Chilean Journal of Statistics
Vol. 3, No. 2, September 2012, 143–155

Probabilistic and Inferential Aspects of Skew-Symmetric Models
Special Issue: "IV International Workshop in honour of Adelchi Azzalini's 60th Birthday"

A dengue fever study in the state of Rio de Janeiro with the use of generalized skew-normal/independent spatial fields

Marcos Oliveira Prates[1,*], Dipak Kumar Dey[2], and Victor Hugo Lachos[3]

[1]Departamento de Estatística, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[2]Department of Statistics, University of Connecticut, Storrs, USA
[3]Departamento de Estatística, Universidade Estadual de Campinas, Campinas, Brazil

(Received: 29 September 2011 · Accepted in final form: 26 June 2012)

Abstract

This paper develops a novel spatial process using generalized skew-normal/independent distributions when the usual Gaussian process assumptions are questionable and transformation to a Gaussian random field is not appropriate. The proposed model provides flexibility in capturing the effects of skewness and heavy tail behavior of the data while maintaining spatial dependence using a conditional autoregressive structure. We use Bayesian hierarchical methods to fit such models and show the validity of our approach. Furthermore, we use Bayesian model selection criteria to choose appropriate models for a real data set on the dengue fever infection in the state of Rio de Janeiro.

Keywords: Bayesian hierarchical methods · Conditional autoregressive · Conditional predictive ordinate · Markov chain Monte Carlo · Scale mixture of skew-normal distributions · Skew-normal/Independent distributions · Spatial association.

Mathematics Subject Classification: Primary 62F15 · Secondary 80M31.

1. Introduction

The field of spatial statistics is very active and has received considerable attention in recent years. With the development of the geographic information systems (GIS), many scientific fields such as agriculture, biology, ecology, geography and geology have been using spatial data analysis to improve overall modeling strategies. Moreover, GIS information has brought to the statistical community a new avenue of collecting data. Spatial methods have become extremely important and necessary to accommodate spatial dependence when performing data analysis. Due to the spatial characteristics for certain data, it is necessary to correctly incorporate spatial dependence in modeling.

[*]Corresponding author. Marcos Oliveira Prates, Departamento de Estatística, ICEX/UFMG, Av. Antônio Carlos, 6627, Pampulha, Belo Horizonte, MG, Brasil, 31270-901. Email: marcosop@est.ufmg.br

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

To reduce unrealistic distributional assumptions, e.g., symmetry and/or thin tails, there is a tendency in spatial data analysis towards more flexible spatial methods that are capable of representing the data features in a more realistic way. For example, under a geostatistical point of view, different asymmetric models are proposed to better adjust to data sets where the normal assumption is not appropriate; see Kim and Mallick (2004); Karimi et al. (2010).

In this paper, we focus on an approach based on the generalized linear mixed models (GLMM) (see Breslow and Clayton, 1993), which is an important class of statistical models that are widely used to describe dependent data, such as is the case of spatial data. For this type of data, GLMMs introduce dependence through random effects. Under the GLMM framework, scientists have been using simultaneous autoregressive (SAR) (see Whittle, 1954) as well as conditional autoregressive (CAR) random fields (see Besag, 1974), as tools to accommodate spatial dependence for modeling of data. However, the Gaussian assumption in the random effect implies symmetry and thin tail, which may not be appropriate for many applications; see, e.g., Prates et al. (2011b).

When the Gaussian assumption for the random effects is not adequate, e.g., when data are skewed or present a heavy tail behavior, we need alternative distributions to realistically represent the data; see, e.g., Genton (2004) and Arellano-Valle and Genton (2010). Sahu et al. (2003) (see also Arellano-Valle and Azzalini, 2006) defined the multivariate skew-normal (SN) distribution as follows. A random vector $\boldsymbol{Y}$ is said to follow a p-variate SN distribution with location vector $\boldsymbol{\mu} \in \mathbb{R}^p$, scale matrix $\boldsymbol{\Sigma}$ positive definite and skewness $p \times p$ matrix $\boldsymbol{\Lambda} = \operatorname{diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)^\top$, if its density is

$$f(\boldsymbol{y}) = 2^p \phi_p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}) \Phi_p(\boldsymbol{\Lambda}^\top \boldsymbol{\Omega}^{-1}(\boldsymbol{y} - \boldsymbol{\mu})|\boldsymbol{\Delta}), \quad \boldsymbol{y} \in \mathbb{R}^p, \tag{1}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, $\boldsymbol{\Delta} = (\boldsymbol{I} + \boldsymbol{\Lambda}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\Lambda})^{-1} = \boldsymbol{I} - \boldsymbol{\Lambda}^\top \boldsymbol{\Omega}^{-1} \boldsymbol{\Lambda}$ (with $\boldsymbol{I}$ being a p-dimensional identity matrix), $\phi_p(\cdot|\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and $\Phi_p(\cdot|\boldsymbol{\Sigma})$ is the $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ density and the $\mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ distribution function, respectively. We write $\mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ to indicate that $\boldsymbol{Y}$ has density as given in Equation (1). For $\boldsymbol{\Lambda} = 0$, Equation (1) reduces to the usual $\mathrm{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The SN distribution as defined in Equation (1) can be stochastically represented as

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{\Lambda}|\boldsymbol{T}_1| + \boldsymbol{T}_2, \tag{2}$$

where $\boldsymbol{T}_1 \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{I})$ is independent of $\boldsymbol{T}_2 \sim \mathrm{N}_p(\boldsymbol{0}, \boldsymbol{\Sigma})$ and $|\boldsymbol{T}_1|$ denotes the component wise absolute value of $\boldsymbol{T}_1$. Thus, $|\boldsymbol{T}_1|$ follows a p-dimensional standard half-normal distribution denoted by $\mathrm{HN}_p(\boldsymbol{0}, \boldsymbol{I})$. Note that the expression given in Equation (2) provides a representation which is a useful tool for generation of random observations from Equation (1). This representation is also appropriate for developing various theoretical properties of the SN distribution. According to Sahu et al. (2003), the expectation and covariance matrix of $\boldsymbol{Y} \sim \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ are respectively given by

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}_{\mathrm{SN}} = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}} \boldsymbol{\lambda} \tag{3}$$

and

$$\mathrm{Var}[\boldsymbol{Y}] = \boldsymbol{\Sigma}_{\mathrm{SN}} = \boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right) \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top = \boldsymbol{\Sigma} + \left(1 - \frac{2}{\pi}\right) \operatorname{diag}(\lambda_1^2, \ldots, \lambda_p^2).$$

Because the matrix $\boldsymbol{\Lambda}$ is diagonal, the introduction of skewness increases the variance elements $\text{Var}[Y_i]$ without increasing the covariance elements $\text{Cov}(Y_i, Y_j)$ when $i \neq j$, thus decreasing the correlation between Y_i and Y_j. Moreover, the skewness parameters do not affect the spatial dependence structure provided by $\boldsymbol{\Sigma}$. When $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, then the density in Equation (1) gives independent marginal distributions. Different types of distributions have also been proposed to overcome the limitation of Gaussian processes to handle skewed or heavy tailed data; see, e.g., Allard and Naveau (2007); Prates et al. (2011b).

The generalized skew-normal/independent (GSNI) distribution used in this paper is developed primarily from the multivariate SN density given in Equation (1) proposed by Sahu et al. (2003) for Bayesian regression problems, and it is not equivalent from the multivariate skew-normal/independent (SNI) densities developed in Lachos et al. (2010), which was motivated by the SN version proposed in Azzalini and Dalla Valle (1996). Moreover, the GSNI distribution has the advantage that the covariance matrix is partitioned in two components: a spatial component and a skewness component. Furthermore, the GSNI family has as members the normal and the SN distributions.

The DATASUS system, provided by the Brazilian Ministry of Heath, possesses a variety of information in epidemiological diseases around Brazil. In our analysis, we focus on the dengue fever in the counties of Rio de Janeiro state. Dengue fever is a disease that occurs all over the world, mainly in tropical regions. Around the world 2.5 billion people live in tropical areas where the probability of infection is high. Epidemiologists are aware that the dengue fever has become a serious public health problem all over Brazil. We collected explanatory variables as income and percentage of treated water in each county in Rio de Janeiro to perform our analysis. In our study, we investigated the characteristics of the counties with high risk of dengue fever incidence and which explanatory variables can be used to better understand the contamination risk at each county.

The article is organized as follows. In Section 2 we introduce the GSNI distribution and its properties. Using this distribution, we describe how to accommodate spatial dependence. Then, we define a new generalized skew-Gaussian spatial field in Section 3. In Section 4, we introduce a variety of model comparison criteria. We illustrate the new proposed methodology with a real data analysis on the dengue fever infection in Rio de Janeiro in Section 5. In Section 6, we conclude the paper with a discussion.

2. Generalized Skew-Normal/Independent Distributions

Following Zeller (2009), we define the p-dimensional generalized SNI (GSNI) vector $\boldsymbol{Y}$, denoted from now on as $\boldsymbol{Y} \sim \text{GSNI}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, H_p(\cdot;\nu))$, as a multivariate mixture of SN distribution, where $\boldsymbol{U} = (U_1, \ldots, U_p)^\top$ is a random vector instead of a random variable used in Bandyopadhyay et al. (2010). Thus, $\boldsymbol{Y}$ can be represented as

$$\pi(\boldsymbol{y}) = \int_{\mathbb{R}_+^p} f(\boldsymbol{y}|\boldsymbol{u}) dH_p(\boldsymbol{u};\nu),$$

where $f(\cdot|\boldsymbol{u})$ is the conditional density of the random vector $\boldsymbol{Y}$ given $\boldsymbol{U} = \boldsymbol{u}$, and $\boldsymbol{U}$ is a positive random vector with distribution function $H_p(\cdot;\nu) = \prod_{i=1}^p H_i(\cdot;\nu)$, with $H_i(\cdot;\nu) = H(\cdot;\nu)$ being the distribution of the mixture variable U_i, for $i = 1, \ldots, p$, and ν is a parameter indexing of the distribution H.

Similarly to Equation (2), $\boldsymbol{Y}$ can be stochastically represented as

$$\boldsymbol{Y} = \boldsymbol{\mu} + \boldsymbol{U}^{-1/2} \odot \boldsymbol{Z}, \tag{4}$$

where $\boldsymbol{Z} \sim \mathrm{SN}_p(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, $\boldsymbol{U}^{-1/2} = (U_1^{-1/2}, \ldots, U_p^{-1/2})^\top$, and U_i's are positive, independent and identically distributed random variables, independent of $\boldsymbol{Z}$. In Equation (4), $\odot$ represents the Hadamard product, that is, $\boldsymbol{U} \odot \boldsymbol{Z} = (U_1 Z_1, \ldots, U_p Z_p)^\top$ if both $\boldsymbol{U}$ and $\boldsymbol{Z}$ are of dimension p, and $\boldsymbol{U} \odot Z = (U_1 Z, \ldots, U_d Z)^\top$ if Z is scalar. Clearly, when $\boldsymbol{U}^{-1/2}$ is set to be a scalar, the GSNI distribution is equivalent to the SNI distribution proposed by Bandyopadhyay et al. (2010). As in proposition 2.4 presented in Arellano-Valle and Genton (2005), we obtain the fact that given $\boldsymbol{U} = \boldsymbol{u}$, $\boldsymbol{Y}$ has a SN distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}_u = \mathrm{diag}(\boldsymbol{u}^{-1/2})\boldsymbol{\Sigma}\mathrm{diag}(\boldsymbol{u}^{-1/2})$, and skewness parameter matrix $\boldsymbol{\Lambda}_u = \mathrm{diag}(\boldsymbol{u}^{-1/2})\boldsymbol{\Lambda}$, that is, $\boldsymbol{Y}|\boldsymbol{U} = \boldsymbol{u} \sim \mathrm{SN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}_u, \boldsymbol{\Lambda}_u)$. Hence, the density of $\boldsymbol{Y}$ is

$$f(\boldsymbol{y}) = 2^p \int_{\mathbb{R}_+^p} \phi_p(\boldsymbol{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}_u)\Phi_p(\,\mathrm{diag}\,(\boldsymbol{u}^{1/2})\boldsymbol{\Lambda}^\top\boldsymbol{\Omega}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})|\boldsymbol{\Delta})dH_p(\boldsymbol{u};\nu)d\boldsymbol{u},$$

where $\boldsymbol{\Omega}_u = \,\mathrm{diag}\,(\boldsymbol{u}^{-1/2})\boldsymbol{\Omega}\,\mathrm{diag}\,(\boldsymbol{u}^{-1/2})$.

From Equations (3) and (4), it is clear that

$$\mathbb{E}[\boldsymbol{Y}] = \boldsymbol{\mu}_{\mathrm{GSNI}} = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\kappa_1(\nu)\boldsymbol{\lambda}$$

and

$$\mathrm{Var}[\boldsymbol{Y}] = \boldsymbol{\Sigma}_{\mathrm{GSNI}} = \kappa_2(\nu)\left(\boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right) + \left(\kappa_2(\nu) - \kappa_1^2(\nu)\right)\frac{2}{\pi}\boldsymbol{\lambda}\boldsymbol{\lambda}^\top,$$

where $\kappa_\alpha(\nu) = \mathbb{E}[U^{-\alpha/2}]$, $\alpha \in \{1, 2\}$, and the moments are well defined.

As in Bandyopadhyay et al. (2010), this class of asymmetric GSNI distributions contains a variety of skewed distributions based on different choices of the distribution of the mixture $\boldsymbol{U}$, such as, for $j = 1, \ldots, p$:

(1) Multivariate SN: $H_j = 1$.
(2) Multivariate skew-t (ST): $H_j = \Gamma(\nu/2, \nu/2)$.
(3) Multivariate skew-slash (SSL): $H_j = \mathrm{Beta}(\nu, 1)$.
(4) Multivariate contaminated normal (SCN): $H_j = \begin{cases} \nu_2 & \text{with prob} \quad \nu_1; \\ 1 & \text{with prob} \quad 1-\nu_1. \end{cases}$

The normal, Student-t, slash and contaminated normal distributions are obtained by setting $\boldsymbol{\Lambda} = 0$. All these distributions have heavier tails than that of the SN one and can be used for robust inference. In order to have a zero-mean vector ($\boldsymbol{\mu}_{\mathrm{GSNI}} = \mathbf{0}$), we should assume the location parameter $\boldsymbol{\mu} = -\sqrt{2/\pi}\kappa_1\boldsymbol{\lambda}$, which is what we assume throughout this article.

3. Generalized Skew-Gaussian Spatial Field

Suppose that we observe $(Y_i, \boldsymbol{X}_i)$ at sites $i = 1, \ldots, n$, where Y_i is the response variable and $\boldsymbol{X}_i$ a $q \times 1$ vector of covariates that corresponds to the response Y_i at ith site. Let $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$ be a vector of unobserved random effects with joint distribution F, which introduces spatial dependence. A spatial GLMM assumes that, given $(\boldsymbol{X}_i, e_i)$, the observations Y_i's are independent with density $\varphi(y_i; \xi_i)$ belonging to a one-parameter exponential family

$$\varphi(y;\theta_i) = \exp\left(\frac{\xi_i y - \psi(\xi_i)}{a(\phi)} + c(y,\phi)\right),$$

where $a(\cdot)$, $\psi(\cdot)$ and $c(\cdot)$ are known functions, ϕ is a scale or dispersion parameter, ξ_i is the canonical parameter, and the support of the distribution does not depend on ξ_i. Let $\mu_i = \mathbb{E}[Y_i|\boldsymbol{X}, \boldsymbol{e}]$, where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)^\top$ is the matrix of covariates. The conditional expectation μ_i is connected to the covariate $\boldsymbol{X}_i$ and random effect e_i through a fixed link function g given by

$$g(\mu_i) = \eta_i + e_i,$$

where $\eta_i = \boldsymbol{X}_i^\top \boldsymbol{\beta}$ is the fixed effect and $\boldsymbol{\beta}$ is a $q \times 1$ vector of regression coefficients of covariates $\boldsymbol{X}_i$. The dependence among random effects $\boldsymbol{e}$ determines the spatial dependence among conditional means $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$. Therefore, to fully specify a spatial GLMM, it is necessary to specify both the link function g and the joint distribution F of $\boldsymbol{e}$. Commonly, F is chosen to be a multivariate normal with mean zero and covariance matrix $\boldsymbol{\Sigma}$.

Instead of defining F as a multivariate normal distribution function, we propose to use the GSNI distribution for the random effects in order to model areal dependence. Besag (1974) proposed the CAR model as an alternative to capture dependence within areas. The CAR model defines the following covariance matrix:

$$\boldsymbol{\Sigma} = \sigma^2(\boldsymbol{I} - \rho \boldsymbol{W})^{-1}\boldsymbol{M},$$

where $\boldsymbol{W}$ is an $n \times n$ matrix with zeros on the diagonal and the neighbor weights (w_{ij}) in the off-diagonal positions, if i is neighbor of j, and 0 otherwise, and $\boldsymbol{M}$ is an $n \times n$ diagonal matrix, that is, $\boldsymbol{M} = \operatorname{diag}(\tau_1^2, \ldots, \tau_n^2)$. To assure that $\boldsymbol{\Sigma}$ is positive definite, we need some constraints $w_{ij}\tau_j^2 = w_{ji}\tau_i^2$ and $\rho \in (1/\gamma_{\min}, 1/\gamma_{\max})$, where γ's are the eigen values of $\boldsymbol{M}^{-1/2}\boldsymbol{W}\boldsymbol{M}^{1/2}$.

A random vector $\boldsymbol{\phi}$ is defined to follow a generalized skew-Gaussian spatial field (GSGSF) when $\boldsymbol{\phi} \sim \text{GSNI}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda}, H_n(\cdot; \nu))$, where $\boldsymbol{\Sigma}$ has a spatial dependence generated by a CAR (SAR) structure and $H_n(\cdot; \nu)$ is one of the distributions presented in Section 2.

Using a generalized linear mixed model approach, we can define a spatial random effect to follow a GSGSF and therefore capture the skewness and/or heavy tail behavior of the data. Suppose we have n responses $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^\top$ that come from a one-parameter exponential family distribution with density or mass φ. Thus, we can model the response $\boldsymbol{Y}$ as

$$\begin{aligned}
&Y_i \sim \varphi(\mu_i),\ i = 1, \ldots, n,\\
&g(\mu_i) = \boldsymbol{X}_i^\top \boldsymbol{\beta} + \boldsymbol{\phi},\\
&\boldsymbol{\phi} \sim \text{GSGSF}_n\left(-\sqrt{\frac{2}{\pi}}\kappa_1 \lambda \boldsymbol{1}_n, \boldsymbol{\Sigma}, \lambda \boldsymbol{I}, H_n(\cdot; \nu)\right),
\end{aligned}$$

where g is a link function, $\boldsymbol{X}_i$ is the vector of covariates for $i = 1, \ldots, n$, $\boldsymbol{\beta}$ contains the regression coefficients, $\boldsymbol{1}_n = (1, \ldots, 1)^\top$, $\boldsymbol{\Sigma}$ is the spatial dependence matrix generated by the covariance structure of a CAR or SAR model, H_n is one of the positive distributions presented in Section 2 and the skewness parameter λ is an only scalar to avoid overparametrization and identifiability problems. With this representation, our approach provides a flexible way to incorporate multivariate asymmetric spatial random effects into modeling.

4. Model Comparison

Here, we describe a variety of Bayesian criteria to perform model selection. Model comparison measures based on the posterior predictive densities are often easier to work on MCMC settings. MCMC methods are able to produce these measures without much extra effort.

The CPO model comparison is a Bayesian cross-validation approach; see, e.g., Geisser (1993) and Dey et al. (1997). Let $\boldsymbol{y}$ be the observed responses and $\boldsymbol{y}_{-i}$ denote the observed response vector excluding the ith observation. The CPO statistic associated with the ith observation, conditioning on $\boldsymbol{y}_{-i}$, is defined as the marginal posterior predictive density of y_i given by

$$\mathrm{CPO}_i = \int f(y_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-i})\mathrm{d}\boldsymbol{\theta}, \tag{5}$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \rho, \lambda, \nu, \sigma^2)$ is the vector of parameters of the distribution with density f, $f(y_i|\boldsymbol{\theta})$ is the conditional density or mass of y_i given $\boldsymbol{\theta}$, and $\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-i})$ is the posterior density of $\boldsymbol{\theta}$ based on data $\boldsymbol{y}_{-i}$. The intuition behind the CPO criterion is to choose a model with higher predictive power measured in terms of predictive density. The idea is similar to that of a leave-one-out cross validation in that the predictive density of each data point is evaluated at a density fitted from all other data points.

Although a closed form of Equation (5) is not available, Dey et al. (1997) showed that CPO_i can be estimated from a Monte Carlo integration approach could be approximated by a harmonic mean

$$\widehat{\mathrm{CPO}_i} = B\left(\sum_{j=1}^{B}\left[\frac{1}{f(y_i|\boldsymbol{\theta}^{(j)})}\right]\right)^{-1},$$

where B denotes the size of a MCMC sample of the posterior distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ and $\boldsymbol{\theta}^{(j)}$ is the parameter vector $\boldsymbol{\theta}$ in the jth MCMC sample. This approximation is valid when Y_i's are assumed to be conditionally independent given $\boldsymbol{\theta}$. Because the approximation is based on the posterior given all the observations, its calculation is straightforward.

To compare different models, we define a single measure for each one of them, the logarithm of the pseudo-marginal likelihood (LPML), defined by $\mathrm{LPML} = \sum_{i=1}^{n} \log \widehat{\mathrm{CPO}}_i$. The model with the largest LPML is the best one. For any two competing models, the comparison can be graphically displayed by plotting the log-ratio of CPO_i from the two models against the ith observation number. Points supporting either one of the models are above and below the zero lines, respectively; see, e.g., Prates et al. (2011a). In addition to the CPO, we are also going to consider other Bayesian model selection criteria, such as the deviance information criterion (DIC; see Spiegelhalter et al., 2002), the expected Akaike information criteria (EAIC; see Carlin and Louis, 2000) and the expected Bayesian information criteria (EBIC; see Brooks, 2002).

DIC, EAIC and EBIC are based on the posterior mean of the deviance, that is, $\mathbb{E}\left[D(\boldsymbol{\theta})\right]$, which is also a measure of fit and can be approximated by using the MCMC output, considering the value of

$$\bar{D} = \frac{1}{B}\sum_{j=1}^{B} D(\boldsymbol{\theta}^{(j)}),$$

where B represents the number of iterations, and

$$D(\boldsymbol{\theta}) = -2\log(f(\boldsymbol{y}|\boldsymbol{\theta})) = -2\sum_{i=1}^{n}\log(f(y_i|\boldsymbol{\theta})),$$

where $f(y_i|\boldsymbol{\theta})$ is the conditional density or mass of y_i given $\boldsymbol{\theta}$. EAIC, EBIC and DIC can be estimated using MCMC output by considering

$$\widehat{\text{EAIC}} = \bar{D} + 2p, \quad \widehat{\text{EBIC}} = \bar{D} + p\log(n), \quad \text{and} \quad \widehat{\text{DIC}} = \bar{D} + \widehat{\rho_D} = 2\bar{D} - \widehat{D},$$

respectively, where p is the number of parameters in the model and n is the total number of observations. The measure ρ_D is the effective number of parameters as described in Spiegelhalter et al. (2002), and is defined as

$$\rho_D = \; \mathbb{E}\left[D(\boldsymbol{\theta})\right] - D(\mathbb{E}[\boldsymbol{\beta}], \mathbb{E}[\rho], \mathbb{E}[\lambda], \mathbb{E}[\nu], \mathbb{E}[\sigma^2]).$$

The term $D(\mathbb{E}[\boldsymbol{\beta}], \mathbb{E}[\rho], \mathbb{E}[\lambda], \mathbb{E}[\nu], \mathbb{E}[\sigma^2])$ is the deviance of the posterior mean obtained when considering the mean values of the generated posterior means of the model parameters, which is estimated by

$$\widehat{D} = D\left(\frac{1}{B}\sum_{j=1}^{B}\boldsymbol{\beta}^{(j)}, \frac{1}{B}\sum_{j=1}^{B}\rho^{(j)}, \frac{1}{B}\sum_{j=1}^{B}\lambda^{(j)}, \frac{1}{B}\sum_{j=1}^{B}\nu^{(j)}, \frac{1}{B}\sum_{j=1}^{B}(\sigma^2)^{(j)}\right).$$

Unlike the LPML, smaller values of the EAIC, EBIC and DIC imply better fit of the model.

5. Application to Dengue Fever in Rio de Janeiro

Dengue is an arbovirus that has become a serious public health problem. Around the world around 2.5 billion people live on areas with higher infection risk of dengue fever. Tropical regions provide a susceptible habitat for the main disease vector, the mosquito *Aedes aegypti* because of its temperature and humidity.

Brazil is a country localized in South America where most of its land is under a tropical climate. Therefore, dengue fever has become a main public health problem in the country. Rio de Janeiro is a state in Brazil with a high incidence of dengue fever cases. We collected data for the number of cases of dengue fever by counties of Rio de Janeiro in the year of 2011; see `http://www2.datasus.gov.br/DATASUS`.

To study the dengue fever incidence, demographic information of each county was collected and incorporated into the analysis to improve the fit of the model to the data. The explanatory variables could provide possible explanations to dengue fever incidence in Rio de Janeiro state. It is believed that social conditions as treated water supply and county average income can be important explanatory variables to explain the occurrence of cases. We use the CENSUS 2010 (see `http://www.ibge.gov.br`) to collect the population, percentage of homes living with more than one minimum salary (income), and percentage of homes with treated water (water) by county.

5.1 Exploratory analysis

It is believed that regions with high incidence of dengue fever should affect and/or be affected by neighbors regions, because the *Aedes Aegypti* mosquito can migrate and infect people in areas nearby. This observation makes reasonable to incorporate a spatial component in the models. Moreover, we would like to check whether the residuals must be modeled by a distribution with heavy tail and/or asymmetry.

To investigate the need of spatial data analysis, we compare the dengue fever incidence and the expected number of infected (ENI) distributions. The ENI is calculated by

$$\text{ENI}_i = \text{pop}_i \frac{\text{cases}_+}{\text{pop}_+},\ i = 1, \ldots, n,$$

where pop_i is the population risk at the ith region, pop_+ is the total population in Rio de Janeiro and cases_+ is the total number of infected people in the state. Basically, the ENI measures the expected number of infections assuming that all regions have the same relative risk of a person getting infected. From Figure 1, it is clear that the spatial distribution of dengue fever does not seem to be randomly distributed in space and, moreover, it seems that the distributions of the incidence and ENI are not the same, which can produce asymmetry and heavy tails when controlling by the population risk.

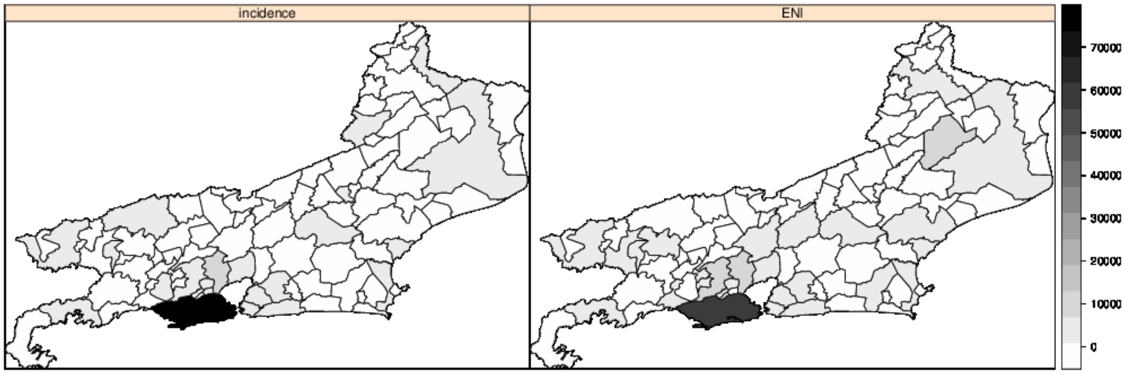

Figure 1. Left panel: incidence dengue fever counts for each county in Rio de Janeiro. Right panel: expected number of infected for each county in Rio de Janeiro

After demonstrating the necessity of spatial effect in the data, we continue to investigate the appropriateness of asymmetric and/or heavy tail models instead of the commonly used normal regression. Let Y_i be the incidence of dengue fever in each Rio de Janeiro county, for $i = 1, \ldots, 92$. We fit a Poisson regression as

$$\begin{aligned} Y_i &\sim \text{Poisson}(\text{ENI}_i \times \delta_i), \\ \log(\delta_i) &= \boldsymbol{X}_i^\top \boldsymbol{\beta}, \end{aligned}$$

where δ_i is the relative risk of each area, $\boldsymbol{X}_i$ are the covariates (income and water) and $\boldsymbol{\beta}$ are the fixed effects, following which, we perform model fitting on the residuals of the Poisson analysis to verify if the symmetry and thin tails assumptions are appropriate. To do so, we use the `mixsmsn` R package (see Prates et al., 2011) for fitting the residual with distributions of the SNI class and check what distribution is more adequate to fit the residuals. Under the AIC measure the skew-t distribution is the one which best fits the residuals. This can be easily verified using the histogram in Figure 2 and it is hence evident that the normal assumption for the residuals is not the most appropriate. In the next section, we study the dengue fever incidence using the GSGSF to improve fitting.

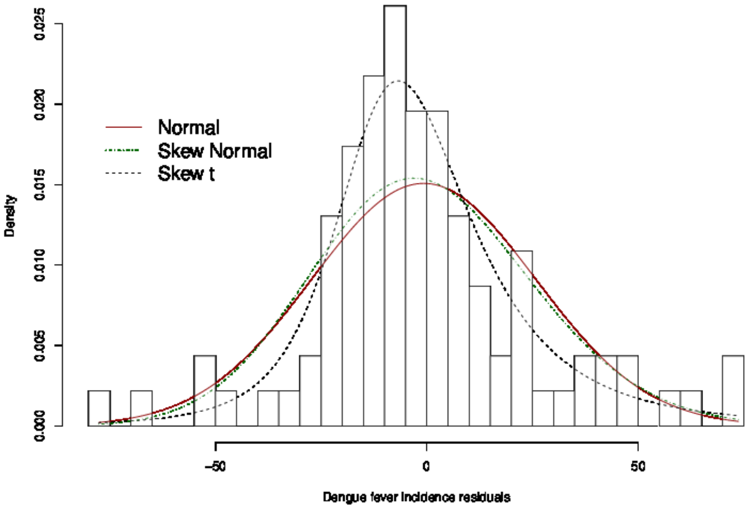

Figure 2. Fitting of the normal, SN and skew-t distributions to the residuals of the dengue fever Poisson GLM analysis.

5.2 Dengue fever with GSGSF

To accommodate spatial dependence between the neighboring counties, we use the proper precision CAR matrix specified in Section 3, selecting $\boldsymbol{M} = \operatorname{diag}(1/n_1, \ldots, 1/n_n)$ and $w_{ij} = 1/n_i$, if i is a neighbor of j, and zero otherwise, where n_i is the number of neighbors of the ith region. Given σ^2 and ρ, we can define $\boldsymbol{\Sigma}$ and use it in GSGSF representation, with λ and ν to calculate $\boldsymbol{\Sigma}_{\text{GSGSF}}$.

Suppose that, for each of the 92 counties in Rio de Janeiro, we observe the incidence of dengue fever (Y_i), for $i = 1, \ldots, 92$, and the set of covariates $\boldsymbol{X}_i$, which are income and water. Then, we model the counts of dengue fever by county as

$$\begin{aligned} Y_i &\sim \text{Poisson}(\text{ENI}_i \times \delta_i), \;\; i = 1, \ldots, 92, \\ \log(\delta_i) &= \boldsymbol{X}_i^\top \boldsymbol{\beta} + \phi_i, \\ \boldsymbol{\phi} &\sim \text{GSGSF}_{92}\left(-\sqrt{\frac{2}{\pi}}\kappa_1 \lambda \mathbf{1}_{92}, \boldsymbol{\Sigma}, \lambda \boldsymbol{I}, H_{92}(\cdot;\nu)\right), \end{aligned}$$

where δ_i is the relative risk of the ith county, for $i = 1, \ldots, 92$, $\boldsymbol{\beta}$ contains the regression coefficients, and $\boldsymbol{\phi}$ is a skew spatial field that accommodates the underlying spatial dependency in $\boldsymbol{\Sigma}$ with a CAR structure, asymmetry in λ and heavy tails in ν, which is a parameter of one of the possible $H_{92}(\cdot;\nu)$ presented in Section 2. To fully specify the model, vague hyper-priors are chosen. The skewness parameter, λ, is set to follow a $\text{N}(0, 10)$ distribution, the spatial dependence parameter, ρ, follows a $\text{U}(-1, 1)$ distribution, and the overall precision parameter is set to follow a $\Gamma(0.05, 0.05)$ distribution. We use the OpenBUGS software (see Lunn et al., 2009) to fit the MCMC, where for each model, 250,000 samples were collected and after a burning period of 50,000 and thinning by 200 iteration resulted in a valid sample of 1,000 observations.

Within the new family proposed in Section 3, we fit $\boldsymbol{\phi}$ with 5 different GSGSF (SN, Student-t, skew-t, contaminated normal and skew contaminated normal) as an alternative to the normal spatial field to analyze the incidence of dengue fever at the Rio de Janeiro counties.

Table 1. The selection criteria for the different proposed models.

Model	LPML	DIC	EAIC	EBIC
Normal	-453.27	844.03	675.55	688.16
Skew-normal	-450.34	843.70	672.16	687.29
Student-t	-451.82	844.52	671.70	686.83
Skew-t	-449.01	843.07	668.81	686.46
Contaminated normal	-451.90	844.47	670.40	688.05
Skew contaminated normal	-449.50	843.75	669.29	688.28

From Table 1, we can see that the LPML, DIC, EBIC and EAIC statistics agreed that the skew-t distribution is the preferred model. From these results, it is possible to observe that the data present skewness, because all the skewed models performed better results than their non-skewed versions. Moreover, the normal model has a poorer fit, indicating the existence of asymmetry and heavy tails in the data, which can be accommodated by the skew-t model.

From Table 1, the skew-t model is preferred, and we hence continue our analysis focusing on the skew-t model. The results presented in Table 2 complements the ones obtained in the skew-t and normal regressions. In the skew-t analysis, we can see that the higher income of the region has less chance of dengue fever contamination, while in the normal regression, the higher treated water has the higher chance of contamination. This is a counter intuitive result because people who receive treated water do not have to keep water in containers, which is the preferred reproduction location of the mosquito *Aedes aegypti*. This result is probably due to the fact that the normal distribution cannot correctly accommodate the regions that have unexpected counts of dengue fever (outliers). Moreover, we can see that the 95% highest posterior density (HPD) intervals for the skew-t distribution are smaller than for the normal case. From Table 1, we can see that the estimates of the tail parameter $\nu = 5.3$ and the skewness parameter $\lambda = -0.77$, with tight HPD intervals, strongly evidence the necessity of more flexible models.

Table 2. The estimates for the skew-t and normal models.

Coefficients	skew-t		normal	
	Estimates	95% HPD interval	Estimates	95% HPD interval
Intercept	-0.63	(-0.78,-0.40)	-0.80	(-1.10,-0.50)
Income	-0.38	(-0.69,-0.08)	-0.31	(-0.64,0.01)
Water	0.07	(-0.13,0.27)	0.38	(0.05,0.73)
λ	-0.77	(-1.33,-0.35)		
ν	5.30	(2.05,12.12)		

The standardized mortality ratio (SMR) provides a point estimate of the infection relative risk, δ_i, of each county. The SMR is defined as

$$\text{SMR}_i = \frac{Y_i}{\text{ENI}_i},\ 1,\ldots,92,$$

for the state of Rio de Janeiro.

Figure 3 presents the SMR and the mean posterior relative risk (RR) of each county in Rio de Janeiro. It is clear that the mean RR is very close to the empirical point estimate. This indicates a good fit of the skew-t model to the data and allows epidemiologists to make inference on the RR of each county using the posterior sample.

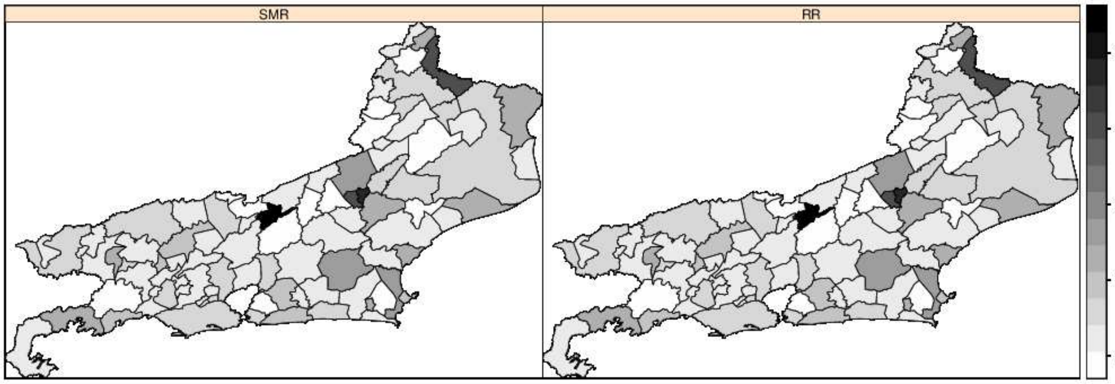

Figure 3. Left panel: SMR estimates for the counties of Rio de Janeiro. Right panel: the posterior estimates of the RR for the counties of Rio de Janeiro.

6. Conclusions

In this paper, we have presented the class of generalized skew-normal/independent distributions. The proposed class extends the skew-normal/independent class proposed by Bandyopadhyay et al. (2010). With the use of the presented class and generalized linear mixed models, we have incorporated the notion of asymmetric spatial fields with the creation of the generalized skew-Gaussian spatial fields. To illustrate an application of these spatial fields, we have presented a dengue fever incidence study for the counties in the state of Rio de Janeiro. From our analysis, it is clear that the skewed distributions outperformed the symmetric distributions and there have been also a need for heavy-tail models to improve the fit to the data. The skew-t distribution had the best fit among the possible options, providing that counties with lower income have higher risk of dengue fever infection. The presented models are easily implemented in the OpenBUGS software, by means of which these models can be used to analyze data when the symmetry or normality assumptions are not appropriate for empirical data with spatial dependence.

Acknowledgments

The first and third authors acknowledge the partial financial support from FAPESP. The first author also acknowledge FAPEMIG for partial financial support.

References

Allard, D., Naveau, P., 2007. A new spatial skew-normal random field model. Communications in Statistics - Theory and Methods, 36, 1821–1834.

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

Arellano-Valle, R.B., Genton, M.G., 2005. On fundamental skew distributions. Journal of Multivariate Analysis, 96, 93–116.

Arellano-Valle, R.B., Genton, M.G., 2010. Multivariate unified skew-elliptical distributions. Chilean Journal of Statistics, 1, 17–33.

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

Bandyopadhyay, D., Lachos, V.H., Abanto-Valle, C.A., Ghosh, P., 2010. Linear mixed models for skew-normal/independent bivariate responses with an application to periodontal disease. Statistics in Medicine, 29, 2643–2655.

Besag, J., 1974. Spatial interaction and the statistical analysis of lattice data systems (with discussion). Journal of The Royal Statistical Society Series B - Statistical Methodology, 36, 192–225.

Breslow, N.E., Clayton, D.G., 1993. Approximate inference in generalized linear mixed models. Journal of the American Statistical Association, 88, 9–25.

Brooks, S.P., 2002. Discussion on the paper by Spiegelhalter, Best, Carlin and van der Linde (2002). Journal of The Royal Statistical Society Series B - Statistical Methodology, 64, 616–618.

Carlin, B.P., Louis, T.A., 2000. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall, Boca Raton, FL.

Dey, D.K., Chen, M.H., Chang, H., 1997. Bayesian approach for nonlinear random effects models. Biometrics, 53, 1239–1252.

Geisser, S., 1993. Predictive Inference: An Introduction. Chapman & Hall, London.

Genton, M.G. (ed.), 2004. Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality. Chapman & Hall, Boca Raton, FL.

Karimi, O., Omre, H., Mohammadzadeh, M., 2010. Bayesian closed-skew Gaussian inversion of seismic AVO data for elastic material properties. Geophysics, 75, 185–198.

Kim, H.-M., Mallick, B.K., 2004. A Bayesian prediction using the skew Gaussian distribution. Journal of Statistical Planning and Inference, 120, 85–101.

Lachos, V.H., Ghosh, P., Arellano-Valle, R.B., 2010. Likelihood based inference for skew–normal independent linear mixed models. Statistica Sinica, 20, 303–322.

Lunn, D., Spiegelhalter, D., Thomas, A., Best, N., 2009. The BUGS project: Evolution, critique and future directions (with discussion). Statistics in Medicine, 28, 3049–3082.

Prates, M.O., Dey, D.K., Willig, M.R., Yan, J., 2011a. Intervention analysis of hurricane effects on snail abundance in a tropical forest using long-term spatiotemporal data. Journal of Agricultural, Biological, and Environmental Statistics, 16, 142–156.

Prates, M.O., Dey, D.K., Willig, M.R., Yan, J., 2011b. Transformed Gaussian markov random fields and spatial modeling. Technical report. University of Connecticut, Statistics Department.

Prates, M.O., Lachos, V.H., Barbosa-Cabral, C.R., 2011. mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions. Available at `http://CRAN.R-project.org/package=mixsmsn`.

Sahu, S.K., Dey, D.K., Branco, M.D., 2003. A new class of multivariate skew distributions with applications to Bayesian regression models. Canadian Journal of Statistics, 31, 129–150.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, A., 2002. Bayesian measures of model complexity and fit. Journal of The Royal Statistical Society Series B - Statistical Methodology, 64, 583–639.

Whittle, P., 1954. On stationary processes in the plane. Biometrika, 41, 434–439.

Zeller, C.B., 2009. Distribuições de misturas da escala skew-normal: Estimação e diagnóstico em modelos lineares. Ph.D. Universidade Estadual de Campinas, São Paulo, Brazil.