Generelized Skew-Normal/Independent Fields with Applications

Marcos O. Prates[1], Dipak K. Dey[2,3,4], and Victor H. Lachos[1]

[1]Departmento de Estatística, Universidade Estadual de Campinas

[2]Department of Statistics, University of Connecticut

[3]Institute for Public Health Research, University of Connecticut Health Center

[4]Center for Environmental Sciences & Engineering, University of Connecticut

November 11, 2011

Abstract

The last decade has witnessed major developments in Geographical Information Systems (GIS) technology resulting in the need for Statisticians to develop models that account for spatial clustering and variation. Study of spatial patterns are very important in epidemiological and environmental problems. Due to spatial characteristics it is extremely important to correctly incorporate spatial dependence in modeling. This paper develops a novel spatial process using generalized skew–normal/independent distributions when the usual Gaussian process assumptions are invalid and transformation to a Gaussian random field is not appropriate. Our proposed method incorporates skewness as well as heavy tail behavior of the data while maintaining spatial dependence using a Conditional Auto Regressive (CAR) structure. We use Bayesian hierarchical methods to fit such models. Consequently we use a Bayesian model selection approach to choose appropriate models for a empirical data set.

KEY WORDS: Bayesian hierarchical methods, Conditional Auto Regressive (CAR), Conditional predictive ordinate, Markov Chain Monte Carlo (MCMC), skew normal distributions, skew normal independent distributions, spatial association.

1. INTRODUCTION

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

To reduce unrealistic assumptions, there is a tendency in spatial data analysis to incorporate the spatial dependence to represent features of the data as realistic as possible. Such assumption drive spatial data analysis towards more flexible methods.

Generalized linear mixed models (GLMM) (Breslow and Clayton 1993) are an important class of statistical models that are widely used to model dependent data. For dependent data, such as spatial data, GLMMs introduce dependence through random effects. Under the GLMM framework, scientists have been using Simultaneous Auto-regressive (SAR) (Whittle 1954) as well as Conditional Auto-regressive (CAR) (Besag 1974), as tools to accommodate spatial dependence for modeling areal data. However, the Gaussian assumption in the random effect implies symmetry and thin tail which may not be appropriate for many applications (e.g., Prates et al. 2011a).

When the Gaussian assumption for the random effects is not adequate, e.g., when data is skewed or present a heavy tail behavior, we need alternative distribution to realistically represent the data (Branco and Dey 2001). Sahu et al. (2003) and Arellano-Valle and Azzalini (2006) define the skew–normal (SN) distribution as follows. A random vector $\mathbf{Y}$ is said to follow a d–variate SN distribution with location vector $\boldsymbol{\mu} \in \mathbb{R}^d$, scale matrix $\boldsymbol{\Sigma}_{d\times d}$ positive definite (p.d.) and skewness matrix $\boldsymbol{\Lambda} = \text{Diag}(\boldsymbol{\lambda})$, where $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)^\top$, if its probability density function (pdf) is

$$f(\mathbf{y}) = 2^d\phi_d(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega})\Phi_d(\boldsymbol{\Lambda}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})|\boldsymbol{\Delta}), \quad \mathbf{y} \in \mathbb{R}^d, \tag{1}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top$, $\boldsymbol{\Delta} = (\mathbf{I} + \boldsymbol{\Lambda}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{\Lambda})^{-1} = \mathbf{I} - \boldsymbol{\Lambda}^\top\boldsymbol{\Omega}^{-1}\boldsymbol{\Lambda}$ ($\mathbf{I}$ is a d– dimensional identity matrix), $\phi_d(.|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\Phi_d(.|\boldsymbol{\Sigma})$ denote the pdf of $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and cumulative density function (CDF) of $N_d(\mathbf{0}, \boldsymbol{\Sigma})$, respectively. We write $\text{SN}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ to indicate that $\mathbf{Y}$ has density (1). For $\boldsymbol{\Lambda} = 0$, equation (1) reduces to the usual symmetric multivariate $N_d(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution. The SN distribution as defined in (1) can be stochastically represented as

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

where $\mathbf{T}_1 \sim N_d(0, \mathbf{I})$ is independent of $\mathbf{T}_2 \sim N_d(0, \boldsymbol{\Sigma})$ and $|\mathbf{T}_1|$ denotes the component wise absolute value of $\mathbf{T}_1$, thus $|\mathbf{T}_1|$ follows a d–dimensional standard half-normal distribution denoted by $HN_d(\mathbf{0}, \mathbf{I})$. Note that the expression (2) provides a representation which is a useful tool for a generation of random observation from (1). This representation is also useful for developing various theoretical properties of the SN distribution. According to Arellano-Valle and Azzalini (2006), the expectation and covariance matrix of $\mathbf{Y} \sim \text{SN}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$ are given by

$$E[\mathbf{Y}] = \boldsymbol{\mu}_{SN} = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\boldsymbol{\lambda} \text{ and } Var[\mathbf{Y}] = \boldsymbol{\Sigma}_{SN} = \boldsymbol{\Sigma} + (1 - \frac{2}{\pi})\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top.$$

Since the matrix $\boldsymbol{\Lambda}$ is diagonal, the introduction of skewness does not affect the spatial dependence structure. When $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$ then the density (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 (e.g., Prates et al. 2011a).

In this paper, we extend the definition of the multivariate Skew-normal/independent (SNI) distribution (Lachos et al. 2010) to introduce a new class of Spatial processes. We propose a generalized SNI (GSNI) process in which the covariance matrix can be partitioned in two components: a spatial component and a skewness component. The class of GSNI distribution has as specific members the normal and skew-normal distributions as will be presented in Section 2.

The National Cancer Institute offers public use county-level lung cancer summaries for 99 counties of Iowa, U.S.A. Although cigarettes is the main cause of lung cancer worldwide it is not the only one. The radon is a radioactive, colorless, odorless, tasteless noble gas, occurring naturally in nature as a decay product of uranium or thorium. Radon gas levels were measured per counties in the state of Iowa. The Radon is considered a health hazard due to its radioactivity. The United States Environmental Protection Agency believes that radon is the second most frequent cause of lung cancer (U.S. Environmental Protection Agency 2009). Epidemiological studies have shown that high levels of indoor concentrations of radon increases the risks of lung cancer. We present a study the of the death by lung cancer in the state of Iowa including radon and other covariates.

The article is organized as follows, in Section 2 we introduce GSNI distribution and its properties. Using the GSNI we describe how to accommodate spatial dependence. Consequently, we define a new generalized Skew–Gaussian spatial field in Section 3. We introduce the conditional predictive ordinate (CPO) as a model selection criterion in Section 4. In Section 5 we illustrate the new proposed methodology with empirical data analysis on lung cancer in the state of Iowa by counties. We conclude the paper in Section 6 with a discussion.

2. GENERALIZED SKEW-NORMAL/INDEPENDENT DISTRIBUTIONS

The SNI distribution is defined as a scale mixture of the skew-normal distribution as presented in (1). A density $\pi(x)$ is a scale mixture of densities when it can be represented as

$$\pi(x) = \int f(x|u) dH(u),$$

where $f(\cdot|u)$ is the conditional density of the random vector $\mathbf{X}$ given $\mathbf{U} = u$, called the scale factor, following a distribution function H, where H is called the mixing distribution. The pdf $\pi(x)$ belongs to normal independent (NI) family when $f(\cdot|u)$ is a normal density with mean vector $\boldsymbol{\mu}$ and covariance matrix $u^{-1}\boldsymbol{\Sigma}$ and $\mathbf{U}$ is a positive random variable (Andrews and Mallows 1974). Thus, the SNI distribution is generated by scale mixtures of skew–normal distributions (see Lachos et al. 2010, for more details).

Following Zeller (2009), we define the d–dimensional Generalized SNI (GSNI) vector $\mathbf{Y}$, denoted from now on as $Y \sim \mathrm{GSNI}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, as a multivariate mixture of skew–normal distribution, where $\mathbf{U} = (U_1, \ldots, U_d)^\top$ is a vector instead of a scalar. Thus, $\mathbf{Y}$ can be constructed as

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{U}^{-1/2} \odot \mathbf{Z}, \tag{3}$$

where $\mathbf{Z} \sim \mathrm{SN}_d(\mathbf{0}, \boldsymbol{\Sigma}, \boldsymbol{\Lambda})$, $\mathbf{U}^{-1/2} = (U_1^{-1/2}, \ldots, U_d^{-1/2})^\top$, and U_i's are positive, independent and identically distributed random variables, independent of $\mathbf{Z}$, with CDF $H_d(\cdot; \nu) = \prod_{i=1}^d H(\cdot; \nu)$ and ν is a parameter of the distribution H. In (3), $\odot$, represents the Hadamard product, i.e., $\mathbf{Y} \odot \mathbf{Z} = (Y_1 Z_1, \ldots, Y_d Z_d)^\top$ if both $\mathbf{Y}$ and $\mathbf{Z}$ are of dimension d, and $\mathbf{Y} \odot \mathbf{Z} = (Y_1 Z, \ldots, Y_d Z)^\top$ if Z is scalar. Clearly, when $\mathbf{U}^{-1/2}$ is set to be a scalar the GSNI is equivalent to the SNI distribution proposed by Lachos et al. (2010). Similar to the proposition 2.4 presented in Arellano-Valle and Genton (2005), we obtain that given $\mathbf{U} = \mathbf{u}$, $\mathbf{Y}$ has a SN distribution with location vector $\boldsymbol{\mu}$, scale matrix $\boldsymbol{\Sigma}_u = \mathrm{Diag}(\mathbf{u}^{-1/2})\boldsymbol{\Sigma}\mathrm{Diag}(\mathbf{u}^{-1/2})$ and skewness parameter matrix $\boldsymbol{\Lambda}_u = \mathrm{Diag}(\mathbf{u}^{-1/2})\boldsymbol{\Lambda}$, i.e., $\mathbf{Y}|\mathbf{U} = \mathbf{u} \sim \mathrm{SN}_d(\boldsymbol{\mu}, \boldsymbol{\Sigma}_u, \boldsymbol{\Lambda}_u)$. Hence, the density function of $\mathbf{Y}$ is given by

$$f(\mathbf{y}) \quad = \quad 2^d \int_{\mathbb{R}_+^d} \phi_d(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Omega}_u)\Phi_d(Diag(\mathbf{u}^{1/2})\boldsymbol{\Lambda}^\top\boldsymbol{\Omega}^{-1}(\mathbf{y} - \boldsymbol{\mu})|\boldsymbol{\Delta})dH_d(\mathbf{u})d\mathbf{u}, \tag{4}$$

where $\boldsymbol{\Omega}_u = Diag(\mathbf{u}^{-1/2})\boldsymbol{\Omega}Diag(\mathbf{u}^{-1/2})$.

From (1) and (3) it is clear that

$$E[\mathbf{Y}] = \boldsymbol{\mu}_{GSNI} = \boldsymbol{\mu} + \sqrt{\frac{2}{\pi}}\kappa_1\boldsymbol{\lambda}$$

and

$$Var[\mathbf{Y}] = \mathbf{\Sigma}_{GSNI} = \kappa_2(\nu)\mathbf{\Sigma} + (\kappa_2(\nu) - \frac{2}{\pi}\kappa_1^2(\nu))\mathbf{\Lambda}\mathbf{\Lambda}^\top, \tag{5}$$

where $\kappa_\alpha(\nu) = E[U^{-\alpha/2}]$, $\alpha \in \mathbb{R}$ and the moments are well defined. As observed in the skew–normal case, $\mathbf{\Sigma}$ is responsible for accommodating dependence in $\mathbf{\Sigma}_{GSNI}$ because $\mathbf{\Lambda}$ is diagonal.

Cabral et al. (2012) show that the asymmetrical class of SNI distributions contains a variety of skewed distributions with different choices of the mixture $\boldsymbol{U}$, such as:

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

All these distributions have heavier tails than that of the SN one and can be used for robust inference. Note that, like the SN distribution, the components of the SNI class are uncorrelated, when $\mathbf{\Sigma}$ is a diagonal matrix. Further, in order to have a zero-mean vector ($\boldsymbol{\mu}_{SNI} = 0$), we should assume the location parameter $\boldsymbol{\mu} = -\sqrt{\frac{2}{\pi}}\kappa_1\boldsymbol{\lambda}$, 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 correspond to response Y_i at site i. Let $\boldsymbol{e} = (e_1, \ldots, e_n)^\top$ be a vector of unobserved random effects with joint distribution H, which introduces spatial dependence. A spatial GLMM assumes that, given $(\boldsymbol{X}_i, e_i)$, the observations Y_i's are independent with density $h(y_i;\theta_i)$ belonging to a one parameter exponential family

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

where $a(\cdot)$, $\psi(\cdot)$ and $c(\cdot)$ are known functions, ϕ is a scale or dispersion parameter, θ_i is the canonical parameter, and the support of the distribution does not depend on θ_i. Let $\mu_i = E(Y_i|\boldsymbol{X},\boldsymbol{e})$, 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:

$$g(\mu_i) = \eta_i + e_i, \tag{6}$$

where $\eta_i = \boldsymbol{X}_i^\top\beta$ is the fixed effect, and β 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 H of $\boldsymbol{e}$. Commonly, H is chosen to be a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}$.

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

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

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, $\boldsymbol{M} = \text{diag}\,(\tau_1^2,\ldots,\tau_n^2)$. To assure that Σ is positive definite we need some constraints $w_{ij}\tau_j^2 = w_{ji}\tau_i^2$ and $\rho \in (1/\lambda_{\min}, 1/\lambda_{\max})$ where λ's are the eigen values of $\boldsymbol{M}^{-1/2}\boldsymbol{W}\boldsymbol{M}^{1/2}$.

A random vector Z is defined to follow a Generalized Skew–Gaussian Spatial Field (GS-GSF) when $Z \sim \text{GSNI}_n(\boldsymbol{\mu},\boldsymbol{\Sigma},\boldsymbol{\Lambda})$ and $\boldsymbol{\Sigma}$ has a spatial dependence generated by CAR (SAR) structure. Thus, $\mathbf{\Sigma}_{GSNI}$ will depend on its neighbors since the dependence structure of $\mathbf{\Sigma}_{GSNI}$ comes entirely from $\mathbf{\Sigma}$.

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 $Y = (Y_1, \ldots, Y_n)^\top$ that comes from a one parameter exponential family distribution with probability density/mass function (pdf) h. Thus, we can model the response Y as

$$\begin{aligned} &Y_i \sim h(\mu_i), i = 1, \ldots, n, \\ &g(\mu_i) = X_i\beta + \phi, \\ &\phi \sim GSGSF_n(-\sqrt{\frac{2}{\pi}}\kappa_1\lambda\mathbf{1}_n, \mathbf{\Sigma}, \lambda\boldsymbol{I}), \end{aligned}$$

where g is a link function, X_i is covariates for $i = 1, \ldots, n$, β are the regression coefficients, $\mathbf{1}_n = (1, \ldots, 1)^\top$, $\mathbf{\Sigma}$ is spatial dependence matrix generated by the covariance structure of a CAR or SAR model and λ is the skewness parameter to avoid overparametrization and identifiability problems. With this representation this approach provides a flexible way to incorporate multivariate asymmetric spatial random effects into modeling.

4. CONDITIONAL PREDICTIVE ORDINATE

The CPO model comparison is a Bayesian cross-validation approach (e.g., Gelfand et al. 1992; Geisser 1993; Dey et al. 1997). Let $\boldsymbol{y}$ be the observed responses $\{Y_i : i = 1, \ldots, n\}$. Let $\boldsymbol{y}_{-i}$ denote the observed response vector excluding the i^{th} observation. The CPO statistic associated with the i^{th} observation is defined as the marginal posterior predictive density of y_i, conditioning on $\boldsymbol{y}_{-i}$:

$$\text{CPO}_i = \int f(y_i|\boldsymbol{\theta})\pi(\boldsymbol{\theta}|\boldsymbol{y}_{-i})\text{d}\boldsymbol{\theta}, \tag{7}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$ is the vector of parameters of the distribution f, $f(y_i|\boldsymbol{\theta})$ is the conditional probability mass/density function 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 (7) is not available, Dey et al. (1997) showed that CPO_i can be estimated from a Monte Carlo integration approach and is approximated by a harmonic mean:

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

where B denotes the size of a MCMC sample of the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$, $\boldsymbol{\theta}^{(j)}$ is the parameter vector $\boldsymbol{\theta}$ in the j^{th} MCMC sample. This approximation is valid when Y_i are assumed to be conditionally independent given $\boldsymbol{\theta}$. Since 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 model, the logarithm of the pseudo-marginal likelihood (LPML), LPML $= \sum_{i=1}^{n} \log \widehat{\text{CPO}}_i$. 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 observation number i; points supports either one of the models are above and below the zero lines, respectively (e.g., Prates et al. 2011b).

5. IOWA LUNG CANCER

The National Cancer Institute is perhaps the most complete source of cancer data in the US, offering public use county-level summaries for several states in various parts of the country. We collected the death counts of Iowa lung cancer. The Iowa lung cancer data consist of the average death count of lung cancer patients from the 99 counties of the state of Iowa who have died between 2003 to 2007 (http://factfinder.census.gov/).

To study the lung cancer death counts over Iowa counties, we include potential important covariates to describe the death counts. First, we include the Radon average per county over the years between 2003 to 2007. The Radon is radioactive, colorless, odorless, tasteless noble gas, occurring naturally as the decay product of radium. It is one of the densest substances that remains a gas under normal conditions and is considered to be a health hazard due to its radioactivity (http://www.idph.state.ia.us/eh/radon.asp). Another important factor to account for cancer death rates is the population (in thousands) at risk per county. From the National Census Bureau, demographic information were collected and may help understanding and interpreting the risk factors of lung cancer deaths. For each county we collected the percentage of people with a bachelor or higher degree, percentage of seniors and average house income (in thousand dollars).

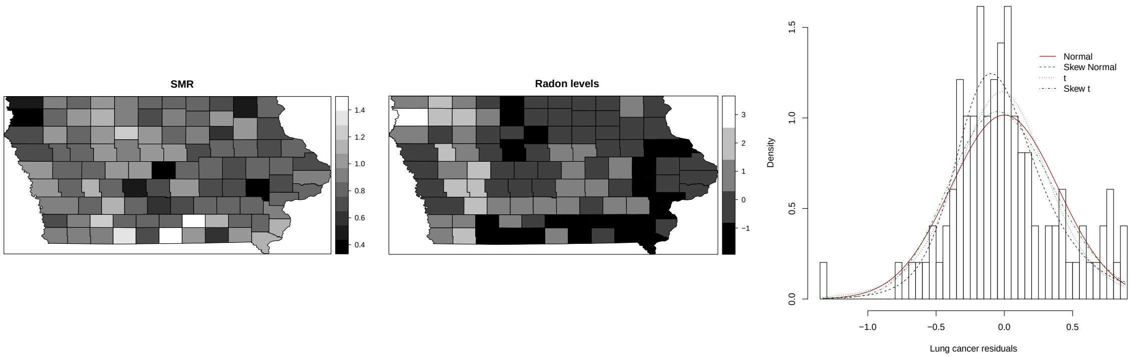
5.1 Exploratory Analysis

We initiate our study checking for the necessity of spatial random effects. Moreover, we would like to check that the residuals must be modeled by a distribution with heavy tail and/or asymmetry. Intuitively, the use of spatial data analysis is important since as we can see in Figure 1 the Standardized Mortality Ratio (SMR) of lung cancer and the radon levels do not seem be randomly distributed in the space. We use the Moran's I statistic (Moran 1950) to validate our observations. We conclude that both SMR and radon levels are spatially dependent with p-values 0.001 and 0.015 respectively.

After demonstrating the necessity of spatial effect in the data we continue to investigate the appropriateness of assymetric and/or heavy tail models instead of the common used normal regression. Let Y_i be the number of lung cancer deaths in each Iowa county, $i = 1, \ldots, 99$. In order to do so, we fit a CAR regression in the

$$\log(Y+1) = X\beta + e,$$

where e has a CAR distribution. After, we perform a modeling fitting in the residuals of the analysis to verify what type of distribution better fits the residuals. Using the *mixsmsn* R



(a) (b) (c)

Figure 1: (a) SMR levels for the counties of Iowa. (b) Radon levels for the counties of Iowa. (c) Histogram of the of the residuals of the CAR regression with different SNI distributions fitting.

package (Prates et al. 2011c), for fitting the residual with distributions of the class SNI, we get that under the AIC criterion the skew-t distribution is the one the best fit the residuals. This can also be verified in the histogram in Figure 1(c). Therefore we conclude that the normal assumption for the residuals are not the most appropriate. In the next section, we study the lung cancer death using the GSGSF to improve fitting.

5.2 Lung Cancer with GSGSF

In order, to accommodate spatial dependence between the neighboring counties, we use the proper precision matrix specified in the CAR model. The precision matrix can be represented as $\boldsymbol{Q} = \frac{1}{\sigma^2}\boldsymbol{M}^{-1}(\boldsymbol{I} - \rho\boldsymbol{W})$, where $\boldsymbol{M}$ is diagonal matrix with each M_{ii} corresponding to the number of neighbors of site i, n_i, $\boldsymbol{W}$ is a normalized weight matrix, σ^2 is a parameter to measure the overall variability in the model and ρ measures the spatial association. Recall from Section 3 that the GSGSF covariance can be partitioned by its scale matrix $\boldsymbol{\Sigma}$ and the skewness parameter λ. Therefore, given σ^2 and ρ, we can define $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ and use it in GSGSF representation to calculate $\boldsymbol{\Sigma}_{GSGSF}$ which is basically a re-scaled sum of $\boldsymbol{\Sigma}$ and the skewness parameter.

Suppose that for each county, $i = 1, \ldots, 99$, we observe the lung cancer death (Y_i) and the set of covariates X_i. Then we model the death by county as

$$\begin{aligned} Y_i &\sim \text{Poisson}(\mu_i),\ i = 1, \ldots, 99, \\ \log(\mu_i) &= X_i\beta + \phi \\ \phi &\sim \text{GSGSF}_{99}(-\sqrt{\frac{2}{\pi}}\kappa_1\lambda\mathbf{1}_{99}, \mathbf{\Sigma}, \lambda\boldsymbol{I}), \end{aligned}$$

where β are the regression coefficients and ϕ accommodates the underlying spatial dependency with $\mathbf{\Sigma}$ with a CAR structure. To fully specify the model, vague hyper–priors are chosen. The skewness parameter, λ, is set to follow a $N(0, 100)$, the spatial dependence parameter, ρ, follows $U(0, 1)$ and the overall precision parameter is set to follow a $\Gamma(0.01, 0.01)$.

Within the new family proposed in Section 3, we propose 5 new models (skew–normal, Student–t, skew–t, slash and skew–slash) as alternative to the normal spatial field to analyze the number of death by lung cancer at the Iowa counties. The LPML statistics is used to perform our model selection. The estimated LPML values are -304.30, -296.62, -291.59, -292.54, -290.26, and -291.37 for the normal, skew normal, Student-t, skew-t, slash and skew slash processes respectively. From this results it is possible to observe that data does not present skewness since there is no significant difference between the skew-t (skew slash) and Student-t (slash) distributions by the LPML criterion. Because both normal and skew normal have a poorer fit there is an indication that the data have heavy tails, which can be accommodated by both student-t and slash models. If we focus in analyzing the normal class, it is possible to see improvement between the normal and the skew normal models, this can be due to the fact that the skewness parameter can somehow capture the heavy tail behavior.

From the LPML analysis we have the slash model has the bigger LPML, and so, the corresponding results are presented in Table 1. Observing the HPD intervals we can conclude that population and the covariate bachelor percentage are significant factor in predicting lung cancer deaths while the others seems not significant. The population factor apparently

Table 1: The estimates for the lung cancer death counts for slash model.

Coefficients	Estimates	95% HPD interval
Intercept	2.42	(0.23,4.64)
Radon	-0.02	(-0.05,0.05)
Bachelor percentage (10^{-2})	-2.14	(-4.12,-0.17)
Senior percentage	-1.08	(-4.45,2.45)
Population (10^{-2})	2.04	(1.44,2.71)
House Income (10^{-3})	0.97	(-1.37,3.08)
ν	1.33	(1.10,1.82)

increases the lung cancer death for bigger population, this could be due to the fact that the urban areas tend to have more people at risk and more risk factors for lung cancer than rural areas. The bachelor percentage factor tends to decrease the cancer deaths which could be due to the fact that people with bachelor degree have a better information and condition to early detect lung cancer improving chance of cure. Initially we expected that radon levels should be an important predictor to lung cancer deaths. However, such assumption is not validated in our analysis. An explanation for this apparent contradiction is that epidemiologists studies shown that high indoor radon levels increases the chance of lung cancer, but our radon levels are not indoors but outdoors. The slash distribution is symmetric ($\lambda = 0$) but ν captures the tail behavior. Since ν is estimated to be 1.33 we can see, as expected, that there is a heavy-tail underlying spatial process that could not be captured by the common normal model.

6. CONCLUSIONS

In this paper we presented a the class of distributions called GSNI. The proposed class successfully extend the SNI class proposed by Lachos et al. (2010). With the use of the

GSNI in GLMM we incorporated the notion of asymmetric spatial fields with the creation of the GSGSF.

To illustrate applications of the GSGSF, we presented lung cancer death for the state of Iowa. From our analysis it is clear that the Slash distribution have better fitting than the traditional CAR model. Also, we were able to show that distributions that can handle heavy tail highly improved the fitting when compared with thin tail distributions.

The presented models are easily implemented in WinBUGS and can be used to analyze data when the symmetric or the thin tails assumptions are not appropriate for empirical data with a spatial dependence.

ACKNOWLEDGMENTS

The first and third author acknowledges the partial financial support from FAPESP.

REFERENCES

Andrews, D. F. and Mallows, C. L. (1974), “Scale Mixtures of Normal Distributions,” *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, 36, 99–102.

Arellano-Valle, R. B. and Azzalini, A. (2006), “On the Unification of Families of Skew-Normal Distributions,” *Scandinavian Journal of Statistics*, 33, 561–574.

Arellano-Valle, R. B. and Genton, M. G. (2005), “On fundamental skew distributions,” *Journal of Multivariate Analysis*, 96, 93–116.

Besag, J. (1974), “Spatial Interaction and the Statistical Analysis of Lattice Data Systems (with discussion),” *Journal of the Royal Statistical Society, Series B*, 36, 192–225.

Branco, M. D. and Dey, D. K. (2001), “A General Class of Multivariate Skew-Elliptical Distributions,” *Journal of Multivariate Analysis*, 79, 99–113.

Breslow, N. E. and Clayton, D. G. (1993), “Approximate Inference in Generalized Linear Mixed Models,” *Journal of the American Statistical Association*, 88, 9–25.

Cabral, C. R. B., Lachos, V. H., and Prates, M. O. (2012), “Multivariate mixture modeling using skew-normal independent distributions,” *Computational Statistics & Data Analysis*, 56, 126 – 142.

Dey, D. K., Chen, M. H., and Chang, H. (1997), “Bayesian Approach for Nonlinear Random Effects Models,” *Biometrics*, 53, 1239–1252.

Geisser, S. (1993), *Predictive Inference: An Introduction*, London: Chapman & Hall Ltd.

Gelfand, A. E., Dey, D. K., and Chang, H. (1992), “Model Determination Using Predictive Distributions, with Implementation Via Sampling-based Methods (Disc: P160-167),” in *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, eds. Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M., Clarendon Press [Oxford University Press], pp. 147–159.

Lachos, V. H., Ghosh, P., and Arellano-Valle, R. B. (2010), “Likelihood Based Inference For Skew–normal Independent Linear Mixed Models,” *Statistica Sinica*, 20, 303–322.

Moran, P. A. P. (1950), “Notes on Continuous Stochastic Phenomena,” *Biometrika*, 37, 17–33.

Prates, M. O., Dey, D. K., R., W. M., and Yan, J. (2011a), “Transformed Gaussian Markov Random Fields and Spatial Modeling,” Tech. Rep. 18, University of Connecticut, Statistics Department.

Prates, M. O., Dey, D. K., Willig, M. R., and Yan, J. (2011b), “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, 10.1007/s13253-010-0039-1.

Prates, M. O., Lachos, V. H., and Cabral, C. R. B. (2011c), *mixsmsn: Fitting finite mixture of scale mixture of skew-normal distributions*, r package version 0.2-9.

Sahu, S. K., Dey, D. K., and Branco, M. D. (2003), "A New Class of Multivariate Skew Distributions with Applications to Bayesian Regression Models," *The Canadian Journal of Statistics*, 31, 129–150.

U.S. Environmental Protection Agency (2009), "A Citzens Guide to Randon," .

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. thesis, Universidade Estadual de Campinas.