`GeoModels` Tutorial: simulation, estimation and prediction of spatial data with heavy tails using Tukey-h random fields

Christian Caamaño
Moreno Bevilacqua

September 8, 2021

Introduction

In this tutorial we show how to analyze spatial data with heavy tails using Tukey-h random fields (Xua and Genton, 2017) with the R package `GeoModels` (Bevilacqua et al. (2018)). The Tukey-h distribution is a flexible parametric model, which is able to accommodate flexible tail behaviour and, in particular, heavier tails than the ones induced by Gaussian random fields.

We first load the R libraries needed in 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);
model="Tukeyh";  # model name in the  GeoModels package
set.seed(818);
```

Simulation of Tukey-h random fields

The definition of a Tukey-h random field starts by considering a 'parent' Gaussian random field $G = \{G(\boldsymbol{s}), \boldsymbol{s} \in S\}$, where $\boldsymbol{s}$ represents a location in the domain S. In this tutorial we consider the spatial case *i.e.* $S \subseteq \mathbb{R}^2$. However, the package `GeoModels` allows to work also with spatio-temporal data or data defined on a sphere of arbitrary radius. The Gaussian field G is assumed weakly stationary with zero mean, unit variance and correlation function $\rho(\boldsymbol{h}) = \text{cor}(G(\boldsymbol{s}+\boldsymbol{h}), G(\boldsymbol{s}))$.

Let $T_h^* = \{T_h^*(\boldsymbol{s}), \boldsymbol{s} \in A\}$ be a RF with standard Tukey-h marginal distribution defined trough a monotonic transformation $\tau_h(x) = xe^{\frac{hx^2}{2}}, \quad x \in \mathbb{R}$ of a standard Gaussian RF G as:

$$T_h^*(\boldsymbol{s}) =: \tau_h(G(\boldsymbol{s})) = G(\boldsymbol{s})e^{\frac{h(G(\boldsymbol{s}))^2}{2}}. \tag{1}$$

The inverse transformation $\tau_h^{-1}(x)$ can be expressed in terms of the Lambert function *i.e.*, $\tau_h^{-1}(x) = sign(x)\left(\frac{W(ht^2)}{h}\right)^{1/2}$ where $W(\cdot)$ is the Lambert-W function (Goerg, 2015).

This kind of RF has marginal symmetric distributions and the parameter $h \in [0, 1/2)$ governs the tail behavior of the RF, with a larger value of h indicating a heavier tail.

Specifically the marginal distribution has (asymptotically) a Pareto-heavy tailed distribution with tail index equal to $1/h$ (Morgenthaler and Tukey, 2000). If $h = 0$ the Gaussian RF G is obtained as special limit case.

In addition, the Tukey-h RF has marginal distribution given by (Goerg, 2015):

$$f_{T_h^*}(t) = \frac{\tau_h^{-1}(t)}{t(1 + W(ht^2))}\phi(\tau_h^{-1}(t), 0, 1), \tag{2}$$

with $\mathbb{E}(T_h^*(\boldsymbol{s})) = 0$, $Var(T_h^*(\boldsymbol{s})) = (1-2h)^{-3/2}$ and the correlation function is given by:

$$\rho_{T_h^*}(\boldsymbol{h}) = \frac{\rho(\boldsymbol{h})(1-2h)^{3/2}}{[(1-h)^2 - h^2\rho^2(\boldsymbol{h})]^{3/2}}. \tag{3}$$

Then, we define the location-scale transformation process $T_h = \{T_h(\boldsymbol{s}), \boldsymbol{s} \in A\}$ as:

$$T_h(\boldsymbol{s}) := \mu(\boldsymbol{s}) + \sigma T_h^*(\boldsymbol{s}) \tag{4}$$

with $\mathbb{E}(T_h(\boldsymbol{s})) = \mu(\boldsymbol{s})$ and $Var(T_h(\boldsymbol{s})) = \sigma^2(1-2h)^{-3/2}$ and a spatial regression model can be specified by assuming that $\mu(\boldsymbol{s}) = X(\boldsymbol{s})^T\boldsymbol{\beta}$ where $X(\boldsymbol{s})$ 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 assume $k = 2$.

To obtain a simulation from T_h we need to specify regression mean, tail and variance parameters *i.e.* β_1, β_2, h, σ^2 respectively. Additionally, we need to specify a parametric correlation $\rho(\boldsymbol{h})$ for the 'parent' Gaussian random field. We first set the spatial coordinates. In this example we consider 1500 locations uniformly distributed in the unit square:

```
N=1500;
coords=cbind(runif(N),runif(N));
plot(coords ,pch=20,xlab="",ylab="");
```

For the correlation function $\rho(\boldsymbol{h})$ of the 'parent' Gaussian random field G we assume an isotropic Matérn model (Matérn, 1986):

$$\rho_{\alpha,\gamma}(\boldsymbol{h}) = \frac{2^{1-\gamma}}{\Gamma(\gamma)}\left(\frac{\|\boldsymbol{h}\|}{\alpha}\right)^{\gamma}\mathcal{K}_{\gamma}\left(\frac{\|\boldsymbol{h}\|}{\alpha}\right), \qquad \|\boldsymbol{h}\| \geq 0. \tag{5}$$

where $\mathcal{K}_{\gamma}$ is a modified Bessel function of the second kind of order γ, $\gamma > 0$ is the smoothness parameter and $\alpha > 0$ the spatial scale parameter. Then, we set the parameter associated to this correlation model:

```
corrmodel = "Matern";          ## correlation model
scale = 0.2/3;                 ## scale parameter
smooth=0.5;                    ## smooth parameter
nugget=0;                      ## nugget parameter
```

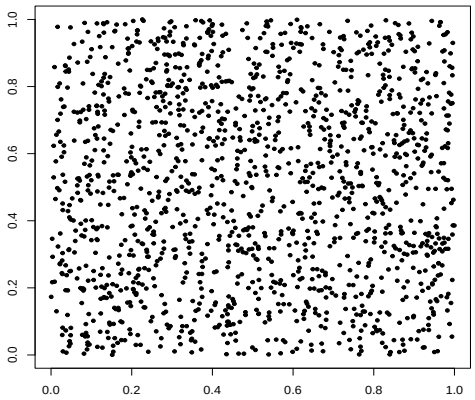

Figure 1: Spatial location sites used in the tutorial.

and we set the tail and variance parameters of the Tukey-h random field:

```
tail=0.1;       # tail parameter
sill= 1;        # variance parameter
```

Finally we set the mean regression parameters and the regression matrix:

```
mean = 0.5; mean1= -1;# regression paramteres
a0=rep(1,N);a1=runif(N,-1,1);
X=cbind(a0,a1) ## regression matrix
```

We are now ready to simulate a realization of the Tukey-h random field T_h using the function *GeoSim*. Simulation is performed exploiting the stochastic representation (1), where the underlying the Gaussian field is generated with Cholesky decomposition:

```
param=list(nugget=nugget,mean=mean,mean1=mean1, scale=scale,
                smooth=smooth, sill=sill,tail=tail);
data = GeoSim(coordx=coords,corrmodel=corrmodel,
                 param=param,model=model,X=X)$data
```

Estimation of Tukey-h random fields

Estimation of regression, tail and correlation parameters of the Tukey-h random field T_h can be performed using conditional pairwise likelihood estimation. Let $f_{\boldsymbol{T}^*_{h;ij}}(t_i, t_j)$ the density of the bivariate random vector $(T^*_h(\boldsymbol{s}_i), T^*_h(\boldsymbol{s}_j))^T$ given by:

$$f_{\boldsymbol{T}^*_h}(\boldsymbol{t}) = \frac{\tau_h^{-1}(t_i)\tau_h^{-1}(t_j)}{t_i t_j (1 + W(ht_i^2))(1 + W(ht_j^2))} \phi_2(\tau_h^{-1}(\boldsymbol{t}), \boldsymbol{0}, R_2) \tag{6}$$

with $R_2 = [\rho(\boldsymbol{s}_i - \boldsymbol{s}_j)]_{i,j=1}^2$ the correlation matrix associated to the underlying correlation function and the transformation $\tau_h^{-1}(\boldsymbol{x})$ applies pointwise for a given vector $\boldsymbol{x}$.

Given a partial realization $(t(\boldsymbol{s}_1), \ldots, t(\boldsymbol{s}_N))^T$ of the Tukey-h random process T_h defined in equation (4), the density of the bivariate random vector $(T_h(\boldsymbol{s}_i), T_h(\boldsymbol{s}_j))^T$ can be obtained from (6) as:

$$f_{\boldsymbol{T}_{h;ij}}(t(\boldsymbol{s}_i), t(\boldsymbol{s}_j)) = \frac{1}{\sigma^2} f_{\boldsymbol{T}^*_{h;ij}}\left(\frac{t(\boldsymbol{s}_i) - \mu(\boldsymbol{s}_i)}{\sigma}, \frac{t(\boldsymbol{s}_j) - \mu(\boldsymbol{s}_j)}{\sigma}\right). \tag{7}$$

Then, the conditional pairwise likelihood function is defined as:

$$cpl(\boldsymbol{\theta}) = \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} log\left(\frac{f_{\boldsymbol{T}_{h;ij}}(t(\boldsymbol{s}_i), t(\boldsymbol{s}_j))}{f_{T_h}(t(\boldsymbol{s}_j))}\right) w_{ij} \tag{8}$$

where w_{ij} are non-negative weights, not depending on $\boldsymbol{\theta}$. An efficient way to specify the (non symmetric) weights from computational and efficient viewpoint is based on neighborhoods:

$$w_{ij}(k) = \begin{cases} 1 & \mathbf{s}_i \in N_k(\mathbf{s}_j) \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

Here $N_k(\mathbf{s}_l)$ is the set of the neighbors of order $k = 1, 2, \ldots$ of the point $\mathbf{s}_l$ and in this case $\boldsymbol{\theta} = (\beta_1, \beta_2, h, \sigma^2, \alpha, \gamma)^T$. The conditional pairwise likelihood estimator $\hat{\boldsymbol{\theta}}_{cpl}$ is obtained maximizing (8) with respect to $\boldsymbol{\theta}$. In the `GeoModels` package, we can choose the fixed parameters and the parameters that can be estimated.

Conditional pairwise likelihood estimation can be performed using the function *GeoFit* or *GeoFit2*. In particular, the function *GeoFit2* first computes maximum likelihood estimation (under the hypothesis of independence) of the marginal parameters and then computes conditional pairwise likelihood estimation using the estimates at the first step as starting values in the optimization algorithm.

In this example, we perform optimization of (8) using the function `nlminb` that allows box-constrained optimization using PORT routines. However other type of optimization algorithms available in R can be used (`BFGS` or `Nelder-Mead` for instance).

```
optimizer="nlminb";
fixed1<-list(nugget=nugget,smooth=smooth);
start1<-list(mean=mean, mean1=mean1,scale=scale,sill=sill,tail=tail);
I=Inf;
lower1<-list(mean=-I, mean1=-I,scale=0,sill=0,tail=0);
```

```
upper1<-list(mean=I, mean1=I,scale=I,sill=I,tail=0.5);



fit2 <- GeoFit2(data=data,coordx=coords,corrmodel=corrmodel,
optimizer=optimizer,lower=lower1,upper=upper1,
type="Pairwise",likelihood="Conditional",
neighb=4,X=X,start=start1,fixed=fixed1, model = model);
```

Note that the option *neighb=4* set the neighborhood order of the weight function (9) i.e. $k = 4$.

The object `fit2` include informations about the conditional pairwise likelihood estimation:

```
fit2
##########################################################################
Maximum  Composite-Likelihood Fitting of Tukeyh Random Fields
Setting: Conditional Composite-Likelihood
Model: Tukeyh
Type of the likelihood objects: Pairwise
Covariance model: Matern
Optimizer: nlminb
Number of spatial coordinates: 1500
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: -6848.66
Estimated parameters:
    mean      mean1      scale       sill       tail
 0.43734   -1.00500    0.06962    1.02365    0.11640
##########################################################################
```

Checking model assumptions

Given the estimation of the mean regression and sill parameters, the estimated residuals

$$\widehat{T_h^*(\boldsymbol{s}_i)} = \frac{t(\boldsymbol{s}_i) - X(\boldsymbol{s}_i)^T \widehat{\boldsymbol{\beta}}}{(\widehat{\sigma}^2)^{\frac{1}{2}}} \quad i = 1, \dots N$$

can be viewed as a realization of the process T_h^*. The residuals can be computed using the *GeoResiduals* function:

```
res=GeoResiduals(fit2); # computing residuals
```

The marginal distribution assumption on the residuals can be graphically checked trough a qq-plot using the *GeoQQ* function (see Figure 2, left part):

```
### checking model  residuals assumptions: marginal distribution
GeoQQ(res)
```

The covariance model assumption can be checked comparing the empirical and the estimated semi-variogram using the *GeoVariogram* and *GeoCovariogram* functions (see Figure 2, right part):

```
### checking model  residuals assumptions: covariance model
vario <- GeoVariogram(data=res$data,coordx=coords,maxdist=0.4);
GeoCovariogram(res,show.vario=TRUE, vario=vario,pch=20);
```

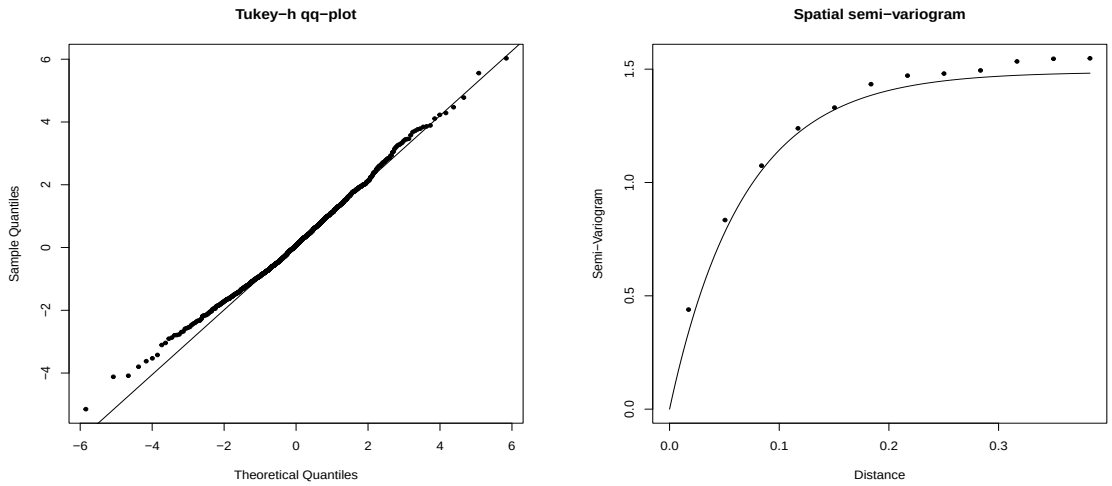

Figure 2: From left to right: qq-plot of the residuals using the $T_{0.02}^*$ distribution and empirical vs estimated semi-variogram for the residuals.

The semi-variogram is computed using the correlation function (3).

Prediction of Tukey-h random fields

For a given spatial location $\boldsymbol{s}_0$ with associated covariates $X(\boldsymbol{s}_0)$, the optimal linear prediction of a Tukey-h random field is given by:

$$\widehat{T}_h(\boldsymbol{s}_0) = X(\boldsymbol{s}_0)^T\boldsymbol{\beta} + \sum_{i=1}^{N} \lambda_i[t(\boldsymbol{s}_i) - X(\boldsymbol{s}_i)^T\boldsymbol{\beta}] \tag{10}$$

where the vector of weights $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_N)'$ is given by $\boldsymbol{\lambda} = R_h^{-1}\boldsymbol{c}_h$ and

- $\boldsymbol{c}_h = (cor(T_h(\boldsymbol{s}_0), T_h(\boldsymbol{s}_1)), \ldots, cor(T_h(\boldsymbol{s}_0), T_h(\boldsymbol{s}_N)))^T$.
- $R_h = [cor(T_h(\boldsymbol{s}_i), T_h(\boldsymbol{s}_j))]_{i,j=1}^N$ is the correlation matrix.

Moreover the associated mean square error (MSE) is given by:

$$MSE(\widehat{T}_h(\boldsymbol{s}_0)) = \sigma^2(1-2h)^{-3/2}(1 - \boldsymbol{c}_h^T R_h^{-1}\boldsymbol{c}_h). \tag{11}$$

The predictor can be viewed as an optimal Gaussian predictor assuming (3) as correlation function. If the parameters are unknown, both (10) and (11) can be computed replacing the parameters with the conditional pairwise likelihood estimates. In particular, R_h and $\boldsymbol{c}_h$ can be computed using (3) coupled with the estimates of the Matérn correlation function and of the degrees of freedom.

Kriging and associated MSE can be obtained using the `GeoKrig` function. We first need to specify the spatial locations to predict and, 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 (10), using the estimated parameters, can be performed using the `GeoKrig` function (computation can be time consuming):

```
param_est=as.list(c(fit2$param,fixed1));
pr=GeoKrig(data=data, coordx=coords,loc=loc_to_pred, X=X,Xloc=Xloc,
     corrmodel=corrmodel,model=model,mse=TRUE,param= param_est)
```

Finally, a kriging map with associate mean square error (Figure 3) can be obtained with the following code:

```
par(mfrow=c(1,3));
colour = rainbow(100);
#### map of  data
quilt.plot(coords[,1], coords[,2], data,col=colour,main="Data");
# linear kriging
map=matrix(pr$pred,ncol=length(xx));
image.plot(xx,xx,map,col=colour,xlab="",ylab="",main="SimpleKriging");
#associated mean squared error
map_mse=matrix(pr$mse,ncol=length(xx));
image.plot(xx,xx,map_mse,col=colour,xlab="",ylab="",main="MSE")
```

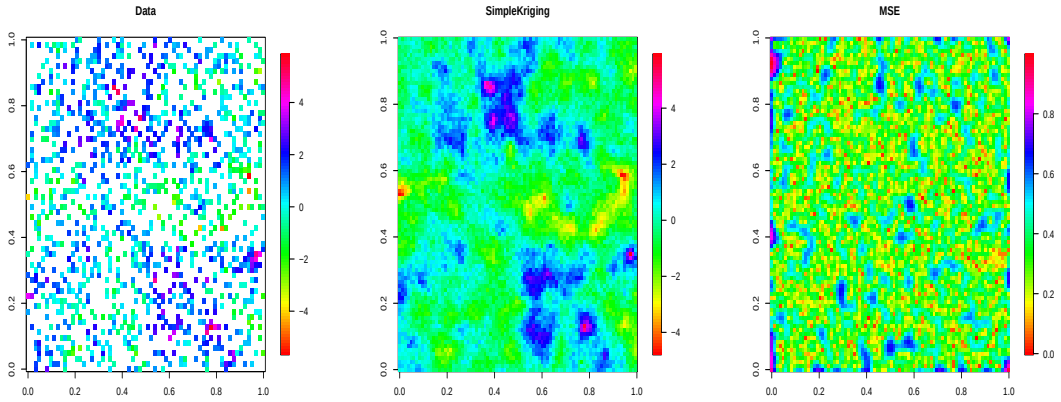

Figure 3: From left to right: observed spatial data, associated kriging map and mean square error map.

References

Bevilacqua, M., V. Morales-Oñate, and C. Caamaño-Carrillo (2018). *GeoModels: A Package for Geostatistical Gaussian and non Gaussian Data Analysis*. R package version 1.0.3-4.

Goerg, G. M. (2015). The lambert way to gaussianize heavy-tailed data with the inverse of tukey's h transformation as a special case. *The Scientific World Journal 2015*, 1–16.

Matérn, B. (1986). *Spatial Variation: Stochastic Models and their Applications to Some Problems in Forest Surveys and Other Sampling Investigations* (2nd ed.). Heidelberg: Springer.

Morgenthaler, S. and J. W. Tukey (2000). Fitting quantiles: Doubling, hr, hq, and hhh distributions. *Journal of Computational and Graphical Statistics 9*(1), 180–195.

Xua, G. and M. G. Genton (2017). Tukey g-and-h random fields. *Journal of the American Statistical Association 112*, 1236 –1249.