GeoModels Tutorial: analysis of spatial data using skew-Gaussian random fields

Moreno Bevilacqua
Víctor Morales-Oñate
Christian Caamaño

October 30, 2018

Introduction

In this tutorial we show how to analyze geo-referenced spatial data using skew-Gaussian random fields (RFs), as depicted in Zhang and El-Shaarawi (2010) and Alegria et al. (2017) with the `R` package `GeoModels`.

We first load the R libraries needed for this tutorial and set the name of the model in the `GeoModels` package:

```
>rm(list=ls())
>require(devtools)
>install_github("vmoprojs/GeoModels")
>require(GeoModels)
>require(fields)
>require(sn)
>model="SkewGaussian"  # model name in the  GeoModels package
>set.seed(89)
```

Simulation of a skew Gaussian random field

Let us consider a spatial Gaussian RF $Z = \{Z(\boldsymbol{s}), \boldsymbol{s} \in S\}$, where $\boldsymbol{s}$, in this tutorial, represents a location in a spatial domain $S \subset R^2$. We assume that Z is stationary with zero mean, unit variance and correlation function given by $\rho(\boldsymbol{h}) = \mathrm{cor}(Z(\boldsymbol{s}+\boldsymbol{h}), Z(\boldsymbol{s}))$.

We consider the RF $Y = \{Y(\boldsymbol{s}), \boldsymbol{s} \in S\}$ defined in Zhang and El-Shaarawi (2010) :

$$Y(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \eta|Z_1(\boldsymbol{s})| + \sqrt{\sigma^2} Z_2(\boldsymbol{s}) \tag{1}$$

where Z_1 and Z_2 are indipendent copies of Z, $\eta \in \mathbb{R}$ is an asymmetry parameter and $\sigma^2 > 0$.

Moreover $\mu(\boldsymbol{s}) = X(\boldsymbol{s})^T\boldsymbol{\beta}$ and $X(\boldsymbol{s}$ is a k-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ is a k-dimensional vector of (unknown) parameters (in this tutorial we fix $k = 2$).

Then, marginally $(Y(\boldsymbol{s})-\mu(\boldsymbol{s}))/(\sigma^2+\eta^2)^{\frac{1}{2}}$ follows a standard skew-Gaussian distribution $SN(0, 1, \alpha)$ (Azzalini and Capitanio, 2014) *i.e*:

$$f_Y(y) = 2\phi(y)\Phi(\alpha y)$$

with $\alpha = \eta/\sigma$.

Under this setting $\mathbb{E}(Y(\boldsymbol{s})) = \mu(\boldsymbol{s}) + \eta(2/\pi)^{1/2}$, $\text{var}(Y(\boldsymbol{s})) = \sigma^2 + \eta^2(1-2/\pi)$ and the correlation function is given by:

$$\rho_U(\boldsymbol{h}, \eta, \sigma^2) = \frac{2\eta^2}{\pi\sigma^2 + \eta^2(\pi-2)}\Big((1-\rho^2(\boldsymbol{h}))^{1/2} + \rho(\boldsymbol{h})\arcsin(\rho(\boldsymbol{h})) - 1\Big) + \frac{\sigma^2\rho(\boldsymbol{h})}{\sigma^2 + \eta^2(1-2/\pi)}. \tag{2}$$

Suppose we want to simulate a realization of Y at $N = 1200$ spatial locations uniformly distributed in the unit square. We first set the spatial coordinates with associated covariates.

```
>N=1200  # number of locations sites
# Define the spatial-coordinates
>x = runif(N); y = runif(N)
>coords=cbind(x,y)
>X=cbind(rep(1,N),runif(N))    # matrix regression
```

We then specify the regression mean, variance, asymmetry and nugget parameters

```
>mean= 0.5; mean1=-0.5    # regression parameters
>sill=1.5
>nugget=0
>skew=-3
```

where `mean`, `mean1` , `sill`, `skew` are respectively β_1, β_2, σ^2 and η

For the correlation function we assume the isotropic Generalized Wendland class:

$$\rho(\boldsymbol{h}; \alpha, \delta, \kappa) = \begin{cases} \frac{1}{B(2\kappa,\delta+1)} \int_{||\boldsymbol{h}||/\alpha}^{1} u(u^2 - (||\boldsymbol{h}||/\alpha)^2)^{\kappa-1}(1-u)^{\delta}\,du, & 0 \le ||\boldsymbol{h}|| < \alpha, \\ 0, & ||\boldsymbol{h}| \ge \alpha, \end{cases} \tag{3}$$

In the spatial case, the model is valid if $\delta \ge 1.5 + \kappa$ and $\kappa \ge 0$.

Using asymptotic arguments Bevilacqua et al. (2018) show that this correlation model exibits the same features of the Matern correlation model from prediction point of view. Additionally, it is compactly supported an interesting feature from computational point of view.

The names of the parameters associated to a given correlation model can be obtained with the `CorrParam` function:

```
>corrmodel = "GenWend"    ## correlation model
```

```
>CorrParam(corrmodel)
[1] "power2" "scale"  "smooth"
```

Then we set the correlation parameters of the Generalized Wendland model:

```
>scale = 0.2
>smooth=1
>power2=5
```

where the `scale` and `smooth` parameters corresponds to α (the compact support of the correlation model) and κ the smoothness parameter. Under this setting the RF is once mean square differentiable

We are now ready to simulate the space time Gaussian RF using the function `GeoSim`:

```
>param=list(mean=mean,mean1=mean1,sill=sill,nugget=nugget,scale=scale,
     skew=skew,power2=power2,smooth=smooth)
>data = GeoSim(coordx=coords, corrmodel=corrmodel,model=model,X=X,
              sparse=TRUE,param=param)$data
```

Note that the option `sparse=TRUE` allows to consider algorithms for sparse matrices when performing Cholesky decomposition, as described in the package `spam` (Gerber et al. (2017)). Informations about the sparsity of the covariance matrix can be obtained though the function `GeoCovmatrix` with the following code:

```
>cc= GeoCovmatrix(coordx=coords, corrmodel=corrmodel,model=model,
             sparse=TRUE,X=X, param=param)$covmatrix
>cc$nozero
[1] 0.1044653
```

This means that (approximatively) 90% of the covariance matrix are zeros $i.e$ the matrix is highly sparsed.

Estimation of a skew Gaussian random field

Estimation of the regression and correlation parameters of the skew Gaussian random field Y can be performed using pairwise likelihood estimation. Given a realization $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), \dots, Y(\boldsymbol{s}_N)$ let $f_Y(y_i, y_j)$ the density of the bivariate random vector $\boldsymbol{Y}_{ij} = (Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))$ given by (Alegria et al., 2017):

$$f_{\mathbf{Y}_{ij}}(y_i, y_j) = 2\sum_{l=1}^{2} \phi_n(\boldsymbol{Y} - \boldsymbol{\mu}; \boldsymbol{A}_l)\Phi_n(\boldsymbol{c}_l; \boldsymbol{0}, \boldsymbol{B}_l) \tag{4}$$

where

$$\begin{aligned}
\boldsymbol{A}_l &= \sigma^2 R + \eta^2 R_l \\
\boldsymbol{c}_l &= \eta R_l(\sigma^2 R + \eta^2 R_l)^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}) \\
\boldsymbol{B}_l &= R_l - \eta^2 R_l(\sigma^2 R + \eta^2 R_l)^{-1} R_l \\
\boldsymbol{\mu} &= [\mu(\boldsymbol{s}_i)]_{i=1}^{n}
\end{aligned}$$

where $R = [\rho(\boldsymbol{s}_i - \boldsymbol{s}_j)]_{i=j=1}^{N}$ is the correlation matrix of the latent Gaussian RF and R_1 and R_2 depend on the correlation matrix R (Alegria et al., 2017).

Then, the pairwise likelihood function is defined as:

$$pl(\boldsymbol{\theta}) = \sum_{i=i}^{N-1} \sum_{j=i+1}^{N} log(f_{\mathbf{Y}_{ij}}(y_{il}, y_{jk}))w_{ij} \tag{5}$$

where w_{ij} are non-negative weights, not depending on $\boldsymbol{\theta}$, specified as:

$$w_{ij} = \begin{cases} 1 & ||\boldsymbol{s}_i - \boldsymbol{s}_j|| < d \\ 0 & \text{otherwise} \end{cases}. \tag{6}$$

and in this case $\boldsymbol{\theta} = (\beta_1, \beta_2, \sigma^2, \eta, \alpha, \delta, \kappa)^T$. The pairwise likelihood estimator $\hat{\boldsymbol{\theta}}_{pl}$ is obtained maximizing (5) with respect to $\boldsymbol{\theta}$. In the `GeoModels` package we can choose the fixed parameters and the parameters that must be estimated. Pairwise likelihood estimation is performed with the function `GeoFit`:

```
## estimation with pairwise likelihood
>start=list(sill=sill,mean=mean,mean1=mean1,scale=scale,skew=skew)
>fixed=list(power2=power2,nugget=nugget,smooth=smooth)
>fit=GeoFit(data=data,coordx=coords,corrmodel=corrmodel,X=X,
               maxdist=0.04,model=model,
               start=start,fixed=fixed)
```

Note that the option `maxdist=0.04` set the (arbitrary) compact supports of the weight function (6) i.e. $d = 0.04$. A suitable choice of the weights allows to improve both the statistical and computational efficiency (Alegria et al., 2017).

The object `fit` include informations about the pairwise likelihood estimation:

```
>fit
##################################################################
Maximum Composite-Likelihood Fitting of SkewGaussian Random Fields
Setting: Marginal Composite-Likelihood
Model associated to the likelihood objects: SkewGaussian
Type of the likelihood objects: Pairwise
Covariance model: Wend0
Number of spatial coordinates: 1200
Number of dependent temporal realisations: 1
Type of the random field: univariate
Number of estimated parameters: 5
Type of convergence: Successful
Maximum log-Composite-Likelihood value: -14159.45
Estimated parameters:
   mean    mean1    scale     sill     skew
 0.4902  -0.4494   0.1918   1.5353  -2.9251
##################################################################
```

Checking model assumptions

Given the estimation of the mean regression, variance and skewness parameters, the estimated residuals

$$\widehat{\epsilon(\boldsymbol{s}_i)} = \frac{Y(\boldsymbol{s}_i) - X(\boldsymbol{s}_i)^T \widehat{\boldsymbol{\beta}}}{(\widehat{\sigma}^2 + \eta^2)^{\frac{1}{2}}} \qquad i = 1, \dots, N$$

can be viewed as a realization of a zero mean stationary RF with marginal distribution $SN(0, 1, \alpha)$ with $\alpha = \eta/\sigma$ and with correlation function $\rho_U(\boldsymbol{h}, \eta/(\eta^2 + \sigma^2)^{\frac{1}{2}}, \sigma^2/(\eta^2 + \sigma^2))$. The estimated residuals can be computed using the `GeoResiduals` function:

```
>res=GeoResiduals(fit)  # computing residuals
```

Then the marginal distribution assumption on the residuals can be graphically checked for instance with a qq-plot (Figure 1, left part):

```
### checking model assumptions: marginal distribution
>probabilities = (1:N)/(N+1)
>skgauss.quantiles = qsn(probabilities, xi=0, omega=1,
     alpha=as.numeric(fit$param['skew']/fit$param['sill']^0.5))
```

```
>plot(sort(skgauss.quantiles), sort(c(res$data)),
   xlab="",ylab="",main="Skew-Gaussian␣qq-plot")
>abline(0,1)
```

The correlation model assumption can be checked comparing the empirical and the estimated space-time semivariogram functions using the `GeoVariogram` and `GeoCovariogram` functions (Figure 1, right part):

```
### checking model assumptions: ST semi-variogram model
>vario = GeoVariogram(data=res$data,coordx=coords,maxdist=0.5)
>GeoCovariogram(res,show.vario=TRUE, vario=vario,pch=20)
```

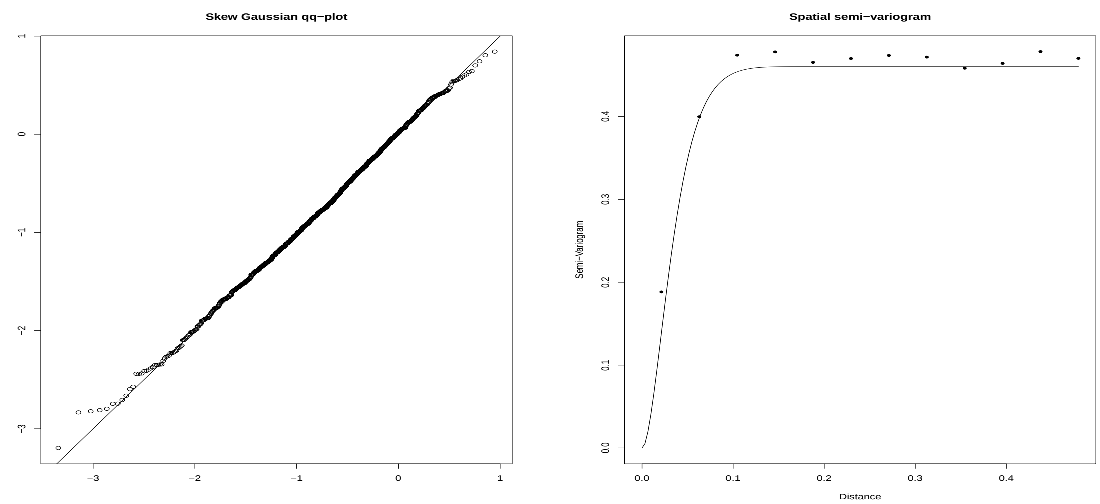

Figure 1: Left: QQ-plot for the residuals of the skew Gaussian RF. Right: space-time empirical vs estimated semi-variogram function for the residuals

Prediction of skew Gaussian random fields

For a given space time location $(\boldsymbol{s}_0)$ with associated covariates $X(\boldsymbol{s}_0)$, the optimal linear prediction of a skew Gaussian RF is given by:

$$\widehat{Y}(\boldsymbol{s}_0) = \mu(\boldsymbol{s}_0) + \sum_{i=1}^{N} \lambda_i [Y(\boldsymbol{s}_i) - \mu(\boldsymbol{s}_i)] \tag{7}$$

where $\mu(\boldsymbol{s}) = X(\boldsymbol{s})^T \widehat{\boldsymbol{\beta}}$, the vector of weights $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ is given by $\boldsymbol{\lambda} = C^{-1}\boldsymbol{c}$ where

- $\boldsymbol{c} = (cor(Y(\boldsymbol{s}_0), Y(\boldsymbol{s}_1)), \ldots, cor(Y(\boldsymbol{s}_0), Y(\boldsymbol{s}_N)))^T$.

- $C = [cor(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j))]_{i=j=1}^{N}$ is the correlation matrix.

Both can be computed using (2). Optimal linear prediction can be performed using the `GeoKrig` function. We first set the spatial locations to predict and the associated covariates. In this example we consider a spatial regular grid:

```
>xx=seq(0,1,0.012)
>loc_to_pred=as.matrix(expand.grid(xx,xx))
>Nloc=nrow(loc_to_pred)
>Xloc=cbind(rep(1,Nloc),runif(Nloc))
```

Then the optimal linear prediction (7), using the estimated parameters, can be performed using the `GeoKrig` function:

```
>param_est=as.list(c(fit$param,fixed))
>pr=GeoKrig(data=data, coordx=coords,loc=loc_to_pred,corrmodel=corrmodel,
             model=model,mse=TRUE,X=X,Xloc=Xloc,
            sparse=TRUE,param= param_est)
```

and we can compare the map of simulated data with the predictions (and associated mean square error) with the following code:

```
>colour = rainbow(100)
>par(mfrow=c(1,3))
#### map of simulated data
>quilt.plot(x, y, data,col=colour,main="Data")
>map=matrix(pr$pred,ncol=length(xx))
## prediction map
>image.plot(xx, xx, map,col=colour,xlab="",ylab="",main="Simple-Kriging")
## mse prediction map
>map_mse=matrix(pr$mse,ncol=length(xx))
>image.plot(xx, xx, map_mse,col=colour,xlab="",ylab="",main="mse␣")
```

References

Alegria, A., S. Caro, M. Bevilacqua, E. Porcu, and J. Clarke (2017). Estimating covariance functions of multivariate skew-gaussian random fields on the sphere. *Spatial Statistics 22*, 388 – 402.

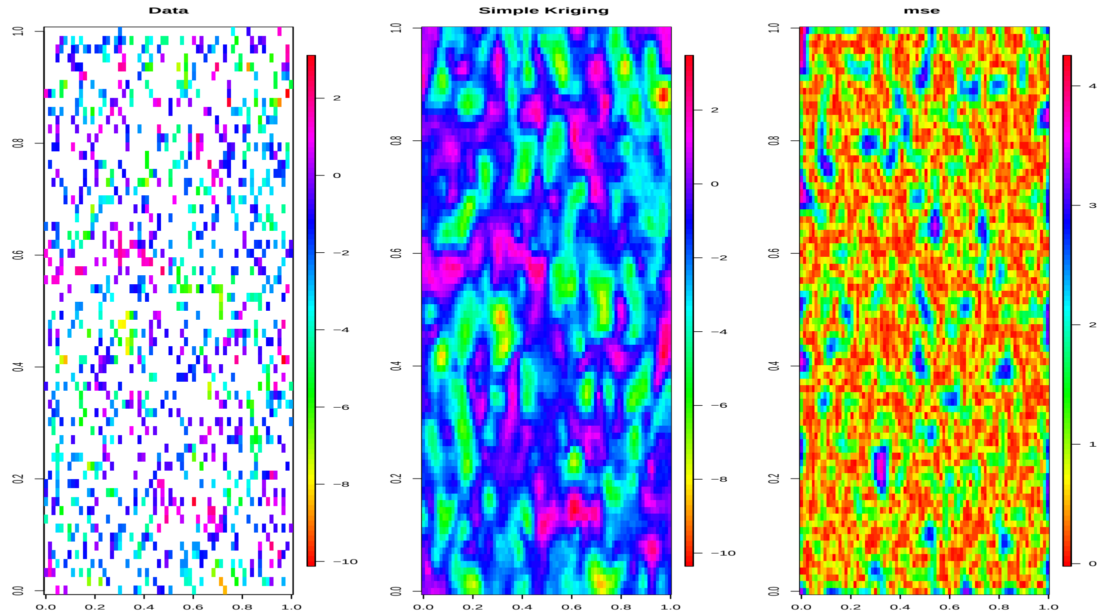

Figure 2: From left to right: observed spatial data, associated kriging map and mean square error map for the skew Gaussian RF.

Azzalini, A. and A. Capitanio (2014). *The Skew-Normal and Related Families.* United States of America by Cambridge University Press, New York.

Bevilacqua, M., T. Faouzi, R. Furrer, and E. Porcu (2018). Estimation and prediction using generalized Wendland functions under fixed domain asymptotics. *The Annals of Statistics*. to appear.

Gerber, F., K. Moesinger, and R. Furrer (2017). Extending R packages to support 64-bit compiled code: An illustration with spam64 and GIMMS NDVI3g data. *Computer & Geoscience 104*, 109–119.

Zhang, H. and A. El-Shaarawi (2010). On spatial skew-Gaussian processes and applications. *Environmetrics 21(1)*, 33–47.