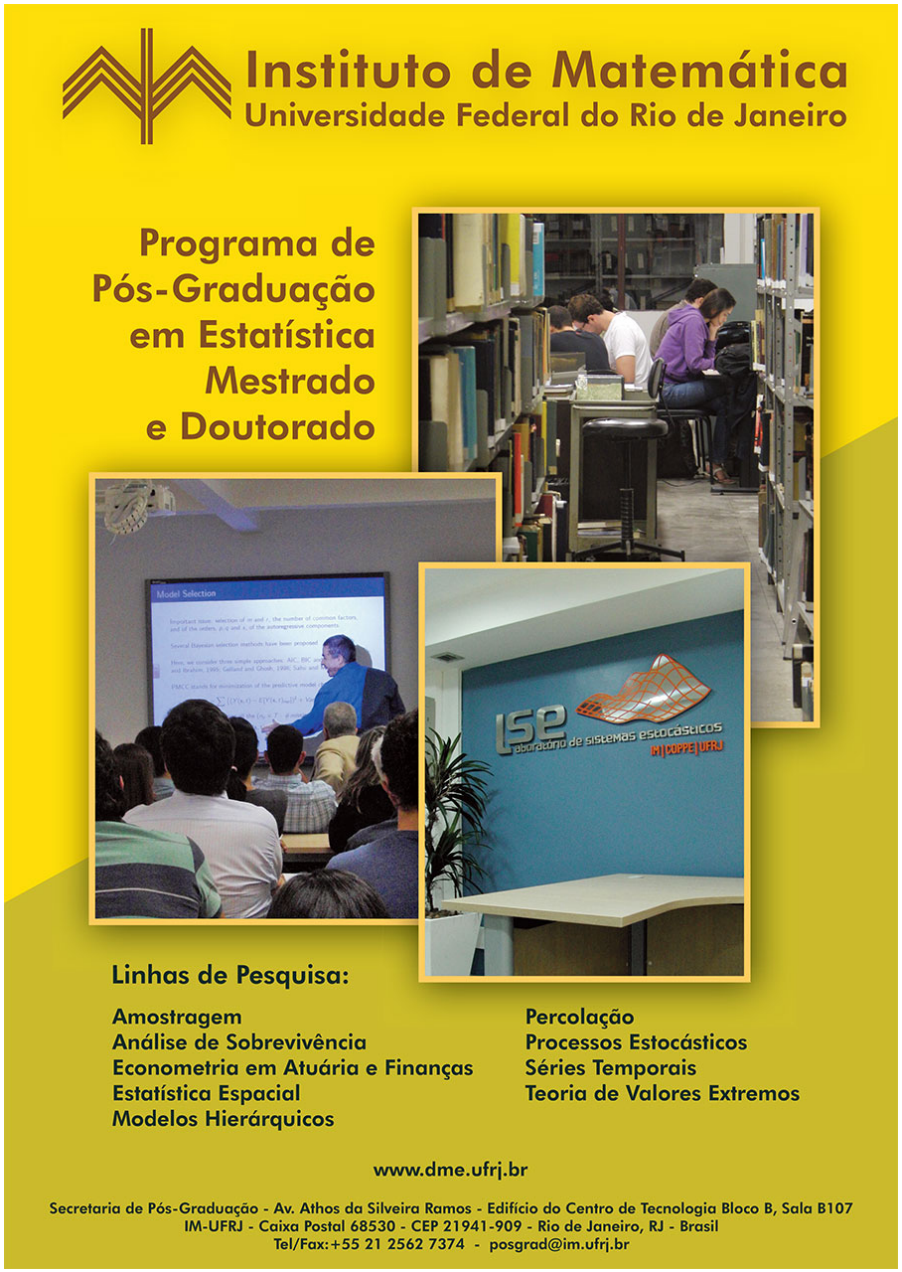
Practicum : Spatial Regression

Alexandra M. Schmidt
Instituto de Matemática
UFRJ - Brazil
www.dme.ufrj.br/∼alex

PASI 2014
Búzios, RJ, Brazil
June 2014

www.dme.ufrj.br



Exploratory (Spatial) Data Analysis

Alex Schmidt
IM-UFRJ
Brazil

1. Non-spatial summaries
   - Numerical summaries: Mean, median, standard deviation, range, etc.
     Not useful for spatial data, as they ignore the location information. Notice that in spatial statistics data should not be regarded as having come from a single population, usually they are NOT i.i.d.
   - Stem-and-leaf display:
     Better than a numerical summary as gives a complete picture of the data. Again, the location is ignored, it gives no indication of the data's spatial structure.

Exploratory (Spatial) Data Analysis

Alex Schmidt
IM-UFRJ
Brazil

PASI
June 2014

2. Methods used to explore large-scale spatial variation
   - 3-D scatter plot: a plot of Y_i versus location ($d = 2$);
   - Plot of Y_i versus each marginal coordinate (latitude and longitude);
   - Contour plot of Y_i (assuming $d = 2$). Requires some kind of smoothing operation.

Exploratory (Spatial) Data Analysis

Alex Schmidt
IM-UFRJ
Brazil

PASI
June 2014

3. Methods to explore small-scale variation:
   (a) dependence
   - Variogram cloud
     - Plot $(Z(\mathbf{s}_i) - Z(\mathbf{s}_j))^2$ versus $(\mathbf{s}_i - \mathbf{s}_j)^{1/2}$ for all possible pairs of observations;
     - The plot is often an unintelligible mess, hence it may often be advisable to bin the lags and plot a boxplot for each bin;
     - Note that this implicitly assumes isotropy (does not differentiate any directions).
     - The square-root differences are more resistant to outliers.

Exploratory (Spatial) Data Analysis

Alex Schmidt
IM-UFRJ
Brazil

- Empirical semivariogram (or variogram)
  - Plot one-half the average squared difference (or, for the variogram, merely the average squared difference) of observations lagged the same distance and direction apart, versus the lag
  - Assume here that the data are regularly spaced
  - Formally we plot $\hat{\gamma}(\mathbf{h}_u)$ versus $\mathbf{h}_u$, where

$$\hat{\gamma}(\mathbf{h}_u) = \frac{1}{2N(\mathbf{h}_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}_u} \{Z(\mathbf{s}_i) - Z(\mathbf{s}_j)\}^2,$$
$$(u = 1, \cdots, k),$$

    $\mathbf{h}_1, \cdots, \mathbf{h}_k$ are the distinct values of $\mathbf{h}$ represented in the data set, and $N(\mathbf{h}_u)$ is the number of times that lag $\mathbf{h}_u$ occurs in the data set
  - Note that this implicitly assumes stationarity of some kind
  - If $d = 2$, you can display as a 3-D plot or you can superimpose a few selected directions (e.g.: N-S,NW-SE, E-W, and NE-SW) on the same 2-D graph

Exploratory (Spatial) Data Analysis

Alex Schmidt
IM-UFRJ
Brazil

- Sample autocovariance function
  - Similar in many ways to the sample semivariogram
  - Plot of $\hat{C}(\mathbf{h}_u)$ versus $\mathbf{h}_u$, where

$$\hat{C}(\mathbf{h}_u) = \frac{1}{N(\mathbf{h}_u)} \sum_{\mathbf{s}_i - \mathbf{s}_j = \mathbf{h}_u} (Z(\mathbf{s}_i) - \overline{Z})(Z(\mathbf{s}_j) - \overline{Z})$$

  - This is a spatial generalization of an important tool used by time series analysts
- 3-D plot of correlation range versus spatial location, computed from a moving window. This method can detect nonstationarity in dependence

Exploratory (Spatial) Data Analysis

Alex Schmidt
IM-UFRJ
Brazil

PASI
June 2014

4. Methods used mainly to explore small-scale variation: (b) variability
   - 3-D plot of standard deviation versus spatial location, computed from a moving window. May reveal nonstationarity in variability, and indicate which portion(s) of A is (are) different from the rest
   - Scatterplot of standard deviation versus mean, computed from a moving window. May also reveal nonstationarity in variability, but differently from the previous method. Strictly this method is non-spatial

```
require("geoR")
prices=read.table("house_prices.txt",header=TRUE)
prices1=prices[1:500,]
names(prices1)
prices=na.exclude(prices1)
names(prices)
#standardizing covariates
stdsqft=as.numeric(scale(prices$sqft,scale=TRUE))
stdage=as.numeric(scale(prices$age,scale=TRUE))
stddistfree=as.numeric(scale(prices$dist_freeway,scale=TRUE))
#using the standardizing variables
prices$sqft=stdsqft
prices$age=stdage
prices$dist_freeway=stddistfree
#transforming into a geoR object
geoprices=as.geodata(prices2,coords.col=c(8,9),covar.col=2:7,
    data.col=1)
plot(geoprices)
plot(geoprices,scatter3d=TRUE)
```

```
#transforming into a geoR object
geoprices=as.geodata(prices2,coords.col=c(8,9),covar.col=2:7,
plot(geoprices)
plot(geoprices,scatter3d=TRUE)

#checking for duplicated coordinates
dup.coords(geoprices)
#jittering the duplicated locations
coord=geoprices$coords
coordjitter=jitter2d(geoprices$coords,max=0.001)
dup.coords(coordjitter)
cbind(geoprices$coords,coordjitter)
geoprices$coords=coordjitter
plot(geoprices)
```

Introduction

Alex Schmidt
IM-UFRJ
Brazil

Basic Model: Data ($\mathbf{Y}$) are a (partial) realization of a random process (*stochastic process* or *random field*)

$$\{Y(\mathbf{s}) : \mathbf{s} \in D\}$$

where D is a fixed subset of R^d with positive d-dimensional volume. In other words, the spatial index $\mathbf{s}$ varies *continously* throughout the region D.

NOTE:

A stochastic process is a collection of random variables $X(t), t \in T$ defined on a common probability space indexed by t which is in the index set T which describes the evolution of some system. For example, $X(t)$ could be the number of people in line at time t, or $Y(\mathbf{s})$ the amount of rainfall at location $\mathbf{s}$.

GOAL:

- Want a method of predicting $Y(\mathbf{s}_0)$ for any $\mathbf{s}_0$ in D.
- Want this method to be optimal (in some sense).

What do we need?

- Want $Y(\mathbf{s}) : \mathbf{s} \in D$ to be continuous and "smooth enough" (local stationarity)
- description of spatial covariation
- once we obtain the spatial covariation how to get predicted values

Basic Approach: given variance structure, predict.

Alex Schmidt
IM-UFRJ
Brazil

Gaussian Processes

Alex Schmidt
IM-UFRJ
Brazil

Before we proceed let us define *Gaussian Processes*

Definition:

A function $Y(.)$ taking values $y(\mathbf{s})$ for $\mathbf{s} \in D$ has a Gaussian process distribution with mean function $m(.)$ and covariance function $c(.,.)$, denoted by

$$Y(.) \sim GP(m(.), c(.,.))$$

if for any $\mathbf{s}_1, \cdots, \mathbf{s}_n \in D$, and any $n = 1, 2, \cdots$, the joint distribution of $Y(\mathbf{s}_1), \cdots, Y(\mathbf{s}_n)$ is multivariate Normal with parameters given by

$E\{Y(\mathbf{s}_j)\} = m(\mathbf{s}_j)$ and
$Cov(Y(\mathbf{s}_i), Y(\mathbf{s}_j)) = c(\mathbf{s}_i, \mathbf{s}_j)$.

Intrinsic Stationarity

Alex Schmidt
IM-UFRJ
Brazil

It is defined through first differences:

$$\begin{aligned} \mathrm{E}(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) &= 0, \\ \mathrm{Var}(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s})) &= 2\gamma(\mathbf{h}) \end{aligned}$$

The quantity $2\gamma(\mathbf{h})$ is known as the **variogram**.
$\gamma(.)$ is known as the *semi-variogram*.
In geostatistics, $2\gamma(.)$ is treated as a *parameter* of the random process $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$ (because it describes the covariance structure).

Second Order Stationarity

Alex Schmidt
IM-UFRJ
Brazil

Statistically speaking, some further assumptions about Y have to be made. Otherwise, the data represent an *incomplete* sampling of a single realization, making inference impossible.
A random function $Y(.)$ satisfying:

$$\begin{aligned} \mathrm{E}(Y(\mathbf{s})) &= \mu \quad \forall \mathbf{s} \in D \\ \mathrm{Cov}(Y(\mathbf{s}) - Y(\mathbf{s}')) &= C(\mathbf{s} - \mathbf{s}') \quad \forall \mathbf{s}, \mathbf{s}' \in D \end{aligned}$$

is defined to be **second-order stationary**. Furthermore, if $C(\mathbf{s} - \mathbf{s}')$ is a function only of $|| \mathbf{s} - \mathbf{s}' ||$ (it is not a function of the locations), then $C(.)$ is said to be **isotropic**.

Notice that a process which is second order stationary is also intrinsic stationary, the inverse is not necessarily true.
There is a stronger type of stationarity which is called **strict stationarity** (joint probability distribution of the data depends only on the relative positions of the sites at which the data were taken).
If the random process $Y(.)$ is Gaussian, we need only to specify its first-order and second-order properties, namely its *mean function* and its *covariance function*.
In practice, an assumption of second-order stationarity is often sufficient for inference purposes and it will be one of our basic assumptions.

Alex Schmidt
IM-UFRJ
Brazil

Alex Schmidt
IM-UFRJ
Brazil

Covariogram and Correlogram

If

$$Cov(Y(\mathbf{s}), Y(\mathbf{s}')) = C(\mathbf{s} - \mathbf{s}')$$

for all $\mathbf{s}, \mathbf{s}' \in D$, $C(.)$ is called the *covariogram*.
If $C(\mathbf{0}) > 0$, we can define

$$\rho(\mathbf{h}) = C(\mathbf{h})/C(\mathbf{0})$$

as the *correlogram*.
Properties:

- $C(\mathbf{h}) = C(-\mathbf{h})$
- $\rho(\mathbf{h}) = \rho(-\mathbf{h})$
- $\rho(\mathbf{0}) = 1$
- $C(\mathbf{0}) = Var(Y(\mathbf{s}))$ if $Y(.)$ is second order stationary.

Relationships between covariogram and variogram

Alex Schmidt
IM-UFRJ
Brazil

Consider

$$\begin{aligned} Var(Y(\mathbf{s}) - Y(\mathbf{s}')) &= Var(Y(\mathbf{s})) + Var(Y(\mathbf{s}')) \\ &\quad -2Cov(Y(\mathbf{s}), Y(\mathbf{s}')) \end{aligned}$$

If $Y(.)$ is second order stationary,

$$Var(Y(\mathbf{s}) - Y(\mathbf{s}')) = 2\{C(\mathbf{0}) - C(\mathbf{s} - \mathbf{s}')\}$$

If $Y(.)$ is intrinsically stationary,

$$2\gamma(\mathbf{h}) = 2\{C(\mathbf{0}) - C(\mathbf{h})\}$$

If $C(\mathbf{h}) \to 0$ as $|| \mathbf{h} || \to \infty$ then $2\gamma(\mathbf{h}) \to 2C(\mathbf{0})$. ($C(\mathbf{0})$ is the *sill* of the variogram).
The variogram estimation is to be preferred to covariogram estimation. (See Cressie, p.70 for more details)

Features of the Variogram

Alex Schmidt
IM-UFRJ
Brazil

- $\gamma(-\mathbf{h}) = \gamma(\mathbf{h})$
- $\gamma(\mathbf{0}) = 0$
- If $\lim_{\mathbf{h}\to 0} \gamma(\mathbf{h}) = c_0 \neq 0$, then c_0 is called the **nugget effect**.
- Mathematically a nugget effect means:
  - If Y is L_2 continuous(processes $Y(.)$ for which $E(Y(\mathbf{s}+\mathbf{h}) - Y(\mathbf{s}))^2 \to 0$, as $|| \mathbf{h} ||\to 0)$, nugget effects cannot happen!
    - So if continuity is assumed at the microscale (very small h) in the $Y(.)$ process, the only possible reason for $c_0 > 0$ is measurement error. (Recall $2\gamma(\mathbf{0})$ is the variance of the difference between two measurements taken at *exactly* the same place).
  - In practice we only have data $\{y(\mathbf{s}_i) : i = 1, \cdots, n\}$ so we can't say much for lags $h < min\{|| \mathbf{s}_i - \mathbf{s}_j ||: 1 \leq i < j \leq n\}$

- Often in spatial prediction we assume nugget effect is entirely due to measurement error. In other words, it is reasonable to envisage that a measured value at location $\mathbf{x}$ could be replicated, and that the resulting multiple values would not be identical. In this case, an alternative is to model a "white noise" zero-mean process that adds extra variation to each observation, that is

$$Y(\mathbf{s}_i) = S(\mathbf{s}_i) + Z_i,$$

where $S(\mathbf{x})$ follows a Gaussian process with covariance function $\gamma(u) = \sigma^2\rho(u)$ such that $\rho(0) = 1$ and the Z_i are mutually independent, $N(0, \tau^2)$ random variables. And the nuggect effect would be

$$\rho_Y(u) = \sigma^2\rho(u)/(\sigma^2 + \tau^2) \to \sigma^2/(\sigma^2 + \tau^2) < 1$$

as $u \to 0$

Properties of the variogram

Alex Schmidt
IM-UFRJ
Brazil

(i) $2\gamma(.)$ continuous at origin implies $Y(.)$ is L_2 continuous

(ii) $2\gamma(\mathbf{h})$ does not approach 0 as $\mathbf{h} \to$ origin implies $Y(.)$ is not L_2 continuous and is highly irregular

(iii) $2\gamma(.)$ is a positive constant (except at the origin where it is zero). Then $Y(\mathbf{s})$ and $Y(\mathbf{s}')$ are uncorrelated for any $\mathbf{s} \neq \mathbf{s}'$, regardless of how close they are; $Y(.)$ is often called *white noise*.

Properties of the variogram (continuing)

Alex Schmidt
IM-UFRJ
Brazil

PASI
June 2014

(iv) $2\gamma(.)$ must be conditionally negative-definite, i.e.

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a_i a_j 2\gamma(\mathbf{s}_i - \mathbf{s}_j) \leq 0$$

for any finite number of locations $\{\mathbf{s}_i : i = 1, \cdots n\}$ and real numbers $\{a_1, \cdots, a_n\}$ satisfying $\sum_{i=1}^{n} a_i = 0$.

(v) $2\gamma(\mathbf{h})/ \mid\mid \mathbf{h} \mid\mid^2 \rightarrow 0$ as $\mid\mid \mathbf{h} \mid\mid \rightarrow \infty$, i.e. $2\gamma(\mathbf{h})$ can't increase too fast with $\mid\mid \mathbf{h} \mid\mid^2$.

Some Isotropic Parametric Covariance Functions

Alex Schmidt
IM-UFRJ
Brazil

Most parametric variogram models used in practice will include a nugget effect, and in the stationary case will therefore be of the form

$$2\gamma(h) = \tau^2 + \sigma^2(1 - \rho(h))$$

$\rho(h)$ must be a positive definite function. Also we would usually require the model for the correlation function $\rho(h)$ to incorporate the following features:

1. $\rho(.)$ is monotone non-increasing in h;
2. $\rho(h) \to 0$ as $h \to \infty$;
3. at least one parameter in the model controls the rate at which $\rho(h)$ decays to zero.

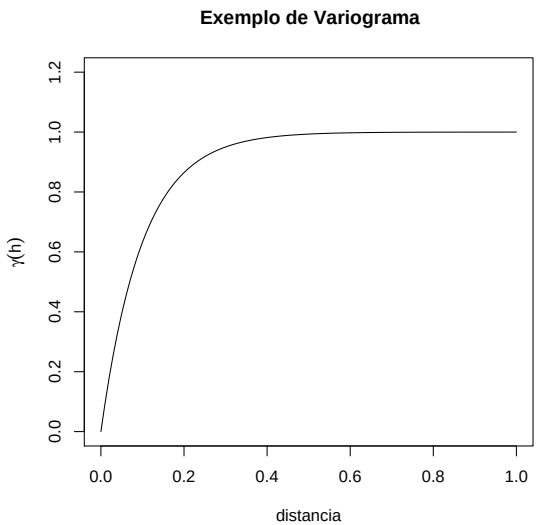
In addition we may wish to include in the model some flexibility in the overall shape of the correlation function. Hence, a parametric model for the correlation function can be expected to have one or two parameters, and a model for the variogram three or four (the two correlation parameters plus the two variance components).

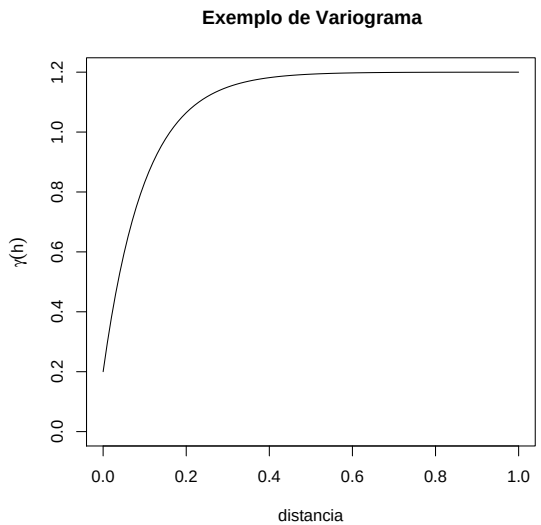
Alex Schmidt
IM-UFRJ
Brazil

How does a variogram usually look like?

Alex Schmidt
IM-UFRJ
Brazil

- *nugget effect* - represents micros-cale variation or measurement error. It can be estimated from the empirical variogram as the value of $\gamma(h)$ for $h = 0$.
- *sill* - the $\lim_{h \to \infty} \gamma(h)$ representing the variance of the random field.
- *range* - the distance (if any) at which data are no longer autocorrelated.





- **The spherical family**
  This one parameter family of correlation function is defined by

$$\rho(h;\phi) = \begin{cases} 1 - \frac{3}{2}(h/\phi) + \frac{1}{2}(h/\phi)^3 & ;0 \leq h \leq \phi \\ 0 & ;h > \phi \end{cases}$$

  Because the family depends only on a scale parameter ϕ, it gives no flexibility in shape.
  The spherical correlation function is continuous and twice-differentiable at the origin.
  Therefore corresponds to a mean-square differentiable process $Y(\mathbf{s})$.

- **The powered exponential family**
  This two parameter family is defined by

$$\rho(h) = \exp\{-(h/\phi)^{\kappa}\},$$

  with $\phi > 0$ and $0 < \kappa \leq 2$.
  The corresponding process $Y(\mathbf{s})$ is mean-square continuous (but non differentiable) if $\kappa < 2$, but becomes mean-square infinitely differentiable if $\kappa = 2$.
  The *exponential correlation function* corresponds to the case where $\kappa = 1$.
  The case $\kappa = 2$ is called the *Gaussian correlation function*.

Alex Schmidt
IM-UFRJ
Brazil

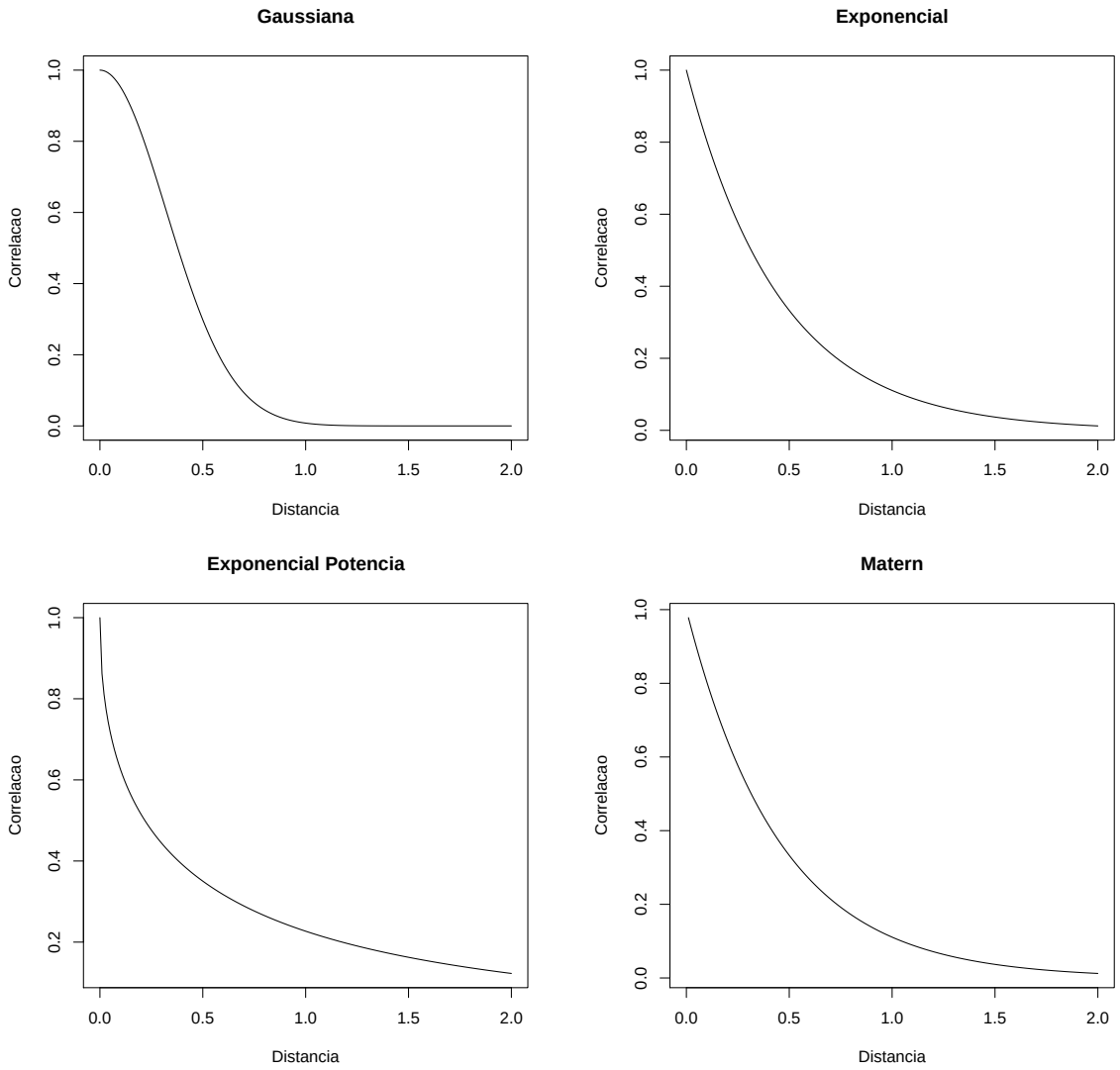
- **The Matérn family**
  It is defined by

$$\rho(h;\phi;\kappa) = \{2^{\kappa-1}\Gamma(\kappa)\}^{-1}(h/\phi)^{\kappa}K_{\kappa}(h/\phi)$$

  where (ϕ, κ) are parameters and $K_{\kappa}(.)$ denotes the modified Bessel function of the third kind of order κ.
  This family is valid for any $\phi > 0$ and $\kappa > 0$.
  The case $\kappa = 0.5$ is the same as the exponential correlation function,
  $\rho(h) = \exp(-h/\phi)$.
  The Gaussian correlation function is the limiting case as $\kappa \to \infty$.

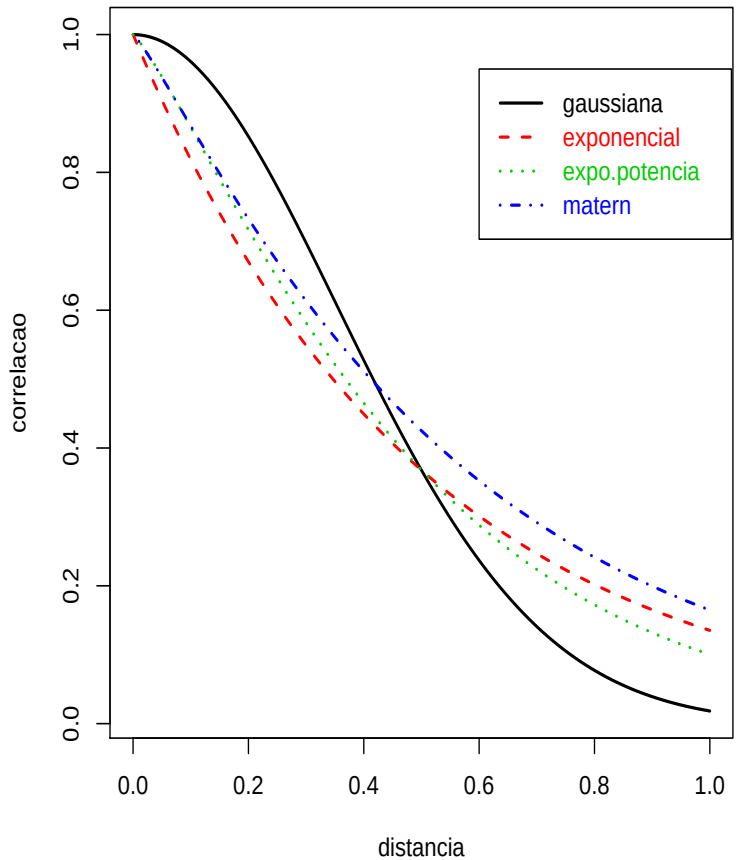
One attractive feature of this family is that the parameter κ controls the differentiability of the underlying process $Y(\mathbf{s})$ in a very direct way; the integer part of κ gives the number of times that $Y(\mathbf{s})$ is mean-square differentiable. The Matérn family is probably the best choice as a flexible, yet simple (only two parameters) correlation function for general case.

Alex Schmidt
IM-UFRJ
Brazil



Alex Schmidt
IM-UFRJ
Brazil

PASI
June 2014



Alex Schmidt
IM-UFRJ
Brazil

PASI
June 2014

Spatial Regression

Alex Schmidt
IM-UFRJ
Brazil

Generalized least squares (GLS) with *known* covariance matrix.

- Model:
  $\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad E(\boldsymbol{\epsilon}) = \mathbf{0}, \quad Var(\epsilon) = \mathbf{V},$
  where $\mathbf{V}$ is a completely specified positive definite matrix. (e.g. $V_{ij} = \sigma^2 \exp\{-\phi d_{ij}\}$, exponential family with ϕ and σ^2 known).
- GLS estimator of $\boldsymbol{\beta}$:
  $\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{F}'\mathbf{V}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{V}^{-1}\mathbf{Y}.$

But in practice ϕ is unknown and consequently $\mathbf{V}$ cannot be completely specified. A natural solution is to replace ϕ in the evaluation of $\mathbf{V}$ by an estimator ϕ, thereby obtaining $\hat{\mathbf{V}} = \mathbf{V}(\hat{\phi})$ (Estimated GLS).
EGLS estimator of $\boldsymbol{\beta}$:
$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}$

Alex Schmidt
IM-UFRJ
Brazil

Classical Procedure:

1. Estimate $\boldsymbol{\beta}$ using ordinary least squares;
2. Estimate residuals from this $\boldsymbol{\beta}$ estimate;
3. Calibrate the semi-variogram from the residuals;
4. Use the calibration of the semi-variogram to estimate $\mathbf{V}$;
5. Re-estimate $\boldsymbol{\beta}$ using $\boldsymbol{\beta}_{GLS}$.

```
#removing some observed values for prediction
predloc=c(2,10,85,91,102)
prices2=prices[-predloc,]

#standard regression
reg=lm(prices2[,1]~prices2[,2]+prices2[,3]
   +prices2[,4]+prices2[,6])
fitreg=summary(reg)
fitreg


#Residual Analysis
resid=matrix(c(prices2[,8],prices2[,9],fitreg$residuals),
    ncol=3,byrow=FALSE)
georesid=as.geodata(resid,coords.col=c(1,2),data.col=3)
plot(georesid)
```

```
#variogram of the residuals
# binned variogram
vario.b <- variog(georesid, max.dist=1)
# variogram cloud
vario.c <- variog(georesid, max.dist=1, op="cloud")
#binned variogram and stores the cloud
vario.bc <- variog(georesid, max.dist=1, bin.cloud=TRUE)
# smoothed variogram
#vario.s <- variog(georesid, max.dist=1, op="sm", band=0.2)
# plotting the variograms:
par(mfrow=c(2,2))
plot(vario.c, main="variogram cloud")
plot(vario.bc, bin.cloud=TRUE,
       main="clouds for binned variogram")
plot(vario.b, main="binned variogram")
#plot(vario.s, main="smoothed variogram")
```

```
#OLS estimate of the variogram
varioresid=variog(georesid,max.dist=0.08)
#ols estimates of the variogram parameters
olsvari=variofit(varioresid,ini=c(0.02,0.1))
summary(olsvari)
par(mfrow=c(1,1))
plot(variog(georesid,max.dist=0.10))
lines(olsvari)
```

```
#Envelopes for an empirical variogram by simulating
#data for given model parameters.
#Computes bootstrap paremeter estimates
resid.mc.env <- variog.mc.env(georesid, obj.variog
      = varioresid)
plot(varioresid,ylim=c(0,0.1))
lines(resid.mc.env)



#fitting the parameters of a variogram "by eye"
eyefit(varioresid)
```

```
#fitting a spatial regression model
meantrend=trend.spatial(~sqft+age+bedrooms+dist_freeway,
    geoprices)



#a set of initial values
#spatialreg1=likfit(geoprices,trend=meantrend,
      ini.cov.pars=c(0.15,0.1),nospatial=TRUE)
load(file="spatialreg1.RData")
#save("spatialreg1", file="spatialreg1.RData")
summary(spatialreg1)
```

```
#spatial interpolation for locations left out from inference
names(prices)
xp=prices[predloc,c(2:4,7)]
meantrend.loc <- trend.spatial(~sqft+age+bedrooms+
            dist_freeway, xp)

kc <- krige.conv(geoprices, loc=locationpred,
          krige=krige.control(trend.d=meantrend,
                              trend.l=meantrend.loc,
                    obj.model=spatialreg1),
          output=output.control(n.pred=1000))
names(kc)
dim(kc$simul)
par(mfrow=c(3,2), mar=c(3,3,0.5,0.5))
for(i in 1:5){
  hist(kc$simul[i,], main="", prob=T)
  lines(density(kc$simul[i,]))
  abline(v=prices[predloc,1][i], col=2)
}
```

Review of Kriging

Alex Schmidt
IM-UFRJ
Brazil

Recall that

1. $S(.)$ is a stationary Gaussian process with $E[S(\mathbf{s})] = 0$, $Var(S(\mathbf{s})) = \sigma^2$ and correlation function $\rho(\phi; \mathbf{h}) = Corr(S(\mathbf{s}), S(\mathbf{s} - \mathbf{h}))$;
2. The conditional distribution of Y_i given $S(.)$ is Gaussian with mean $\mu(\mathbf{x}_i) + S(\mathbf{x}_i)$ and variance τ^2;
3. $\mu(\mathbf{s}) = \sum_{j=1}^{p} \beta_j f_j(\mathbf{s})$ for known explanatory variables $f_j(.)$;
4. $Y_i : i = 1, 2, \cdots, n$ are mutually independent, conditional on $S(.)$.

These assumptions imply that the joint distribution of $\mathbf{Y}$ is multivariate Normal,

$$\mathbf{Y} \mid \boldsymbol{\theta} \sim N_n(\mathbf{F}\boldsymbol{\beta}, \sigma^2\mathbf{R} + \tau^2\mathbf{I})$$

where:

$\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \tau^2)$,

$\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)$,

$\mathbf{F}$ is the $n \times p$ matrix with j^{th} column $f_j(\mathbf{x}_1), \cdots, f_j(\mathbf{x})$,

$\mathbf{I}$ is the $n \times n$ identity matrix,

Parameter Estimation

Alex Schmidt
IM-UFRJ
Brazil

In our model the vector of parameters is given by $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \tau^2)$.
From Bayes' Theorem,

$$\pi(\boldsymbol{\theta}) \propto p(\boldsymbol{\theta})\, l(\boldsymbol{\theta})$$

It follows that

$$\pi(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \tau^2) \propto p(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \tau^2) \mid \mathbf{V} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y} - \mathbf{F}\boldsymbol{\beta})\right\} \quad (1)$$

where $\mathbf{V} = \sigma^2\mathbf{R} + \tau^2\mathbf{I}$. Usually we assume that the parameters are independent *a priori*, then

$$p(\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\gamma}) = p(\boldsymbol{\beta})\, p(\sigma^2)\, p(\tau^2)\, p(\boldsymbol{\gamma})$$

Predictive Inference

Alex Schmidt
IM-UFRJ
Brazil

Notice that

$$p(\mathbf{S} \mid \mathbf{Y}) = \int p(\mathbf{S} \mid \mathbf{Y}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{Y}) d\boldsymbol{\theta}$$

The Bayesian predictive distribution is an average of classical predictive distributions for particular values of $\boldsymbol{\theta}$, weighted according to the posterior distribution of $\boldsymbol{\theta}$.
The effect of the above averaging is typically to make the predictions more conservative, in the sense that the variance of the predictive distribution is usually expected to be larger than the variance of the distribution $(\mathbf{S} \mid \mathbf{Y}, \hat{\boldsymbol{\theta}})$, obtained by plugging an estimate of $\boldsymbol{\theta}$ into the classical predictive distribution.

Prediction with uncertainty only in the mean parameter

Alex Schmidt
IM-UFRJ
Brazil

Assume σ^2 and ϕ are known and $\tau^2 = 0$.
Assume a Gaussian prior mean for the parameter $\boldsymbol{\beta} \sim N(\mathbf{m}_{\boldsymbol{\beta}}, \sigma^2 \mathbf{V}_{\boldsymbol{\beta}})$, where σ^2 is the (assumed known) variance of $S(\mathbf{x})$, the posterior is given by

$$\boldsymbol{\beta} \mid \mathbf{Y} \sim N(\hat{\boldsymbol{\beta}}; \hat{\mathbf{V}}_{\boldsymbol{\beta}})$$

where
$\hat{\boldsymbol{\beta}} = (\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1}(\mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}} + \mathbf{F}'\mathbf{R}^{-1}\mathbf{y})$
and
$\hat{\mathbf{V}}_{\boldsymbol{\beta}} = \sigma^2(\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1}$

Prediction of $Y_0 = S(\mathbf{x}_0)$ at an arbitrary location $\mathbf{x}_0$,

$$p(Y_0 \mid \mathbf{Y}, \sigma^2, \phi) =$$

$$\int p(Y_0 \mid \mathbf{Y}, \boldsymbol{\beta}, \sigma^2, \phi) p(\boldsymbol{\beta} \mid \mathbf{Y}, \sigma^2, \phi) d\boldsymbol{\beta}$$

This predictive distribution is Gaussian with mean and variance given by

$$\begin{aligned} E[Y_0 \mid \mathbf{Y}] &= (\mathbf{F}_0 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{F})(\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1} \\ &\quad \mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}} + \\ &\quad [\mathbf{r}'\mathbf{R}^{-1} + (\mathbf{F}_0 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{F}) \\ &\quad (\mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{R}^{-1}]\mathbf{Y} \\ V[Y_0 \mid \mathbf{Y}] &= \sigma^2[\mathbf{R}_0 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{r} + \\ &\quad (\mathbf{F}_0 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{F})(\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1} \\ &\quad (\mathbf{F}_0 - \mathbf{r}'\mathbf{R}^{-1}\mathbf{F}]. \end{aligned}$$

where $\mathbf{r}$ is the vector of correlations between $S(\mathbf{x}_0)$ and $S(\mathbf{x}_i) : i = 1, \cdots, n$.

The predictive variance has 3 components: the first and second components represent the marginal variance for Y_0 and the variance reduction after observing $\mathbf{Y}$, whilst the third component accounts for the additional uncertainty due to the unknown value of $\boldsymbol{\beta}$. This last component reduces to zero if $V_{\boldsymbol{\beta}} = 0$, since this formally corresponds to $\boldsymbol{\beta}$ being known beforehand.
In the limit as all diagonal elements of $\mathbf{V}_{\boldsymbol{\beta}} \to \infty$, these formulae for the predictive mean and variance correspond exactly to the universal kriging predictor and its associated kriging variance, which in turn reduce to the formulae for ordinary kriging if the mean value surface is assumed constant.
$\Rightarrow$ ordinary and universal kriging can be interpreted as Bayesian prediction under prior ignorance about the mean (and known σ^2).

Alex Schmidt
IM-UFRJ
Brazil

Prediction with uncertainty in all model parameters

In this case all the quantities in the model are considered unknown $\Rightarrow$ need to specify their prior distribution.

- Likelihood is like in equation (1);
- Bayesian Model completed with prior distribution for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \tau^2)$ and assume, for example, $V_{ij} = \sigma^2 \exp(-\phi(d_{ij})^\kappa)$; in this case $\boldsymbol{\gamma} = (\phi, \kappa)$;
- $p(\boldsymbol{\beta}) \sim N_p(\mathbf{0}, \sigma^2 \mathbf{V}_{\boldsymbol{\beta}})$, for some fixed diagonal matrix $\mathbf{V}_{\boldsymbol{\beta}}$;
- ϕ must be positive, then one possibility is $p(\phi) \sim Ga(a_2, b_2)$;
- κ could have an uniform prior on the interval $[0.05, 2]$.
- Usually we use $p(\boldsymbol{\sigma}^2) \sim IG(a_1, b_1)$ and $p(\tau^2) \sim IG(a_3, b_3)$ (conjugacy of the normal-gamma) .

Alex Schmidt
IM-UFRJ
Brazil

Posterior distribution

Alex Schmidt
IM-UFRJ
Brazil

Assume $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2, \phi, \tau^2, \kappa)$ then

$$\begin{aligned}
\pi(\boldsymbol{\theta}) \quad \propto \quad & \mid \mathbf{V}(\boldsymbol{\gamma}) \mid^{-1/2} \exp\left\{ -\frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})' \mathbf{V}(\boldsymbol{\gamma})^{-1} \right. \\
& \left. (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \right\} \\
& \times (\sigma^2)^{-n/2} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{i=1}^{p} (\beta_i - \mu_i)^2 \right\} \\
& \times (\phi)^{(a_2 - 1)} \exp\{-b_2 \phi\} \\
& \times (\sigma^2)^{-(a_1+1)} \exp\{-b_1/\sigma^2\} \\
& \times (\tau^2)^{-(a_3+1)} \exp\{-b_3/\tau^2\}
\end{aligned}$$

$\Rightarrow$ Unknown distribution $\Rightarrow$ need of approximation.
Usually, we use MCMC methods.

Predictive Inference

Alex Schmidt
IM-UFRJ
Brazil

- Recall that

$$p(\mathbf{S} \mid \mathbf{Y}) = \int p(\mathbf{S} \mid \mathbf{Y}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{Y}) d\boldsymbol{\theta}$$

- Suppose we can draw samples from the joint posterior distribution for $\boldsymbol{\theta}$, i.e

$$\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}, \cdots, \boldsymbol{\theta}^{(L)} \sim \pi(\boldsymbol{\theta})$$

- Then

$$\begin{aligned} p(\mathbf{S} \mid \mathbf{Y}) &= \int p(\mathbf{S} \mid \mathbf{Y}, \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \mathbf{Y}) d\boldsymbol{\theta} \\ &\approx \frac{1}{L} \sum_{i=1}^{L} p(\mathbf{S} \mid \mathbf{Y}, \boldsymbol{\theta}^{(i)}) \end{aligned}$$

$\downarrow$

This is Monte Carlo integration

Predictive Inference (Bayesian kriging)

Alex Schmidt
IM-UFRJ
Brazil

- Prediction of $Y(\mathbf{s}_0)$ at a new site $\mathbf{s}_0$ associated covariates $\mathbf{x}_0 = \mathbf{x}(\mathbf{s}_0)$
- Predictive distribution:

$$p(y(\mathbf{s}_0) \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0) = \int p(y(\mathbf{s}_0, \boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0) d\boldsymbol{\theta}$$

$$= \int p(y(\mathbf{s}_0) \mid \mathbf{y}, \boldsymbol{\theta}, \mathbf{X}, \mathbf{X}_0) p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}$$

- $p(y(\mathbf{s}_0) \mid \mathbf{y}, \boldsymbol{\theta}, \mathbf{X}, \mathbf{X}_0)$ is normal since $p(y(\mathbf{s}_0), \mathbf{y} \mid \boldsymbol{\theta}, \mathbf{X}, \mathbf{X}_0)$ is!
- $\Rightarrow$ easy Monte Carlo estimate using composition with Gibbs draws $\boldsymbol{\theta}^{(1)}, \cdots, \boldsymbol{\theta}^{(L)}$: for each $\boldsymbol{\theta}^{(l)}$ drawn from $p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X})$, draw $Y(\mathbf{s}_0)^{(l)}$ from $p(y(\mathbf{s}_0) \mid \mathbf{y}, \boldsymbol{\theta}^{(l)}, \mathbf{X}, \mathbf{X}_0)$

Predictive Inference

Alex Schmidt
IM-UFRJ
Brazil

- Suppose we want to predict at a set of m sites, say $S_0 = \{\mathbf{s}_{01}, \cdots, \mathbf{s}_{0m}\}$
- We could individually predict each site "independently" using method of the previous slide
- BUT joint prediction may be of interest, e.g., bivariate predictive distributions to reveal pairwise dependence, to reflect posterior associations in the realized surface
- Form the unobserved vector $\mathbf{Y}_0 = (Y(\mathbf{s}_{01}, \cdots, Y(\mathbf{s}_{0m}))$, with $\mathbf{X}_0$ as covariate matrix for S_0, and compute

$$p(\mathbf{y}_0 \mid \mathbf{y}, \mathbf{X}, \mathbf{X}_0) = \int p(\mathbf{y}_0 \mid \mathbf{y}, \boldsymbol{\theta}, \mathbf{X}, \mathbf{X}_0) p(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{X}) d\boldsymbol{\theta}$$

- Again, posterior sampling using composition sampling

Main References

Alex Schmidt
IM-UFRJ
Brazil

PASI
June 2014

- Banerjee, S., Carlin, B. P. and Gelfand, A. E. (2004) Hierarchical Modeling and Analysis of Spatial Data. New York: Chapman and Hall
- Cressie, N.A. (1993) Statistics for Spatial Data. John Wiley & son Inc.
- Diggle, P.J. e Ribeiro Jr., P.J. (2007) Model based Geostatistics. Springer Series in Statistics