_i$, and monitor the posterior distributions, focusing in particular on that of $\boldsymbol{\zeta}_i$. As a simple measure of how the resulting posterior has changed, we compute the absolute value of the induced change in the marginal posterior mean of $\boldsymbol{\zeta}_i$ and divide it by the standard deviation computed under the benchmark prior. We shall call this the "relative change". If $\boldsymbol{\zeta}_i$ is a vector, we calculate this relative change for each element and we report the maximum relative change across elements. Table 2 lists the various values used for the alternative prior hyperparameters and possible alternative prior distributions, as well as the maximum (across priors) relative change recorded for each parameter using the GLG model. Of course, this measure is not invariant with respect to reparameterization. Nevertheless, we feel this can give a rough idea of where some extra effort in prior elicitation (or sensitivity analyses) may pay off. On the

$\boldsymbol{\zeta}_i$	hyperpar.	Benchmark	Alternative values or cases	max. rel. change	
				(19)	(20)
$\boldsymbol{\beta}$	c_1	10^4	$10^2, 10^6$; prior (18) with $\tilde{c}_1 = 10$	0.04	0.07
σ	(c_2, c_3)	$(10^{-6}, 10^{-6})$	$(10^{-9}, 10^{-3})$, $(10^{-3}, 10^{-9})$	0.07	0.13
ω^2	(c_4, c_5)	$(0.66, 1)$	$(0.87, 3)$, $(0.63, 0.25)$; Exp. with mean 0.5	0.40	0.42
ν	(c_6, c_7)	$(0.5, 2)$	$(0.75, 6)$, $(0.45, 0.5)$; Exp. with mean 0.2	5.49	3.99
θ_1	c_8	0.92	$0.22, 2.3$	0.50	0.46
θ_2	c_9	0.5	$0.25, 2$	0.79	0.83

Table 2: Prior sensitivity analysis: Setup of the experiment and maximum relative change.

basis of the summary in Table 2 and the complete results of the sensitivity analysis, we can partition the parameters in three groups.

Firstly, the prior on $(\boldsymbol{\beta}, \sigma)$ does not appear to play a very critical role. For $\boldsymbol{\beta}$ we note that the maximum relative change occurs in the intercept β_1, while the other elements display even more robustness

with respect to the prior changes. In addition, the inference on other parameters is virtually unaffected by changes in the prior on $(\boldsymbol{\beta}, \sigma)$.

The second group of parameters is (ω^2, ν). Here the priors are more important in driving the results, and the prior on ω^2 also affects the inference on σ. By far the largest influence of prior changes occurs for ν. The extreme case corresponds to the dispersed prior, which leads to a large increase in both the posterior mean and the posterior standard deviation (a factor 5.8 times the standard deviation in the benchmark case under (19) and a factor 4.5 under (20)). Clearly, there is relatively little direct information in the data on ν, so that the prior is quite important. Of course, we have used rather extreme alternative priors and, if anything, our benchmark prior is conservative in that it concentrates a lot of mass around small values of ν. There is a marked positive relation between induced changes in ω^2 and ν and a negative relation between those in ω^2 and σ. The latter might suggest that information in the data is more readily available in terms of τ^2 than ω^2.

The third group are the parameters of the Matèrn correlation function (θ_1, θ_2). The priors clearly have a role to play and the data do not seem to be highly informative on the individual parameters, especially the smoothness parameter θ_2. There is a clear negative relation between changes in θ_1 and θ_2, which is somewhat less when we parameterize in terms of ρ (as defined in Section 3), but remains to some extent even under prior independence between ρ and θ_2.

In the direction where the data are least informative (*i.e.* on ν), we find that the results under (20) are slightly less sensitive to the prior than when we use the prior based on (19). This and the fact that the parameterization in terms of ρ is easier to interpret (see Stein, 1999) leads us to adopt the prior with prior independence between ρ and θ_2 as our benchmark prior in the sequel.

Overall, there is moderate sensitivity to changes in the prior, especially that on ν. It has to be stressed that we are considering rather dramatic deviations in the prior, while only using a fairly small dataset. However, the results are not very robust to the particular prior chosen, so the latter clearly matters. The evidence suggests that the best direction for extra effort in the prior elicitation is primarily towards ν but also θ_2 and (ω^2, θ_1), as these seem the most critical dimensions of the prior. Our sensitivity analysis tells us that we really need to put some structure into the problem through the prior since the data information is not that strong in certain directions. Thus, we need to think carefully about our priors and try to use as much information as we have available in eliciting reasonable prior distributions. In this particular context, we feel that this strategy is preferable to relying on automatic noninformative priors like the reference prior (if such priors are at all available; see also Section 3).

5.2 Identifiability

The GLG model introduces the extra parameter ν beyond the parameterisation of the usual Gaussian model, and it is natural to examine to what extent information on the parameters can be recovered from data. From the expression for the kurtosis following (9) it is clear that ν is identifiable from the data, so we are not faced with a nonidentifiable parameter as in Neath and Samaniego (1997), where the prior is the only source of information, conditionally on the other (identified) parameters in the model. Nevertheless, we may wonder how much information can be gained from data sets of realistic sizes.

In order to address this issue, we generate data from our GLG model with a constant mean surface (*i.e.* $k = 1$ and $\mathbf{f}(\mathbf{x}) = 1$) and locations that are randomly drawn from a uniform distribution on the unit square. We conduct inference on all parameters, using the benchmark prior established in Subsection 5.1. We focus particular interest on ν, ω^2 and the correlation parameters in $\boldsymbol{\theta}$, as we would expect inference to be most challenging for these parameters. The latter fact is corroborated by the prior sensitivity analysis in Subsection 5.1.

Throughout, we use a sample size of $n = 100$ with $\beta_1 = 10$ and $\sigma^2 = 1$. The left panel of Figure 1 displays the posterior distributions for ν for five data sets generated with $\omega^2 = 0.1$ and $\theta_1 = \theta_2 = 0.5$ and with different values of ν, ranging from $\nu = 0.01$ (virtually Gaussian) to $\nu = 3$ (very fat tails, see Table 1). These are marginal posterior densities for ν while conducting inference on all parameters in the model and keeping the same prior, and they clearly indicate that the data do allow for meaningful inference on ν, even with this quite moderate sample size. Inevitably, there is some influence from the prior (which has a mode at 0.1 and a mean of 0.36), but the posterior distributions assign generous mass to neighbourhoods of the values used to generate the data. The same applies for posterior inference on ω^2, presented in the right panel of Figure 1 for data sets generated with $\nu = 2$, $\theta_1 = \theta_2 = 0.5$ and ω^2 ranging from 0.01 to 2.

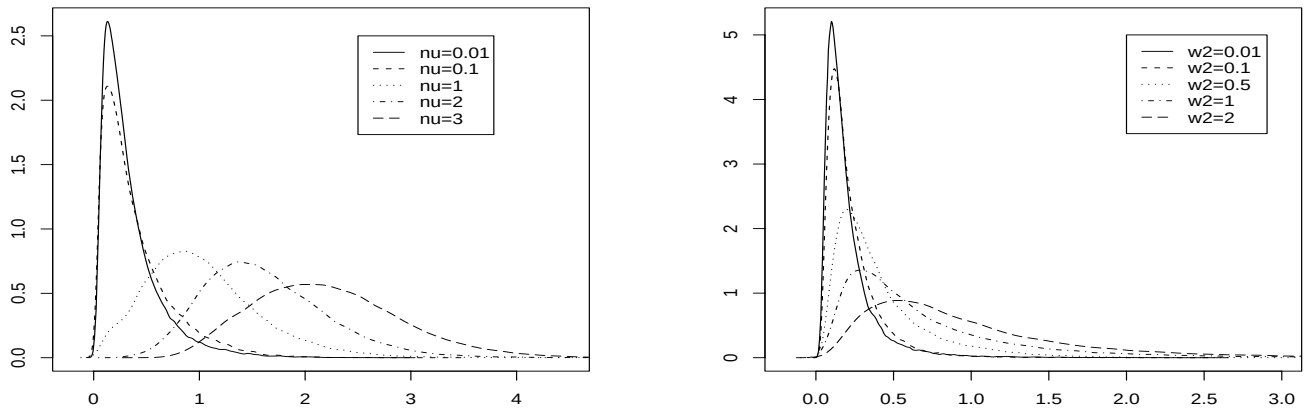

Figure 1: Simulated GLG data: Posterior distributions for ν (left panel) and ω^2 (right panel). The legends indicate the values of ν and ω^2 used to generate the data.

We have also experimented with different values of $\boldsymbol{\theta}$. The data are informative on $\boldsymbol{\theta}$, but sometimes it is hard to separately identify θ_1 and θ_2. This is partly due to a negative correlation in the posterior between θ_1 and θ_2, which leads to a correlation coefficient of -0.60 and is clearly illustrated by the scatter plot of draws (thinned to one tenth of the recorded draws for ease of presentation) in the left panel of Figure 2. This is based on the GLG model fitted to data generated from the GLG model with $\nu = 2$,

$\theta_1 = \theta_2 = 0.5$ and $\omega^2 = 0.1$, using the benchmark prior. In line with the discussion in Stein (1999) as mentioned in Section 3, changing the parameterisation to $(\rho = 2\theta_1\sqrt{\theta_2}, \theta_2)$ somewhat reduces the posterior correlation (to -0.50), as illustrated in the right panel of Figure 2.

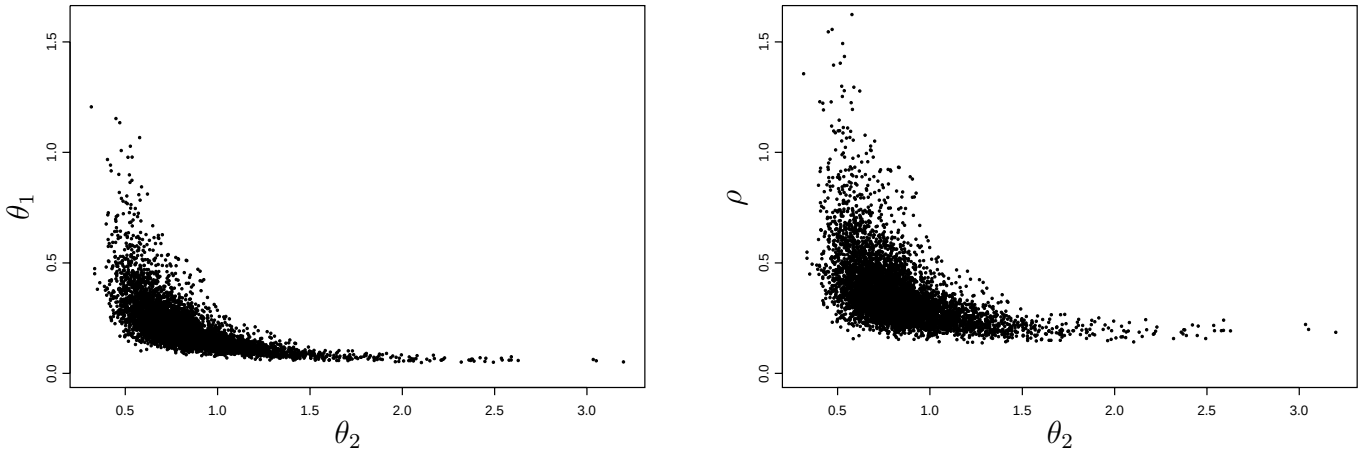

Figure 2: Simulated GLG data: Scatter plots of thinned posterior draws of θ_1 versus θ_2 (left panel) and ρ versus θ_2 (right panel).

Finally, we investigated the potential of the GLG model to identify and accommodate observations that would be outliers for the Gaussian model. We simulated $n = 100$ data points as above, but now from the Gaussian model with $\beta_1 = 10, \sigma^2 = 1, \omega^2 = 0.1$, and $\theta_1 = \theta_2 = 0.5$. This leads to a Bayes factor of the GLG model versus the Gaussian model of about three to one, indicating that even Gaussian data can lead to slightly more support for the GLG model. We then select two observations at random, observations 3 and 38, and we add one (equal to the standard deviation of the spatial process) to observation 3 and subtract two from observation 38. This is the situation we will denote as "moderate outliers". We make the identification slightly more challenging by not choosing the outlying observations close together, so that the model needs to identify two separate (small) areas of inflated variance. To create a more extreme case ("strong outliers") we add two to observation 3 and subtract three from observation 38. With moderate outliers the Bayes factor in favour of the GLG model increases to $4.8 \cdot 10^9$, indicating overwhelming support for the GLG model, and for the case of strong outliers, the support for the GLG model is even higher with a Bayes factor of $8.2 \cdot 10^{18}$. The posterior mean (standard deviation) of ν is $1.12(0.33)$ for the moderate outliers and $1.29(0.18)$ for the more extreme case. In addition, the actual outlying observations are easily identified as the ones with smallest posterior mean values of λ_i: even with moderate outliers the posterior mean of λ_i for observation 3 (the least extreme outlier) is less than half of the third smallest posterior mean. Table 3 presents the Bayes factors in favour of $\lambda_i = 1$ for observations 3 and 38, along with two other, randomly chosen (unperturbed) observations. We also

present separately the results for (10) and (11), which illustrates that the use of (10) alone (which is simpler to obtain) without the correction factor in (11) would underestimate the evidence against $\lambda_i = 1$ for the outliers, especially the more extreme ones. Clearly, in both cases the Bayes factors strongly indicate that observations 3 and 38 are outlying observations, whereas no such evidence occurs for the other two observations, in line with the way we have perturbed the data.

	moderate outliers			strong outliers		
obs.#	(10)	(11)	BF for $\lambda_i = 1$	(10)	(11)	BF for $\lambda_i = 1$
3	0.011	0.91	0.010	0.000	0.43	0.000
16	1.10	0.97	1.06	0.90	1.03	0.93
23	0.89	0.96	0.86	0.95	1.03	0.98
38	0.000	0.61	0.000	0.000	0.048	0.000

Table 3: Simulated Gaussian data with outliers: Bayes factors in favour of $\lambda_i = 1$ for selected observations. The entries 0.000 indicate values less than 0.0005.

6 Applications

In this section, we apply our GLG model to two data sets and compare the results with those obtained from the Gaussian model in (1). For both applications, we choose a quadratic form for the trend or mean function $\mathbf{f}(\mathbf{x})$, where $\mathbf{x}$ lies in a subset of $\Re^2$. As before, our results will be based on every 5th draw from an MCMC chain of length 2,000,000 with a burn-in of 500,000, which was more than sufficient for convergence. All results in this section are computed under the benchmark prior from Section 3.

6.1 Topographic Data

We use the same data on 52 topographic elevations as in Subsection 5.1 and now present some results derived with the benchmark prior based on (20). In particular, we will contrast results from the Gaussian model with those obtained using the non-Gaussian GLG model.

Table 4 presents some posterior results for the elements of $\boldsymbol{\beta}$, as defined through (26). Clearly, the

Model	β_1	β_2	β_3	β_4	β_5	β_6
Gaussian	8.10(0.16)	-0.17(0.17)	0.66(0.26)	-0.53(0.18)	0.05(0.21)	-0.04(0.28)
GLG	7.89(0.36)	-0.18(0.20)	0.78(0.23)	-0.49(0.23)	0.08(0.26)	0.001(0.25)

Table 4: Topographic data: Posterior means (standard deviation) of the trend parameters.

inference on the mean surface is similar for both models. Some of the regressors seem to have little

effect, but we will not specifically consider the specification of the mean function in this example, for comparability with other papers where a full quadratic trend was used. A summary of the posterior inference on the other parameters in the models is provided in Table 5. Besides the raw parameters, we also present two quantities that are more directly comparable between the models: since the marginal sampling variance of z_i is $\text{var}(z_i) = \sigma^2 \exp(\nu) + \tau^2$ under the GLG model, we can not directly compare *e.g.* σ, but we can compare $\sigma^2 \exp(\nu)$ as the variance of the spatially correlated process (with $\nu = 0$ for the Gaussian model). Thus, $\tau/[\sigma \exp(\nu/2)]$ is the relative standard deviation of the process inducing the nugget effect. Clearly, the Gaussian model attributes a larger fraction of the sampling variance to the nugget effect. We also present moments for the alternative range parameter $\rho = 2\theta_1\sqrt{\theta_2}$. From the latter and the results for $\boldsymbol{\theta}$ it seems that the Gaussian model finds slightly less spatial correlation than the GLG model and makes the field a bit smoother. As the Gaussian model finds it harder to deal with extreme or "outlying" observations, it puts relatively more emphasis on measurement error in explaining the data than on the smoothly varying field. That will tend to increase prediction uncertainty, as illustrated later in the context of the next application.

	Gaussian	GLG
σ	0.34 (0.08)	0.28 (0.11)
ω^2	0.25 (0.16)	0.33 (0.37)
θ_1	0.80 (0.62)	1.11 (1.16)
θ_2	3.13 (2.15)	2.77 (1.93)
ρ	2.26 (0.90)	2.85 (1.83)
ν	0 (0)	0.72 (0.54)
$\sigma^2 \exp(\nu)$	0.12 (0.06)	0.18 (0.20)
$\tau/[\sigma \exp(\nu/2)]$	0.48 (0.14)	0.36 (0.14)

Table 5: Topographic data: Posterior means (standard deviation) for some non-trend parameters.

In comparison with the results of Berger *et al.* (2001), who use a Gaussian model without nugget effect and with a fixed value for θ_2 under a reference prior, we find evidence of more smoothness (they try $\theta_2 = 0.5, 1$ and 1.5 and assign most posterior mass to $\theta_2 = 1$), and less spatial correlation. In this example, the main cause for the differences is not the non-Gaussian modelling (as they also appear for our Gaussian model), but rather the incorporation of the nugget effect and, to a lesser extent, the difference in prior (mostly on $\boldsymbol{\theta}$). Results without a nugget effect (for either Gaussian or GLG model) are much closer to those of Berger *et al.* (2001). However, the data favour the presence of a nugget effect: for the Gaussian model, the Bayes factor in favour of the model with nugget effect is 5.6, and it is 14 for the GLG model.

Figure 3 shows the posterior densities for some of the parameters, overplotted with the corresponding prior density. We also find that θ_1 and θ_2 are negatively correlated in the GLG posterior with a correlation

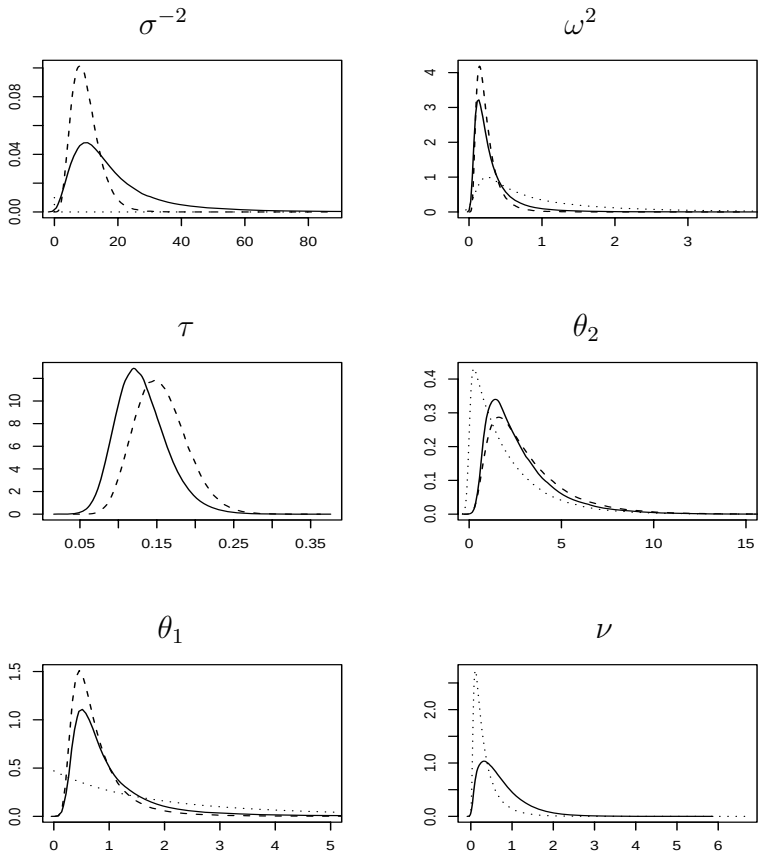

Figure 3: Topographic data: Posterior densities for non-trend parameters for the Gaussian model (dashed line) and the GLG model (solid line). Dotted lines indicate the prior densities.

coefficient of -0.46. From the evidence in Stein (1999) we would expect this correlation to be weaker when we take ρ as the range parameter. The posterior correlation between ρ and θ_2 is indeed a bit smaller, and amounts to -0.36. In the Gaussian case, the correlation coefficient of θ_1 and θ_2 is -0.52, while $\text{corr}[\rho, \theta_2|\mathbf{z}] = -0.33$. Clearly, the posterior distribution on ν is not very concentrated, but there is considerable mass away from zero, suggesting some support for the GLG model.

As a formal comparison of both models, we compute Bayes factors. We find that the Bayes factor in favour of the GLG model is 350, which amounts to considerable data support for this non-Gaussian model. In order to identify the main observations that drive this result, we have singled out the observations with the smallest mean values for λ_i, and we present in Table 6 the Bayes factors in favour of $\lambda_i = 1$. It is not surprising that the observations with smallest mean λ_i also lead to the lowest Bayes factors. There is substantial evidence that we need the GLG model to accommodate some of the observations, especially observations 48 and 37. The latter observations could safely be labelled as "outliers" in the Gaussian model. All observations mentioned in Table 6 are clustered together, so we have really identified one area of relatively large variance. Again, we find that the use of (10) alone without the correction factor in (11) would underestimate the evidence against $\lambda_i = 1$, especially for the more extreme

outliers. It is worth noting that over a very wide variety of prior specifications, based either on (19) or (20), the same observations were found to correspond to the smallest posterior mean values of λ_i.

obs.#	$\mathrm{E}[\lambda_i\|\mathbf{z}]$	S.Dev.$[\lambda_i\|\mathbf{z}]$	(10)	(11)	BF for $\lambda_i = 1$
37	0.42	0.32	0.74	0.26	0.19
47	0.53	0.42	1.05	0.59	0.62
48	0.30	0.25	0.25	0.24	0.061
49	0.47	0.37	0.94	0.48	0.44

Table 6: Topographic data: Bayes factors in favour of $\lambda_i = 1$ for selected observations.

Figure 4 presents the predictive mean and standard deviation surfaces for the GLG case. These are computed from the predictive in Section 2.2.3 over a regular grid of 20×20 points covering the observed region. The peak in predictive uncertainty corresponds to the area of inflated variance which contains the outlying observations in Table 6.

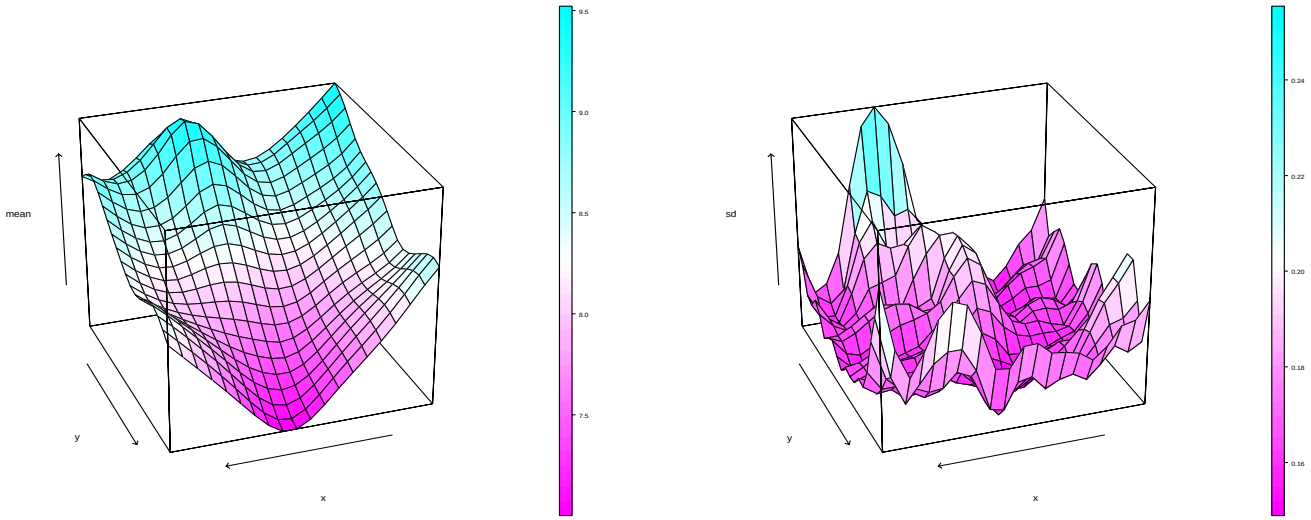

Figure 4: Topographic data: Predictive mean surface and standard deviation.

6.2 Temperature Data

The second dataset refers to the maximum temperatures recorded in an unusually hot week in May 2001 in 63 locations within the Spanish Basque country. As this region is quite mountainous (with

the altitude of the monitoring stations in between 16 and 1188 meters), altitude is added as an extra explanatory variable in the mean function (corresponding to regression coefficient β_7). Table 7 presents some posterior results for $\boldsymbol{\beta}$, where we have used the benchmark prior but with prior (18) for $\boldsymbol{\beta}$. Results with the full benchmark prior (*i.e.* using (17)) are virtually identical. Clearly, the difference between the

Model	β_1	β_2	β_3	β_4	β_5	β_6	β_7
Gaussian	3.19(0.22)	-0.20(0.23)	0.19(0.31)	-0.20(0.23)	0.37(0.45)	-0.24(0.28)	-0.40(0.18)
GLG	3.23(0.06)	-0.08(0.11)	0.12(0.13)	-0.19(0.09)	0.09(0.24)	-0.17(0.14)	-0.42(0.07)

Table 7: Temperature data: Posterior means (standard deviation) of the trend parameters.

two models is now larger than in the previous example. In particular, the Gaussian model tends to higher absolute values for β_2 and β_5 and the inference on $\boldsymbol{\beta}$ is generally a lot less concentrated for this model. The effect of altitude is clear for both models and higher altitude tends to lower the mean temperature reading, as expected. In addition, we note that the inclusion of some of the components in the trend function is perhaps questionable. Since we have used prior (18), we can immediately compute Bayes factors (through Savage-Dickey density ratios) of the model with $\beta_j = 0, j = 1, \ldots, k$ versus the full model. The largest Bayes factors for both the Gaussian and the GLG model are for β_3 (Bayes factors of 13.9 and 23.6, respectively), which is evidence that the mean surface does not really vary in the square of the Easting coordinate (x_1^2). Thus, the subsequent results are obtained for the specification with $\beta_3 = 0$ and using the benchmark prior with (17). Posterior results for the remaining elements of $\boldsymbol{\beta}$ are very close to those in Table 7.

	Gaussian	GLG
σ	0.32 (0.11)	0.09 (0.03)
ω^2	1.22 (0.79)	1.27 (1.12)
τ	0.31 (0.06)	0.08 (0.02)
θ_1	5.71 (10.33)	4.02 (12.70)
θ_2	1.87 (2.03)	0.61 (0.98)
ρ	8.64 (8.20)	2.35 (2.97)
ν	0 (0)	2.51 (0.76)
$\sigma^2 \exp(\nu)$	0.11 (0.09)	0.12 (0.15)
$\tau/[\sigma \exp(\nu/2)]$	1.05 (0.35)	0.30 (0.12)

Table 8: Temperature data: Posterior means (standard deviation) for some non-trend parameters.

Posterior inference on the other parameters in the models is presented in Table 8. Even more so than in the previous example, the Gaussian model assigns a larger importance to the nugget effect (see the difference in $\tau/[\sigma \exp(\nu/2)]$), while making the surface a lot smoother than the GLG model. In order to accommodate the outlying observations (see later), the Gaussian model needs to dramatically

increase the values of both σ and τ. Since most of the posterior mass for ν is well away from zero, it is not surprising that the evidence in favour of the GLG model is very strong indeed. In particular, the Bayes factor in favour of the GLG model is $3.4 \cdot 10^{20}$, a lot of which is attributable to three very extreme observations: observations 20, 36 and 40, which are all close together. Table 9 presents the Bayes factors in favour of $\lambda_i = 1$ for the four observations with smallest mean λ_i. Clearly, all observations listed in

obs.#	$E[\lambda_i \mid \mathbf{z}]$	S.Dev.$[\lambda_i \mid \mathbf{z}]$	(10)	(11)	BF for $\lambda_i = 1$
20	0.020	0.024	0.000	0.74	0.000
36	0.015	0.020	0.000	0.79	0.000
40	0.016	0.020	0.000	0.62	0.000
41	0.059	0.085	0.006	0.57	0.004

Table 9: Temperature data: Bayes factors in favour of $\lambda_i = 1$ for selected observations. The entries 0.000 indicate values less than 0.0005.

Table 9 are outliers, especially the first three (with smallest $E[\lambda_i \mid \mathbf{z}]$). This indicates two regions with inflated variance, the region covering observations 20, 36 and 40 (in the centre ofÁlava, the southernmost province of the Basque Country), and the area around observation 41 (Andoain, in the North-East). Again, the correction factors in (11) differ appreciably from one, illustrating the need for their inclusion. Interestingly, if instead we use the temperatures recorded in a particularly cold week in December 2001, we find that the Gaussian model performs slightly better than the GLG model with a Bayes factor around 10, both for minimum and maximum temperatures.

Figure 5 displays the predictive densities, computed as in Section 2.2.3 for five unobserved locations, ranging in altitude from 53 to 556 meters. The GLG model leads to heavier extreme tails than the Gaussian model as a consequence of the scale mixing. Nevertheless, in the (relevant) central mass of the distribution, the GLG predictives clearly are more concentrated than the Gaussian ones, illustrating that the added uncertainty due to the scale mixing is more than offset by changes in the inference on other aspects of the model. In particular, the nugget effect is much less important for the non-Gaussian model. From (14) it is clear that the predictive standard deviation is bounded from below by τ (in order to interpret the numbers in Table 8 in terms of observables measured in degrees centigrade, we need to multiply τ by a factor 10, due to scaling of the data). Clearly, a lot of the predictive uncertainty in the Gaussian case is due to the nugget effect.

Finally, Figure 6 contrasts the empirical semivariogram with its model-based theoretical counterpart. We use two different empirical semivariograms: the classical method of moments estimator of Matheron (1962) and a robust estimator introduced by Cressie and Hawkins (1980), based on fourth roots of squared differences and given by (2.4.12) in Cressie (1993, p. 75). These are calculated for the residuals after OLS estimation of the trend function. The OLS results for $\boldsymbol{\beta}$ are very close to the posterior means in Table 7. The posterior median of the semivariogram induced by the models is presented, along with the 2.5th and 97.5th percentiles. Given the stationarity of $Z(\mathbf{x}) - \mathbf{f}'(\mathbf{x})\boldsymbol{\beta}$, the fitted semivariogram as a function of distance d is simply given by $\sigma^2 \exp(\nu) + \tau^2 - \mathrm{cov}_Z(d)$ with $\mathrm{cov}_Z(d)$ as in (8). The substantial difference between the classical and the robust empirical semivariograms is consistent with the presence

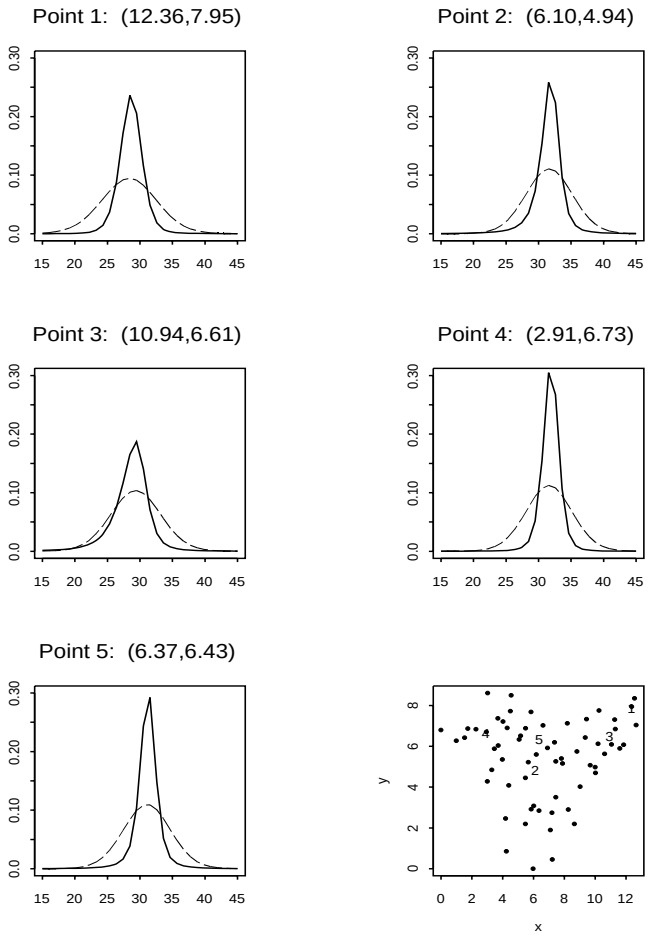

Figure 5: Temperature data: Predictive densities at five unobserved locations. The observables are measured in degrees centigrade, and the elevations at the predicted sites range from 53 (point 2) to 556 (point 3) meters. Dashed line: Gaussian; solid line: GLG. The lower right panel indicates the locations of the observed sites by dots and the five unobserved sites by their respective numbers.

of extreme outliers. The semivariograms implied by the models are also very different. The large nugget effect of the Gaussian model is clear, and the 95% credible interval for this model totally misses the robust empirical estimate and even the classical estimate lies outside this interval for considerable ranges of distances. The GLG model does a lot better by capturing the empirical semivariograms within the credible bounds for virtually all distances. The credible intervals for the GLG model are wider, but the median is also substantially lower and the shape is different, especially close to zero.

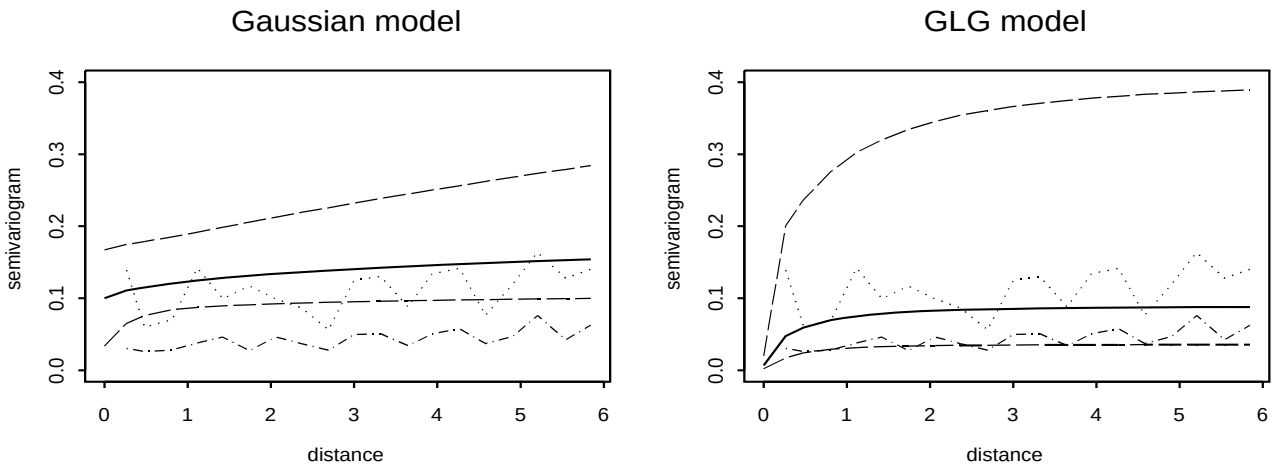

Figure 6: Temperature data: Empirical and model semivariograms. As empirical semivariograms we use the classical estimator indicated by the dotted line and the robust estimator of Cressie and Hawkins (1980) indicated by the dash-dotted line. The posterior median of the model-induced semivariogram is represented by a solid line, whereas the dashed lines indicate the 2.5th and 97.5th posterior percentiles.

7 Concluding Remarks

In this paper, we have proposed a new stochastic specification for geostatistical models in order to accommodate non-Gaussian processes, in particular processes with fat-tailed finite-dimensional distributions. Our non-Gaussian model is consistent with a stochastic process, and retains important properties such as mean square continuity and smoothness. We build our model through scale mixing a Gaussian process, and this immediately leads to a natural way to detect subregions with relatively large observational variance.

Our approach to inference is Bayesian, and we advocate the use of a carefully elicited prior. It is clear from our sensitivity analysis that the prior matters for posterior inference with the small datasets that one typically encounters in practice, so prior elicitation is important in this context. We develop an MCMC sampler and implement it for both posterior and predictive inference. The code (in Fortran) can deal with applications of practically relevant sample sizes and our empirical results illustrate both the practicality of inference with our model as well as the data support in both our applications for the non-Gaussian model proposed here (the GLG model). In the context of maximum temperature data, the GLG model is heavily favoured by the data over the Gaussian model and leads to much more concentrated predictive distributions for unobserved sites. This is a direct consequence of the fact that the non-Gaussian model is able to accommodate the outlying observations without increasing the nugget effect. For these data we also find that the GLG model provides a much better fit to a robust empirical semivariogram.

Whereas the GLG model frees up the tail behaviour, it still imposes symmetry and unimodality on the spatial process. These latter restrictions could *e.g.* be avoided by mixtures of non-zero mean GLG processes. Our approach could easily be embedded in a wider modelling framework, where the Gaussian

process could be replaced as a building block by the more general GLG process. For example, our GLG process could replace the underlying Gaussian process in modelling transformed data as in De Oliveira *et al.* (1997) or in the generalized linear modelling approach of Diggle *et al.* (1998). In addition, we could use a convolution of GLG processes rather than Gaussian processes to induce non-stationarity in an approach similar to Fuentes (2002).

Appendix A: Proofs

Proof of Proposition 1

We need to prove that the finite-dimensional distributions in (3) are consistent with a stochastic process. The Kolmogorov consistency conditions are checked as follows:

Symmetry under permutation: let $\pi_1, \ldots, \pi_n$ be any permutation of $1, \ldots, n$ and define $y_i = z_i - \mathbf{f}(\mathbf{x}_i)'\boldsymbol{\beta}$. From the representation in (4) it is clear that $P_{y_1,\ldots,y_n}(B_1, \ldots, B_n) = P_{y_{\pi_1},\ldots,y_{\pi_n}}(B_{\pi_1}, \ldots, B_{\pi_n})$, if and only if the same holds for the distribution of $\epsilon_1/\sqrt{\lambda_1}, \ldots, \epsilon_n/\sqrt{\lambda_n}$. Since this symmetry property is satisfied by $\boldsymbol{\epsilon}$ and $\boldsymbol{\epsilon}$ is independent of $\boldsymbol{\lambda}$, the condition $P_{\lambda_1,\ldots,\lambda_n}(A_1, \ldots, A_n) = P_{\lambda_{\pi_1},\ldots,\lambda_{\pi_n}}(A_{\pi_1}, \ldots, A_{\pi_n})$ is necessary and sufficient.

Dimensional consistency: consider defining the model for an $(n+1)$-dimensional vector given by $\mathbf{z}_{n+1} = (\mathbf{z}', z_{n+1})'$, and define

$$\boldsymbol{\Lambda}_{n+1} = \begin{pmatrix} \boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0}' & \lambda_{n+1} \end{pmatrix}, \mathbf{C}_{\boldsymbol{\theta}}^{n+1} = \begin{pmatrix} \mathbf{C}_{\boldsymbol{\theta}} & \mathbf{b}_{\boldsymbol{\theta}} \\ \mathbf{b}_{\boldsymbol{\theta}}' & c_{\boldsymbol{\theta}}^{n+1} \end{pmatrix}, \text{ and } \mathbf{X}_{n+1} = \begin{pmatrix} \mathbf{X} \\ \mathbf{f}'(\mathbf{x}_{n+1}) \end{pmatrix},$$

conformably with $\mathbf{z}_{n+1}$. Then (3) will imply

$$p(\mathbf{z}_{n+1}|\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\theta}, \boldsymbol{\Lambda}_{n+1}) = f_N^{n+1}\left(\mathbf{z}_{n+1}|\mathbf{X}_{n+1}\boldsymbol{\beta}, \sigma^2(\boldsymbol{\Lambda}_{n+1}^{-\frac{1}{2}}\mathbf{C}_{\boldsymbol{\theta}}^{n+1}\boldsymbol{\Lambda}_{n+1}^{-\frac{1}{2}}) + \tau^2\mathbf{I}_{n+1}\right),$$

which we can integrate with respect to $\boldsymbol{\Lambda}_{n+1}$ and marginalise with respect to z_{n+1} to obtain

$$\begin{aligned} p(\mathbf{z}|\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\theta}) = \; & \int_{\Re}\int_{\Re_+^{n+1}} p(\mathbf{z}_{n+1}|\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\theta}, \boldsymbol{\Lambda}_{n+1})\mathrm{d}P_{\lambda_1}\ldots\mathrm{d}P_{\lambda_{n+1}}\mathrm{d}z_{n+1} = \\ & \int_{\Re_+^n} f_N^n\left(\mathbf{z}|\mathbf{X}\boldsymbol{\beta}, \sigma^2(\boldsymbol{\Lambda}^{-\frac{1}{2}}\mathbf{C}_{\boldsymbol{\theta}}\boldsymbol{\Lambda}^{-\frac{1}{2}}) + \tau^2\mathbf{I}_n\right)\mathrm{d}P_{\lambda_1}\ldots\mathrm{d}P_{\lambda_n}, \end{aligned}$$

by a simple use of Fubini's Theorem and the fact that $P_{\lambda_{n+1}}$ is proper. Thus, the induced marginal sampling model for $\mathbf{z}$ is exactly the same as if we would have started from n locations, implying dimensional consistency.

Proof of Proposition 2

Using Leibniz identity we know that the derivative of order $2q$ of (8) will exist and be finite at zero if and only if the same applies for the $2q$th derivatives of (a) $C_{\boldsymbol{\theta}}(d)$ and of (b) $f(d) \equiv \exp\left(\nu\left\{1 + \frac{1}{4}[C_{\boldsymbol{\theta}}(d) - 1]\right\}\right)$. The conditions on (a) are exactly the conditions of the proposition. For (b) we note that the $2q$th derivative of $f(d)$ is equal to

$$\frac{\nu}{4}f(d)\{C_{\boldsymbol{\theta}}(d)^{(2q)} + g(\nu, d)\},$$

where $C_{\boldsymbol{\theta}}(d)^{(2q)}$ denotes the $2q$th derivative of $C_{\boldsymbol{\theta}}(d)$ and $g(\nu, d)$ exists and is finite at zero if and only if the same holds for the derivatives of $C_{\boldsymbol{\theta}}(d)$ up to and including order $2q-1$.

Appendix B: the Generalized Inverse Gaussian Distribution

GIG$(\lambda, \delta, \gamma)$ will denote a Generalized Inverse Gaussian (GIG) distribution with density function

$$f(x) = \frac{(\gamma/\delta)^{\lambda}}{2\mathcal{K}_{\lambda}(\delta\gamma)}\, x^{\lambda-1} \exp\left\{-\frac{1}{2}(\delta^2 x^{-1} + \gamma^2 x)\right\}, \qquad x \in \Re_+,$$

where $\mathcal{K}_{\lambda}$ is the modified Bessel function of the third kind and the parameters δ and γ take positive real values, whereas $\lambda \in \Re$. For $\lambda < 0$ we also allow $\gamma = 0$ (leading to the inverse gamma distribution) and for $\lambda > 0$ we can have $\delta = 0$ (which leads to the gamma distribution). Bibby and Sørensen (2003) provide a detailed discussion of this distribution (and its uses in financial modelling).

The mode of the GIG distribution is $\delta^2/\{2(1-\lambda)\}$ when $\gamma = 0$ and

$$\frac{\lambda - 1 + \sqrt{(\lambda-1)^2 + \delta^2\gamma^2}}{\gamma^2} \qquad \text{for } \gamma > 0.$$

Defining $\psi = \delta\gamma$, we can express the mean as

$$\frac{\delta}{\gamma}\frac{\mathcal{K}_{\lambda+1}(\psi)}{\mathcal{K}_{\lambda}(\psi)}$$

and the variance is

$$\left(\frac{\delta}{\gamma}\right)^2 \left(\frac{\mathcal{K}_{\lambda+2}(\psi)}{\mathcal{K}_{\lambda}(\psi)} - \frac{\mathcal{K}^2_{\lambda+1}(\psi)}{\mathcal{K}^2_{\lambda}(\psi)}\right).$$

References

Banerjee, S. and Gelfand, A.E. (2003). On smoothness properties of spatial processes. *Journal of Multivariate Analysis*, **84**, 85–100.

Berger, J.O., and Bernardo, J.M. (1992). On the development of reference priors (with discussion), in: J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, eds., *Bayesian Statistics 4.* Oxford University Press, Oxford, pp. 35-60.

Berger, J.O., De Oliveira, V. and Sansó, B. (2001). Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, **93**, 1361–1374.

Bernardo, J.M., and Smith, A.F.M. (1994), *Bayesian Theory.* Wiley, Chichester.

Bibby, B.M. and Sørensen, M. (2003). Hyperbolic processes in Finance, in S.T. Rachev (ed.): *Handbook of Heavy Tailed Distributions in Finance*, Elsevier, New York, pp. 211–248.

Cressie, N. (1993). *Statistics for Spatial Data.* John Wiley & Sons, New York.

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

De Oliveira, V. and Ecker, M.D. (2002). Bayesian hot spot detection in the presence of a spatial trend: Application to total nitrogen concentration in Chesapeake Bay. *Environmetrics*, **13**, 85–101.

De Oliveira, V., Kedem, B. and Short, D.A. (1997). Bayesian prediction of transformed Gaussian random fields. *Journal of the American Statistical Association*, **92**, 1422–1433.

Davis, J.C. (1973). *Statistics and Data Analysis in Geology*. Wiley, New York.

Diggle, P.J. Tawn, J.A. and Moyeed, R.A. (1998). Model-based geostatistics (with discussion). *Applied Statistics*, **47**, 299–326.

Ecker, M.D. and Gelfand, A.E. (1997). Bayesian variogram modelling for an isotropic spatial process. *Journal of the Agricultural, Biological and Environmental Statistics*, **2**, 347–369.

Fang, K.-T., Kotz, S. and Ng, K.-W. (1990) *Symmetric Multivariate and Related Distributions*, Chapman and Hall, London.

Fernández, C., Osiewalski, J. and Steel, M.F.J. (1995). Modelling and inference with υ-spherical distributions, *Journal of the American Statistical Association*, **90**, 1331–1340.

Fernández, C., Osiewalski, J. and Steel, M.F.J. (1997). Classical and Bayesian inference robustness in multivariate regression models, *Journal of the American Statistical Association*, **92**, 1434–1444.

Fuentes, M. (2002). Spectral methods for nonstationary spatial processes, *Biometrika*, **89**, 197–210.

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

Handcock, M.S. and Wallis, J.R. (1994). An approach to statistical-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association*, **84**, 368–390.

Kass, R.E. and Raftery, A.E. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Kelker, D. (1970) Distribution theory of spherical distributions and a location-scale parameter generalization, *Sankhya*, A, **32**, 419–430.

Kim, H.-M. and Mallick, B.K. (2003). A note on Bayesian spatial prediction using the elliptical distribution. *Statistics and Probability Letters*, **64**, 271–276.

Kitanidis, P.K. (1986) Parameter uncertainty in estimation of spatial functions: Bayesian analysis, *Water Resources Research*, **22**, 499–507.

Matheron, G. (1962). *Traité de Géostatistique Appliqueé, Tome I*. Mémoires du Bureau de Recherches Géologiques et Minières, No. 14, Editions Technip, Paris.

Neath, A.A. and Samaniego, F.J. (1997). On the efficacy of Bayesian inference for nonidentifiable models. *American Statistician*, **51**, 225–232.

Newton, M.A. and Raftery, A.E. (1994). Approximate Bayesian inference by the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Association, Series B*, **3**, 3–48.

Paulo, R. (2005). Default priors for Gaussian processes. *Annals of Statistics*, forthcoming.

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

Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors by using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, **90**, 614–618.

Wakefield, J.C., Smith, A.F.M, Racine-Poon, A. and Gelfand, A.E. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, **43**, 201–221.