Journal of Multivariate Analysis 114 (2013) 16–28

Contents lists available at SciVerse ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Non-Gaussian modeling of spatial data using scale mixing of a unified skew Gaussian process

Hamid Zareifard, Majid Jafari Khaledi *

Department of Statistics, Tarbiat Modares University, P.O. Box 14115-134, Tehran, Iran

ARTICLE INFO

Article history:
Received 22 July 2010
Available online 16 July 2012

AMS subject classifications:
62H11
62H12

Keywords:
Spatial modeling
Scale mixing
Unified skew Gaussian
Random process
Outlier
EM algorithm
Slice sampling

ABSTRACT

In this paper, we introduce a unified skew Gaussian-log Gaussian model and propose a general class of spatial sampling models that can account for both heavy tails and skewness. This class includes some models proposed previously in the literature. The likelihood function involves analytically intractable integrals and direct maximization of the marginal likelihood is numerically difficult. We obtain maximum likelihood estimates of the model parameters, using a stochastic approximation of the EM algorithm (SAEM). The predictive distribution at unsampled sites is approximated based on Markov chain Monte Carlo samples. The identifiability of the parameters and the performance of the proposed model is investigated by a simulation study. The usefulness of our methodology is demonstrated by analyzing a Pb data set in a region of north Iran.

1. Introduction

Spatial modeling can provide a statistically sound approach for explaining a response variable observed over a region. The traditional approaches to model spatial data are based on the Gaussian distribution. This assumption might be overly restrictive to represent the data. The real data could be highly non-Gaussian and may show features like heavier tails or skewness.

Considerable interest has been focused on approaches that allow the Gaussian assumption to be relaxed in spatial models. For modeling skewed distributions, De Oliveira et al. [9] extended the Bayesian transformed Gaussian model applying the Box–Cox family of power transformations. To handle some of the potential weaknesses associated with this method, Kim and Mallick [21] developed the skew Gaussian random field based on the skew normal distribution (Azzalini and Capitanio [5]). Even if their model has an appealing construction, Genton and Zhang [17] demonstrated that it has an identifiability problem. In a different way, Zhang and El-Shaarawi [28] recently introduced a class of stationary process that have skewed marginal distributions. They obtained maximum likelihood estimates of model parameters based on a Monte Carlo EM algorithm. Dominguez-Molina et al. [10] and Gonzalez-Farias et al. [18] proposed a multivariate closed skew normal distribution which is closed under marginalization and conditioning. Accordingly, an n-dimensional random vector $\mathbf{Y}$ is said to have a multivariate closed skew-normal distribution, denoted by $CSN_{n,m}(\boldsymbol{\mu}, \Sigma, D, \boldsymbol{\nu}, \Theta)$, if its density is

$$\phi_n(\mathbf{y}; \boldsymbol{\mu}, \Sigma)\Phi_m(D(\mathbf{y}-\boldsymbol{\mu}); \boldsymbol{\nu}, \Theta)/\Phi_m(0; \boldsymbol{\nu}, \Theta + D\Sigma D'), \qquad (1)$$

* Corresponding author.
E-mail addresses: zareifard@modares.ac.ir (H. Zareifard), jafari-m@modares.ac.ir, khaledi1800@yahoo.com (M. Jafari Khaledi).

0047-259X/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2012.07.003

where $\boldsymbol{\mu} \in \Re^n$, $\boldsymbol{\nu} \in \Re^m$ and $\Sigma \in \Re^{n\times n}$ and $\Theta \in \Re^{m\times m}$ are both covariance matrices, $D \in \Re^{m\times n}$, $\phi_n(\mathbf{y}; \boldsymbol{\mu}, \Sigma)$, and $\Phi_n(\mathbf{y}; \boldsymbol{\mu}, \Sigma)$ are the probability density function (*pdf*) and cumulative distribution function (*cdf*), respectively, of the n-dimensional normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ, and D' is the transpose of the matrix D. A more detailed description of these and other skewed models may be found in the book edited by Genton [15].

If $D = \mathbf{0}$ this density reduces to the multivariate normal one and if $m = 1$ this clearly reduces to the skew-normal distribution [4]. Allard and Naveau [2] used the multivariate closed skew normal and introduced a spatial skewed Gaussian process. To increase the amount of skewness in the vector $\mathbf{Y}$ as well as to simplify the interpretation of this density, Allard and Naveau [2] assumed that $m = n$, $\boldsymbol{\nu} = 0$, $\Theta = \Sigma$ and $D = \delta I_n$ in which $\delta \in \Re$ is a single parameter controlling skewness and I_n is the identity matrix of dimension n. This model is also referred to as the homotopic model.

Arellano-Valle and Azzalini [3] presented a family of skew-normal distributions which unifies a plethora of SN distributions including CSN with no over-parametrization problem. According to their approach, an n-dimensional random vector $\mathbf{Y}$ has a multivariate unified skew-normal distribution, denoted by $SUN_{n,m}(\boldsymbol{\mu}, \Sigma, \Gamma, \boldsymbol{\nu}, \Delta)$, if its density is

$$\phi_n(\mathbf{y}; \boldsymbol{\mu}, \Sigma)\Phi_m(\Gamma'\Sigma^{-1}(\mathbf{y} - \boldsymbol{\mu}); \boldsymbol{\nu}, \Delta - \Gamma'\Sigma^{-1}\Gamma)/\Phi_m(0; \boldsymbol{\nu}, \Delta), \tag{2}$$

where $\Delta \in \Re^{m\times m}$ is a correlation matrix, $\Gamma \in \Re^{n\times m}$. The SUN and CSN classes are equivalent when $\Delta = \Theta + D\Sigma D'$ and $\Gamma = \Sigma D'$.

Note that if $\mathbf{Y}$ is partitioned as $\mathbf{Y} = (\mathbf{Y}_1', \mathbf{Y}_2')'$, then the marginal distribution of k-dimensional vector $\mathbf{Y}_1$ is $SUN_{k,m}(\boldsymbol{\mu}_1, \Sigma_{11}, \Gamma_1, \boldsymbol{\nu}, \Delta)$, where parameters correspond to the partition of $\mathbf{Y}$ as

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \qquad \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} \Gamma_1 \\ \Gamma_2 \end{pmatrix}.$$

Similar in spirit with [2], we will now introduce a new spatial skewed Gaussian process. Let $\{G(\mathbf{s}), \mathbf{s} \in R \subseteq \Re^d\}$, $d \geq 1$, be a spatial, ergodic, stationary, zero-mean Gaussian process with stationary covariance function $c(\mathbf{h}) = Cov(G(\mathbf{s} + \mathbf{h}), G(\mathbf{s}))$ and denoted the covariance matrix of the random vector $\mathbf{G} = (G(\mathbf{s}_1), \dots, G(\mathbf{s}_n))'$ by Σ. In order to link this spatial structure with the skew-normal distribution, we assume that the m-dimensional vector $\mathbf{T}$ has a normal distribution, such that

$$\begin{pmatrix} \mathbf{T} \\ \mathbf{G} \end{pmatrix} \sim N_{m+n}\left(\mathbf{0}, \begin{pmatrix} \Delta & \Gamma' \\ \Gamma & \Sigma \end{pmatrix}\right). \tag{3}$$

Now, we define a SUN random process $\{Y(\mathbf{s})\}$ as

$$Y(\mathbf{s}) \overset{d}{=} \mu(\mathbf{s}) + [G(\mathbf{s})|\mathbf{T} > \boldsymbol{\nu}].$$

For any $(\mathbf{s}_1, \dots, \mathbf{s}_n)$, $\mathbf{Y} = (Y(\mathbf{s}_1), \dots, Y(\mathbf{s}_n))'$ has the multivariate unified skew-normal distribution (2). We denote this process the unified skew Gaussian (SUN) process. For $m = n$, $\Delta = \Sigma$, $\Gamma = \frac{\delta}{\sqrt{1+\delta^2}}\Sigma$ and $\boldsymbol{\nu} = 0$, the SUN process corresponds to the homotopic model.

The observed data may contain outliers which have extreme values compared to their neighboring observation values. Outlying observations can be error measurements or they may belong to a region in the space with larger observational variance relative to the rest. Hence, the spatial outliers may be isolated or grouped. Based on a process with fat-tailed finite dimensional distributions, Palacios and Steel [26] introduced the Gaussian log-Gaussian (GLG) process based on a ratio of a Gaussian and a log-Gaussian process. In fact, they replaced the Gaussian stochastic process $\epsilon(\mathbf{s})$ by a ratio of independent stochastic processes, $\epsilon(\mathbf{s})/\sqrt{\lambda(\mathbf{s})}$, where the mixing term $\lambda(\mathbf{s})$ is a log-Gaussian stochastic process. This requires a smooth $\lambda(\mathbf{s})$ process, and this means that observations with particularly small values of mixing variables will tend to cluster together. However, they called observations with small $\lambda(\mathbf{s})$ "outliers", even though these observations belong to a region with larger observational variance, relative to the rest. Additionally, Fonseca and Steel [11] considered a similar mixing in the nugget effect component allowing for individual outliers. They also extended the ideas by Palacios and Steel [26] to processes in space and time. In fact, their approach used mixing at two levels. These approaches are suitable only for symmetric heavier tail distributions will fail to handle skewed data. To deal with this problem, the use of scale mixture of skew normal distributions has recently received attention for non-spatial data (see, e.g., [23,6,12,20,13,27]). Based on this idea, we intend to describe a general spatial model that can take skewness and heavy tails into account which we term the unified skew Gaussian-log-Gaussian (SUGLG) model. This model allows us to accommodate and identify outliers.

Considering the identifiability problems associated with non-Gaussian spatial random fields, Genton and Zhang [17] proposed some remedies to avoid the unidentifiability. Subsequently, the GLG and homotopic models do not have the identifiability problem. Our model is also compatible with their remedies. In the sequel, we develop the likelihood-inference methodology for our model. As the likelihood function involves analytically intractable integrals over the distribution of the mixing random variables and since direct maximization of the marginal likelihood is difficult numerically, we apply a stochastic approximation expectation–maximization (SAEM) algorithm to maximize the likelihood function. In this algorithm, for approximating the relevant conditional expectations, we use a Markov chain Monte Carlo algorithm based on the slice sampling [24,1]. Then, we compute predictions using the estimates of the parameters. Since the predictive distribution cannot be evaluated in closed form, it will be approximated using Markov chain Monte Carlo. Outliers are identified through the highest posterior density (HPD) of random mixing variables. Finally, our approach is illustrated using simulated data, as well as applying it to a real data set.

The organization of the paper is as follows. Section 2 introduces the mixture model and derives some of its properties. Section 3 considers maximum likelihood estimation of the model parameters and describes the *SAEM* algorithm. Using simulated data, the identifiability of the parameters and the performance of the proposed model is examined in Section 4. Section 5 illustrates the use of proposed methodology on a spatial data set. Conclusions are given in Section 6.

2. The model

Let $Z(\mathbf{s})$ be a random process defined for locations $\mathbf{s}$ in some spatial region $R \subseteq \Re^d$. We assume that

$$Z(\mathbf{s}) = \mathbf{f}'(\mathbf{s})\boldsymbol{\beta} + W(\mathbf{s}) + \tau\rho(\mathbf{s}), \tag{4}$$

where the mean surface is assumed to be a linear function of $\mathbf{f}(\mathbf{s}) = (f_1(\mathbf{s}), \ldots, f_k(\mathbf{s}))$, a vector of k known functions of the spatial coordinates, with unknown coefficient vector $\boldsymbol{\beta} \in \Re^k$. Further, the second-order stationary error process $W(\mathbf{s})$ is a unified skew Gaussian random process corresponding to the homotopic model. In fact, if $\mathbf{W} = (W(\mathbf{s}_1), \ldots, W(\mathbf{s}_n))'$ is a vector from this random process at n different locations $\mathbf{s}_1, \ldots, \mathbf{s}_n$, we assume that the distribution of vector $\mathbf{W}$ is

$$\mathbf{W} \sim SUN_{n,n}\left(0, \omega^2 C_{\boldsymbol{\theta}}, \frac{\omega\delta}{\sqrt{1+\delta^2}} C_{\boldsymbol{\theta}}, 0, C_{\boldsymbol{\theta}}\right), \tag{5}$$

where the skewness parameter δ and the scale parameter ω belong to $\Re$ and $\Re^+$, respectively; $C_{\boldsymbol{\theta}}$ is the $n \times n$ correlation matrix with $C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_j\|)$ as its (i, j)th element; $C_{\boldsymbol{\theta}}(d)$ is a valid correlation function of distance d, parameterized by a vector $\boldsymbol{\theta}$. Moreover, $\rho(\mathbf{s})$ denotes an uncorrelated Gaussian process with zero mean and unit variance, modeling the so-called "nugget effect", which allows for measurement error and small-scale variation. The scale parameters τ is defined in $\Re^+$. With regard to the alternative representation of [3] for SUN families, we have

$$\mathbf{W} \overset{d}{=} \omega\frac{\delta}{\sqrt{1+\delta^2}}\mathbf{U} + \omega\frac{1}{\sqrt{1+\delta^2}}\mathbf{V}, \tag{6}$$

where $\mathbf{V} \sim N_n(0, C_{\boldsymbol{\theta}})$ and $\mathbf{U} \sim TN_n(0; 0, C_{\boldsymbol{\theta}})$ are independent. It must be noted that $TN_n(\mathbf{c}; \boldsymbol{\mu}, \Sigma)$ denotes the $N_n(\boldsymbol{\mu}, \Sigma)$ distribution truncated below at a point $\mathbf{c}$. To facilitate the computations, we consider the reparametrization $\alpha = \omega\delta/\sqrt{1+\delta^2}$ and $\sigma = \omega/\sqrt{1+\delta^2}$. Thus, we can replace model (6) with the following model

$$\mathbf{W} \overset{d}{=} \alpha\mathbf{U} + \sigma\mathbf{V}. \tag{7}$$

Now, we use the idea of scale mixing in order to construct processes that imply finite dimensional distributions with heavy tails. In fact, we intend to introduce a model that in addition to skewness, allows for modeling regions in space with larger observational variance relative to the rest of the space. For this purpose, we can add the scaling variables in the skewed spatial model (7) to account for heavy tails as follows

$$\mathbf{W}_\lambda \overset{d}{=} \alpha\Lambda^{-\frac{1}{2}}\mathbf{U} + \sigma\Lambda^{-\frac{1}{2}}\mathbf{V}, \tag{8}$$

where $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ and $\lambda_i = \lambda(\mathbf{s}_i)'s$ are spatially correlated mixing variables which are independent of the other model components. In sequel, we aim to define a random process such that its finite-dimensional distributions are equal in distribution with the linear model (8). For this, we first suppose that $G(\mathbf{s})|\lambda(\mathbf{s})$ is a spatial second-order stationary zero-mean Gaussian process. Let $\mathbf{G} = (G(\mathbf{s}_1), \ldots, G(\mathbf{s}_n))'$ and $\boldsymbol{\lambda} = (\lambda(\mathbf{s}_1), \ldots, \lambda(\mathbf{s}_n))'$, then we denote the covariance matrix of $\mathbf{G}|\boldsymbol{\lambda}$ by $\Sigma_w = (\sigma^2 + \alpha^2)\Lambda^{-\frac{1}{2}}C_{\boldsymbol{\theta}}\Lambda^{-\frac{1}{2}}$. Also, let $\mathbf{T}$ be a normal vector of dimension n and consider the augmented normal vector $(\mathbf{T}, \mathbf{G})'$ conditional on $\boldsymbol{\lambda}$, such that

$$\left.\begin{pmatrix}\mathbf{T}\\ \mathbf{G}\end{pmatrix}\right| \boldsymbol{\lambda} \sim N_{2n}\left(\mathbf{0}, \begin{pmatrix} C_{\boldsymbol{\theta}} & \alpha C_{\boldsymbol{\theta}}\Lambda^{-\frac{1}{2}} \\ \alpha\Lambda^{-\frac{1}{2}}C_{\boldsymbol{\theta}} & \Sigma_w \end{pmatrix}\right). \tag{9}$$

Now, we define a *SUGLG* process $\{W_\lambda(\mathbf{s})\}$ based on the following hierarchical representation:

1. $[W_\lambda(\mathbf{s})|\lambda(\mathbf{s})] \overset{d}{=} [G(\mathbf{s})|\mathbf{T} > \mathbf{0}, \lambda(\mathbf{s})]$.
2. $\lambda(\mathbf{s})$ is a log-Gaussian stochastic process.

Thus, for any vector $\mathbf{W}_\lambda = (W_\lambda(\mathbf{s}_1), \ldots, W_\lambda(\mathbf{s}_n))'$ with $\mathbf{G}$ and $\mathbf{T}$ distributed according to (9), $\mathbf{W}_\lambda$ is equal in distributions to the linear form (8). Hence, the distribution of $\mathbf{W}_\lambda|\boldsymbol{\lambda}$ is $SUN_{n,n}(0, \Sigma_w, \alpha\Lambda^{-\frac{1}{2}}C_{\boldsymbol{\theta}}, 0, C_{\boldsymbol{\theta}})$.

Now we introduce the random process $Z(\mathbf{s})$ as

$$Z(\mathbf{s}) = \mathbf{f}'(\mathbf{s})\boldsymbol{\beta} + W_\lambda(\mathbf{s}) + \tau\rho(\mathbf{s}), \tag{10}$$

where the second-order stationary error process $W_\lambda(\mathbf{s})$ is a *SUGLG* process. We can also represent the random process $Z(\mathbf{s})$ as follows

$$Z(\mathbf{s}) = \mathbf{f}'(\mathbf{s})\boldsymbol{\beta} + \frac{W(\mathbf{s})}{\sqrt{\lambda(\mathbf{s})}} + \tau\rho(\mathbf{s}), \tag{11}$$

where $W(\mathbf{s})$ is a *SUN* random process, such that $\mathbf{W} = (W(\mathbf{s}_1), \dots, W(\mathbf{s}_n))' \sim SUN_{n,n}(0, (\sigma^2 + \alpha^2)C_{\boldsymbol{\theta}}, \alpha C_{\boldsymbol{\theta}}, 0, C_{\boldsymbol{\theta}})$. Thus, if we denote the vector of observations by $\mathbf{Z} = (Z(\mathbf{s}_1), \dots, Z(\mathbf{s}_n))$, then the model at the n sample locations can be written as

$$\mathbf{Z} \overset{d}{=} X\boldsymbol{\beta} + \alpha\Lambda^{-\frac{1}{2}}\mathbf{U} + \sigma\Lambda^{-\frac{1}{2}}\mathbf{V} + \tau\boldsymbol{\rho}, \tag{12}$$

where $X = (\mathbf{f}(\mathbf{s}_1), \dots, \mathbf{f}(\mathbf{s}_n))'$ and $\boldsymbol{\rho} = (\rho(\mathbf{s}_1), \dots, \rho(\mathbf{s}_n))'$. Recall that $\mathbf{V}$ and $\mathbf{U}$ are similar to those in the linear model (6). In this model, the observations with small values of mixing variables λ_i characterize the regions with a relatively large variance around the underlying trend surface. In fact, the mixing process $\lambda(\mathbf{s})$ is spatially correlated, and observations with small λ_i's belong to a region with larger observational variance relative to the rest of space and thus allows for modeling regions in the space with larger observational variance. However, following [26], we will continue to call observations with small λ_i "outliers". As such, we assume that $\eta(\mathbf{s}) = \ln(\lambda(\mathbf{s}))$ is a Gaussian random field with finite-dimensional distributions:

$$\boldsymbol{\eta} = (\ln(\lambda_1), \dots, \ln(\lambda_n))' \sim N_n\left(-\frac{\nu}{2}\mathbf{1}_n, \nu C_{\boldsymbol{\theta}}\right), \tag{13}$$

where $\mathbf{1}_n$ is a vector of 1's of order n. We can easily see that this latter implies a lognormal distribution for λ_i with $E(\lambda_i) = 1$ and $Var(\lambda_i) = e^\nu - 1$. If ν is large, then with large probability, among the observations there exists a region with larger observational variance relative to the rest. Note that we can use different correlation matrices for $\mathbf{W}$ and $\boldsymbol{\eta}$, but in order to reduce complexity of the model, the same correlation matrices are used.

Let $\mathbf{z} = (z_1, \dots, z_n)$ be a single realization of the considered random field, where $z_i = Z(\mathbf{s}_i)$, then the conditional joint distribution of $\mathbf{z}$, given $\boldsymbol{\lambda}$, is

$$p_{\boldsymbol{\psi}}(\mathbf{z}|\Lambda) = SUN_{n,n}(X\boldsymbol{\beta}, \Sigma_z, \alpha\Lambda^{-\frac{1}{2}}C_{\boldsymbol{\theta}}, 0, C_{\boldsymbol{\theta}}), \tag{14}$$

where $\boldsymbol{\psi}' = (\boldsymbol{\beta}, \sigma, \alpha, \tau, \boldsymbol{\theta}', \nu)$ is the vector of all the model parameters and $\Sigma_z = \Sigma_w + \tau^2 I_n$. Given the mixing variables $\boldsymbol{\lambda}$, only the unified skew normal distribution behavior is assumed. We can think of model (12) with the finite-dimensional distributions in (14) as the Gaussian and skew Gaussian model given $\boldsymbol{\lambda}$ and α. If $\Lambda = I_n$ and $\alpha = 0$ the model (12) is Gaussian, and if $\Lambda = I_n$ and $\tau = 0$, it reduces to the homotopic model. Moreover, if $\alpha = 0$, then we arrive at the GLG model. In the sequel, we call the model (11) with the mixing distribution in (13) the SUGLG model.

2.1. The correlation function

To ease the computations, we assume that the process $W(\mathbf{s})$ and $\eta(\mathbf{s})$ have an isotropic power exponential correlation function given by

$$C_{\boldsymbol{\theta}}(d) = \exp(-\xi d^\phi) = \gamma^{d^\phi}, \quad \xi > 0, \gamma = \exp(-\xi) \in (0, 1), \phi \in (0, 2],$$

where d is the Euclidean distance and $\boldsymbol{\theta} = (\gamma, \phi)'$ with γ the range parameter and ϕ the smoothness parameter. The range parameter γ controls how fast the correlation decays with distance, and parameter ϕ controls the smoothness of the random field.

This class which is very flexible and popular, as in the cases $\phi = 1$ and $\phi = 2$, contains the exponential and Gaussian correlation functions, respectively, which are often used in applications.

3. Maximum likelihood estimation

In this section, we consider the problem of computing the maximum likelihood (ML) estimates of the model parameters. The likelihood function for $\boldsymbol{\psi}$ given the observed sample $\mathbf{z}$ can be written as

$$L(\boldsymbol{\psi}; \mathbf{z}) = \int_{\Re_+^n} f_{SUN}^{n,n}(\mathbf{z}|X\boldsymbol{\beta}, \Sigma_z, \alpha\Lambda^{-\frac{1}{2}}C_{\boldsymbol{\theta}}, 0, C_{\boldsymbol{\theta}})dP_{\boldsymbol{\lambda}}.$$

As observed, the likelihood function does not have a simple form, so direct maximization of the likelihood is intractable. In such cases, the EM algorithm is a popular strategy. For our model, the expectation of the complete likelihood involved in the E-step of the algorithm cannot be done in a closed form. A possible solution to this problem is to compute the expectation using a Monte Carlo method, this is known as the Monte Carlo EM (*MCEM*) algorithm. Recently, a stochastic approximation version of *EM* (*SAEM*), has been suggested by Delyon et al. [8] as a powerful alternative to *EM* when expectation step is untractable. This algorithm is computationally more efficient than a classical *MCEM*, due to recycling the simulations from one iteration to the next in the smoothing phase of the algorithm (see [22]). Thus, we will use the *SAEM* algorithm.

In practice, we usually know the sign of α based on whether the distribution of observations is right-skewed or left-skewed. Thus, we assume that $\alpha > 0$. To ease the computations, we rewrite the model as the following hierarchical model

$$\begin{aligned} &\mathbf{Z} = X\boldsymbol{\beta} + \Lambda^{-\frac{1}{2}}(\mathbf{U} + \mathbf{V}) + \tau\boldsymbol{\rho}, \\ &\mathbf{U} \sim TN_n(0; 0, \alpha^2 C_{\boldsymbol{\theta}}), \qquad \mathbf{V} \sim N_n(0, \sigma^2 C_{\boldsymbol{\theta}}), \qquad \boldsymbol{\eta} \sim N_n\left(-\frac{\nu}{2}\mathbf{1}_n, \nu C_{\boldsymbol{\theta}}\right). \end{aligned} \tag{15}$$

With this hierarchical model, we treat $\mathbf{U}$, $\mathbf{V}$ and $\boldsymbol{\eta}$ as latent variables. Hence, if $\mathbf{Z}^+ = (\mathbf{Z}, \mathbf{U}, \mathbf{V}, \boldsymbol{\eta})$ is the complete data, then the complete-data log likelihood is

$$\ell_c(\boldsymbol{\psi}; \mathbf{z}^+) = \log p_{\boldsymbol{\beta},\tau}(\mathbf{z}|\mathbf{u}, \mathbf{v}, \boldsymbol{\eta}) + \log p_{\alpha^2,\boldsymbol{\theta}}(\mathbf{u}) + \log p_{\sigma^2,\boldsymbol{\theta}}(\mathbf{v}) + \log p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta}).$$

It follows, after some simple algebra, that the sum of the three first terms is:

$$\begin{aligned} \ell_c^*(\boldsymbol{\psi}; \mathbf{z}^+) = {} & -\frac{n}{2}\log\tau^2 - \frac{1}{2\tau^2}(\mathbf{z} - X\boldsymbol{\beta} - \mathbf{W}_\lambda)'(\mathbf{z} - X\boldsymbol{\beta} - \mathbf{W}_\lambda) - \frac{n}{2}\log\sigma^2 \\ & - \frac{1}{2\sigma^2}\mathbf{V}'C_{\boldsymbol{\theta}}^{-1}\mathbf{V} - \frac{n}{2}\log\alpha^2 - \frac{1}{2\alpha^2}\mathbf{U}'C_{\boldsymbol{\theta}}^{-1}\mathbf{U} - \log\int_{\Re_+^n}\exp\left(-\frac{1}{2}\mathbf{U}'C_{\boldsymbol{\theta}}^{-1}\mathbf{U}\right)d\mathbf{U} \\ & - \frac{n}{2}\log\nu - \log|C_{\boldsymbol{\theta}}| - \frac{1}{2\nu}\left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right)' C_{\boldsymbol{\theta}}{}^{-1}\left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right)', \end{aligned} \tag{16}$$

where $\mathbf{W}_\lambda = \Lambda^{-\frac{1}{2}}(\mathbf{U} + \mathbf{V})$. The *SAEM* algorithm starts with some initial estimate $\boldsymbol{\psi}^{(0)}$. At the mth iteration, given estimate $\boldsymbol{\psi}^{(m)}$, the new estimate $\boldsymbol{\psi}^{(m+1)}$ is generated in two steps; a stochastic approximation step, and a maximization step. The process is iterated from the starting value $\boldsymbol{\psi}^{(0)}$ to convergence. Based on (16), we have

$$\begin{aligned} Q(\boldsymbol{\psi}|\boldsymbol{\psi}^{(m)}) = {} & E_{\boldsymbol{\psi}^{(m)}}[\ell_c(\boldsymbol{\psi}; \mathbf{z}^+)|\mathbf{z}] = -\frac{n}{2}\log\tau^2 - \frac{1}{2\tau^2}[\mathbf{z}'\mathbf{z} - 2\boldsymbol{\beta}'X'\mathbf{z} \\ & - 2\mathbf{z}'E_{\boldsymbol{\psi}^{(m)}}(\mathbf{W}_\lambda|\mathbf{z}) + \boldsymbol{\beta}'X'X\boldsymbol{\beta} + 2\boldsymbol{\beta}'X'E_{\boldsymbol{\psi}^{(m)}}(\mathbf{W}_\lambda|\mathbf{z}) \\ & + E_{\boldsymbol{\psi}^{(m)}}(\mathbf{W}_\lambda'\mathbf{W}_\lambda|\mathbf{z})] - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}tr[C_{\boldsymbol{\theta}}^{-1}E_{\boldsymbol{\psi}^{(m)}}(\mathbf{V}\mathbf{V}'|\mathbf{z})] \\ & - \frac{n}{2}\log\alpha^2 - \frac{1}{2\alpha^2}tr[C_{\boldsymbol{\theta}}^{-1}E_{\boldsymbol{\psi}^{(m)}}(\mathbf{U}\mathbf{U}'|\mathbf{z})] - \log\int_{\Re_+^n}\exp\left(-\frac{1}{2}\mathbf{U}'C_{\boldsymbol{\theta}}^{-1}\mathbf{U}\right)d\mathbf{U} \\ & - \frac{n}{2}\log\nu - \log|C_{\boldsymbol{\theta}}| - \frac{1}{2}\mathbf{1}_n'[C_{\boldsymbol{\theta}}^{-1}E_{\boldsymbol{\psi}^{(m)}}(\boldsymbol{\eta}|\mathbf{z})] - \frac{\nu}{8}\mathbf{1}_n'C_{\boldsymbol{\theta}}^{-1}\mathbf{1}_n - \frac{1}{2\nu}tr[C_{\boldsymbol{\theta}}^{-1}E_{\boldsymbol{\psi}^{(m)}}(\boldsymbol{\eta}\boldsymbol{\eta}'|\mathbf{z})]. \end{aligned} \tag{17}$$

At iteration m of the *SAEM* algorithm, if $\{(\mathbf{U}_i, \mathbf{V}_i, \boldsymbol{\eta}_i)\}_{i=1}^l$ are samples from the joint posterior distribution $p_{\boldsymbol{\psi}^{(m)}}(\mathbf{U}, \mathbf{V}, \boldsymbol{\eta}|\mathbf{z})$, then an stochastic approximation of $E_{\boldsymbol{\psi}^{(m)}}(g(\mathbf{U}, \mathbf{V}, \boldsymbol{\eta})|\mathbf{z})$ is

$$E^{m+1} = E^m + \gamma_{m+1}\left(\frac{1}{l}\sum_{i=1}^{l} g(\mathbf{U}_i, \mathbf{V}_i, \boldsymbol{\eta}_i) - E^m\right)$$

where γ_m is a smoothing parameter, i.e. a decreasing sequence of positive numbers.

At the maximizing step, new estimates are obtained by maximizing (17) where the conditional expectations have been replaced by corresponding stochastic approximations. Thus, if $E_i^{m+1}, i = 1, \ldots, 9$ are stochastic approximations of the conditional expectations corresponding to each component in Eq. (17), then the updates of the parameters at the mth iteration can be produced in the following closed form expressions:

$$\begin{aligned} \boldsymbol{\beta}^{(m+1)} &= (X'X)^{-1}X'[\mathbf{z} - E_1^{m+1}], \\ \tau^{2(m+1)} &= \frac{1}{n}[\mathbf{z}'\mathbf{z} - \boldsymbol{\beta}^{(m+1)'}(X'X)\boldsymbol{\beta}^{(m+1)} - 2\mathbf{z}'E_1^{m+1} + E_2^{m+1}], \\ \boldsymbol{\theta}^{(m+1)} &= Arg\min_{\boldsymbol{\theta}}\left\{2\log|C_{\boldsymbol{\theta}}| + n\log\ tr(C_{\boldsymbol{\theta}}^{-1}E_3^{m+1}) + n\log\ tr(C_{\boldsymbol{\theta}}^{-1}E_4^{m+1})\right. \\ &\qquad \left. + 2\log\int_{\Re_+^n}\exp\left(-\frac{1}{2}\mathbf{U}'C_{\boldsymbol{\theta}}^{-1}\mathbf{U}\right)d\mathbf{U} + n\log\nu_{\boldsymbol{\theta}} + \mathbf{1}_n'C_{\boldsymbol{\theta}}^{-1}E_5^{m+1} + \frac{1}{\nu_{\boldsymbol{\theta}}}tr[C_{\boldsymbol{\theta}}^{-1}E_6^{m+1}] + \frac{\nu_{\boldsymbol{\theta}}}{4}\mathbf{1}_n'C_{\boldsymbol{\theta}}^{-1}\mathbf{1}_n\right\}, \\ \sigma^{2(m+1)} &= \frac{1}{n}tr(C_{\boldsymbol{\theta}^{(m+1)}}^{-1}E_3^{m+1}), \\ \alpha^{2(m+1)} &= \frac{1}{n}tr(C_{\boldsymbol{\theta}^{(m+1)}}^{-1}E_4^{m+1}), \\ \nu^{(m+1)} &= 2[\mathbf{1}_n'C_{\boldsymbol{\theta}^{(m+1)}}{}^{-1}\mathbf{1}_n]^{-1}\left(-n + \sqrt{n^2 + \mathbf{1}_n'C_{\boldsymbol{\theta}^{(m+1)}}{}^{-1}\mathbf{1}_n\ tr(C_{\boldsymbol{\theta}^{(m+1)}}{}^{-1}E_6^{m+1})}\right), \end{aligned}$$

where

$$\nu_{\boldsymbol{\theta}} = 2\frac{-n+\sqrt{n^2+\mathbf{1}_n' C_{\boldsymbol{\theta}}{}^{-1}\mathbf{1}_n\, tr(C_{\boldsymbol{\theta}}{}^{-1}E_6^{m+1})}}{\mathbf{1}_n' C_{\boldsymbol{\theta}}{}^{-1}\mathbf{1}_n}.$$

The convergence of the SAEM algorithm depends on the choice of the smoothing parameter γ_m and the number of simulations from the joint posterior distribution. In principle, the number of simulations should not affect the convergence of the *SAEM* algorithm but a good choice of it can improve the performance of the algorithm. Because the starting point may not be in the neighborhood of maximum likelihood estimate, we use a large value for γ_m, so that the parameters move quickly into vicinity of maximum likelihood estimate. After moving the estimates $\boldsymbol{\psi}^{(m)}$ around the maximum likelihood estimate, we use a small value for γ_m to stabilize the algorithm in the neighborhood of maximum likelihood estimate. Thus, to improve convergence in the *SAEM* algorithm, we choose γ_m as

$$\gamma_m = \begin{cases} 1, & \text{for } 1 \le m \le K; \\ \dfrac{1}{m-K}, & \text{for } m > K, \end{cases}$$

where K was determined graphically by plotting the values of the *SAEM* estimates against the iteration number. In the *MCEM* algorithm, γ_k would be equal to 1 for all iterations. Because the *SAEM* algorithm is a stochastic algorithm, a deterministic convergence criterion is not appropriate. We recommend implementing the *SAEM* algorithm with a sufficiently large number of iterations and checking the convergence by plotting the values of the *SAEM* estimates against the iteration number.

In the sequel, we will describe an *MCMC* algorithm for sampling from the joint posterior distribution $p_{\boldsymbol{\psi}}(\mathbf{U}, \mathbf{V}, \boldsymbol{\eta}|z)$.

3.1. Markov Chain Monte Carlo Simulation

To implement the Markov chain Monte Carlo simulation, we write the complete posterior conditional distributions as

$$p(\mathbf{u}, \mathbf{v}|\mathbf{z}, \boldsymbol{\eta}) = p(\mathbf{v}|\mathbf{z}, \mathbf{u}, \boldsymbol{\eta})p(\mathbf{u}|\mathbf{z}, \boldsymbol{\eta}), \tag{18}$$

$$p(\boldsymbol{\eta}|\mathbf{z}, \mathbf{u}, \mathbf{v}) \propto p_{\boldsymbol{\psi}}(\mathbf{z}|\mathbf{u}, \mathbf{v}, \boldsymbol{\eta})p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta}), \tag{19}$$

where

$$p(\mathbf{u}|\mathbf{z}, \boldsymbol{\eta}) = TN_n(0, \Sigma_u^{-1}\Lambda^{-\frac{1}{2}}\Sigma^{-1}\boldsymbol{\epsilon}, \Sigma_u^{-1}),$$

$$p(\mathbf{v}|\mathbf{z}, \mathbf{u}, \boldsymbol{\eta}) = N_n\left(\frac{1}{\tau^2}\Sigma_v^{-1}\Lambda^{-\frac{1}{2}}(\boldsymbol{\epsilon} - \Lambda^{-\frac{1}{2}}\mathbf{u}), \Sigma_v^{-1}\right), \tag{20}$$

with $\boldsymbol{\epsilon} = \mathbf{z} - X\boldsymbol{\beta}$, $\Sigma = \tau^2 I_n + \sigma^2\Lambda^{-\frac{1}{2}}C_{\boldsymbol{\theta}}\Lambda^{-\frac{1}{2}}$, $\Sigma_u = \frac{1}{\alpha^2}C_{\boldsymbol{\theta}}^{-1} + \Lambda^{-\frac{1}{2}}\Sigma^{-1}\Lambda^{-\frac{1}{2}}$, $\Sigma_v = \frac{1}{\sigma^2}C_{\boldsymbol{\theta}}^{-1} + \frac{1}{\tau^2}\Lambda^{-1}$ and $p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta})$ denotes the distribution in (13). The full conditional posterior of $\mathbf{V}$ is known and is easy to sample from. Although the full conditional of $\mathbf{U}$ defines a standard probability distribution, sampling from this distribution is simply impracticable. In fact, two methods can be used for sampling from this full conditional. If n is small, we propose to use rejection sampling: generate proposals of multivariate normal distribution which are accepted if they are inside the support region otherwise they get rejected. However, the rejection sampling may be inefficient when n is large. In this case, the Gibbs iterative algorithm is preferable in which each component is generated conditional on all other components of $\mathbf{U}$.

The full conditional of $\boldsymbol{\eta}$ does not define a standard probability distribution. The main issue is that the elements of $\boldsymbol{\eta}$ are not conditionally independent given other parameters and data. This complicates the matter in view of the large dimension of $\boldsymbol{\eta}$. To solve this problem, Palacios and Steel [26] partitioned the elements of $\boldsymbol{\eta}$ in blocks, each of which corresponds to a cluster of observations that are relatively close together. Indeed, they wanted to confine most of the dependence between the η_i's to the same cluster. For each cluster, they used a Metropolis–Hastings step in their MCMC algorithm. Their method has some drawbacks. First, the Metropolis–Hastings algorithm is difficult to automatize since it involves tuning tailored to each application. Second, the convergence time of the Markov chain increases with increasing the number of clusters. Third, the inferences can be affected by the clustering algorithm.

Recently, auxiliary variable methods based on slice sampler have been found to provide an attractive strategy by those who used the Markov chain Monte Carlo (MCMC) algorithms to simulate from complex nonnormalized multivariate densities [24]. For sampling from the full conditional of $\boldsymbol{\eta}$, we implement the slice sampling algorithm based on two auxiliary variables [1,28]. For this purpose, if $U_1|\mathbf{z}, \boldsymbol{\eta}, \mathbf{u}, \mathbf{v}$ and $U_2|\boldsymbol{\eta}$ have the uniform distribution on the intervals $[0, p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}, \mathbf{v})]$ and $[0, p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta})]$, respectively, then

$$p(U_1, U_2, \boldsymbol{\eta}|\mathbf{z}, \mathbf{u}, \mathbf{v}) \propto I_{\{U_1 < p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}, \mathbf{v})\}} I_{\{U_2 < p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta})\}}, \tag{21}$$

where I denotes the indicator function. Thus,

$$p(\boldsymbol{\eta}|\mathbf{z}, \mathbf{u}, \mathbf{v}, U_1, U_2) \propto I_{\{U_1 < p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}, \mathbf{v})\}} I_{\{U_2 < p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta})\}}. \tag{22}$$

Now, if e_1 and e_2 represent exponential distribution with mean 1, say $\exp(1)$, then $\log U_1 = \log p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}, \mathbf{v}) - e_1$ and $\log U_2 = \log p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta}) - e_2$, given $\mathbf{z}$, $\mathbf{u}$, $\mathbf{v}$ and $\boldsymbol{\eta}$. Based on these assumptions, we introduce an algorithm for sampling from the joint posterior distribution. Indeed, given $\mathbf{u}^{(t)}$, $\mathbf{v}^{(t)}$ and $\boldsymbol{\eta}^{(t)}$, we can summarize the main steps of the slice sampling algorithm at step $(t+1)$ as

1. Draw $e_1^{(t+1)}$ and $e_2^{(t+1)}$ from $\exp(1)$, and let
   $a_t = \log p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}^{(t)}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}) - e_1^{(t+1)}$, $b_t = \log p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta}^{(t)}) - e_2^{(t+1)}$.
2. Draw $\boldsymbol{\eta}^{(t+1)}$ from a uniform distribution on
   $\{\boldsymbol{\eta};\, a_t < \log p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)})\} \bigcap \{\boldsymbol{\eta};\, b_t < \log p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta})\}$.
3. Draw $(\mathbf{u}^{(t+1)}, \mathbf{v}^{(t+1)})$ from $p(\mathbf{u}, \mathbf{v}|\mathbf{z}, \boldsymbol{\eta}^{(t+1)})$.
4. Iterate above steps until we get the appropriate number of MCMC samples.

For details regarding step 2, see the Appendix.

3.2. Prediction

In many applications, predicting the response values at new locations is an important goal. In this section, predictions at new locations are made using the plug-in method in which we first assume true values of parameters are known and predict the vector $\mathbf{Z_0} = (Z_{0_1}, \ldots, Z_{0_p})$ at unsampled locations $\mathbf{s}_{0_1}, \ldots, \mathbf{s}_{0_p}$. We then carry out the prediction under the parameter estimates. For this, we must first determine the joint distribution of $(\mathbf{W}', \mathbf{W_0}')'$ in which $\mathbf{W}_0 = (W(\mathbf{s}_{0_1}), \ldots, W(\mathbf{s}_{0_p}))'$. According to SUN process, we have

$$(\mathbf{W}', \mathbf{W_0}')' \overset{d}{=} [(\mathbf{G}', \mathbf{G_0}')'|\mathbf{T} > \mathbf{0}], \tag{23}$$

where $\mathbf{G} = (G(\mathbf{s}_{0_1}), \ldots, G(\mathbf{s}_{0_p}))'$ and the joint distribution of $((\mathbf{G}', \mathbf{G_0}')', \mathbf{T}')'$ is

$$\begin{pmatrix} \mathbf{T} \\ \begin{pmatrix} \mathbf{G} \\ \mathbf{G}_0 \end{pmatrix} \end{pmatrix} \sim N_{2n+p}\left(\mathbf{0}, \quad \begin{pmatrix} C_{\boldsymbol{\theta}}^{oo} & \alpha\Gamma^{+'} \\ \alpha\Gamma^{+} & \Sigma^{+} \end{pmatrix}\right), \tag{24}$$

where $\Sigma^+ = (\sigma^2 + \alpha^2) C_{\boldsymbol{\theta}}^+$ and

$$C_{\boldsymbol{\theta}}^+ = \begin{pmatrix} C_{\boldsymbol{\theta}}^{oo} & C_{\boldsymbol{\theta}}^{op} \\ C_{\boldsymbol{\theta}}^{po} & C_{\boldsymbol{\theta}}^{pp} \end{pmatrix}, \qquad \Gamma^+ = \begin{pmatrix} C_{\boldsymbol{\theta}}^{oo} \\ \Gamma_0 \end{pmatrix},$$

with $C_{\boldsymbol{\theta}}^{pp} = [C_{\boldsymbol{\theta}}(\|\mathbf{s}_{0_i} - \mathbf{s}_{0_j}\|)]_{p\times p}$, $C_{\boldsymbol{\theta}}^{op} = [C_{\boldsymbol{\theta}}(\|\mathbf{s}_i - \mathbf{s}_{0_j}\|)]_{n\times p}$ and $\Lambda_p = diag(\exp(\boldsymbol{\eta}_0))$. Finally, Γ_0 is a $p \times n$ matrix modeling the influence of the n truncated normal values on the interpolated points. With regard to (24), we choose $\Gamma_0 = C_{\boldsymbol{\theta}}^{po}$. Consequently, the joint distribution of $(\mathbf{W}', \mathbf{W_0}')'$ will be

$$(\mathbf{W}', \mathbf{W_0}')' \sim SUN_{n+p,n}(0, \Sigma^+, \alpha\Gamma^+, 0, C_{\boldsymbol{\theta}}^{oo}). \tag{25}$$

Also, based on (11) and (25), the joint distribution of $(\mathbf{Z}, \mathbf{Z_0})'$ can be easily obtained as

$$p(\mathbf{z}, \mathbf{z_0}|\boldsymbol{\eta}, \boldsymbol{\eta_0}) = SUN_{n+p,n}(X_{op}\boldsymbol{\beta}, \Lambda^{+^{-\frac{1}{2}}} \Sigma^+ \Lambda^{+^{-\frac{1}{2}}} + \tau^2 I_{n+p}, \alpha\Lambda^{+^{-\frac{1}{2}}} \Gamma^+, 0, C_{\boldsymbol{\theta}}^{oo}), \tag{26}$$

where $\boldsymbol{\eta_0} = (\boldsymbol{\eta}(\mathbf{s}_{0_1}), \ldots, \boldsymbol{\eta}(\mathbf{s}_{0_p}))'$, $X_{op} = (X', X_p')'$ with $X_p = (\mathbf{f}(s_{0_1}), \ldots, \mathbf{f}(s_{0_p}))'$ and $\Lambda^+ = \begin{pmatrix} \Lambda_o & 0 \\ 0 & \Lambda_p \end{pmatrix}$. Now, to make prediction, we need to obtain the full predictive distribution

$$p(\mathbf{z}_0|\mathbf{z}) = \int_{\Re^p} \int_{\Re^n} p(\mathbf{z}_0|\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\eta_0}) p(\boldsymbol{\eta_0}|\boldsymbol{\eta}, \mathbf{z}) p(\boldsymbol{\eta}|\mathbf{z}) d\boldsymbol{\eta} d\boldsymbol{\eta_0}. \tag{27}$$

In this case, since the process is non-Gaussian, the full predictive distribution cannot be evaluated in closed form but can be approximated using Monte Carlo samples, where drawings from $p(\boldsymbol{\eta}, |\mathbf{z})$ are obtained directly from the MCMC algorithm, described in the previous section. Since $p(\boldsymbol{\eta_0}|\boldsymbol{\eta}, \mathbf{z}) = p(\boldsymbol{\eta_0}|\boldsymbol{\eta})$, we can evaluate (27) using samples of $\boldsymbol{\eta_0}$ from

$$p(\boldsymbol{\eta_0}|\boldsymbol{\eta}) = N_n\left(-\frac{\nu}{2}\mathbf{1}_p + C_{\boldsymbol{\theta}}^{po} C_{\boldsymbol{\theta}}^{oo-1}\left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right), \nu(C_{\boldsymbol{\theta}}^{pp} - C_{\boldsymbol{\theta}}^{po} C_{\boldsymbol{\theta}}^{oo-1} C_{\boldsymbol{\theta}}^{op})\right). \tag{28}$$

Also, based on (26), the conditional density of $[\mathbf{z_0}|\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\eta}_0]$ takes the form

$$p(\mathbf{z_0}|\mathbf{z}, \boldsymbol{\eta}, \boldsymbol{\eta_0}) = SUN_{p,n}(\boldsymbol{\mu}_{z_0|z}, \Sigma_{z_0|z}, \alpha\Gamma_{z_0|z}, -\alpha C_{\boldsymbol{\theta}}^{oo} \Lambda_o^{-\frac{1}{2}} \Sigma_z^{-1}(\mathbf{z} - X\boldsymbol{\beta}), \Delta_{z_0|z}), \tag{29}$$

where $\boldsymbol{\mu}_{z_0|z} = X_p\boldsymbol{\beta} + A(\mathbf{z} - X\boldsymbol{\beta})$ and

$$A = (\sigma^2 + \alpha^2)\Lambda_p^{-\frac{1}{2}} C_{\boldsymbol{\theta}}^{po} \Lambda_o^{-\frac{1}{2}} \Sigma_z^{-1},$$

$$\Sigma_{z_0|z} = (\sigma^2 + \alpha^2)\Lambda_p^{-\frac{1}{2}} C_{\boldsymbol{\theta}}^{pp} \Lambda_p^{-\frac{1}{2}} + \tau^2 I_p - A(\Lambda_o^{-\frac{1}{2}} C_{\boldsymbol{\theta}}^{op} \Lambda_p^{-\frac{1}{2}}),$$

$$\Gamma_{z_0|z} = \Lambda_p^{-\frac{1}{2}} C_{\boldsymbol{\theta}}^{po} - A\Lambda_o^{-\frac{1}{2}} C_{\boldsymbol{\theta}}^{oo},$$

$$\Delta_{z_0|z} = C_{\boldsymbol{\theta}}^{oo} - \alpha^2 C_{\boldsymbol{\theta}}^{oo} \Lambda_o^{-\frac{1}{2}} \Sigma_z^{-1} \Lambda_o^{-\frac{1}{2}} C_{\boldsymbol{\theta}}^{oo}.$$

Thus, for each posterior drawing of $\boldsymbol{\eta}$, we generate a drawing from (28) and finally using sampling from density in (29), we can obtain a realization from the predictive distribution (27). Repeating the aforementioned steps as many times as required, we achieve samples from the predictive distribution as $\{\mathbf{z}_{\mathbf{0}}^{(j)}; j = 1, \ldots, J\}$. The estimates of the spatial predictor and prediction variance are then given as

$$E(\mathbf{Z}_{\mathbf{0}}|\mathbf{z}) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{z}_{\mathbf{0}}^{(j)},$$

$$Var(\mathbf{Z}_{\mathbf{0}}|\mathbf{z}) = \frac{1}{J}\sum_{j=1}^{J} \mathbf{z}_{\mathbf{0}}^{(j)} \mathbf{z}_{\mathbf{0}}^{(j)'} - \frac{1}{J^2}\sum_{j=1}^{J} \mathbf{z}_{\mathbf{0}}^{(j)} \sum_{j=1}^{J} \mathbf{z}_{\mathbf{0}}^{(j)'}. \tag{30}$$

3.3. Outlier detection

The model (11) can account for regions with larger observational variance. In other words, this model is able to accommodate the regions with larger observational variance and does not need to identify them for inference. But, it may be useful to have a criterion to identify regions with inflated variance. In this model, for $\lambda_i = 1$, the marginal sampling distribution of observations is the unified skew Gaussian. Also, observations with particularly small values of mixing variables λ_i tend to be away from the mean surface; hence, this observation may be considered as an outlier.

In order to identify the outliers, Palacios and Steel [26] proposed to compute the Bayes factor in favor of the model with $\lambda_i = 1$ (and all other elements of λ_i free), against the model with free λ_i. However, this method becomes computationally very intensive as it requires to generate a new sample of posterior distribution for each observations. To overcome this problem, we apply the highest posterior density (HPD) $100(1-\alpha)\%$ credible intervals of random mixing variables for determining regions with increased variance. Since in this situation, the posterior distribution does not have a closed form, the algorithm of [7] can be used to determine this region. In this algorithm, $\{\lambda_{i_j}\}_{j=1}^{l}$ denotes ergodic MCMC samples from the posterior distribution $p(\lambda_i|\mathbf{z})$ and $\lambda_{i_{(j)}}$ is the jth ordered statistic, an HPD region will be obtained for λ_i as below:

$$R_\alpha^i(n) = (\lambda_{i_{(k^*)}}, \lambda_{i_{(k^*+[(1-\alpha)n])}}), \tag{31}$$

where $[(1-\alpha)n]$ denotes the integer part of $(1-\alpha)n$ and k^* is selected in a way that:

$$\lambda_{i_{(k^*+[(1-\alpha)n])}} - \lambda_{i_{(k^*)}} = \min_{1\le k\le n-[(1-\alpha)n]} (\lambda_{i_{(k+[(1-\alpha)n])}} - \lambda_{i_{(k)}}). \tag{32}$$

4. Simulation study

Information contained in data on certain identifiable parameters is often very limited. Sometimes such parameters can be poorly estimated with practically attainable sample sizes, which can substantially affect the estimates of parameters of primary interest. The SUGLG model introduces the extra parameters α and ν beyond the parametrization of the usual Gaussian model, so it is natural to examine to what extent information on these parameters can be recovered from data. Also, the mixing process have been added in the skew-Gaussian spatial process to account for heavy tails. Therefore, the other fundamental question is about the capability of model to identify outliers. In this section, we first use simulation to assess the identifiability of the parameters and the ability of the SUGLG model to correctly identify outlying observations. Notice that inferences are based on the *SAEM* algorithm with 100 iterations and $K = 80$. Also, the total number of Monte Carlo simulations performed at each iteration was chosen to be $l = 500$.

To address parameters identifiability, we generate data from our SUGLG model with a constant mean surface and spatial sampling points were considered on regular lattice of the unit square. We focus on the extra parameters α and ν, because we would expect inference to be most challenging for these parameters. Throughout, we use a sample size of $n = 100$ with $\beta = 0$, $\sigma^2 = 1$, $\tau^2 = 0.1$, $\phi = 1$ and $\gamma = 0.2$. To consider identifiability of α, four data sets were generated with $\alpha = 0.01$, 1, 1.5 and 2. Then, the maximum likelihood of the model parameters are obtained via the SAEM algorithm. Table 1 displays the estimations for α. This table clearly indicates that the data allow for meaningful inference on α, even with this

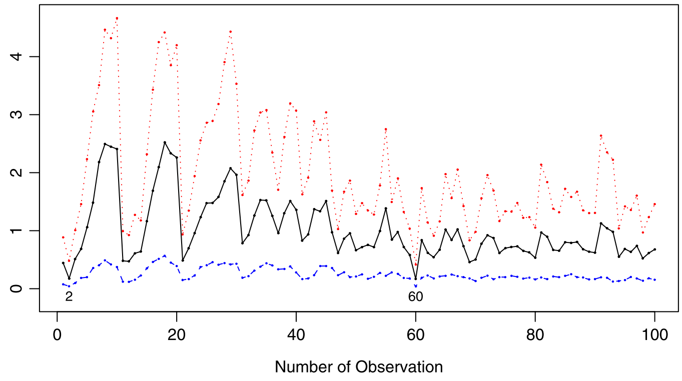

Fig. 1. The HPD regions and the estimated values of $E(\lambda_i|\mathbf{z})$, $i = 1, \ldots, 100$, under the SUGLG model. Solid line: $E(\lambda|\mathbf{z})$. Dotted line: lower bound. Dashed line: upper bound.

Table 1
Identifiability of two parameters α and ν.

α	True value	0.01	1	1.5	2
	Estimated value	0.09	1.13	1.37	1.86
ν	True value	0.1	1	2	3
	Estimated value	0.06	1.12	1.68	2.87

Table 2
Bias and root of mean square error of parameter estimates under various models.

Parameter	True value	Gaussian		GLG		SUN		SUGLG	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
β_0	0	−0.148	0.196	−0.109	0.098	0.088	0.052	−0.053	0.034
β_1	0.1	−0.217	0.147	0.086	0.072	−0.231	0.189	−0.084	0.061
β_2	−0.1	0.193	0.101	0.046	0.032	0.053	0.094	0.022	0.016
σ^2	1	8.217	6.526	3.455	2.058	6.281	5.409	1.011	0.093
τ^2	0.1	2.570	2.235	1.893	1.692	1.449	1.608	0.974	0.102
ϕ	1	0.796	0.881	0.413	0.247	0.421	0.239	0.128	0.0179
γ	0.2	0.299	0.138	0.121	0.092	0.141	0.075	0.099	0.069
α	2	–	–	–	–	1.06	1.033	0.346	0.288
ν	2	–	–	0.630	0.715	–	–	−0.343	0.277

quite moderate sample size. The same applies for inference on ν presented in Table 1 for four data sets generated with $\nu = 0.1$, 1, 2 and 3.

We also investigated the potential of the SUGLG model to identify outliers. For this, we generate $n = 100$ data points based on the model (4) and (5) with $\beta = 0, \sigma^2 = 1, \alpha = 3, \tau^2 = 0.1, \phi = 1$ and $\gamma = 0.2$. We then select two observations (2 and 60). Location 2 is contaminated subtracting 2 units from the simulated data and location 60 is modified adding 2 units to the simulated data. Fig. 1 presents the 95% HPD regions and the posterior mean values of λ_i, $i = 1, \ldots, 100$ under the SUGLG model. We came to the conclusion by this figure that two observations 2 and 60 have the smallest posterior mean values of mixing variables and very small HPD intervals.

Now, in order to examine the performance of the proposed model, we simulated 50 data sets from the SUGLG model with $\sigma^2 = 1, \tau^2 = 0.1, \phi = 1, \gamma = 0.2, \alpha = 2, \nu = 2$. In addition, assuming that $\mathbf{s} = (s_1, s_2)$, we consider the mean function $\beta_0 + \beta_1 s_1 + \beta_2 s_2$ with $\boldsymbol{\beta} = (0, 0.1, -0.1)$. To compare four models Gaussian, GLG, SUN and SUGLG, we computed the average of bias (Bias) and the root of mean-square error (RMSE) of the parameters, calculated in the following way for a parameter κ

$$Bias(\hat{\kappa}) = \bar{\hat{\kappa}} - \kappa, \qquad RMSE(\hat{\kappa}) = \sqrt{\frac{1}{50}\sum_{i=1}^{50}(\hat{\kappa}_i - \kappa)^2},$$

where $\hat{\kappa}_i$ is the estimation of κ from simulated data set i, and $\bar{\hat{\kappa}} = \frac{1}{50}\sum_{i=1}^{50}\hat{\kappa}_i$. The results are summarized in Table 2. The SUGLG model has the best performance, in the sense that the estimates under this model are most appropriate than those of the other models. The bias and RMSE measures for both σ^2 and τ^2 are noticeably large under the three other models. Although the value of Bias is of course large under our model, it can be justified with respect to under estimation for ν and over estimation for ϕ and γ. We conclude that the proposed method gives satisfactory results.

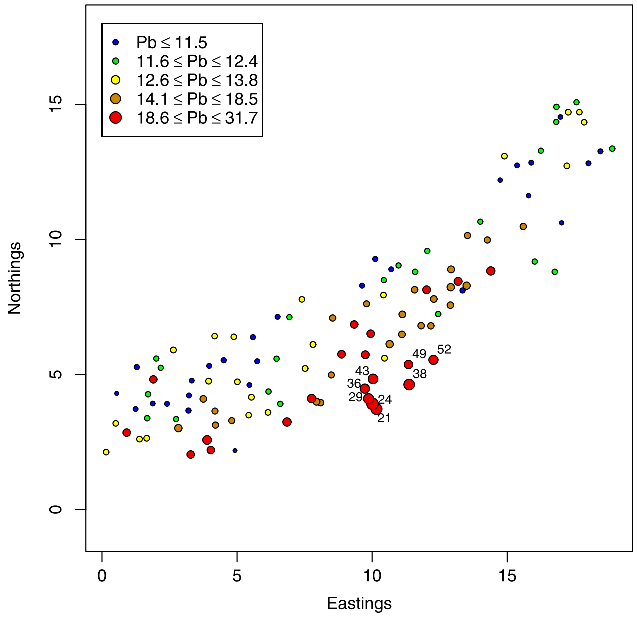

Fig. 2. Sampling locations of the Pb data.

Table 3
Maximum likelihood estimates of the SUGLG model parameters for the Pb data.

Parameter	α	$\boldsymbol{\beta}$	σ^2	τ^2	ϕ	γ	ν
Estimation	1	12.3	4.3	1.7	1.2	0.73	1.5

5. Illustrative example

The analyzed data set contains a total of 117 samples that were collected for mapping the Pb-contaminated areas in soils of a region of north Iran. Sampling locations are shown in Fig. 2. Also, for exploratory purpose, the histogram, ignoring sampling locations, and the highly robust empirical semivariogram (see [16]) of data are plotted in Fig. 3. The histogram suggests that the data have moderately right-skewed distribution. Furthermore, it is clear from the highly robust empirical semivariograms that there exists a strong spatial correlation as well as a nugget effect in the data set. Since explanatory analysis of the data did not show any significant relation between Pb and the spatial locations, the mean function is assumed to be constant, so $k = 1$. At each iteration of the SAEM algorithm, the total number of Monte Carlo simulations performed was chosen to be $l = 500$. Also, the convergence of the algorithm is then monitored by plotting the parameter values at each iteration versus the iteration number. The algorithm was stopped when iterates appeared to fluctuate randomly. Therefore, algorithm is implemented with 200 iterations and $K = 150$. Under these assumptions for the SAEM algorithm, we first compare the influence of the presence of a nugget effect in the SUGLG model through Bayes factor (see [19]). For this, the marginal likelihood of data for the considered SUGLG models was computed using the modified harmonic mean estimator $\hat{p}_4$ of [25]. The Bayes factor in favor of the SUGLG model with the nugget effect versus the SUGLG model without the nugget effect was evaluated as 107 that indicates the data favor the presence of the nugget effect. The estimation of model parameters under the SUGLG model are presented in Table 3. This table clearly indicates the existence of strong spatial correlation. To assess the predictive performance of the SUGLG model and compare it to those of three other models (i.e. Gaussian, SUN and GLG), we use a cross-validation approach based on single-point deletion predictive distribution, as described by Gelfand et al. [14]. We use $MSPR = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{z}_i)^2$ criterion, where z_i is the observed value at location $\mathbf{s}_i$ and $\hat{z}_i$ is the predicted value. MSPR of corresponding the SUGLG, GLG, SUN and Gaussian models are equal to 10.6, 15.1, 17.2 and 21.6, respectively. Thus, the SUGLG model evidently outperforms three other models. As another criterion, we computed Bayes factors between models. The Bayes factor in favor of the SUGLG versus the Gaussian, GLG and SUN models was evaluated as 5.9×10^{27}, 2.8×10^{20} and 3.2×10^{6}, respectively, that indicated overwhelming support for the SUGLG model.

Fig. 4 presents the 95% HPD regions and the posterior mean values of $\lambda_i, i = 1, \ldots, 117$ under the SUGLG model. According to this figure, the smallest posterior mean values of mixing variables were found to correspond to observations 21, 24, 29, 36, 38, 43, 49 and 52 (see Fig. 2 to observe the exact locations). As seen, these mentioned observations are clustered together, so they belong to a region with larger observational variance relative to the rest. Furthermore, the length of HPD interval is considerably smaller for these observations.

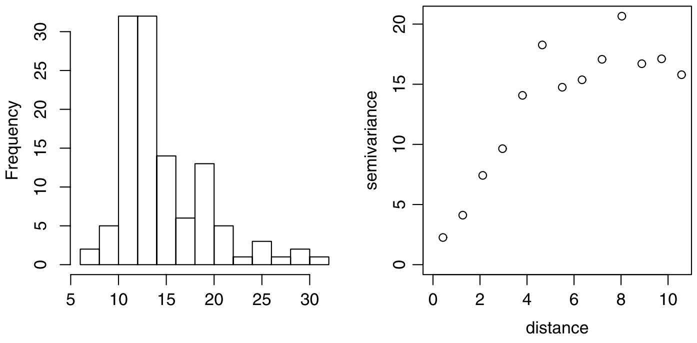

Fig. 3. Histograms (left column) and robust empirical semivariograms (right column) from the simple regression for plumb data.

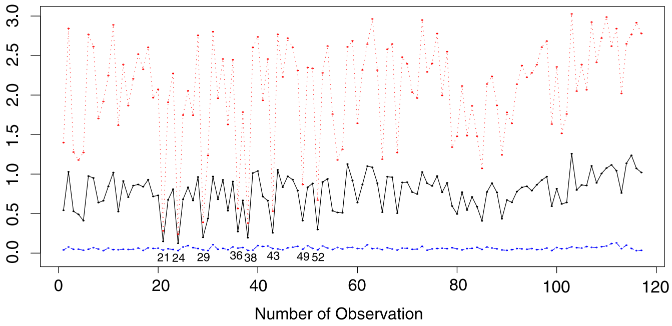

Fig. 4. The HPD regions and the estimated values of $E(\lambda_i|\mathbf{z})$, $i = 1, \ldots, 100$, under SUGLG model. Solid line: $E(\lambda|\mathbf{z})$. Dotted line: lower bound. Dashed line: upper bound.

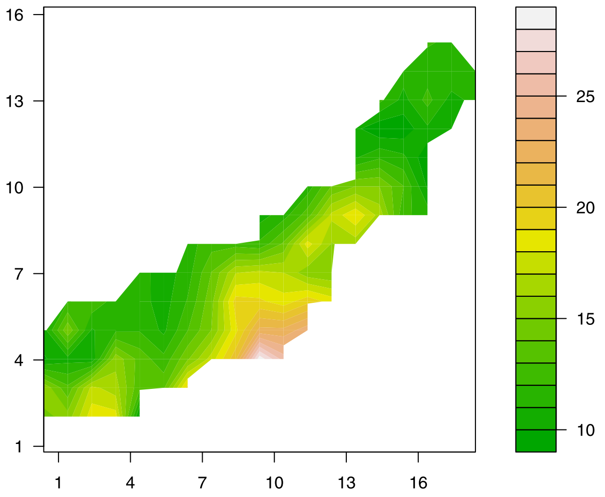

Fig. 5. The contour map corresponding to the predictive mean for the Pb data.

Finally, the contour map corresponding to the predictive mean under the SUGLG case, is shown in Fig. 5. According to this figure, the predictions are highest in the area of inflated variance that contains the aforementioned observations.

6. Conclusions

The unified skew Gaussian-log Gaussian (SUGLG) model presented in this paper provides a new approach to account for both skewness and heavy tails which are two pervasive features of the spatial data. We developed a likelihood-based approach for the inference and a SAEM algorithm for estimating the model parameters. Using simulated data, meaningful inference on parameters was concluded. Also, the simulation results indicate that regions in space with relatively large observational variance can appropriately identified utilizing the proposed model. The numerical example provides a useful tool for illustrating our methodology and the results showed that the Pb data clearly supports our model compared to the three other models.

Finally, it is worth noting that the model considered in this work has the potential to be used in other model frameworks, such as the generalized linear geostatistical models and the spatiotemporal data. Any further development on these issues will be very interesting.

Acknowledgments

The Associate Editor and two referees are gratefully acknowledged. Their precise comments and constructive suggestions have clearly improved the manuscript. We also thank Håvard Rue and Khalil Shafie Holighi for editing the English of this manuscript.

Appendix

We offer a method to draw a sample from the uniform distribution on $\{a_t < \log p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)})\} \bigcap \{b_t < \log p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta})\}$. At first, we have

$$a_t < \log p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)}) \Leftrightarrow (\boldsymbol{\epsilon} - \Lambda^{-\frac{1}{2}}\mathbf{w}^{(t)})'(\boldsymbol{\epsilon} - \Lambda^{-\frac{1}{2}}\mathbf{w}^{(t)}) < a_t^*,$$
$$b_t < \log p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta}) \Leftrightarrow \left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right)' C_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right) < b_t^*, \tag{33}$$

where $\mathbf{w}^{(t)} = \mathbf{u}^{(t)} + \mathbf{v}^{(t)}$, $a_t^* = (\boldsymbol{\epsilon} - \Lambda^{(t)^{-\frac{1}{2}}}\mathbf{w}^{(t)})'(\boldsymbol{\epsilon} - \Lambda^{(t)^{-\frac{1}{2}}}\mathbf{w}^{(t)}) + 2\tau^2 e_1^{(t+1)}$ and $b_t^* = (\boldsymbol{\eta}^{(t)} + \frac{\nu}{2}\mathbf{1}_n)'C_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\eta}^{(t)} + \frac{\nu}{2}\mathbf{1}_n) + 2\nu e_2^{(t+1)}$. Thus,

$$\{\boldsymbol{\eta}; a_t < \log p_{\boldsymbol{\psi}}(\mathbf{z}|\boldsymbol{\eta}, \mathbf{u}^{(t)}, \mathbf{v}^{(t)})\} \bigcap \{\boldsymbol{\eta}; b_t < \log p_{\nu,\boldsymbol{\theta}}(\boldsymbol{\eta})\}$$
$$= \{\boldsymbol{\eta} = \log(\boldsymbol{\lambda}); (\boldsymbol{\epsilon} - \Lambda^{-\frac{1}{2}}\mathbf{w}^{(t)})'(\boldsymbol{\epsilon} - \Lambda^{-\frac{1}{2}}\mathbf{w}^{(t)}) < a_t^*\} \bigcap \left\{\boldsymbol{\eta}; \left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right)' C_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right) < b_t^*\right\}.$$

Now, we define

$$I_i(\boldsymbol{\eta}_{-i}^{(t)}) = \left\{\eta^* \in \Re; \left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right)' C_{\boldsymbol{\theta}}^{-1}\left(\boldsymbol{\eta} + \frac{\nu}{2}\mathbf{1}_n\right) < b_t^* \text{ if } \boldsymbol{\eta} = (\eta_1^{(t)}, \ldots, \eta_{i-1}^{(t)}, \eta^*, \eta_{i+1}^{(t)}, \ldots, \eta_n^{(t)})\right\}.$$

Then, $I_i(\boldsymbol{\eta}_{-i}^{(t)})$ contains all possible values of the ith coordinates in order for $\boldsymbol{\eta}$ to remain in the n dimensional oval while the other $n-1$ coordinates are fixed. Clearly, $I_i(\boldsymbol{\eta}_{-i}^{(t)})$ is a non-empty interval because $\eta_i^{(t)} \in I_i(\boldsymbol{\eta}_{-i}^{(t)})$. Let $(\underline{\mathbf{e}}_1, \ldots, \underline{\mathbf{e}}_n)$ and $(l_1, \ldots, l_n)$ are eigenvectors and eigenvalues of $C_{\boldsymbol{\theta}}$, respectively, where $\underline{\mathbf{e}}_j = (e_{1j}, \ldots, e_{nj})'$. It follows, after some algebra, that if

$$a_1 = \sum_{j=1}^{n} \frac{1}{l_j} e_{ij}^2,$$
$$a_2 = 2\sum_{j=1}^{n}\sum_{\substack{k=1\\k\neq i}}^{n} \frac{e_{ij}}{l_j} e_{kj}\left(\eta_k^{(t)} + \frac{\nu}{2}\right),$$
$$a_3 = \left(\eta_i^{(t)} + \frac{\nu}{2}\right) a_2 + \left(\eta_i^{(t)} + \frac{\nu}{2}\right)^2 a_1 + 2\nu e_2^{(t+1)},$$
$$f_1 = \frac{-a_2 + \sqrt{a_2^2 + 4a_1a_3}}{2a_1},$$
$$f_2 = \frac{-a_2 - \sqrt{a_2^2 + 4a_1a_3}}{2a_1},$$

then

$$I_i(\boldsymbol{\eta}_{-i}^{(t)}) = \left\{\eta^* \in \Re;\ \min\{f_1, f_2\} < \eta^* + \frac{\nu}{2} < \max\{f_1, f_2\}\right\}. \tag{34}$$

Also, let

$$J_i(\boldsymbol{\eta}_{-i}^{(t)}) = \Big\{\eta^* \in \Re^+;\ (\boldsymbol{\epsilon} - \Lambda^{-\frac{1}{2}}\mathbf{w}^{(t)})'(\boldsymbol{\epsilon} - \Lambda^{-\frac{1}{2}}\mathbf{w}^{(t)}) < a_t^* \text{ if}$$
$$\Lambda = diag(\exp(\eta_1^{(t)}), \ldots, \exp(\eta_{i-1}^{(t)}), \exp(\eta^*), \exp(\eta_{i+1}^{(t)}), \ldots, \exp(\eta_n^{(t)}))\Big\}.$$

Clearly, $J_i(\boldsymbol{\eta}^{(t)}_{-i})$ is also a non-empty interval because $\eta_i^{(t)} \in J_i(\boldsymbol{\eta}^{(t)}_{-i})$. Now, if $\xi_2 = \xi_1 - \epsilon_i^2 \geq 0$, where $\xi_1 = (\epsilon_i - \frac{w_i^{(t)}}{\sqrt{\lambda_i^{(t)}}})^2 + 2\tau^2 e_1^{(t+1)}$ then

$$J_i(\boldsymbol{\eta}^{(t)}_{-i}) = \left\{ \eta^* \in \Re^+;\ \eta^* > \log\left(\frac{|w_i^{(t)}|\sqrt{\xi_1} - \epsilon_i w_i^{(t)}}{\xi_2} \right)^2 \right\}. \tag{35}$$

Also, when $\xi_2 < 0$, we have

$$J_i(\boldsymbol{\eta}^{(t)}_{-i}) = \left\{ \eta^* \in \Re^+;\ \log\left(\frac{|w_i^{(t)}|\sqrt{\xi_1} - \epsilon_i w_i^{(t)}}{\xi_2} \right)^2 < \eta^* < \log\left(\frac{|w_i^{(t)}|\sqrt{\xi_1} + \epsilon_i w_i^{(t)}}{-\xi_2} \right)^2 \right\}. \tag{36}$$

Thus, under (34), (35) and (36), ith element $\boldsymbol{\eta}^{(t+1)}$ can be generated uniformly on the interval $I_i(\boldsymbol{\eta}^{(t)}_{-i}) \bigcap J_i(\boldsymbol{\eta}^{(t)}_{-i})$.

References

[1] D.K. Agarwal, A.E. Gelfand, Slice sampling for simulation based fitting of spatial data models, Statistics and Computing 15 (1) (2005) 61–69.

[2] D. Allard, P. Naveau, A new spatial skew-normal random field model, Communications in Statistics 36 (2007) 1821–1834.

[3] R.B. Arellano-Valle, A. Azzalini, On the unification of families of skew-normal distributions, Scandinavian Journal of Statistics 33 (2006) 561–574.

[4] A. Azzalini, The skew-normal distribution and related multivariate families, Scandinavian Journal of Statistics 32 (2005) 159–188.

[5] A. Azzalini, A. Capitanio, Statistical applications of the multivariate skew normal distribution, Journal of the Royal Statistical Association 61 (1999) 579–602.

[6] V. Cancho, D. Dey, V. Lachos, M. Andrade, Bayesian nonlinear regression models with scale mixtures of skew normal distributions: Estimation and case influence diagnostics, Computational Statistics and Data Analysis 55 (2011) 588602.

[7] M.H. Chen, Q.M. Shao, Monte Carlo estimation of Bayesian credible intervals and HPD intervals, Journal of Computational and Graphical Statistics 8 (1) (1999) 69–92.

[8] B. Delyon, M. Lavielle, E. Moulines, Convergence of a stochastic approximation version of the EM algorithm, Annals of Statistics 27 (1999) 94–128.

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

[10] J. Dominguez-Molina, G. Gonzalez-Farias, A.K. Gupta, 2003. The multivariate closed skew-normal distribution. Technical Report 03-12. Department of Mathematics and Statistics, Bowling Green State University.

[11] T. Fonseca, M. Steel, Non-Gaussian spatiotemporal modelling through scale mixing, Biometrika 98 (2011) 761–774.

[12] A. Garay, V. Lachos, C. Abanto-Valle, Nonlinear regression models based on scale mixtures of skew-normal distributions, Journal of the Korean Statistical 40 (2011) 115–124.

[13] C.R.B. Cabral, V.H. Lachos, M.R. Madruga, Bayesian analysis of skew-normal independent linear mixed models with heterogeneity in the random-effects population, Journal of Statistical Planning and Inference 142 (2012) 181–200.

[14] A.E. Gelfand, D.K. Dey, H. Chang, Model determination using predictive distributions with implementation via sampling-based methods, Bayesian Statistics 4 (1992) 147167.

[15] M.G. Genton, Skew-Elliptical Distributions and their Applications, Chapman & Hall, Boca Raton, FL, 2004.

[16] M.G. Genton, Highly robust variogram estimation, Math. Geology 30 (1998) 213–221.

[17] M.G. Genton, H. Zhang, 2012 Identifiability problems in some non-Gaussian spatial random fields. Chilean Journal of Statistics (in press).

[18] G. Gonzalez-Farias, J. Dominguez-Molina, A. Gupta, The closed skew-normal distribution, in: M. Genton (Ed.), Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality, Chapman Hall/CRC, Boca Raton, FL, 2004, pp. 25–42.

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

[20] H.M. Kim, M. Genton, Characteristic functions of scale mixtures of multivariate skew-normal distributions, Journal of Multivariate Analysis 102 (2011) 1105–1117.

[21] H. Kim, B.K. Mallick, A Bayesian prediction using the skew-Gaussian processes, Journal of Statistical Planning and Inference 120 (2004) 85–101.

[22] E. Kuhn, M. Lavielle, Coupling a stochastic approximation version of EM with a MCMC procedure, ESAIM: Probability and Statistics 8 (2004) 115–131.

[23] V.H. Lachos, D.K. Dey, V.G. Cancho, Robust linear mixed models with skew-normal independent distributions from a Bayesian perspective, Journal of Statistical Planning and Inference 139 (2009) 40984110.

[24] R. Neal, Slice sampling, Annals of Statistics 31 (2003) 705–767.

[25] M.A. Newton, A.E. Raftery, Approximate Bayesian Inference by the Weighted Likelihood Bootstrap(with discussion), Journal of the Royal Statistical Association 3 (1994) 3–48.

[26] M. Palacios, M. Steel, Non-Gaussian Bayesian geostatistical modeling, Journal of the American Statistical Association 101 (2006) 604–618.

[27] H. Zareifard, M. Jafari Khaledi, 2012, Empirical Bayes estimation in regression models with generalized skew-slash errors. Communications in Statistics: Theory and Methods (in press).

[28] H. Zhang, A. El-Shaarawi, On spatial skew-Gaussian processes and applications, Environmetrics 21 (2010) 33–47.