Spatial Statistics 35 (2020) 100398

Contents lists available at ScienceDirect

Spatial Statistics

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

A hierarchical bi-resolution spatial skew-*t* model☆

Felipe Tagle [a], Stefano Castruccio [a,*], Marc G. Genton [b]

[a] *Department of Applied and Computational Mathematics and Statistics, University of Notre Dame, Notre Dame, IN 46556, United States*

[b] *Statistics Program, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia*

ARTICLE INFO

Article history:
Received 11 April 2019
Received in revised form 19 November 2019
Accepted 27 November 2019
Available online 9 December 2019

Keywords:
EM algorithm
Hierarchical model
Latent process
Non-Gaussian distribution
Skew-*t* distribution
Spatial statistics

ABSTRACT

Advances in Gaussian methodology for spatio-temporal data have made it possible to develop sophisticated non-stationary models for very large data sets. The literature on non-Gaussian spatio-temporal models is comparably sparser and strongly focused on distributing the uncertainty across layers of a hierarchical model. This choice allows to model the data conditionally, to transfer the dependence structure at the process level via a link function, and to use the familiar Gaussian framework. Conditional modeling, however, implies an (unconditional) distribution function that can only be obtained through integration of the latent process, with a closed form only in special cases. In this work, we present a spatio-temporal non-Gaussian model that assumes an (unconditional) skew-*t* data distribution, but also allows for a hierarchical representation by defining the model as the sum of a small and a large scale spatial latent effect. We provide semi-closed form expressions for the steps of the Expectation–Maximization algorithm for inference, as well as the conditional distribution for spatial prediction. We demonstrate how it outperforms a Gaussian model in a simulation study, and show an example of application to precipitation data in Colorado.

☆ This publication is based upon work supported by the King Abdullah University of Science and Technology (KAUST), Saudi Arabia Office of Sponsored Research (OSR) under Award No: OSR-2015-CRG4-2640.

* Corresponding author.
E-mail addresses: ftagleso@nd.edu (F. Tagle), scastruc@nd.edu (S. Castruccio), marc.genton@kaust.edu.sa (M.G. Genton).

https://doi.org/10.1016/j.spasta.2019.100398
2211-6753/

1. Introduction

Gaussian models have always been at the center of investigations in spatio-temporal statistics, as the normal assumption implies an unparalleled degree of analytical tractability and computational convenience. In many applications, however, the data often exhibit features that are inconsistent with this assumption, and therefore require more flexible non-Gaussian models.

Different solutions for non-Gaussian, spatio-temporal data are available depending on the extent of the departure from the normal assumption. When the non-Gaussianity is not too severe, the simplest and more widespread solution is a non-linear transformation of the data that modifies the marginal distribution to resemble the normal distribution, allowing for the subsequent use of the traditional methodology (Cressie, 1993). These *trans-Gaussian* random fields consist of a simple, marginal transformation of the data, such as a logarithm, and sometimes allow to explicitly express the degree of skewness and kurtosis in the model; see Xu and Genton (2017) and references therein.

For severe departures from Gaussianity (e.g., count data), a more sophisticated solution is to rely on a generalized linear model framework that assumes data independence conditionally on a process, while the process is assumed to be Gaussian and captures the spatio-temporal dependence. This *model-based geostatistics* approach (Diggle et al., 1998), whose cornerstone principle is the stratification of the spatio-temporal uncertainty across layers of a (Bayesian) hierarchical model, has recently seen a surge in popularity with the advent of the Integrated Nested Laplace Approximation (INLA, Rue et al. (2009)) and the R-INLA project (Lindgren and Rue, 2015). In recent years, the INLA approach has received a strong interest from the statistical community (and beyond), because of its ability to sparsify the conditional dependence structure and therefore to achieve inference for large spatial data. Although this modeling approach has arguably many advantages, it suffers, by design, of a lack of explicit parametrizations of the non-Gaussian features. Since the focus of the modeling is on a latent process and the data are only modeled conditionally, the (unconditional) data distribution can be expressed only through integrals, by marginalizing over the latent process as well as the prior uncertainty of its hyperparameters (in the case of a Bayesian model). These integrals rarely have a closed form, and hence hamper model interpretability should one be interested in the unconditional properties of the model.

In this work, we propose a new non-Gaussian model focused on the unconditional properties of the data, which also allows for a convenient hierarchical representation for inference purposes. Therefore, we retain the flexibility, as well as (some of) the computational convenience of hierarchical models while also achieving an explicit expression of the moments, hence improving the model interpretability. The model relies on a domain partition into regions, where a common latent process across each region is assumed and estimated. Domain partition in spatio-temporal statistics is a strategy typically employed as a means to perform approximate inference. Indeed, when the data set is too large the likelihood is often misspecified through a composite approach. In the case of non-Gaussian models such as max-stable processes, composite likelihood approaches are necessary as full likelihood inference for more than eleven points is impossible (Castruccio et al., 2016). In this work, the choice of partitioning was not dictated by computational needs, and instead of misspecifying a model we opted for defining a model with a partition, and defined a global model to allow non-Gaussian modeling in space with explicit expression of the unconditional properties of the process.

The model relies on a perturbation of the Gaussian distribution, introduced by Azzalini (1985) in the univariate case: the skew-normal. This distribution generalizes the normal distribution by explicitly accounting for the skewness with a parameter; it has been actively studied and developed for over three decades. Azzalini and Dalla Valle (1996) proposed a multivariate extension, whereas Azzalini and Capitanio (2003) proposed an equivalent formulation for the Student's t distribution. A full review can be found in Genton (2004) and most recently in Azzalini and Capitanio (2014). While some work has focused on extending these multivariate distributions to processes (Zhang and El-Shaarawi, 2010; Mahmoudian, 2017; Bevilacqua et al., 2018; Kim and Mallick, 2004; Kim et al., 2004; Allard and Naveau, 2007; Schmidt et al., 2017), the existence of a process is not guaranteed in general (Genton and Zhang, 2012; Minozzo and Ferracuti, 2012). Skewed distributions cannot be obtained from a central limit theorem or any other large sample

results, they are obtained by perturbing the normal distribution to account for some degree of skewness and kurtosis, and they also emerge naturally in selective population sampling, when a latent variable is involved in the sampling and does not allow a completely random sample (Arellano-Valle et al., 2006).

Skew-t models have been widely used for applications involving multivariate and sometimes spatial data, but to our knowledge, no work has highlighted the potential of this distribution as a baseline for establishing a spatial hierarchical model. Here we show how this reformulation allows for a bi-resolution model with a partition of the spatial domain into multiple regions, each one with a different stationary behavior, linked with a latent large-scale spatial effect.

We perform a frequentist inference, and develop a Monte Carlo Expectation–Maximization (MCEM) algorithm, a choice also used in the spatial skew-normal model developed by Zhang and El-Shaarawi (2010). MCEM is also often used for inference involving skewed distributions in a non-spatial context (e.g., Lin and Lee (2008), Lin (2010), Lachos et al. (2010); see Lachos et al. (2018) for a review). In this work, we derive the semi-explicit expressions of the E and M step, as well as a closed form of the conditional distribution at unobserved locations, in order to allow spatial prediction.

The remainder of the paper is organized as follows: Section 2 introduces the model, Section 3 presents the E and the M step of the algorithm for inference, as well as the conditional distribution at unsampled locations for prediction. Section 4 presents a simulation study where the proposed model is compared with a Gaussian model. Section 5 shows an application to Colorado precipitation data. Section 6 concludes with a discussion.

2. Model definition

For this section, we assume we have a purely spatial process. We consider a scalar, non-Gaussian random field $\{Y(\mathbf{s}), \mathbf{s} \in D\}$, where $D \subset \mathbb{R}^2$, and a partition $D = \cup_{r=1}^{R} \mathcal{D}_r$ such that $\mathcal{D}_r \cap \mathcal{D}_{r'} = \emptyset$ if $r \neq r'$. The partition should be determined by areas where higher moment characteristics are similar, as we will show in the application. We propose a model for $Y(\mathbf{s})$ for which the process, within each region, is multivariate skew-t, with the parameterization of Azzalini and Capitanio (2003).

For each region $\mathcal{D}_r$, and for each point $\mathbf{s} \in \mathcal{D}_r$, the following model applies

$$Y(\mathbf{s}) = \sigma_r \frac{\rho_r U_{0,r} + \lambda_r U_{1,r} + \eta_r(\mathbf{s})}{\sqrt{Z_r}}, \tag{1}$$

where $U_{1,r}$ has a standard half-normal distribution, Z_r a Gamma$(\nu_r/2, \nu_r/2)$ distribution, where the first argument is the scale, while the second the rate, and $\eta_r(\mathbf{s})$ is a stationary Gaussian process independent across r, with mean zero and correlation function that depends on parameters $\boldsymbol{\psi}_r$, with an associated correlation matrix $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}(\boldsymbol{\psi}_r)$; and $\sigma_r \geq 0$, $\rho_r \geq 0$, $\lambda_r \in \mathbb{R}$. Setting $\rho_r = 0$ reduces the model to that of Tagle et al. (2019), in which each region has a multivariate skew-t distribution, but evolves independently from the others. Here, we introduce the random vector $\mathbf{U}_0 = (U_{0,1}, \ldots, U_{0,R})^\top$, assumed to follow a mean zero multivariate normal distribution with correlation matrix $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}(\boldsymbol{\psi}_0)$, with each component being assigned to each region, with the aim of inducing an inter-regional spatial dependence. The vector $\mathbf{U}_0$ hence takes on the role of a large-scale effect, with components interacting additively with the other terms in the numerator and capturing the fine-scale variability, jointly resulting in a multivariate skew-normal distribution. A location parameter could be added in model (1), and in fact the application in Section 5 allows for a changing latent parameter for each region. However, for the ease of analytical tractability and computational convenience in the simulation studies, we chose not to allow for this parameter at this stage.

Model (1) assumes that, conditional on $U_{0,r}$ and $U_{1,r}$, points across different regions are independent, and hence simulations from this model will have discontinuities at the boundaries and there is conditional independence among points for different regions. While this is arguably a suboptimal feature of the model globally this can however be mitigated by constraining the latent process $U_{0,r}$ to be very smooth, by fixing the smoothness of the corresponding covariance matrix $\boldsymbol{\Sigma}_0$.

The additive form ensures that the numerator of (1) remains a multivariate skew-normal distribution given the closure of the skew-normal distribution under convolutions with a normal random variable (Azzalini and Capitanio, 2014). It is straightforward to show that, for a collection of points $\mathbf{s}_1, \ldots, \mathbf{s}_n \in \mathcal{D}_r$, the numerator has a skew-normal distribution with a probability density function (pdf) given by

$$\rho_r U_{0,r} + \lambda_r U_{1,r} + \eta_r(\mathbf{s}) \sim 2\left(\sqrt{1+\lambda_r^2+\rho_r^2}\right)\phi_n(\mathbf{y}; \boldsymbol{\Omega}_r)\, \Phi(\boldsymbol{\alpha}_r^\top \mathbf{y}), \tag{2}$$

where ϕ_n is the pdf of an n-dimensional normal distribution with mean zero and covariance $\boldsymbol{\Omega}_r$, and Φ is the cumulative distribution function of the univariate standard normal distribution. Furthermore,

$$\boldsymbol{\alpha}_r = \frac{\delta_r}{1-\delta_r^2} \frac{\mathbf{1}_n^\top \boldsymbol{\Omega}_r^{-1}}{\left(1 + \frac{\delta_r^2}{1-\delta_r^2}\mathbf{1}_n^\top \boldsymbol{\Omega}_r^{-1}\mathbf{1}_n\right)^{1/2}},$$

$$\boldsymbol{\Omega}_r = (1-\delta_r^2)\left(\boldsymbol{\Sigma}(\boldsymbol{\psi}_r)^* + \frac{\delta_r^2}{1-\delta_r^2}\mathbf{1}_n\mathbf{1}_n^\top\right),$$

$\boldsymbol{\Sigma}(\boldsymbol{\psi}_r)^*$ is a correlation matrix with terms $\boldsymbol{\Sigma}(\boldsymbol{\psi}_r)^*_{ij} = (\boldsymbol{\Sigma}(\boldsymbol{\psi}_r)_{ij} + \rho_r^2)/(1+\rho_r^2)$, and $\delta_r = \frac{\lambda_r}{\sqrt{1+\lambda_r^2+\rho_r^2}}$. Dividing by a Gamma distribution results in the multivariate skew-t distribution.

Let $\mathbf{s}$ be one point of the collection $\mathbf{s}_1, \ldots, \mathbf{s}_n$ with corresponding skewness parameter α drawn from the vector $\boldsymbol{\alpha}_r$. For simplicity of notation, we further drop the index r across all terms. The proposed skew-t formulation allows to explicitly compute the moments. From (1) and (2), we can write the kth moment as

$$\mathrm{E}[Y(\mathbf{s})^k] = \sigma^k(1+\lambda^2+\rho^2)^{k/2}\mathrm{E}[Z^{-k/2}]\mathrm{E}[Y^*(\mathbf{s})^k], \tag{3}$$

where $Y^*(\mathbf{s}) = \{\rho U_0 + \lambda U_1 + \eta(\mathbf{s})\}/\sqrt{1+\lambda^2+\rho^2}$ which has a standard skew-normal distribution.

Since Z is a Gamma distribution, it can be easily shown that

$$\mathrm{E}[Z^{-k/2}] = \frac{(\nu/2)^{k/2}\Gamma\{(\nu-k)/2\}}{\Gamma(\nu/2)},$$

whereas the kth moment of a standard skew-t distribution can be obtained from Azzalini and Capitanio (2014), pg. 32, as

$$\mathrm{E}[Y^*(\mathbf{s})^k] = \sqrt{\frac{2}{\pi}}\frac{\mathrm{sgn}(\alpha)}{\alpha^{k+1}}B_k(\alpha^{-2}),$$

where, for $h > 0$,

$$B_k(h) = \frac{k-1}{h}B_{k-2}(h) + \frac{\beta_{k-1}}{h(1+h)^{k/2}}, \quad k = 2, 3, \ldots$$

$$B_0(h) = \sqrt{\frac{\pi}{2h}}, \quad B_1(h) = \frac{1}{h\sqrt{1+h}},$$

and β_k is the kth moment of the standard normal distribution, $\beta_k = (k-1)!!$, for $k = 2, 4, 6, \ldots$ and 0 otherwise.

The kth moment formula (3) allows to obtain the mean and variance as

$$\mathrm{E}[Y(\mathbf{s})] = \sigma(1+\lambda^2+\rho^2)^{1/2}\frac{(\nu/2)^{1/2}\Gamma\{(\nu-1)/2\}}{\Gamma(\nu/2)}\mathrm{sgn}(\alpha)\sqrt{\frac{2}{\pi(1+\alpha^{-2})}},$$

$$\mathrm{Var}[Y(\mathbf{s})] = \sigma^2(1+\lambda^2+\rho^2)\mathrm{sgn}(\alpha) - \mathrm{E}[Y(\mathbf{s})]^2.$$

The covariance for points $\mathbf{s}_1$ and $\mathbf{s}_2$ belonging to the same region $\mathcal{D}_r$ can be expressed as

$$\mathrm{Cov}(Y(\mathbf{s}_1), Y(\mathbf{s}_2)) = \sigma^2\frac{\nu}{\nu-2}\left\{\rho^2 + \lambda^2 + C_{\boldsymbol{\psi}}(h)\right\} - \mathrm{E}[Y(\mathbf{s}_1)]^2, \tag{4}$$

where $C_{\boldsymbol{\psi}}(h)$ is the correlation function associated with the matrix $\boldsymbol{\Sigma}$. If, instead, $\mathbf{s}_1 \in \mathcal{D}_{r_1}$ and $\mathbf{s}_2 \in \mathcal{D}_{r_2}$ with $r_1 \neq r_2$

$$\text{Cov}(Y(\mathbf{s}_1), Y(\mathbf{s}_2)) = \sigma_{r_1}\sigma_{r_2}\rho_{r_1}\rho_{r_2}\frac{(\nu_{r_1}/2)^{1/2}\Gamma\{(\nu_{r_1}-1)/2\}}{\Gamma(\nu_{r_1}/2)}\frac{(\nu_{r_2}/2)^{1/2}\Gamma\{(\nu_{r_2}-1)/2\}}{\Gamma(\nu_{r_2}/2)}\boldsymbol{\Sigma}_{0;r_1,r_2}. \quad (5)$$

The representation in (1) involves the latent variables $U_{0,r}$, $U_{1,r}$ and Z_r, $r = 1, \ldots, R$, which we choose to estimate with an EM algorithm (Dempster et al., 1977). This approach aims to maximize the model likelihood in the presence of latent processes by alternating between an expectation step to estimate $\mathbf{U}_0$, and a maximization step to estimate all the model parameters, see McLachlan and Krishnan (2007) for details. During the first step, $(U_{0,r}, U_{1,r}, Z_r)^\top$, $r = 1, \ldots, R$ and functions thereof are replaced by their conditional expectations given the data and parameter estimates $\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1^\top, \ldots, \hat{\boldsymbol{\theta}}_R^\top, \hat{\boldsymbol{\psi}}_0^\top)^\top$, with $\boldsymbol{\theta}_r = (\sigma_r, \rho_r, \lambda_r, \boldsymbol{\psi}_r^\top)^\top$. In the second step, $\hat{\boldsymbol{\theta}}$ is updated in the traditional maximum-likelihood sense based on the maximization of the associated log-likelihood. Since a parameter set is required to conduct the first E-step, an initial set of values is provided, which is then subsequently updated by alternating between both steps until convergence is achieved.

Our choice of an EM approach was mostly computational and dictated by the availability of closed form expressions for some of the parameters during the Maximization step, as the next section shows. A full Bayesian model is a possible alternative, but we could not find any conjugate prior for the parameters, so we would have needed another MCMC step within the Gibbs sampling, thus adding additional strain to the already burdensome inference approach.

3. EM algorithm

Henceforth $\mathbf{Y}_r$, $r = 1, \ldots, R$ represents the vector of random variables corresponding to the observations in region $\mathcal{D}_r$, and $|\mathcal{D}_r| = n_r$, so that $\sum_{r=1}^R n_r = n$. The proposed model in (1) has the following hierarchical representation,

$$\begin{aligned}
\mathbf{Y}_r | U_{0,r} = u_{0,r}, U_{1,r} = u_{1,r}, Z_r = z_r &\overset{iid}{\sim} N_{n_r}\left(\tfrac{\sigma_r}{\sqrt{z_r}}(\rho_r u_{0,r} + \lambda_r u_{1,r})\mathbf{1}_{n_r}, \tfrac{\sigma_r^2}{z_r}\boldsymbol{\Sigma}_r\right),\\
\mathbf{U}_0 &\sim N_R(\mathbf{0}, \boldsymbol{\Sigma}_0),\\
U_{1,R} &\overset{iid}{\sim} HN(0, 1),\\
Z_r &\overset{iid}{\sim} \text{Gamma}(\nu_r/2, \nu_r/2),
\end{aligned}$$

where *HN* refers to the half-normal distribution, and $\mathbf{1}_{n_r}$ is a $n_r \times 1$ vector with all entries being equal to one. We now assume to have independent temporal replicates, as will be the case in the application. We thereby consider the vector $\mathbf{y}_t = (\mathbf{y}_{1,t}^\top, \ldots, \mathbf{y}_{R,t}^\top)^\top$, $t = 1, \ldots, T$, as well as $\mathbf{u}_{0,t}$, $\mathbf{u}_{1,t}$, and $\mathbf{z}_{1,t}$, which are defined conformably (in an abuse of notation, we henceforth use lowercase letters to denote both realizations and random quantities). Their joint distribution, for each t, follows from the above representation:

$$\begin{aligned}
p(\mathbf{y}_t, \mathbf{u}_{0,t}, \mathbf{u}_{1,t}, \mathbf{z}_t | \boldsymbol{\theta}) &= p(\mathbf{y}_t | \mathbf{u}_{0,t}, \mathbf{u}_{1,t}, \mathbf{z}_t, \boldsymbol{\theta}) \times p(\mathbf{u}_{0,t} | \boldsymbol{\theta}) \times p(\mathbf{u}_{1,t} | \boldsymbol{\theta}) \times p(\mathbf{z}_t | \boldsymbol{\theta})\\
&= \prod_{r=1}^R \frac{1}{(2\pi)^{n_r/2}\left|\sigma_r^2/z_r\boldsymbol{\Sigma}_r\right|^{1/2}} \exp\left\{-\frac{z_{r,t}}{2\sigma_r^2}\mathbf{x}_{r,t}^\top\boldsymbol{\Sigma}_r^{-1}\mathbf{x}_{r,t}\right\}\\
&\quad \times \frac{1}{(2\pi)^{R/2}|\boldsymbol{\Sigma}_0|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{u}_{0,t}^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{u}_{0,t}\right\} \times \frac{(\nu_r/2)^{\nu_r/2}}{\Gamma(\nu_r/2)} z_{r,t}^{\frac{\nu_r+n_r}{2}-1}\\
&\quad \times \sqrt{\frac{2}{\pi}} \exp\left\{-\frac{u_{1,r,t}^2}{2}\right\}
\end{aligned} \quad (6)$$

where $\mathbf{x}_{r,t} = \mathbf{y}_{r,t} - \sigma_r/\sqrt{z_{r,t}}(\rho_r u_{0,r,t} + \lambda_r u_{1,r,t})\mathbf{1}_{n_r}$. We can further aggregate the time-t vectors, as $\mathbf{y} = (\mathbf{y}_1^\top, \ldots, \mathbf{y}_T^\top)^\top$, with conformable definitions for the latent variables, allowing us to express the corresponding log-likelihood as

$$\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}) = \sum_{r=1}^R \ell_r(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}) - \frac{T}{2}\log|\Sigma_0| - \frac{1}{2}\sum_{t=1}^T \mathbf{u}_{0,t}^\top\boldsymbol{\Sigma}_0^{-1}\mathbf{u}_{0,t}, \quad (7)$$

where

$$\ell_r(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}) = T\frac{\nu_r}{2}\log\left(\frac{\nu_r}{2}\right) - T\log\Gamma\left(\frac{\nu_r}{2}\right) - \frac{Tn_r}{2}\log(\sigma_r) - \frac{T}{2}\log|\boldsymbol{\Sigma}_r| + \left(\frac{\nu_r + n_r}{2} - 1\right)\sum_{t=1}^{T}\log(Z_{r,t}) - \frac{1}{2}\sum_{t=1}^{T}\left(u_{1,r,t}^2 + \nu_r z_{r,t} + \frac{z_{r,t}}{\sigma_r^2}\mathbf{x}_{r,t}^\top\boldsymbol{\Sigma}_r^{-1}\mathbf{x}_{r,t}\right).$$

The log-likelihood involves the inversion of the matrices $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_0$, as well as the computation of their respective determinants, which are typically problematic when the size of the data set becomes large. However, our approach of regionalizing the spatial domain allows for a judicious choice of n_r to ensure the computation feasibility of said operations on $\boldsymbol{\Sigma}_r$, while the dimension of $\boldsymbol{\Sigma}_0$ is determined by the number of regions R for which $R \ll n$ is assumed to hold.

3.1. E-step

The E-step proceeds by computing the conditional expectation, commonly denoted by the Q function,

$$Q(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{[k]}) = \mathbb{E}\left[\ell(\boldsymbol{\theta}|\mathbf{y}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z})|\mathbf{y}, \hat{\boldsymbol{\theta}}^{[k]}\right]$$

based on the value of $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{[k]}$ at the kth iteration, in order to obtain a new estimate of the latent processes $\mathbf{u}_0$, $\mathbf{u}_1$ and $\mathbf{z}$. Technical details on how to simulate the full conditional of $\mathbf{z}$ and the closed form expressions for $\mathbf{u}_0$ and $\mathbf{u}_1$ are provided in the Supplementary Material.

3.2. M-step

Once the estimates for the latent processes $\hat{\mathbf{u}}_0^{[k]}$, $\hat{\mathbf{u}}_1^{[k]}$ and $\hat{\mathbf{z}}^{[k]}$ are available, they are plugged in (7) and a new parameter vector $\hat{\boldsymbol{\theta}}^{[k+1]}$ is obtained as

$$\hat{\boldsymbol{\theta}}^{[k+1]} = \arg\max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta}|\mathbf{y}, \hat{\mathbf{u}}_0^{[k]}, \hat{\mathbf{u}}_1^{[k]}, \hat{\mathbf{z}}^{[k]}).$$

Details of the maximization of each of the parameters, i.e. ν_r, σ_r, ρ_r, λ_r, $\boldsymbol{\Sigma}_r$ and $\boldsymbol{\Sigma}_0$, are provided in the Supplementary Material.

3.3. Prediction

We carry out prediction in the classical sense; first estimating the parameter set $\boldsymbol{\theta}$ based on a collection of observations and subsequently, making use of the plug-in predictive distribution that treats the estimated parameters as truth to generate predictions at another set of locations (Diggle and Ribeiro, 2007). Let us first fix some notation. Let $\mathbf{y} = (\mathbf{y}_{\text{pr}}^\top, \mathbf{y}_{\text{ob}}^\top)^\top$, denote the vector that stacks the vector of desired predictions $\mathbf{y}_{\text{pr}}$ at n_{pr} locations, and a vector of $\mathbf{y}_{\text{ob}}$ observations at n_{ob} locations. The plug-in predictive distribution of $\mathbf{y}_{\text{pr}}$ is then given by

$$p(\mathbf{y}_{\text{pr}}|\mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}}) = \int p(\mathbf{y}_{\text{pr}}|\mathbf{y}_{\text{ob}}, \mathbf{u}_0, \mathbf{u}_1, \mathbf{z}, \hat{\boldsymbol{\theta}})p(\mathbf{u}_0, \mathbf{u}_1, \mathbf{z}|\mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}})\mathrm{d}\mathbf{u}_0\mathrm{d}\mathbf{u}_1\mathrm{d}\mathbf{z}.$$

Denote with $\boldsymbol{\Sigma}$ the covariance matrix of $\mathbf{y}$, and its partition for the observed and predicted part of the model

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{\text{ob,ob}} & \boldsymbol{\Sigma}_{\text{ob,pr}} \\ \boldsymbol{\Sigma}_{\text{pr,ob}} & \boldsymbol{\Sigma}_{\text{pr,pr}} \end{pmatrix}.$$

Applying standard results from multivariate normal theory we obtain

$$p(\mathbf{y}_{\text{pr}}|\mathbf{y}_o, u_0, u_1, z, \hat{\boldsymbol{\theta}}) \sim N_{n_{\text{pr}}}\left\{\frac{\rho u_0 + \lambda u_1}{\sqrt{z}}\mathbf{1}_{n_{\text{pr}}} + \boldsymbol{\Sigma}_{\text{pr,ob}}\boldsymbol{\Sigma}_{\text{ob,ob}}^{-1}\left(\mathbf{y}_{\text{ob}} - \frac{\rho u_0 + \lambda u_1}{\sqrt{z}}\mathbf{1}_{n_{\text{ob}}}\right), \frac{\sigma^2}{z}\left(\boldsymbol{\Sigma}_{\text{pr,pr}} - \boldsymbol{\Sigma}_{\text{pr,ob}}\boldsymbol{\Sigma}_{\text{ob,ob}}^{-1}\boldsymbol{\Sigma}_{\text{ob,pr}}\right)\right\}. \tag{8}$$

On the other hand, the density $p(\mathbf{u}_0, \mathbf{u}_1, \mathbf{z}|\mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}})$ coincides with that of the E-step of the EM algorithm, replacing here the kth parameter set with the parameter set $\hat{\boldsymbol{\theta}}$. Thus, in practice, we may draw samples from $p(\mathbf{u}_0, \mathbf{u}_1, \mathbf{z}|\mathbf{y}_{\text{ob}}, \hat{\boldsymbol{\theta}})$ for as many regions as required, as indicated in Section 3.1, followed by samples for each regional set of prediction locations, according to (8).

4. Simulation study

Here we examine the predictive performance of the proposed skew-t model, which we hereby denote as SKT, against a skew-normal (SN) with $Z_r = 1$ in (1), a bi-resolution Gaussian model (biGau) with $Z_r = 1$ and $\lambda_r = 0$ in (1), and a Gaussian model (GAU) with $Z_r = 1$, $\lambda_r = 0$ and $\rho_r = 0$. We test these four models against samples generated from a Tukey g-and-h random field (Xu and Genton, 2017), a trans-Gaussian process with transformation

$$\tau_{g,h}(z) = \begin{cases} g^{-1}\{\exp(gz) - 1\}\exp\left(\frac{hz^2}{2}\right), & g \neq 0, \\ z\exp(hz^2/2), & g = 0. \end{cases}$$

The transformation $\tau_{g,h}(z)$ allows to control skewness and kurtosis separately using the two parameters g-and-h, and are here sampled randomly for each region from a range of 0.3 to 0.5 for the former, and 0.03 to 0.06 for the latter. Using the biGau model as the underlying model to which the Tukey g-and-h transformation was subsequently applied, the resulting parameter values yielded skewness and kurtosis values ranging from approximately 1.25 to 3, and 6.5 to 25, respectively. We further assume $\phi_0 = 2$, for the range parameter governing the large-scale correlation function, for which region centroids are used to determine intra-region distances. The range parameter for each region is fixed at $\phi_r = 0.1$. We also fix $\sigma_r = 1$, and assume $\nu_r = 7$. A total of ten simulations is performed, each with $T = 150$ spatial replicates. We consider a spatial design composed of $R = 15$ regions, each consisting of a 1×1 square, arranged in a 4-by-4 grid. Within each region, 15 points are chosen at random, for a total of 225 spatial locations (see Fig. S1 in Supplementary Material). After the Tukey g-and-h has been simulated, all four models are fit and we compare differences in prediction accuracy as measured by the continuous ranked probability score (CRPS, Gneiting et al. (2007)).

Inference for the SKT, SN and biGau model is achieved using the EM algorithm outlined in Section 3 (or simplified versions thereof in the case of SN and biGau), with $M = 1000$ MCEM iterations for each spatial replicate t. Inference for the GAU model is achieved by maximum likelihood. We consider multiple initial parameter values, and do not observe convergence to local minima. Figure S2 displays the evolution of $\hat{\boldsymbol{\theta}}^{[k]}$ and the Q-function for 500 iterations of the average of the ten simulations; parameter estimates can be seen stabilizing beyond the 200-th iteration, and close to their true values. The estimation of ν_r appears to be particularly noisy. A quantification of the (asymptotic) standard deviation would require EM-specific approaches such as the Supplemented Expectation–Maximization (SEM, Meng and Rubin (1991)) algorithm. However, this task was computationally infeasible given the large number of temporal replicates and the number of repeated simulations. SEM was however performed in the application from the next section, and the results highlight how the standard deviation in the estimation of ν_r is more than one order of magnitude larger than all other parameters.

Table 1 provides a comparison of the out-of-sample average CRPS values for the four models for each of the 15 regions. The SKT model outperforms all other three models. An increase in model performances is highlighted when allowing two resolutions, i.e. from GAU (0.61) to biGau and SN (0.57), and an additional improvement is achieved when allowing for a skewed behavior for the SKT model (0.56).

Table 1
Out-of-sample average continuous ranked probability scores based on 10 spatial replicates for the SKT, SN, biGau and GAU models for the 15 regions, and their overall mean.

Model	Region															Mean
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
SKT	0.49	0.56	0.47	0.55	0.56	0.62	0.67	0.56	0.60	0.54	0.50	0.65	0.62	0.47	0.57	0.56
SN	0.54	0.57	0.50	0.55	0.59	0.66	0.52	0.55	0.62	0.65	0.58	0.56	0.55	0.65	0.52	0.57
biGau	0.54	0.57	0.50	0.55	0.59	0.66	0.52	0.55	0.62	0.65	0.58	0.57	0.55	0.65	0.52	0.57
GAU	0.55	0.60	0.51	0.61	0.55	0.74	0.61	0.61	0.80	0.68	0.59	0.55	0.63	0.63	0.44	0.61

5. Application

We use the Colorado climatological data obtained from the Geophysical Statistics Project at the National Center for Atmospheric Research (www.cgd.ucar.edu/stats/Data/US.monthly.met). The data set provides monthly data for temperature maxima and minima, as well as precipitation, for the years 1895 to 1997. Here, we focus solely on annual total precipitation amounts (in millimeters), and consider only the 137 stations from the original 397 that have observations over the final 40 years, see Fig. 1. We chose a subset of stations to allow an uninterrupted 40 years long record of annual data. We could in principle have added more, but that would have required adding missing data whose estimation is beyond the scope of this work. We refrain from any transformation of the data (e.g., log-transformed), keeping them to their original scale, except for a prior standardization. The lack of transformation implies an always positive quantity, while all the models we consider allow for values across the entire real line. Figure S1 in the supplementary material shows the aggregated histogram across all locations and years. As apparent from the plot, no data point was exactly equal to zero, and the vast majority of the data are far away from zero as well, hence allowing us to approximate the probability of a negative event to zero. The median skewness across locations is -0.22 (Q1 $= -0.69$, Q3 $= 0.27$), while the median kurtosis is 3.30 $(-3.34, 3.95)$, hence an overall non-Gaussian, skewed and heavy tailed behavior that justifies the use of a ST distribution introduced in the previous sections. We also found no indication of temporal dependence. We calculated the autocorrelation function at each site, along with its asymptotic standard deviation, and we found that no site had a significant lag-1 value. Thus, at annual level it is perfectly reasonable to assume temporal independence. A modified model that would account for temporal dependence and possibly lack of space–time separability would likely be necessary for lower levels of aggregation, i.e. daily or sub-hourly temporal scales.

We extend the model in (1) to account for the non-zero location parameter, which we assume to be the same across all points belonging to the same region. The location transformation adds the term ξ_r to the conditional mean of the sampling distribution, which implies replacing the $\mathbf{y}_{r,t}$ terms with $\mathbf{y}_{r,t} - \xi_r \mathbf{1}_{n_r}$ in the formulas of the EM algorithm, and the additional update as part of the M-step,

$$\xi_r^{(k)} = \frac{\operatorname{tr}\left(\boldsymbol{\Sigma}_r^{-1}\left[1/\sigma_r^2 \sum_{t=1}^{T}\langle z_{r,t}\rangle\left(\mathbf{1}_{n_r}\mathbf{y}_{r,t}^\top + \mathbf{y}_{r,t}\mathbf{1}_{n_r}^\top\right) - 2/\sigma_r \sum_{t=1}^{T}\left(\langle\sqrt{z_{r,t}}u_{0,r,t}\rangle\rho_r + \langle\sqrt{z_{r,t}}u_{1,r,t}\rangle\lambda_r\right)\mathbf{1}_{n_r}\mathbf{1}_{n_r}^\top\right]\right)}{2\operatorname{tr}\left(\boldsymbol{\Sigma}_r^{-1}\left[1/\sigma_r^2 \sum_{t=1}^{T}\langle z_{r,t}\rangle\mathbf{1}_{n_r}\mathbf{1}_{n_r}^\top\right]\right)}$$

We distinguish six regions, shown color-coded in Fig. 1, covering the Eastern plains, the foothills, the high peaks and the adjacent Southern and Western areas. As part of an exploratory analysis, we fit a multivariate skew-t distribution, with the help of the R package `sn` (Azzalini, 2018), independently to each region, and find that the resulting parameter estimates exhibit consistent non-zero skewness in all regions, as well as heavy-tails in most, lending justification to the use of the SKT model. We fit the SKT model, but given the small number of regions ($R = 6$), the EM algorithm was unstable in the identification of the large-scale range parameter ϕ_0, as it tended towards increasingly larger values, in order to reproduce the inter-regional dependence structure. We thus fixed it at a suitably large value, $\phi_0 = 10{,}000$. Estimates for the remaining parameters are shown in Table S1, while those for the GAU model fit are shown in Table S2. The evolution of the Q-function and parameter estimates of each M-step iteration are shown in Fig. S3.

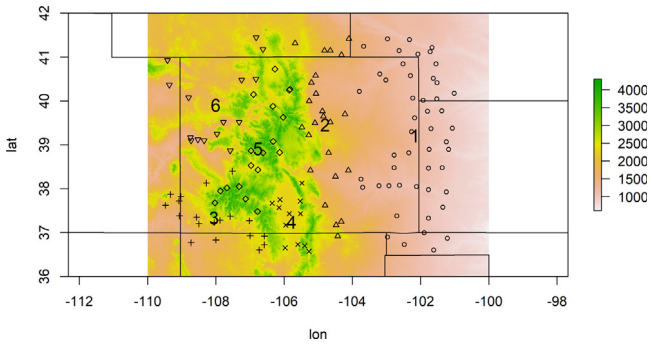

Fig. 1. Location of precipitation stations, each region being assigned a different marker, with region number plotted at the centroid. The overlaid elevation map (in meters) is from the United States Geological Survey GTOPO30 digital elevation model.

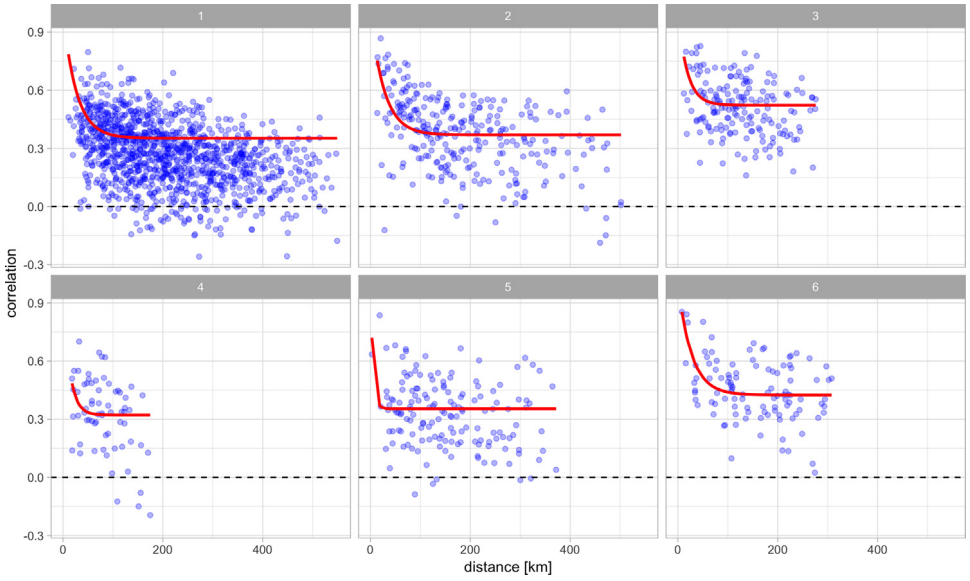

Fig. 2. Empirical correlations (blue dots) for each of the six regions denoted in Fig. 1, and the respective correlation functions (red curves) of the SKT model (4) evaluated using parameter estimates.

Fig. 2 shows the empirical correlation for each region as a function of the distance and the model-implied correlation function evaluated, using estimated parameters. We see overall good agreement between the fitted and empirical correlation, although the terminal value of the correlation function in region 1 seems sensibly too high, underscoring the tension between inducing inter-regional spatial dependence and skewness, and non-vanishing intra-regional spatial correlation. Fig. 3 examines the inter-regional spatial dependence, contrasting the median of the correlations between points located in different regions and the model implied inter-regional correlation. From this figure it is apparent how the model overall adequately captures the inter-regional dependence as well.

We contrast the in-sample predictive performance between both models, using the first 10 spatial replicates of the 40 year period. Table 2 presents the CRPS values for each model, where it is seen that the SKT model offers better predictive accuracy. The results are dependent on the

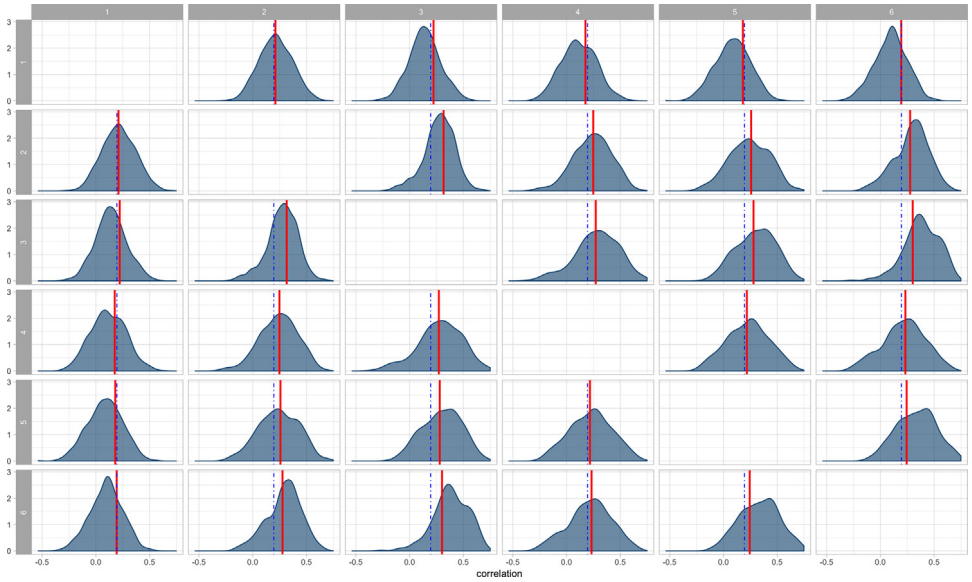

Fig. 3. Matrix of kernel density estimates of the correlations between points belonging to the respective regions, along with median correlation (dashed blue) and the inter-regional correlation (solid red) based on Eq. (5), evaluated at the parameter estimates.

Table 2
Predictive performances of GAU and SKT model in terms of average continuous ranked probability scores for the first ten replicated of the 40 year period.

Model	Region						Mean
	1	2	3	4	5	6	
SKT	0.39	0.36	0.33	0.45	0.41	0.36	0.38
GAU	0.51	0.50	0.47	0.60	0.50	0.60	0.53

number R and location of the clusters, so a simulation study assessing the sensitivity against these factors could be performed. However, such a task is currently too computationally intensive as it would require re-running the model for different configurations. As a general rule of thumb, the regions should not be too small to have unstable estimates, and not too large to avoid losing information about local spatial structure. Moreover, if one is interested in a particular subregion, a clustering approach with more points in that region can be envisioned, and as long as prediction is not performed too close to the edges of the regions, no artifacts from the discontinuity should be apparent.

6. Conclusions

In this work, we proposed a new model for non-Gaussian spatial data that does not rely on the classical conditional independence assumption of hierarchical models. We showed how the skew-t has a representation that can be used as a baseline for a new class of hierarchical models, and is suitable for spatial data that have similar high-order moments.

The proposed model has a closed form expression for all moments, as well as a covariance between points. We chose to perform inference in a frequentist setting via an MCEM algorithm, with semi-closed form expressions for both the E and the M steps. The skew-t model also allows for a continuous underlying process within each region, and hence prediction can be performed

conditionally on the observed data using the estimated parameters from the aforementioned MCEM algorithm. The proposed MCEM algorithms require simulations of $\mathbf{u}_{0,t}$, $\mathbf{u}_{1,t}$ and $\mathbf{z}_t$ for all t in the M-step, so it is computationally burdensome. However, full conditionals for $\mathbf{u}_{0,t}$, $\mathbf{u}_{1,t}$ are available and full MCMC is necessary only for $\mathbf{z}_t$. In order to fit larger data sets with this model, a promising direction of research to achieve approximate but faster computation is the use of Stochastic Approximation EM (SAEM, Jank (2006), Ordoez et al. (2018)), which relies on a Taylor approximation of the Q function.

The model could be generalized to account for spatial covariates. In the application we have provided a common location parameter ξ_r across each region, so the model would also allow for a changing mean. The location parameter could be generalized to allow for spatial covariates, i.e. $\xi_r = \mathbf{x}_r^\top \boldsymbol{\beta}$, for some design vector $\mathbf{x}_r$ and some parameters $\boldsymbol{\beta}$ to be estimated. This specification could be made more flexible with a location specific scale parameter, but that would require a very large sample size to avoid identifiability issues.

The extent of the generalizability of an (additive) hierarchical representation of the skew-t is an open question. In principle, a similar representation could be sought for skew-elliptical models, but the algebra of the E and M steps, or full conditionals in a Bayesian version of this model, are expected to be non-trivial. Additional extensions to scale mixtures of skew-normal (Cabral et al., 2012) would be even more challenging as a closed form expression is not guaranteed, hence the computation would be considerably more burdensome. Besides, there would be subjectivity in the choice of the mixing distribution and for an application a sensitivity study should be performed.

Scalability is also a potential issue when applying such a model to considerably larger spatial data sets. Additional work is needed to seek solutions to either reparametrize the latent parameter space from the current dimension vector of size $3R$, or a Bayesian version of the model with either a Laplace or Hamiltonian approximation of the posterior.

Lastly, the time component here is used only to provide replicates and improve the inference, thus implicitly assuming space–time separability. This assumption is likely not realistic for data sets at high temporal resolution, and will thus require a new formulation that allows a temporally nonstationary model with a marginal structure identical or similar to the one presented in this work.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.spasta.2019.100398.

References

Allard, D., Naveau, P., 2007. A new spatial skew-normal random field model. Comm. Statist. Theory Methods 36, 1821–1834.
Arellano-Valle, R.B., Branco, M.D., Genton, M.G., 2006. A unified view on skewed distributions arising from selections. Can. J. Stat. / Rev. Can. Stat. 34 (4), 581–601.
Azzalini, A., 1985. A class of distributions which includes the normal ones. Scand. J. Stat. 12, 171–178.
Azzalini, A., 2018. The R Package sn: The Skew-Normal and Related Distributions such as the Skew-t (version 1.5-3).. Università di Padova, Italia, URL http://azzalini.stat.unipd.it/SN.
Azzalini, A., Capitanio, A., 2003. Distributions generated by perturbation of symmetry with emphasis on a multivariate skew t-distribution. J. R. Stat. Soc. Ser. B Stat. Methodol. 65 (2), 367–389.
Azzalini, A., Capitanio, A., 2014. The Skew-Normal and Related Families. Cambridge University Press, IMS Monograph Series.
Azzalini, A., Dalla Valle, A., 1996. The multivariate skew-normal distribution. Biometrika 83 (4), 715–726.
Bevilacqua, M., Caamaño, C., Arellano-Valle, R.B., Morales-Onñate, V., 2018. On spatial (skew) t processes and applications. arXiv preprint arXiv:1812.06310.
Cabral, C.R.B., Lachos, V.H., Prates, M.O., 2012. Multivariate mixture modeling using skew-normal independent distributions. Comput. Statist. Data Anal. 56 (1), 126–142.
Castruccio, S., Huser, R., Genton, M.G., 2016. High-order composite likelihood inference for max-stable distributions and processes. J. Comput. Graph. Statist. 25 (4), 1212–1229.
Cressie, N.A., 1993. Statistics for Spatial Data. Wiley.
Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. Ser. B Stat. Methodol. 39, 1–38.
Diggle, P., Ribeiro, P., 2007. Model-based Geostatistics. Springer, New York.

Diggle, P.J., Tawn, J., Moyeed, R., 1998. Model-based geostatistics. J. R. Stat. Soc. Ser. C. Appl. Stat. 47 (3), 299–350.
Genton, M.G., 2004. Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality. Chapman and Hall / CRC, Boca Raton, FL.
Genton, M.G., Zhang, H., 2012. Identifiability problems in some non-gaussian spatial random fields. Chil. J. Stat. 3, 171–179.
Gneiting, T., Balabdaoui, F., Raftery, A.E., 2007. Probabilistic forecasts, calibration and sharpness. J. R. Stat. Soc. Ser. B Stat. Methodol. 69 (2), 243–268.
Jank, W., 2006. Implementing and diagnosing the stochastic approximation EM algorithm. J. Comput. Graph. Statist. 15 (4), 803–829.
Kim, H., Ha, E., Mallick, B., 2004. Spatial prediction of rainfall using skew-normal processes. In: Genton, M.G. (Ed.), Skew-Elliptical Distributions and their Applications: A Journey beyond Normality. CRC Press, Boca Raton, FL, pp. 279–289.
Kim, H., Mallick, B.K., 2004. A Bayesian prediction using the skew-Gaussian processes. J. Statist. Plann. Inference 120, 85–101.
Lachos, V.H., Cabral, C.R.B., Zeller, C.B., 2018. Scale mixtures of skew-normal distributions. In: Finite Mixture of Skewed Distributions. Springer, pp. 15–36.
Lachos, V.H., Ghosh, P., Arellano-Valle, R.B., 2010. Likelihood based inference for skew-normal independent linear mixed models. Statist. Sinica 20, 303–322.
Lin, T.I., 2010. Robust mixture modeling using multivariate skew t distributions. Stat. Comput. 20 (3), 343–356.
Lin, T.I., Lee, J.C., 2008. Estimation and prediction in linear mixed models with skew-normal random effects for longitudinal data. Stat. Med. 27 (9), 1490–1507.
Lindgren, F., Rue, H., 2015. Bayesian spatial modelling with r-inla. J. Stat. Softw. 63 (19), 1–25.
Mahmoudian, B., 2017. A skewed and heavy-tailed latent random field model for spatial extremes. J. Comput. Graph. Statist. 26 (3), 658–670.
McLachlan, G., Krishnan, T., 2007. The EM Algorithm and Extensions, Vol. 382. John Wiley & Sons.
Meng, X.-L., Rubin, D.B., 1991. Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. J. Amer. Statist. Assoc. 86 (416), 899–909.
Minozzo, M., Ferracuti, L., 2012. On the existence of some skew-normal stationary processes. Chil. J. Stat. 3, 157–170.
Ordoez, J.A., Bandyopadhyay, D., Lachos, V.H., Cabral, C.R., 2018. Geostatistical estimation and prediction for censored responses. Spat. Stat. 23, 109–123.
Rue, H., Martino, S., Chopin, N., 2009. Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. 71 (2), 319–392.
Schmidt, A.M., Gonçalves, K., Velozo, P.L., 2017. Spatio-temporal models for skewed processes. Environmetrics (with discussion) 28 (6), 1821–1834.
Tagle, F., Castruccio, S., Crippa, P., Genton, M.G., 2019. A non-Gaussian spatio-temporal model for daily wind speeds based on a multi-variate skew-t distribution. J. Time Series Anal. 40, 312–326.
Xu, G., Genton, M.G., 2017. Tukey g-and-h random fields. J. Amer. Statist. Assoc. 112, 1236–1249.
Zhang, H., El-Shaarawi, A., 2010. On spatial skew-gaussian processes and applications. Environmetrics 21 (1), 33–47.