`GeoModels` Tutorial: simulation, estimation and prediction of spatial data using Gaussian random fields with a compactly supported flexible covariance function

Moreno Bevilacqua, Christian Caamaño-Carrillo

Introduction

In this tutorial we show how to analyze spatial data using Gaussian random fields with a flexible covariance model proposed in Bevilacqua et al. (2020) using the `R` package `GeoModels` (Bevilacqua et al., 2018).

We first load the R libraries needed for the analysis using the `GeoModels` package:

```
rm(list=ls())
require(devtools)
install_github("vmoprojs/GeoModels")
require(GeoModels)
require(fields)
require(spam)
set.seed(89)
```

Simulation of a spatial Gaussian random field with a flexible compactly supported correlation model

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

Then we consider a random field $Y = \{Y(\boldsymbol{s}), \boldsymbol{s} \in S\}$ defined by the location and scale transformation:

$$Y(\boldsymbol{s}) = m(\boldsymbol{s}) + \sigma Z(\boldsymbol{s}) \tag{1}$$

where the spatial mean in the package `GeoModels` can be specified trough a regression model $m(\boldsymbol{s}) = X(\boldsymbol{s})^T\boldsymbol{\beta}$. Here $X(\boldsymbol{s})$ is a k-dimensional vector of covariates and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)^T$ is a k-dimensional vector of (unknown) parameters. Then $\mathbb{E}(Y(\boldsymbol{s})) = X(\boldsymbol{s})^T\boldsymbol{\beta}$, $\text{var}(Y(\boldsymbol{s})) = \sigma^2$ and $\text{cov}(Y(\boldsymbol{s}+\boldsymbol{h}), Y(\boldsymbol{s})) = \sigma^2\rho(\boldsymbol{h})(1-\tau^2)$ where $0 \leq \tau^2 < 1$ is the nugget parameter. In this tutorial, for simplicity, we assume a spatial constant mean, unit variance and zero nugget that is $m(\mathbf{s}) = m$ and $\sigma^2 = 1$, $\tau^2 = 0$. Additionally we assume that $\rho(\boldsymbol{h})$ is an isotropic parametric correlation model that is the correlation depends on the spatial euclidean distance $\|\mathbf{h}\|$, $\mathbf{h} \in \mathbb{R}^2$ and hereafter we set $r = \|\mathbf{h}\|$.

The globally supported Matérn family of covariance functions (Stein, 1999) has played a central role in spatial statistics for decades, being a flexible parametric class with one parameter determining the smoothness of the paths of the underlying spatial field. It is defined as follows:

$$\mathcal{M}_{\nu,\beta}(r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{r}{\beta}\right)^{\nu} \mathcal{K}_{\nu}\left(\frac{r}{\beta}\right), \qquad r \geq 0,$$

for $\nu > 0, \beta > 0$, and it is positive definite in any dimension $d = 1, 2, \ldots$. Here, Γ is the gamma function and $\mathcal{K}_{\nu}$ is the modified Bessel function of the second kind (Abramowitz and Stegun, 1970) of the order ν. The correlation is globally supported because $\mathcal{M}_{\nu,\beta}(r) > 0$, for $r > 0$. The parameter ν indexes the mean squared differentiability of a Gaussian random field having a Matérn correlation function and its associated sample paths. In particular, for a positive integer k, the sample paths are k times differentiable, in any direction, if and only if $\nu > k$ (Stein, 1999; Banerjee et al., 2004).

We now introduce a generalization of the Matérn model proposed in Bevilacqua et al. (2020). Let:

$$\delta_{\nu,\mu,\beta} = \beta \left(\frac{\Gamma(\mu + 2\nu + 1)}{\Gamma(\mu)}\right)^{\frac{1}{1+2\nu}}, \tag{2}$$

where $\nu \geq 0$, $\beta > 0$ and $\mu \geq (d+1)/2+\nu$ and let ${}_2F_1$ the hypergeometric Gaussian function defined as

$${}_2F_1(a, b, c; x) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k}{(c)_k} \frac{x^k}{k!}. \tag{3}$$

The correlation model proposed in Bevilacqua et al. (2020) is a specific compact support reparametrization of the generalized Wendland model (Bevilacqua et al., 2019; Gneiting, 2002). It has a series representation in terms of Hypergeometric Gaussian function ${}_2F_1$ as follows:

$$\mathcal{GW}_{\nu,\mu,\delta_{\nu,\mu,\beta}}(r) = \begin{cases} K\left(1 - \left(\frac{r}{\delta_{\nu,\mu,\beta}}\right)^2\right)^{\nu+\mu} {}_2F_1\left(\frac{\mu}{2}, \frac{\mu+1}{2}; \nu+\mu+1; 1 - \left(\frac{r}{\delta_{\nu,\mu,\beta}}\right)^2\right) & 0 \leq r \leq \delta_{\nu,\mu,\beta} \\ 0 & r > \delta_{\nu,\mu,\beta}, \end{cases} \tag{4}$$

with $K = \frac{\Gamma(\nu)\Gamma(2\nu+\mu+1)}{\Gamma(2\nu)\Gamma(\nu+\mu+1)2^{\mu+1}}$.

Similarly to the Matern case, this model allows for parameterization in a continuous fashion of the mean squared differentiability of the underlying Gaussian random field and its associated sample paths. Specifically, the sample paths of the generalized-Wendland model are k times differentiable, in any direction, if and only if $\nu > k - 0.5$.

The correlation model (4) is very flexible, as it allows us to consider, under the same umbrella, compactly and globally supported correlation functions. In fact Bevilacqua et al. (2020) show that the Matérn family $\mathcal{M}_{\nu+1/2,\beta}$ is a special case of the $\mathcal{GW}_{\nu,\mu,\delta_{\nu,\mu,\beta}}$ model when $\mu \to \infty$. Hence, the parameter μ is crucial to fix the sparseness of the associated correlation matrix ad it allows to switch from the world of flexible compactly supported covariance functions to the world of flexible globally supported correlation functions. Note that the compact support $\delta_{\nu,\mu,\beta}$ depends on ν, β and μ and it is an increasing function of μ.

In Table 1 we report the $\mathcal{GW}_{\nu,\mu,\delta_{\nu,\mu,\beta}}$ correlation model for the special cases $\nu = 0, 1, 2, 3$ and its associated limit case when $\mu \to \infty$ *i.e.* the Matérn correlation model $\mathcal{M}_{\nu+1/2,\beta}$.

Suppose we want to simulate a realization of a random field Y with constant mean equal to five, unit variance and with correlation model (4) at $N = 500$ spatial locations uniformly distributed in the unit square. We first set the spatial coordinates:

```
NN=500          # number of spatial locations
x = runif(NN, 0, 1); y = runif(NN, 0, 1)
coords=cbind(x,y)
```

Then we specify the mean, variance and nugget parameters for the Gaussian random field:

```
NuisParam ("Gaussian")
[1] "mean"   "nugget" "sill"
mean = 5; sill=1; nugget=0
```

where `mean`, `sill` and `nugget` are respectively m, σ^2 and τ^2. Then we set the name of the correlation model used in the `GeoModels` package and the associated correlation parameters:

```
corrmodel="GenWend_Matern"
CorrParam (corrmodel)
[1] "power2" "scale"  "smooth"
mu=3
scale=0.05; power2=1/mu;  smooth=0
```

where `scale` , `power2` and `smooth` correspond to β, $1/\mu$ and ν respectively. Note that the inverse of the μ parameter is used as parametrizaton as suggested in Bevilacqua et al. (2020). The compact support $\delta_{\nu,\mu,\beta}$ of the correlation model in this special case ($\nu = 0$) is given by :

```
scale*(1/power2)
```

ν	$\mathcal{GW}_{\nu,\mu,\delta_{\nu,\mu,\beta}}(r)$	$\mathcal{M}_{\nu+1/2,\beta}(r)$
0	$\left(1-\frac{r}{\delta_{0,\mu,\beta}}\right)_+^{\mu}$	$e^{-\frac{r}{\beta}}$
1	$\left(1-\frac{r}{\delta_{1,\mu,\beta}}\right)_+^{\mu+1}\left(1+\frac{r}{\delta_{1,\mu,\beta}}(\mu+1)\right)$	$e^{-\frac{r}{\beta}}(1+\frac{r}{\beta})$
2	$\left(1-\frac{r}{\delta_{2,\mu,\beta}}\right)_+^{\mu+2}\left(1+\frac{r}{\delta_{2,\mu,\beta}}(\mu+2)+\left(\frac{r}{\delta_{2,\mu,\beta}}\right)^2(\mu^2+4\mu+3)\frac{1}{3}\right)$	$e^{-\frac{r}{\beta}}(1+\frac{r}{\beta}+\frac{r^2}{3\beta^2})$
3	$\left(1-\frac{r}{\delta_{3,\mu,\beta}}\right)_+^{\mu+3}(1+\frac{r}{\delta_{3,\mu,\beta}}(\mu+3)+\left(\frac{r}{\delta_{3,\mu,\beta}}\right)^2(2\mu^2+12\mu+15)\frac{1}{5}$ $+\left(\frac{r}{\delta_{3,\mu,\beta}}\right)^3(\mu^3+9\mu^2+23\mu+15)\frac{1}{15})$	$e^{-\frac{r}{\beta}}(1+\frac{r}{2\beta}+\frac{6r^2}{15\beta^2}+\frac{r^3}{15\beta^3})$

Table 1: The $\mathcal{GW}_{\nu,\mu,\delta_{\nu,\mu,\beta}}$ model with compact support $\delta_{\nu,\mu,\beta}$ (see Equation 2) for $\nu = 0, 1, 2, 3$ and the associated limit case when $\mu \to \infty$ *i.e.*, the Matérn model $\mathcal{M}_{\nu+1/2,\beta}$.

```
[1] 0.15
```

We are now ready to simulate the spatial Gaussian random field using the function `GeoSim`:

```
param= list(nugget=nugget,mean=mean,
     scale=scale, sill=sill, power2=power2,smooth=smooth)
sim = GeoSim(coordx=coords,  corrmodel=corrmodel,
             sparse=FALSE, model="Gaussian",param=param)$data
```

Note that the option `sparse` allows to consider or not algorithms for sparse matrices when performing Cholesky decomposition, using package `spam` (Gerber et al. (2017)). Informations about the sparsity of the covariance matrix can be obtained using the function `GeoCovmatrix` that returns an object associated with covariance matrix:

```
cc = GeoCovmatrix(coordx=coords, corrmodel=corrmodel,  sparse=TRUE,
model="Gaussian", param=param)
is.spam(cc$covmatrix)
[1] TRUE
cc$nozero
[1] 0.062168
```

This means that (approximatively) 94% of the covariance matrix are zeros *i.e* the matrix is highly sparsed.

Maximum likelihood estimation

Given a realization $\boldsymbol{Y} = \{y(\boldsymbol{s}_i)\}$, $i = 1, \ldots, N$ of the random field Y, maximum likelihood estimation that is the maximization of the log-likelihood function

$$l(\boldsymbol{\theta}) = -0.5log(|\sigma^2 R|) - 0.5\frac{(\boldsymbol{Z} - m\mathbf{1})^T R^{-1}(\boldsymbol{Z} - m\mathbf{1})}{\sigma^2}$$

with respect to $\boldsymbol{\theta} = (m, \sigma^2, \beta, \nu, \mu)^T$ where $R = [\mathcal{GW}_{\nu,\mu,\delta_{\nu,\mu,\beta}}(||\boldsymbol{s}_i - \boldsymbol{s}_j||)]_{i,j=1}^N$ is the correlation matrix, can be performed using the `GeoFit` function. The parameter μ can be estimated or fixed. In this case we estimate μ in its reparametrized version i.e. $1/\mu$. We use the optimization method `nlminb` implemented in `R` that allows for box constrained optimization.

```
optimizer="nlminb"
start=list(mean=mean,scale=scale, sill=sill, power2=power2)
I=Inf
lower=list(mean=-I,scale=0, sill=0, power2=0)
upper=list(mean=I,scale=I, sill=I, power2=1/(1.5))
fixed=list(mean=mean,nugget=nugget,smooth=smooth)
fitML <- GeoFit(data=sim,coordx=coords,corrmodel=corrmodel,
                    model="Gaussian", varest=TRUE,
                    optimizer=optimizer,lower=lower,upper=upper,
                    likelihood="Full",type="Standard",
                    start=start,fixed=fixed)
```

The object `fitML` includes informations about the maximum likelihood estimation:

```
fit
#########################################################################
Maximum  Likelihood Fitting of Gaussian Random Fields
Setting: Full Likelihood
Model: Gaussian
Type of the likelihood objects: Standad
Covariance model: GenWend_Matern
Optimizer: nlminb
Number of spatial coordinates: 500
Number of dependent temporal realisations: 1
```

```
Type of the random field: univariate
Number of estimated parameters: 4
Type of convergence: Successful
Maximum log-Likelihood value: -593.19
AIC : 1194
BIC : 1211
Estimated parameters:
   mean    power2     scale      sill
5.04396   0.43166   0.04587   0.96836
Standard errors:
    mean     power2      scale       sill
0.075632   0.103227   0.004547   0.078177
#####################################################################
```

In this example the estimated compact support is:

```
(1/fitML$param["power2"])*fitML$param["scale"]
[1]0.1062642
```

Checking model assumptions

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

$$\widehat{Z}(\boldsymbol{s}_i) = \frac{Y(\boldsymbol{s}_i) - \widehat{m}}{(\widehat{\sigma}^2)^{\frac{1}{2}}} \quad i = 1, \ldots, N$$

can be viewed as a realization of a standard Gaussian random field with correlation function (4). The residuals can be computed using the `GeoResiduals` function:

```
res=GeoResiduals(fitML)  # computing residuals
```

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

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

The correlation model assumption can be checked comparing the empirical and the estimated 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=coordx,maxdist=0.6)
GeoCovariogram(res,vario=vario, show.vario=TRUE,pch=20)
```

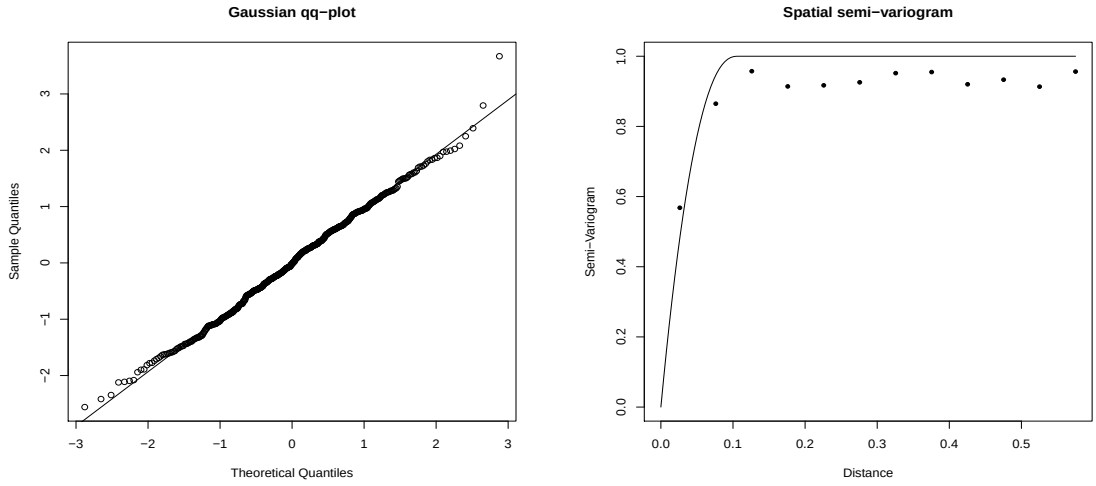

Figure 1: Left: QQ-plot for the residuals of the Gaussian random field. Right: empirical vs estimated semi-variogram function for the residuals

Prediction

For a given space time location $\boldsymbol{s}_0$, the optimal prediction is computed as:

$$\widehat{Y}(\boldsymbol{s}_0) = \widehat{m} + \boldsymbol{c}^T R^{-1}[\boldsymbol{Y} - \widehat{m}] \tag{5}$$

where $\boldsymbol{c} = (\text{cor}(Y(\boldsymbol{s}_0), Y(\boldsymbol{s}_1)), \ldots, cor(Y(\boldsymbol{s}_0), Y(\boldsymbol{s}_N)))^T$ and $R = [\text{cor}(Y(\boldsymbol{s}_i), Y(\boldsymbol{s}_j)]]_{i,j=1}^N$ are associated to the correlation model (4) and can be computed using the maximum likelihood estimates of the correlation parameters.

Kriging can be performed using the `GeoKrig` function. We need to specify the spatial locations to predict. In this example we consider a spatial regular grid:

```
xx=seq(0,1,0.015)
loc_to_pred=as.matrix(expand.grid(xx,xx))     # locations to predict
```

Then the optimal linear prediction (5), using the estimated parameters, can be performed using the `GeoKrig` function:

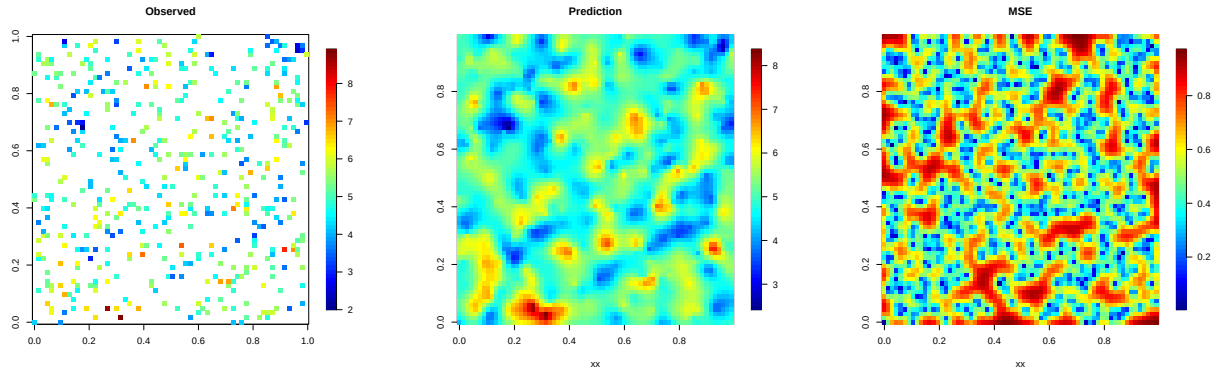

Figure 2: From left to right: observed spatial data, estimated kriging map and associate mean squared error kriging map.

```
param_est=as.list(c(fitML$param,fixed))
pr = GeoKrig(data=sim,coordx=coords,  corrmodel=corrmodel,
 sparse=TRUE,model="Gaussian",mse=TRUE,loc=loc_to_pred,param=param_est)
```

A kriging map with associate mean squared error (Figure 2) can be obtained with the following code:

```
quilt.plot(coords[,1],coords[,2],sim,main ="Observed")
## kriging map
image.plot(xx, xx, matrix(pr$pred,ncol=length(xx)),
           main = paste("Prediction" ),ylab="")
## MSE kriging map
image.plot(xx, xx, matrix(pr$mse,ncol=length(xx)),
           main = paste("MSE"),ylab="")
```

References

Abramowitz, M. and I. A. Stegun (Eds.) (1970). *Handbook of Mathematical Functions*. New York: Dover.

Banerjee, S., B. P. Carlin, and A. E. Gelfand (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Boca Raton: FL: Chapman & Hall/CRC Press.

Bevilacqua, M., C. Caamaño-Carrillo, and E. Porcu (2020). Unifying compactly supported and matérn covariance functions in spatial statistics. *ArXiv e-prints*.

Bevilacqua, M., T. Faouzi, R. Furrer, and E. Porcu (2019). Estimation and prediction using Generalized Wendland functions under fixed domain asymptotics. *The Annals of Statistics 47*, 828–856.

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.

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.

Gneiting, T. (2002). Compactly supported correlation functions. *Journal of Multivariate Analysis 83*(2), 493 – 508.

Stein, M. (1999). *Interpolation of Spatial Data. Some Theory of Kriging*. New York: Springer-Verlag.