On Spatial Skew-Gaussian Processes and Applications

Hao Zhang

Department of Statistics

Purdue University

West Lafayette, IN 47907

U.S.A

Phone: (765)496-9548

zhanghao@purdue.edu

Abdel El-Shaarawi

National Water Research Institute

Environment Canada

867 Lakeshore Road

Burlington, ON L7R 4A6

Canada

El-Shaarawi@squ.edu.om

On Spatial Skew-Gaussian Processes and Applications

October 17, 2008

Abstract

In many applications, observed spatial variables have skewed distributions. It is often of interest to model the shape of the skewed marginal distribution as well as the spatial correlations. We propose a class of stationary processes that have skewed marginal distributions. The covariance function of the process can be given explicitly. We study maximum likelihood inference through a Monte Carlo EM algorithm, and develop a method for the minimum mean-square error prediction. We also present two applications of the process.

Keywords: EM algorithm, Matérn covariogram, Skew-normal distribution, Skew-Gaussian process, Slice sampling

1 Introduction

In many applications, the observed spatial variables are known to have skewed distributions. Although spatial correlation structure always remains to be an interesting modeling aspect, it is of interest to also model the skewed marginal distributions. Frequently in environmental, hydrological and ecological studies, the shape of the distribution is of primary interest. On the other hand, when the process is non-Gaussian, linear prediction such as kriging generally may be outperformed by the minimum mean-square error (MMSE) prediction. The latter prediction requires the full distribution of the process to be modeled.

There are very few models for stationary processes that have skewed marginal distributions and have a covariance function of a manageable parametric form. However, a relatively large number of models have been developed in the literature for univariate and multivariate skewed distributions. Genton (2004) describes inferences about many of the models, one of which is the skew-normal distributions. The first systematic treatment of univariate skew-normal distributions was given by Azzalini (1985) and the multivariate skew-normal distribution was introduced by Azzalini and Dalla Valle (1996) and studied by Azzalini and Capitanio (1999). Recently, Ferreira and Steel (2006) proposed a new approach to the construction of univariate skewed distributions. However, it is not always obvious how to extend a class of univariate skewed distribution to multivariate distributions with skewed marginals.

When spatial data are collected at, say, n sampling sites and there is evidence that they have skewed distributions, one may naturally seek an existing multivariate distribution to fit the data. However, this approach has two obvious problems. First, some existing multivariate skewed distributions are better used when multiple multivariate samples are available. However, a single sample is typically available in spatial statistics and this sample must contain the information about skewness of the marginal distributions. As will be explained in next section, a single spatial sample modeled by the multivariate skew-normal distribution does not necessarily have the information about the skewness. Second, for predicting $Y(\boldsymbol{s})$ at a location $\boldsymbol{s}$ given the observed $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \cdots, Y(\boldsymbol{s}_n))'$, the conditional distribution of $Y(\boldsymbol{s})$ given $\boldsymbol{Y}$ is needed for MMSE prediction, which means that the $(n+1)$ dimensional distribution of $(\boldsymbol{Y}', Y(\boldsymbol{s}))$ is needed. A common approach is to assume that the observation $\boldsymbol{Y}$ is a partial realization of an underlying spatial process and model the process directly. De Oliveira et al. (1997) developed Bayesian transformed Gaussian model based on the Box-Cox transformations. Recently, Palacios and Steel (2006) proposed Gaussian-log-Gaussian (GLG) model by using scale mixing of Gaussian processes. The GLG model has heavier tails than Gaussian models but sill has a symmetric distribution.

The objective of this paper is to define a class of spatial stationary processes that have skewed marginal distributions, and study likelihood-based inferences. The process has a skew-normal marginal distribution and all finite distributions are completely determined by the mean and parameters in covariance. The observed spatial data are assumed to be a partial realization of the process. We call such a process a skew-Gaussian process. It includes the stationary Gaussian processes as a special case. The skew-Gaussian process defined in this work differs from that defined in Kim and Mallick (2004) that bears a similar name, as will be discussed in the next section.

The paper is organized as follows. In Section 2, we define the stationary skew-Gaussian process and provide its covariance function. In Sections 3 and 4 we consider maximum likelihood estimation of parameters and optimal prediction. A Markov chain Monte Carlo method is employed for both estimation and prediction. An example of application of skew-Gaussian process is presented in Section 5 and possible extensions of skew-Gaussian process are discussed in the last section.

2 Stationary Skew-Gaussian Processes

We first provide a brief review of skew-normal distribution. Let X_1 and X_2 be two i.i.d standard normal random variables. For any $\delta \in [-1, 1]$, the distribution of

$$Z = \delta|X_1| + (1 - \delta^2)^{0.5} X_2 \tag{1}$$

is called a skew-normal distribution. The distribution of Z is right-skewed if $\delta > 0$, left-skewed if $\delta < 0$ and is standard normal if $\delta = 0$. Its probability density function is $2\phi(z)\Phi(\alpha z)$ where $\alpha = \delta/(1 - \delta^2)^{0.5}$, and $\phi(z)$ and $\Phi(z)$ are the pdf and cdf of the standard normal distribution. Z^2 has a chi-square distribution with 1 degree of freedom, which is a property shared by the standard normal distribution.

A multivariate extension of (1) is given by Azzalini and Dalla Valle (1996). Consider a k-dimensional normal variable $\mathbf{X} = (X_1, \cdots, X_k)'$ with standardized marginals, independent of $X_0 \sim N(0, 1)$. For $\delta_j \in [-1, 1]$, $j = 1, \cdots, k$, define

$$Z_j = \delta_j|X_0| + (1 - \delta_j^2)^{0.5} X_j. \tag{2}$$

Then the joint distribution of $\boldsymbol{Z} = (Z_1, \cdots, Z_k)'$ is called a multivariate skew-normal distribution and each marginal distribution is skew-normal.

We now consider extending of the skew-normal distribution to a stationary process $Z(\boldsymbol{s})$ for $\boldsymbol{s} \in R^d$ for some $d > 0$ so that each $Z(\boldsymbol{s})$ has a skew-normal distribution and the process $Z(\boldsymbol{s})$ is second-order stationary. Let $X_0(\boldsymbol{s})$ be a stationary Gaussian process with standardized merginals, and $X(\boldsymbol{s})$ be another stationary Gaussian process also with standardized marginals. The two processes are independent and may have different covariance functions. Define

$$Z(\boldsymbol{s}) = \delta |X_0(\boldsymbol{s})| + (1 - \delta^2)^{0.5} X(\boldsymbol{s}) \tag{3}$$

Then this process $Z(\boldsymbol{s})$ is strictly stationary with skew-normal marginals. The covariance function of $Z(\boldsymbol{s})$ can be explicitly expressed in terms of those of $X_0(\boldsymbol{s})$ and $X(\boldsymbol{s})$, as will be given later in this section.

We make a few remarks before discussing more properties of the process $Z(\boldsymbol{s})$. Although each $Z(\boldsymbol{s})$ has a skew-normal distribution, the finite dimensional distribution of $Z(\boldsymbol{s}_1), \cdots, Z(\boldsymbol{s}_n)$ is not the multivariate skew-normal distribution of Azzalini and Dalla Valle (1996) because we let $X_0(\boldsymbol{s})$ vary with $\boldsymbol{s}$. Obviously, if any finite dimensional distribution of the skew-normal process is multivariate skew-normal, $X_0(\boldsymbol{s})$ does not depend on $\boldsymbol{s}$ and (2) becomes

$$Z(\boldsymbol{s}) = \delta |X_0| + (1 - \delta^2)^{0.5} X(\boldsymbol{s}),$$

where X_0 is a standard normal random variable, independent of the process $X(\boldsymbol{s})$.

However, there are drawbacks of such a skew-normal process. First, for any given realization of the process $Z(\boldsymbol{s})$, $Z(\boldsymbol{s})$ behaves just like a Gaussian process with mean $\delta|X_0|$. This imposes a problem on model diagnosing as well as for parameter estimation because in general only one realization is observed in practice. The process so defined is not ergodic in that any single realization contains only incomplete information about the distribution of the process. Second, the degree of skewness is mingled with the spatial correlation so that the stronger the skewness, the stronger the correlation. Specifically, for any $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$, if $X(\boldsymbol{s}_1)$ and $X(\boldsymbol{s}_2)$ are positively correlated, the correlation coefficient between $Z(\boldsymbol{s}_1)$ and

$Z(\boldsymbol{s}_2)$ is greater than

$$\frac{\delta^2 \mathrm{Var}(|X_0|)}{\mathrm{Var}(Z(\boldsymbol{s}))} = \frac{\delta^2(1-2/\pi)}{1-2\delta^2/\pi},$$

which is close to 1 if $Z(\boldsymbol{s})$ is extremely skewed (i.e., $\delta \approx 1$), no matter how far away the two points $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ are apart.

The process $Z(\boldsymbol{s})$ by (3) overcomes these drawbacks, which differs from the skew-Gaussian process proposed by Kim and Mallick (2004). In their work, the observed vector $\boldsymbol{Y}$ was modeled by a multivariate skew-normal distribution and was assumed to be a partial realization of a stationary process whose finite dimensional distributions are all multivariate skew-normal.

Location and scale transformations can be applied to (3) to yield any mean and variance of the observed variables. We therefore propose the following stationary process, which we refer to as skew-Gaussian process.

$$Y(\boldsymbol{s}) = m(\boldsymbol{s}) + \sigma_1|X_1(\boldsymbol{s})| + \sigma_2 X_2(\boldsymbol{s}) + \sigma_0\epsilon(\boldsymbol{s}), \tag{4}$$

where $\sigma_0 \geq 0$, $\sigma_2 \geq 0$ and σ_1 is real, $m(\boldsymbol{s})$ is constant depending only on the location $\boldsymbol{s}$, $X_i(\boldsymbol{s})$ $(i = 1, 2)$ is stationary Gaussian processes with standard marginals and covariogram $\rho_i(\boldsymbol{h})$, $\epsilon(\boldsymbol{s})$ is Gaussian white noise with mean 0 and variance 1. The three processes $X_i(\boldsymbol{s}),\ i = 1, 2$ and $\epsilon(\boldsymbol{s})$ are independent of each other. $m(\boldsymbol{s})$ can be modeled as a linear combination of some explanatory variables, $m(\boldsymbol{s}) = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{s})\beta_j$, for some observable explanatory variables $g_j(\boldsymbol{s})$.

It is obvious that $(Y(\boldsymbol{s}) - m(\boldsymbol{s}))/(\sigma_0^2 + \sigma_1^2 + \sigma_2^2)^{1/2}$ has a p.d.f $2\phi(y)\Phi(\alpha y)$ for $\alpha = \sigma_1/(\sigma_0^2 + \sigma_2)^{1/2}$. The mean is

$$EY(\boldsymbol{s}) = m(\boldsymbol{s}) + \sigma_1(2/\pi)^{1/2}.$$

The covariogram of the process $Y(\boldsymbol{s})$ can be easily given using the following fact, which will be proved in the Appendix.

If two random variables X and Y are jointly normal with standardized marginals and correlation coefficient $\rho \geq 0$, then

$$\mathrm{Cov}(|X|, |Y|) = \frac{2}{\pi}\left(\sqrt{1-\rho^2} + \rho \arcsin(\rho) - 1\right). \tag{5}$$

The covariogram of $Y(\boldsymbol{s})$ is therefore expressed in terms of those of the processes $X_0(\boldsymbol{s})$ and $X_1(\boldsymbol{s})$

$$\begin{aligned} C(\boldsymbol{h}) &= \frac{2\sigma_1^2}{\pi}\left(\sqrt{1-\rho_1(\boldsymbol{h})^2}+\rho_1(\boldsymbol{h})\arcsin(\rho_1(\boldsymbol{h}))-1\right) \\ &+ \sigma_2^2\rho_2(\boldsymbol{h})+\sigma_0^2 1_{\{\boldsymbol{h}=0\}}. \end{aligned} \tag{6}$$

3 Maximum Likelihood Estimation

In this section we consider the maximum likelihood estimation of the parameters in model (4). We assume that $Y(\boldsymbol{s})$ is observed at n sites $\boldsymbol{s}_i$, $i = 1, \cdots, n$ along with p explanatory variables $g_j(\boldsymbol{s}_i), j = 1, \cdots, p$. Furthermore, we assume that $m(\boldsymbol{s}) = \beta_0 + \sum_{j=1}^{p} g_j(\boldsymbol{s})\beta_j$ and the Gaussian process $X_i(\boldsymbol{s})$ $(i = 1, 2)$ has an isotropic Matérn correlation function $\rho(h, \nu_i, \phi_i)$ where

$$\rho(h,\nu,\phi)=\frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{3\nu^{1/2}h}{\phi}\right)^{\nu}\mathcal{K}_\nu\left(\frac{3\nu^{1/2}h}{\phi}\right),\ \nu>0,\phi>0,$$

and $\mathcal{K}_\nu$ is the modified Bessel function of order ν as discussed by Abramowitz and Stegun (1967). This parameterization is similar to that in Handcock and Wallis (1994), but the parameter ϕ here has a more intuitive interpretation that $\rho(h)$ is approximately 0.12 at $h = \phi$ regardless of ν.

Since the process is non-Gaussian, the likelihood function does not have a simple form though it can be explicitly expressed as a weighted sum of 2^n multivariate normal density functions. Direct maximization of the likelihood seems intractable. On the other hand, since we can treat the process $X(\boldsymbol{s})$ as latent variables, the EM algorithm seems to be a natural choice. We will describe an implementation of EM algorithm next.

3.1 EM Algorithm

We now consider and implement an EM algorithm. In practice, we usually know the sign of σ_1 based on whether the distribution of $Y(\boldsymbol{s})$ is right-skewed $(\sigma_1 > 0)$ or left-skewed $(\sigma_1 < 0)$.

We consider the case that $\sigma_1 > 0$ so that $Y(\boldsymbol{s})$ has a right-skewed distribution. If $Y(\boldsymbol{s})$ has a left skewed distribution, we would consider modelling $-Y(\boldsymbol{s})$.

To ease the maximization in the M-step of the EM algorithm, we rewrite the model as follows. Define

$$X(\boldsymbol{s}) = \sigma_1 X_1(\boldsymbol{s}),\ W(\boldsymbol{s}) = \sigma_2 X_2(\boldsymbol{s}) + \sigma_0 \epsilon(\boldsymbol{s}),$$

and write $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \cdots, Y(\boldsymbol{s}_n))'$, $|\mathbf{X}| = (|X(\boldsymbol{s}_1)|, \cdots, |X(\boldsymbol{s}_n)|)'$, $\boldsymbol{W} = (W(\boldsymbol{s}_1), \cdots, W(\boldsymbol{s}_n))'$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_p)'$, and denote by G the $n \times p$ matrix whose ith row is $(1, g_1(\boldsymbol{s}_i), \cdots, g_p(\boldsymbol{s}_i))'$. Then the model can be written as

$$\boldsymbol{Y} = G\boldsymbol{\beta} + |\mathbf{X}| + \boldsymbol{W},$$

where $\mathbf{X} \sim N(0, V_1)$, $\boldsymbol{W} \sim N(0, V_2)$. For simplicity, we write $\psi_k = (\nu_k, \phi_k)$, $k = 1, 2$, $\tau_i = \sigma_i^2, i = 0, 1, 2$. Then the covariance matrices V_i can be written as

$$V_1 = \tau_1 R_1,\ V_2 = \tau_2 R_2 + \tau_0 I,$$

where $R_k = R_k(\psi_k) = (\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|, \psi_k))_{i,j=1}^n,\ k = 1, 2.$

We treat $\mathbf{X}$ as unobservable latent variables. The complete-data log likelihood function for $(\mathbf{X}, \boldsymbol{Y})$ is

$$\log L_c(\theta) = \log f(\mathbf{X}, \sigma_1, \psi_1) + \log f(\boldsymbol{Y}|\mathbf{X}, \beta, \sigma_0, \sigma_2, \psi_2).$$

The EM algorithm runs iteratively. It starts with some initial estimate $\theta^{(0)}$. At the mth iteration, given estimate $\theta^{(m)}$, the new estimate $\theta^{(m+1)}$ maximizes the conditional expectation

$$Q(\theta|\theta^{(m)}) = E_{\theta^{(m)}}(\log Lc(\theta, \boldsymbol{Y}, \mathbf{X})|\boldsymbol{Y})$$

where the expectation is evaluated under $\theta^{(m)}$. This function can be written as

$$Q(\theta|\theta^{(m)}) = E_{\theta^{(m)}}(\log f(\mathbf{X}, \tau_1, \psi_1)|\boldsymbol{Y}) + E_{\theta^{(m)}}(\log f(\boldsymbol{Y}|\mathbf{X}, \boldsymbol{\beta}, \tau_0, \tau_2, \psi_2)|\boldsymbol{Y}) \tag{7}$$

The first term equals

$$-\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau_1 - \frac{1}{2}\log|R_1(\psi_1)| - \frac{1}{2\tau_1}E_{\theta^{(m)}}(\mathbf{X}'R_1^{-1}(\psi_1)\mathbf{X}|\boldsymbol{Y}).$$

Maximizing this function, we get

$$\psi_1^{(m+1)} = \mathrm{ArgMin}(n \log E_{\theta^{(m)}}(\mathbf{X}' R_1^{-1}(\psi_1)\mathbf{X}|\boldsymbol{Y}) + \log |R_1(\psi_1)|) \quad (8)$$

$$\tau_1^{(m+1)} = \frac{1}{n} E_{\theta^{(m)}}(\mathbf{X}' R_1^{-1}(\psi_1^{(m+1)})\mathbf{X}|\boldsymbol{Y}) \quad (9)$$

The other estimates $(\boldsymbol{\beta}^{(m+1)}, \tau_0^{(m+1)}, \tau_2^{(m+1)}, \psi_2^{(m+1)})'$ maximize the second term in (7)

$$E_{\theta^{(m)}}\{\log f(\boldsymbol{Y}|\mathbf{X}, \boldsymbol{\beta}, \tau_0, \tau_2, \psi_2)|\boldsymbol{Y})\}$$
$$= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|V_2| - \frac{1}{2}E_{\theta^{(m)}}\{(\boldsymbol{Y} - G\boldsymbol{\beta} - |\mathbf{X}|)' V_2^{-1}(\boldsymbol{Y} - G\boldsymbol{\beta} - |\mathbf{X}|)|\boldsymbol{Y}\}. \quad (10)$$

Clearly, these latter estimates are related as

$$\boldsymbol{\beta}^{(m+1)} = (G'V_{2,m+1}^{-1}G)^{-1}G'V_{2,m+1}^{-1}(\boldsymbol{Y} - E_{\theta^{(m)}}\{|X|\ |\boldsymbol{Y}\}) \quad (11)$$

where $V_{2,m+1}^{-1}$ is the inverse of matrix V_2 corresponding to the estimates $(\tau_0^{(m+1)}, \tau_2^{(m+1)}, \psi_2^{(m+1)})'$.

An iterative procedure such as the Newton-Raphson algorithm is necessary to maximize (10). The EM algorithm so implemented has the advantage that there are separate maximization (8) to (10) to update estimates.

The conditional expectations above cannot be evaluated in closed form but can be approximated by a Markov chain Monte Carlo method, which will be introduced in the next subsection. Such a Monte Carlo EM algorithm has been developed in literature for the analysis of correlated data (see, e.g., Wei and Tanner, 1990, McCulloch, 1997, Zhang, 2002).

3.2 Slice Sampling for Skew-Gaussian Processes

We now assume that the model parameters are known and consider generating a Markov chain $X^{(t)}$ so that for any function h

$$\lim_{T\to\infty}(1/T)\sum_{t=1}^{T} h(\mathbf{X}^{(t)}) = E(h(\mathbf{X})|\boldsymbol{Y}).$$

This chain then can be used to approximate the conditional expectations in (8) to (10). Metropolis-Hastings algorithm is an obvious choice here because the acceptance probability takes a simple form if the proposal distribution is chosen to be that of $\mathbf{X}$. However, our numerical studies show that the acceptance probability could be as low as 1/4500 for a Metropolis-Hastings algorithm so implemented. Low acceptance probabilities result in slow mixing of the Markov chain.

In this work, we will introduce auxiliary variables and employ the slice sampling method. Slice sampling is a technique of generating from an arbitrary distribution by introducing an auxiliary variable and sampling from two or more uniformly distributions (Neal, 2003). For our problem, it iterates as follows. To simplify notations, we assume that $\boldsymbol{\beta} = \mathbf{0}$ (only for the remaining of this section). Let U have the uniform distribution on the interval $[0, f(\boldsymbol{Y}|\mathbf{X})]$ conditional on $\mathbf{X}$ and $\boldsymbol{Y}$. Then the distribution of $\mathbf{X}$ conditional on U and $\boldsymbol{Y}$ is proportional to

$$f(\mathbf{X})1_{\{U<f(\boldsymbol{Y}|\mathbf{X})\}}.$$

Therefore, Gibbs sampler can be employed in theory to generate $\mathbf{X}$ given $\boldsymbol{Y}$ as follows. Start at some $\mathbf{X}^{(0)}$. Given $\mathbf{X}^{(t)}$, generate $U^{(t+1)} \sim U[0, f(\boldsymbol{Y}|\mathbf{X}^{(t)})]$, $\mathbf{X}^{(t+1)} \sim f(\mathbf{X})1_{\{U^{(t+1)}<f(\boldsymbol{Y}|\mathbf{X})\}}$ and iterate. However, to generate $\mathbf{X}$ from the constrained distribution is not trivial. The naive approach, which generates from $f(\mathbf{X})$ until it satisfies the constraint $f(\boldsymbol{Y}|\mathbf{X}) > U^{(t+1)}$, is neither efficient nor practical. Here, to generate from the constrained distribution $f(\mathbf{X})1_{\{U<f(\boldsymbol{Y}|\mathbf{X})\}}$, we use the idea of slice sampling method by introducing the second auxiliary variable. Given $\mathbf{X}^{(t)}$, generate $\widetilde{U} \sim U[0, f(\mathbf{X}^{(t)})]$, and generate $\mathbf{X}^{(t+1)}$ uniformly on the set

$$\{\mathbf{x} : f(\mathbf{x})1_{\{U^{(t+1)}<f(\boldsymbol{Y}|\mathbf{x})\}} > \widetilde{U}\} = \{\mathbf{x} : f(\mathbf{x}) > \widetilde{U}\} \cap \{\mathbf{x} : f(\boldsymbol{Y}|\mathbf{x}) > U^{(t+1)}\}. \qquad (12)$$

Note that $f(\mathbf{x}) > \widetilde{U}$ if and only if $\log f(\mathbf{x}) > \log(\widetilde{U}/f(\mathbf{X}^{(t)})) + \log(f(\mathbf{X}^{(t)}))$, and if and only if

$$\mathbf{x}'V_1^{-1}\mathbf{x} < -2\log(\widetilde{U}/f(\mathbf{X}^{(t)})) + \mathbf{X}^{(t)'}V_1^{-1}\mathbf{X}^{(t)} = r,$$

which is an n-dimensional oval. For any $\mathbf{x} = (x_1, \cdots, x_n)'$ inside the oval, define

$$I_i(\mathbf{x}, r) = \{x \in R : \mathbf{x}'V_1^{-1}\mathbf{x} < r \text{ if } \mathbf{x} = (x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_n)'\}.$$

Then $I_i(\mathbf{x}, r)$ contains all possible values of the ith coordinates in order for $\mathbf{x}$ to remain in the n-dimensional oval while the other $(i-1)$ coordinates are fixed. Clearly $I_i(\mathbf{x}, r)$ is a non-empty interval because $\mathbf{X}_i^{(t)} \in I_i(\mathbf{x}, r)$.

Similarly $f(\boldsymbol{Y}|\mathbf{x}) > U^{(t+1)}$ if and only if

$$(\boldsymbol{Y} - |\mathbf{x}|)'V_2^{-1}(\boldsymbol{Y} - |\mathbf{x}|) < -2\log(U^{t+1}/f(\boldsymbol{Y}|\mathbf{X}^{(t)})) + (\boldsymbol{Y} - |\mathbf{X}^{(t)}|)'V_2^{-1}(\boldsymbol{Y} - |\mathbf{X}^{(t)}|).$$

Denote the right hand side by a, and define for $i = 1, \cdots, n$,

$$\Delta_i(\mathbf{x}, a) = \{x \in R : (\boldsymbol{Y} - |\mathbf{x}|)'V_2^{-1}(\boldsymbol{Y} - |\mathbf{x}|) < a \text{ if } \mathbf{x} = (x_1, \cdots, x_{i-1}, x, x_{i+1}, \cdots, x_n)'\}.$$

Then $\Delta_i(\mathbf{x}, a)$ is symmetric about 0 and may be one interval or the union of two disjoint intervals.

Therefore for any point in the set (12) and any $i = 1, \ldots, n$, when all coordinates $x_j, j \neq i$ are fixed, x_i varies in one or two intervals that can be explicitly determined. Hence slice sampling can be easily applied again to generate uniformly on the n-dimensional set (12).

To summarize, the slice sampling iterates as follows to generate from $f(\mathbf{X}|\boldsymbol{Y})$. Given $\mathbf{X}^{(t)}$,

```
• Generate $\eta^{(t+1)}, \xi^{(t+1)}$ i.i.d Exp(1), and let
  $r = 2\eta^{(t+1)} + \mathbf{X}^{(t)\prime}V_1^{-1}\mathbf{X}^{(t)}$,    $a = 2\xi + (\boldsymbol{Y} - |\mathbf{X}^{(t)}|)'V_2^{-1}(\boldsymbol{Y} - |\mathbf{X}^{(t)}|)$,
• For(i in 1:n){
  Generate $X$ uniformly distributed on $\boldsymbol{I}_i(\mathbf{X}^{(t)}, r) \cap \Delta_i(\mathbf{X}^{(t)}, a)$ and substitute $X$
for the $i$th element:
  $\mathbf{X}^{(t+1)} = (\mathbf{X}_1^{(t+1)}, \cdots, \mathbf{X}_{i-1}^{(t+1)}, X, \mathbf{X}_{i+1}^{(t)}, \cdots, \mathbf{X}_n^{(t)})'$
  }
• Repeat
```

We note that for any i, the set $\boldsymbol{I}_i(\mathbf{X}^{(t)}, r) \cap \Delta_i(\mathbf{X}^{(t)}, a)$ is not empty because the ith element of $\mathbf{X}^{(t)}$ is in this set.

4 Prediction

In many applications, prediction of values at unsampled locations is necessary. In this section, we assume true parameters are known and predict $Y(\boldsymbol{s})$ at an unsampled location $\boldsymbol{s}$. We then carry out the prediction under the estimates of the parameters. This results in the so-called plug-in prediction. Note that the best linear unbiased prediction or kriging is straightforward because the the explicit expression of the covariogram of the skew-Gaussian process is given in (6). However, because the process is non-Gaussian, the optimal prediction that minimizes the mean squared error is non-linear. This optimal prediction is

$$E(Y(\boldsymbol{s})|\boldsymbol{Y}) = m(\boldsymbol{s}) + E(|X(\boldsymbol{s})|\;|\boldsymbol{Y}) + E(W(\boldsymbol{s})|\boldsymbol{Y}).$$

It cannot be evaluated in closed form but can be approximated using Monte Carlo samples. One trivial approach is to generate an MCMC sample from the conditional distribution of $Y(\boldsymbol{s})$ given $\boldsymbol{Y}$. For each prediction location, an MCMC sample would need to be generated. This approach is computationally cumbersome because a large number of prediction locations is usually used in practice. Next we present a technique that alleviates some of the computational burden. This technique requires generation of Monte Carlo samples only from the n-dimensional multivariate distribution of $\mathbf{X}$ conditional on $\boldsymbol{Y}$, regardless of how many prediction locations to be used. These MC samples are used to calculate $E(Y(\boldsymbol{s})|\boldsymbol{Y})$ for any prediction location $\boldsymbol{s}$. The same technique has been employed in Zhang (2002, 2003) for spatial generalized linear mixed models.

Our method is based on the following fundamental property

$$E(|X(\boldsymbol{s})|\;|\boldsymbol{Y}) = E[E(|X(\boldsymbol{s})|\;|\mathbf{X}, \boldsymbol{W})|\boldsymbol{Y}].$$

Because $X(\boldsymbol{s})$ and $\boldsymbol{W}$ and independent, we have

$$E(|X(\boldsymbol{s})|\;|\mathbf{X}, \boldsymbol{W}) = E(|X(\boldsymbol{s})|\;|\mathbf{X}).$$

Since conditional on $\mathbf{X}$, $X(\boldsymbol{s})$ is normal with conditional mean and variance

$$\mu = \boldsymbol{\lambda}_1' V_1^{-1}\mathbf{X},\; \sigma^2 = \tau_1 - \boldsymbol{\lambda}_1' V_1^{-1}\boldsymbol{\lambda}_1.$$

where $\boldsymbol{\lambda}_1 = \text{Cov}(\mathbf{X}, X(\boldsymbol{s}))$ and $V_1 = \tau_1 R_1(\psi_1)$, simple calculation yields that

$$E(|X(\boldsymbol{s})|\,|\mathbf{X}) = \sigma(2/\pi)^{1/2} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) + \mu\left(2\Phi\left(\frac{\mu}{\sigma}\right) - 1\right) \tag{13}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Then

$$E(|X(\boldsymbol{s})|\,|\boldsymbol{Y}) = E[E(|X(\boldsymbol{s})|\,|\mathbf{X})\boldsymbol{Y}] \approx (1/T)\sum_{t=1}^{M} E(|X(\boldsymbol{s})|\,|\mathbf{X}^{(t)})$$

where $\mathbf{X}^{(t)}$, $t = 1, \cdots, T$ are a Monte Carlo sample from the distribution of $\mathbf{X}$ conditional on $\boldsymbol{Y}$. These MC samples are generated in the MCEM algorithm and can be reused for prediction.

Similarly,

$$E(W(\boldsymbol{s})\,|\boldsymbol{Y}) = E[E(W(\boldsymbol{s})|\boldsymbol{W})\,|\boldsymbol{Y}]$$

and

$$E(W(\boldsymbol{s})|\boldsymbol{W}) = \boldsymbol{\lambda}_2' V_2^{-1}\boldsymbol{W}.$$

where $\boldsymbol{\lambda}_2 = \text{Cov}(\boldsymbol{W}, W(\boldsymbol{s}))$. Hence

$$E(W(\boldsymbol{s})\,|\boldsymbol{Y}) = \boldsymbol{\lambda}_2' V_2^{-1}\widetilde{\boldsymbol{W}},$$

where $\widetilde{\boldsymbol{W}} = E(\boldsymbol{W}|\boldsymbol{Y}) = \boldsymbol{Y} - G\boldsymbol{\beta} - E(|X||\boldsymbol{Y}) \approx \boldsymbol{Y} - G\boldsymbol{\beta} - (1/T)\sum_{t=1}^{T}|\mathbf{X}^{(t)}|$. Therefore, prediction of $W(\boldsymbol{s})$ is performed exactly as in simple kriging if we substitute the estimated $W(\boldsymbol{s}_i)$ for observed $W(\boldsymbol{s}_i)$.

5 An Example

In this section, we present two applications of the skew-Gaussian process. Depth integrated chlorophyll concentrations measured during the 1974 Lake Ontario Surveillance program are used to illustrate the methods of this paper. This Canadian program was established to monitor spatial and temporal changes in water quality of the Great Lakes. The 1974 sampling program consisted of 15 cruises that were conducted between April 16 and November 29. During each cruise, water samples were collected from a number of sites systematically covering the lake and the physical, biological and chemical characteristics of the samples were measured. Chlorophyll was selected because it is a measure of phytoplankton biomass and thus lake productivity. Excessive input of nutrients to the aquatic environment causes excessive productivity and leads to the deterioration of water quality. This was recognized as a major problem for Lakes Erie and Ontario. We used the sounding depth as a covariate since it tends to be negatively correlated with productivity. Shallow regions (near shore) are expected to be high in nutrients and off shore areas low in productivity and nutrients. Here we do not intend to provide a complete analysis for the entire datasets and instead choose to analyze, somewhat arbitrarily, two datasets of chlorophyll collected at different times of the year but at similar locations. Dataset 1 was collected on April 29, 30 and May 1 of 1974 at 40 locations, and dataset 2 was collected on September 30, October 2, 4 and 5 of 1974 at 42 locations. The two sets of observed locations were similar as shown in Figure 1.

[Figure 1 about here.]

[Figure 2 about here.]

One covariate that is also measured along with chlorophyll at each of the locations is the sounding depth. In general, chlorophyll decreases as the sounding depth increases. The linear correlation coefficient between the two variables is -0.7924 and -0.4714. For exploratory purpose, we run simple regression on both datasets and plot in Figure 2 the histogram of residuals and the semivariogram of residuals for each dataset. We see from the histograms that residuals have moderately right-skewed distributions, especially for dataset 1. The semivariograms indicate the existence of strong spatial correlation in dataset 2 and weaker spatial correlation in dataset 1. A nugget effect or measurement error exists in both datasets.

The exploratory data analysis suggests that the skew-Gaussian model is a reasonable choice for the two datasets. We therefore apply model (4) to both datasets. We use the sounding depth as the sole explanatory variable. For each dataset, we scale the distance so that the maximum distance is 1. We assume that $\nu_1 = \nu_2$.

To employ the Monte Carlo EM algorithm for parameter estimation, we apply the slice sampling method to generate 250,000 Monte Carlo samples after burn-in and keep every 50th one. The length of burn-in of the Markov chain generated by the slicing sampling is less than 100. The Monte Carlo sample obtained therefore has a size of 5,000. Our diagnostic analyses show that this sample size is sufficient to ensure convergence of the estimates of conditional expectations. The obtained estimates for dataset 1 are $(\hat{\beta}_0, \hat{\beta}_1) = (11.30, -0.0801)$, $\hat{\nu} = 0.25$ and $(\hat{\tau}_1, \hat{\phi}_1, \hat{\tau}_2, \hat{\phi}_2, \hat{\tau}_0) = (2.7233, 0.3997, 4.1475, 1.9461, 5.2661)$. For dataset 2, the estimates are $(\hat{\beta}_0, \hat{\beta}_1) = (6.6029, -0.0203)$, $\hat{\nu} = 0.5$, and $(\hat{\tau}_1, \hat{\phi}_1, \hat{\tau}_2, \hat{\phi}_2, \hat{\tau}_0) = (0.6765, 0.4048, 1.1579, 1.1631, 1.9944)$. For both datasets, profile likelihood for ν is rather flat, meaning that the data do not have sufficient information to acquire a more precise estimate of ν. It is known that even for stationary Gaussian processes, precise estimation of τ may be difficult. In general, adding sampling sites that are close to each other will reduce the variance of the estimators.

[Figure 3 about here.]

We plot in Figure 3 the histograms of the residuals, $e(\boldsymbol{s}_i) = |X(\boldsymbol{s}_i)| + W(\boldsymbol{s}_i)$, the fitted probability density functions, as well as the empirical semivariograms of the residuals and the fitted semivariograms. The densities fit the histograms reasonably well and the estimated semivariograms also seem to be reasonably close to the empirical ones.

We also note that dataset 1 has a more skewed distribution than dataset 2 as seen from Figure 2. For the skew-Gaussian model, the skewness increases with the ratio $\tau_1/(\tau_2 + \tau_0)$. The estimate of this ratio is 0.289 for dataset 1 and 0.214 for dataset 2.

The error term $e(\boldsymbol{s}) = |X(\boldsymbol{s})| + W(\boldsymbol{s})$ accounts for the spatial variation not accounted for by the sounding depth. To display the spatial variation in the error, we predict $e(\boldsymbol{s})$ at 1578 points evenly distributed in the sampling region. These points were obtained by dividing the sampling region into 2km×2km squares. The contour plots for the two datasets are given in Figure 4. There is a considerable spatial variation in $|X(\boldsymbol{s})| + W(\boldsymbol{s})$ for both datasets with the southwest having the lower values and northeast having the larger values.

[Figure 4 about here.]

For comparison purpose, we also apply the spatial regression with stationary normal error:

$$Y(\boldsymbol{s}_j) = \beta_0 + g(\boldsymbol{s}_j)\beta_1 + \epsilon(\boldsymbol{s}_j),\ j = 1, \cdots, n$$

where $g(\boldsymbol{s}_j)$ is the sounding depth at location $\boldsymbol{s}_j$ and $(\epsilon(\boldsymbol{s}_1), \cdots, \epsilon(\boldsymbol{s}_n))'$ is a partial realization of stationary Gaussian process with mean 0 and a Matérn covariogram that has a nugget effect. Hence the covariance matrix $\sigma_0^2 I_n + \sigma_1^2 R$, where R has the (i, j)th element $\rho(\|\boldsymbol{s}_i - \boldsymbol{s}_j\|, \nu, \phi)$. Maximum likelihood estimation of such model can be obtained through an iterative procedure such as the Newton-Raphson method given by Mardia and Marshall (1984). For dataset 1, we obtained the estimates $(\beta_0, \beta_1) = (12.2670,\ -0.0805)$, $\nu = 0.25$ and $(\sigma_0^2, \sigma_1^2, \phi) = (4.4505, 3.5724, 0.011)$. Because the estimate of the range parameter ϕ is so small, the Gaussian model reveals that the spatial correlation decays very rapidly. For dataset 2, we obtained the estimates $(\beta_0, \beta_1) = (7.0226, -0.0189)$, $\nu = 0.5$ and $(\sigma_0^2, \sigma_1^2, \phi) = (1.9294, 0.3006, 0.3954)$. The spatial correlation does not decay rapidly, which agrees with the empirical semivariogra. However, the estimated partial sill (0.3006) is quite small compared with the nugget effect 1.9294. The empirical semivariogram reveals a much larger partial sill. Therefore, the Gaussian model does not fit either model satisfactorily.

To compare the predictive performance of the two models, we predict $Y(\boldsymbol{s}_i)$ for each sampling site $\boldsymbol{s}_i$ using all $\boldsymbol{s}_j, j \neq i$ and compute the mean squared difference $\sum_{i=1}^{n}(Y(\boldsymbol{s}_i) - \hat{Y}(\boldsymbol{s}_i))^2/n$. For dataset 1, the Gaussian model yields a mean squared error 8.023 and the skew-Gaussian model 6.764. For dataset 2, the mean squared error is 2.179 for the Gaussian model and 1.873 for the skew-Gaussian model. Overall, we believe that the skew-Gaussian model outperforms the Gaussian model for both datasets.

We also applied the log transformation to both datasets and run linear regression on the transformed variable. However, exploratory analysis on the residuals convinced us that log transformation is not appropriate. Hence we do not compare the skew-Gaussian model with the lognormal model.

6 Discussion

The skew-Gaussian process we proposed in this work can be extended to model more skewed distributions. For example, we can replace $|X_1(\boldsymbol{s})|$ in model (4) by $|X_1(\boldsymbol{s})|^q$ where $q > 0$ is an additional parameter and can be estimated, or by an increasing function of $|X_1(\boldsymbol{s})|$. This leads to a wider class of stationary processes with skewed marginals. The EM algorithm and the slice sampling can be applied after slight modification. Prediction can be carried out using the same technique.

Acknowledgment

The research of Zhang was partially supported by the US National Science Foundation grants DMS-0405782 and DMS-0706835.

Appendix

We provide proof of equation (5). It is known that

$$E|X| = E|Y| = \sqrt{2/\pi}.$$

Hence we only need to show that

$$E|XY| = \frac{2}{\pi}\left(\sqrt{1-\rho^2} + \rho \arcsin(\rho)\right).$$

Define $\sigma = (1-\rho^2)^{0.5}$, $I(x,\rho) = E(|Y|1_{\{Y>0\}}|X = x)$. Then

$$E(|Y|1_{\{Y>0\}}|X = -x) = E(|Y|1_{\{Y>0\}}| - X = x) = I(x, -\rho).$$

and simple calculation yields

$$\begin{aligned} I(x,\rho) &= \int_0^\infty \frac{y}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y-\rho x)^2}{2\sigma^2}\right) dy \\ &= \rho x \Phi\left(\frac{\rho x}{\sigma}\right) + \sigma\phi\left(\frac{\rho x}{\sigma}\right) \end{aligned}$$

where $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution.

Using the symmetry of the joint distribution, we write

$$\begin{aligned} E|XY| &= 2E|XY|1_{\{X>0,Y>0\}} + 2E|XY|1_{\{X<0,Y>0\}} \\ &= 2\int_0^\infty x\phi(x)[I(x,\rho) + I(x,-\rho)]dx. \end{aligned}$$

By integration by part, we see the last equation equals

$$\begin{aligned} &-2\int_0^\infty [I(x,\rho) + I(x,-\rho)]d\phi(x) \\ &= \frac{2\sigma}{\pi} + \int_0^\infty \phi(x)[I'(x,\rho) + I'(x,-\rho)]dx \end{aligned}$$

where $I'(x,\rho)$ denotes the partial derivative with respect to x. It is straightforward to show that

$$I'(x,\rho) + I'(x,-\rho) = \rho(2\Phi(\rho x/\sigma) - 1).$$

Then

$$\begin{aligned} E|XY| &= \frac{2\sigma}{\pi} + 2\rho\int_0^\infty \phi(x)[2\Phi(\rho x/\sigma) - 1)]dx \\ &= \frac{2\sigma}{\pi} - \rho + 4\rho\int_0^\infty \phi(x)\Phi(\rho x/\sigma)dx. \end{aligned}$$

Denote the last integral by $J(\rho)$. Then $J(0) = 1/4$ and, because $d(\rho/\sigma)/d\rho = \sigma^{-3}$,

$$J'(\rho) = \frac{1}{2\pi(1-\rho^2)^{0.5}}.$$

Hence

$$J(\rho) = \frac{1}{4} + \frac{\arcsin(\rho)}{2\pi}.$$

The proof is completed.

References

Abramowitz, M. and I. Stegun (1967). *Handbook of Mathematical Functions.* Washington, D.C.: U.S. Government Printing Office. (eds.).

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics 12*, 171–178.

Azzalini, A. and A. Capitanio (1999). Statistical applications of the multivariate skew normal distribution. *Journal of the Royal Statistical Society, Series B: Statistical Methodology 61*, 579–602.

Azzalini, A. and A. Dalla Valle (1996). The multivariate skew-normal distribution. *Biometrika 83*, 715–726.

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

Ferreira, J. T. and M. F. J. Steel (2006). A constructive representation of univariate skewed distributions. *Journal of the American Statistical Association 101*(474), 823–829.

Genton, M. G. (2004). *Skew-elliptical distributions and their applications.* Boca Raton, FL: Chapman & Hall.

Handcock, M. and J. R. Wallis (1994). An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association 89*, 368–390.

Kim, H. and B. K. Mallick (2004). A bayesian prediction using the skew-gaussian processes. *Journal of Statistical Planning and Inference 120*, 85–101.

Mardia, K. V. and R. J. Marshall (1984). Maximum likelihood estimation of models for residual covariance in spatial statistics. *Biometrika 71*, 135–146.

McCulloch, C. E. (1997). Maximum likelihood algorithms for generalized linear mixed models. *Journal of the American Statistical Association 92*, 162–170.

Neal, R. (2003). Slicing sampling. *Annuals of Statistics 31*, 705–767.

Palacios, M. B. and M. F. J. Steel (2006). Non-gaussian bayesian geostatistical modeling. *Journal of the American Statistical Association 101*(474), 604–618.

Wei, G. and M. Tanner (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association 85*, 699–704.

Zhang, H. (2002). On estimation and prediction for spatial generalized linear mixed models. *Biometrics 56*(1), 129–136.

Zhang, H. (2003). Optimal interpolation and the appropriateness of cross-validating variogram in spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics 12*(3), 698–713.

List of Figures

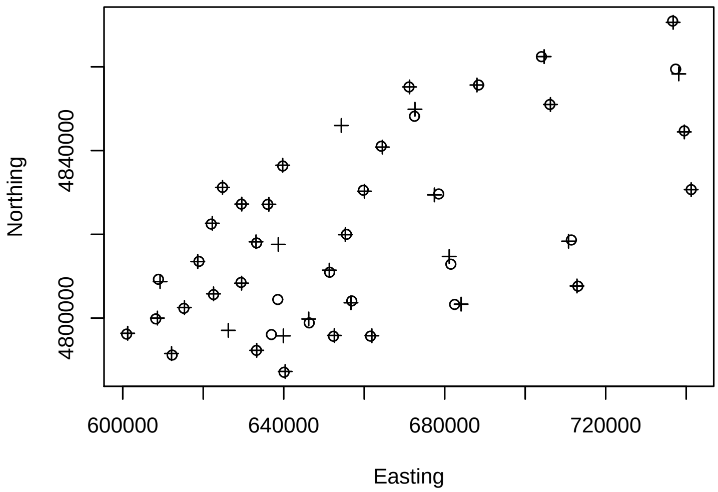


Figure 1: Observed locations: circle (○) for dataset 1 and plus (+) for dataset 2.

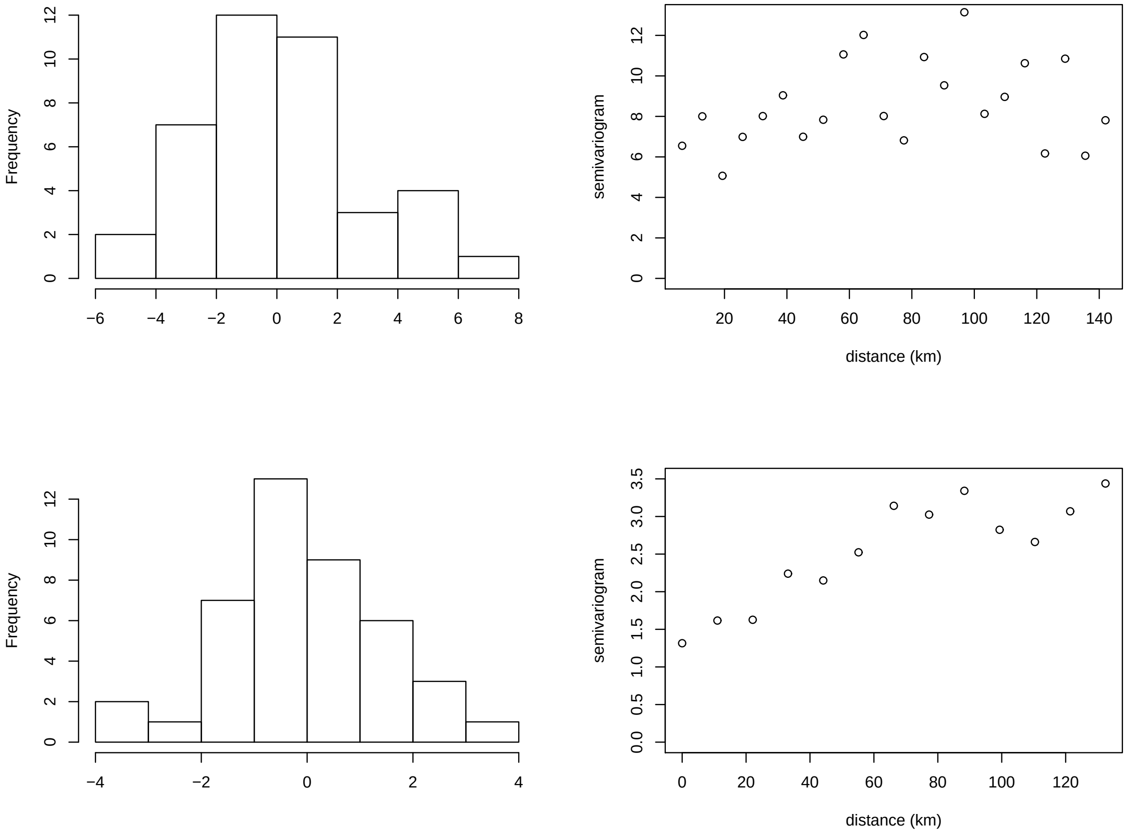


Figure 2: Histograms (left column) and semivariograms (right column) of residuals from the simple regression for dataset 1 (upper row) and dataset 2 (lower row)

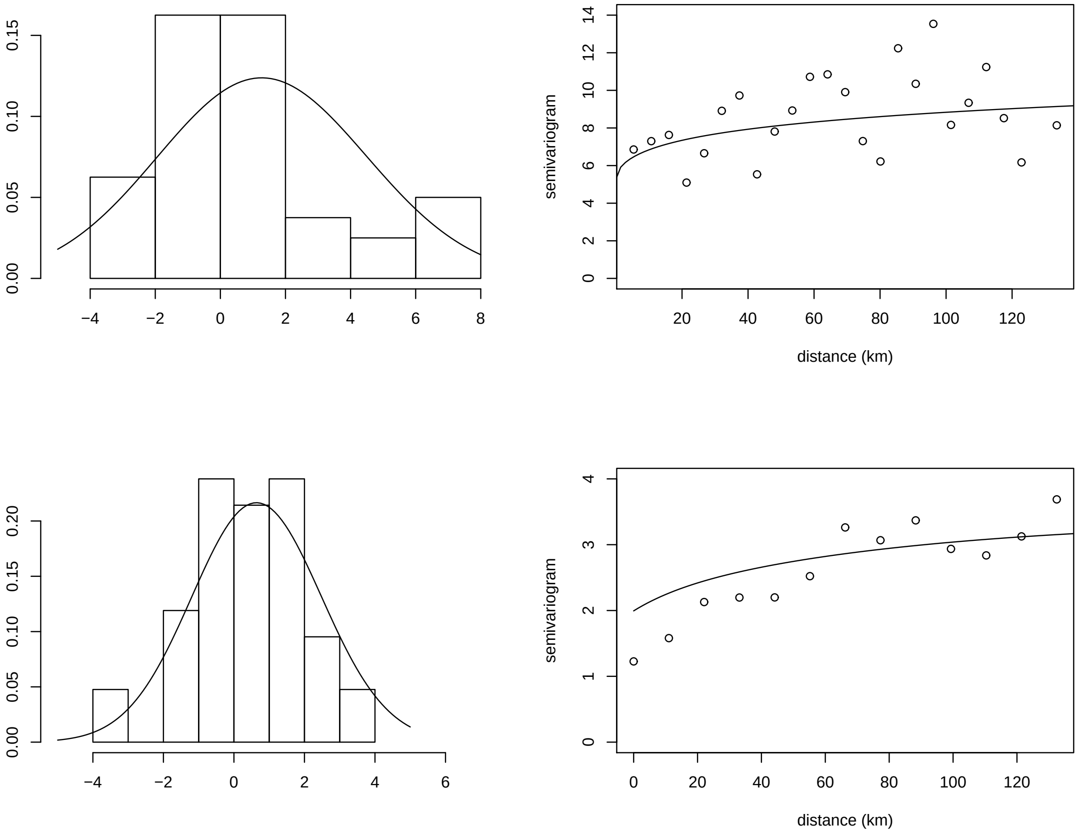
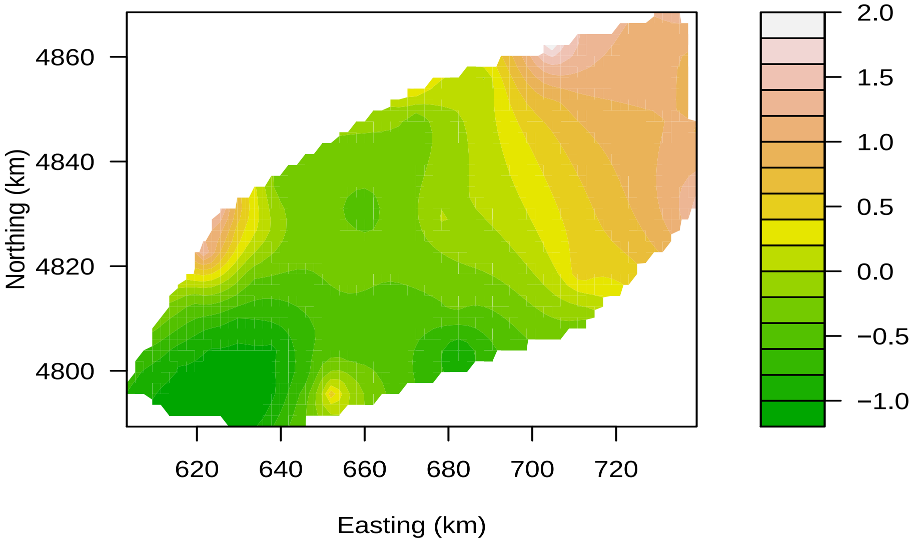


Figure 3: Histograms (left column) and semivariograms (right column) of residuals from the skew-Gaussian model for dataset 1 (upper row) and dataset 2 (lower row)



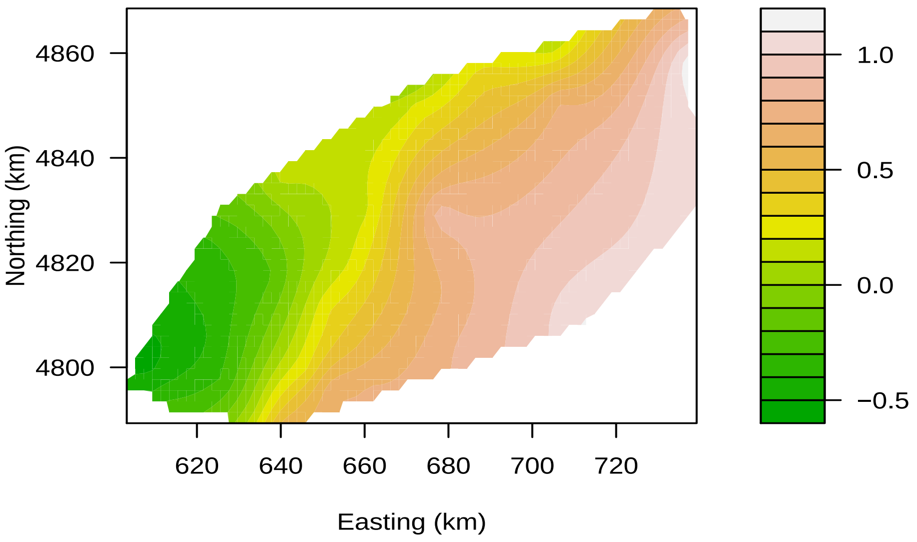


Figure 4: Contour plots of predicted error term for dataset 1 (top) and dataset 2 (bottom)