# Gaussian Process Regression and Emulation

## STAT8810, Fall 2017

M.T. Pratola

September 5, 2017

# Today

More on GP Regression

# The Empirical Variogram

- How do we know if a correlation model is “good” for our data?

# The Empirical Variogram

- How do we know if a correlation model is "good" for our data?
- Suppose $\{Y(s) : s \in D \subset \mathbb{R}^d\}$ is a real-valued process on a domain $D$ of $d$-dimensional Euclidean space, and suppose that differences of variables displaced by $h$ apart depend only on $h$. That is,

$$Var(Y(x+h) - Y(x)) = 2\gamma_Y(h) \forall x, x+h \in D.$$

# The Empirical Variogram

- How do we know if a correlation model is "good" for our data?
- Suppose $\{Y(s) : s \in D \subset \mathbb{R}^d\}$ is a real-valued process on a domain $D$ of $d$-dimensional Euclidean space, and suppose that differences of variables displaced by $h$ apart depend only on $h$. That is,

$$Var(Y(x+h) - Y(x)) = 2\gamma_Y(h) \forall x, x+h \in D.$$

- The quantity $2\gamma_Y(h)$ is a function of only the difference between locations in $D$, and is called the *variogram*.

# The Empirical Variogram

- How do we know if a correlation model is "good" for our data?
- Suppose $\{Y(s) : s \in D \subset \mathbb{R}^d\}$ is a real-valued process on a domain $D$ of $d$-dimensional Euclidean space, and suppose that differences of variables displaced by $h$ apart depend only on $h$. That is,

$$Var(Y(x+h) - Y(x)) = 2\gamma_Y(h) \forall x, x+h \in D.$$

- The quantity $2\gamma_Y(h)$ is a function of only the difference between locations in $D$, and is called the *variogram*.
- A variogram written as a function of $||h||$ is said to be *isotropic*. Otherwise, it is *anisotropic*.

# Example

- The semi-variogram for the powered exponential model is

$$\gamma_Y(h;\theta) = c_Y(0;\theta) - c_Y(h;\theta)$$

where $c_Y(h;\theta) = \sigma^2 exp\left(-\frac{||h||^{\theta_2}}{\theta_1}\right)$.

# Example

- The semi-variogram for the powered exponential model is

$$\gamma_Y(h;\theta) = c_Y(0;\theta) - c_Y(h;\theta)$$

where $c_Y(h;\theta) = \sigma^2 exp\left(-\frac{||h||^{\theta_2}}{\theta_1}\right)$.

- The *empirical* semi-variogram is calculated as

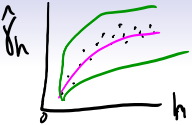
$$\hat{\gamma}_Y(h) = \frac{1}{2}\texttt{ave}\left\{(\overset{Y}{\cancel{Z}}(x_i) - \overset{Y}{\cancel{Z}}(x_j))^2 : ||x_i - x_j|| \in T(h), i,j = 1,\ldots,n\right\}$$

where $T(h)$ is some tolerance region around $h$ (e.g. $h \pm \Delta$, for $\Delta$ small).

# The Empirical Variogram

- How do we know if a correlation model is "good" for our data?

† Cressie & Wikle: Statistics for Spatial-Temporal Data (2011)




# The Empirical Variogram

- How do we know if a correlation model is "good" for our data?

1. Eyeball test: compare the theoretical semivariogram (at various $\theta$'s or maybe $\hat{\theta}_{MLE}$) versus the empirical semivariogram.

† Cressie & Wikle: Statistics for Spatial-Temporal Data (2011)

# The Empirical Variogram

- How do we know if a correlation model is "good" for our data?

1. Eyeball test: compare the theoretical semivariogram (at various $\theta$'s or maybe $\hat{\theta}_{MLE}$) versus the empirical semivariogram.
2. Simple test of $H_0 : Z(\cdot)$ has no spatial dependence.†

† Cressie & Wikle: Statistics for Spatial-Temporal Data (2011)

# Simple Test of Dependence

2. Let $F = \frac{\hat{\gamma}_Y(h_1)}{\hat{\sigma}^2}$ where $\hat{\sigma}^2 = \frac{\sum (Z(x_i) - \hat{\mu})^2}{n-1}$, $\hat{\mu} = \frac{1}{n} \sum Z(x_i)$ and $h_1$ is the smallest lag from all possible lags $h_1, \ldots, h_L$.

Note: this does not test the validity of the correlation model! Only whether correlation is present or not.

# Simple Test of Dependence

2. Let $F = \frac{\hat{\gamma}_Y(h_1)}{\hat{\sigma}^2}$ where $\hat{\sigma}^2 = \frac{\sum(Z(x_i)-\hat{\mu})^2}{n-1}, \hat{\mu} = \frac{1}{n}\sum Z(x_i)$ and $h_1$ is the smallest lag from all possible lags $h_1, \ldots, h_L$.
   - Reject $H_0$ for $|F-1|$ large. How? Permutation-based test is one distribution-free approach:

Note: this does not test the validity of the correlation model! Only whether correlation is present or not.

# Simple Test of Dependence

2. Let $F = \frac{\hat{\gamma}_Y(h_1)}{\hat{\sigma}^2}$ where $\hat{\sigma}^2 = \frac{\sum(Z(x_i)-\hat{\mu})^2}{n-1}, \hat{\mu} = \frac{1}{n}\sum Z(x_i)$ and $h_1$ is the smallest lag from all possible lags $h_1, \ldots, h_L$.
   - Reject $H_0$ for $|F-1|$ large. How? Permutation-based test is one distribution-free approach:
   - i: Permute the data locations $x_{\sigma(1)}, \ldots, x_{\sigma(n)}$

Note: this does not test the validity of the correlation model! Only whether correlation is present or not.

# Simple Test of Dependence

2. Let $F = \frac{\hat{\gamma}_Y(h_1)}{\hat{\sigma}^2}$ where $\hat{\sigma}^2 = \frac{\sum (Z(x_i) - \hat{\mu})^2}{n-1}, \hat{\mu} = \frac{1}{n} \sum Z(x_i)$ and $h_1$ is the smallest lag from all possible lags $h_1, \dots, h_L$.
   - Reject $H_0$ for $|F - 1|$ large. How? Permutation-based test is one distribution-free approach:
   - i: Permute the data locations $x_{\sigma(1)}, \dots, x_{\sigma(n)}$
   - ii: Recompute $F$

Note: this does not test the validity of the correlation model! Only whether correlation is present or not.

# Simple Test of Dependence

2. Let $F = \frac{\hat{\gamma}_Y(h_1)}{\hat{\sigma}^2}$ where $\hat{\sigma}^2 = \frac{\sum (Z(x_i) - \hat{\mu})^2}{n-1}$, $\hat{\mu} = \frac{1}{n} \sum Z(x_i)$ and $h_1$ is the smallest lag from all possible lags $h_1, \ldots, h_L$.
   - Reject $H_0$ for $|F - 1|$ large. How? Permutation-based test is one distribution-free approach:
   - i: Permute the data locations $x_{\sigma(1)}, \ldots, x_{\sigma(n)}$
   - ii: Recompute $F$
   - iii: Repeat steps i,ii many times

Note: this does not test the validity of the correlation model! Only whether correlation is present or not.

# Simple Test of Dependence

2. Let $F = \frac{\hat{\gamma}_Y(h_1)}{\hat{\sigma}^2}$ where $\hat{\sigma}^2 = \frac{\sum(\cancel{Z}Y(x_i)-\hat{\mu})^2}{n-1}$, $\hat{\mu} = \frac{1}{n}\sum \cancel{Z}Y(x_i)$ and $h_1$ is the smallest lag from all possible lags $h_1, \ldots, h_L$.
   - Reject $H_0$ for $|F-1|$ large. How? Permutation-based test is one distribution-free approach:
   - i: Permute the data locations $x_{\sigma(1)}, \ldots, x_{\sigma(n)}$
   - ii: Recompute $F$
   - iii: Repeat steps i,ii many times
   - iv: If the observed $F$ is above the 97.5th percentile or below the 2.5th percentile of the permutation distribution, reject $H_0$ at the 5% level.

Note: this does not test the validity of the correlation model! Only whether correlation is present or not.

# Simple Test of Dependence

**3.** A simulation approach:

† Ribeiro, Paulo and Diggle: `geoR`: a package for geostatistical analysis (2001).

# Simple Test of Dependence

3. A simulation approach:
   - i: Draw independent unconditional realizations from the GP with given correlation parameters.

† Ribeiro, Paulo and Diggle: `geoR`: a package for geostatistical analysis (2001).

# Simple Test of Dependence

3. A simulation approach:
   - i: Draw independent unconditional realizations from the GP with given correlation parameters.
   - ii: Construct empirical variogram at same lags as observed dataset for each GP draw.

† Ribeiro, Paulo and Diggle: `geoR`: a package for geostatistical analysis (2001).

# Simple Test of Dependence

**3.** A simulation approach:

- i: Draw independent unconditional realizations from the GP with given correlation parameters.
- ii: Construct empirical variogram at same lags as observed dataset for each GP draw.
- Repeat i,ii many times.

† Ribeiro, Paulo and Diggle: `geoR`: a package for geostatistical analysis (2001).

# Simple Test of Dependence

**3.** A simulation approach:

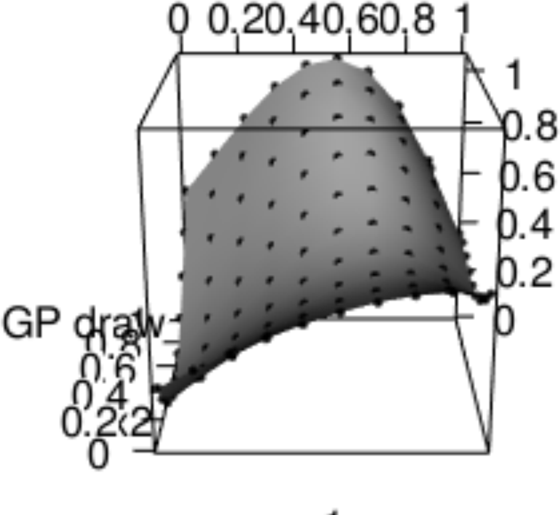
- i: Draw independent unconditional realizations from the GP with given correlation parameters.
- ii: Construct empirical variogram at same lags as observed dataset for each GP draw.
- Repeat i,ii many times.
- Construct variogram "envelopes" from the draws† and see if the observed empirical variogram of the dataset lie within the envelope.

† Ribeiro, Paulo and Diggle: `geoR`: a package for geostatistical analysis (2001).

# Example

```
source("dace.sim.r")
#
# Generate a 2-D realization
#
set.seed(99) #remove for a "real" realization
library(rgl) #for nice plots
n=10
x1=x2=seq(0,1,length=n)
design=as.matrix(expand.grid(x1,x2))
l1=list(m1=abs(outer(design[,1],design[,1],"-")))
l2=list(m2=abs(outer(design[,2],design[,2],"-")))
l.dez=list(l1=l1,l2=l2)
rho=c(0.5,0.5)
alpha=2
s2=1
se2=0
z=sim.field(l.dez,rho,s2,se2=se2,alpha=alpha,
```

# Example



# Example

```
library(geoR)
```

```
## --------------------------------------------------------------
##  Analysis of Geostatistical Data
##  For an Introduction to geoR go to http://www.leg.ufpr.b
##  geoR version 1.7-5.2 (built on 2016-05-02) is now loade
## --------------------------------------------------------------
```

```
gsim=cbind(design,z)
gsim=as.geodata(gsim)
vgram=variog(gsim)
```

```
## variog: computing omnidirectional variogram
```

```
eyefit(vgram)  # Explore different models, etc.
```

# Example

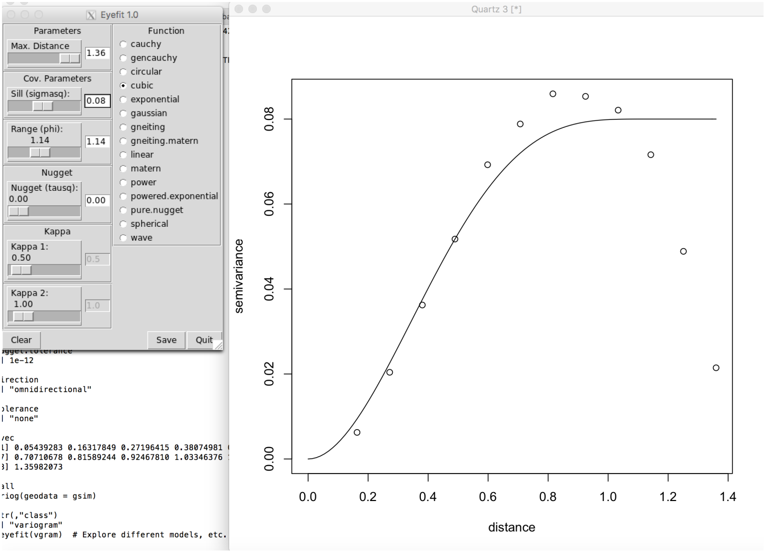



**Figure 2:**

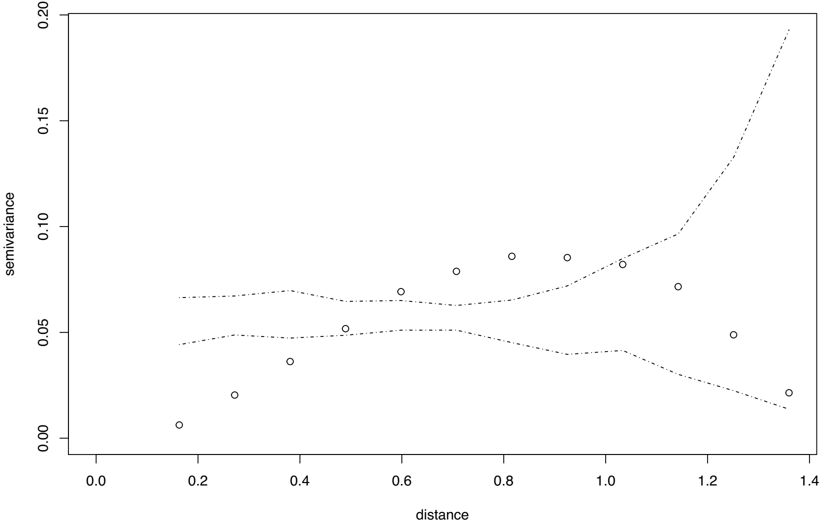
# Example

```
# Construct envelope using permutation test.
vgram.env=variog.mc.env(gsim, obj.var = vgram)
```

```
## variog.env: generating 99 simulations by permutating dat
## variog.env: computing the empirical variogram for the 99
## variog.env: computing the envelops
```

# Example

```
plot(vgram, envelope = vgram.env)
```



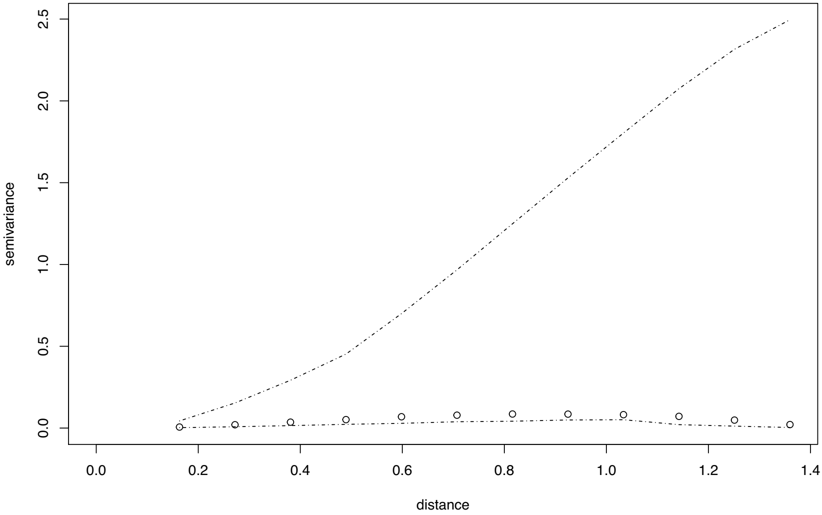
# Example

```
# Construct envelope by simulating from the model
vgram.ml=likfit(gsim, ini.cov.pars = c(2.5, 1),
         cov.model="gaussian", fix.nugget = TRUE,
         nugget=1e-10)
```

```
## kappa not used for the gaussian correlation function
## -----------------------------------------------------
## likfit: likelihood maximisation using the function optir
## likfit: Use control() to pass additional
##          arguments for the maximisation function.
##         For further details see documentation for optim.
## likfit: It is highly advisable to run this function seve
##         times with different initial values for the para
## likfit: WARNING: This step can be time demanding!
## -----------------------------------------------------
## likfit: end of numerical maximisation.
```

# Example

```
plot(vgram, envelope = vgram.env)
```



# Example

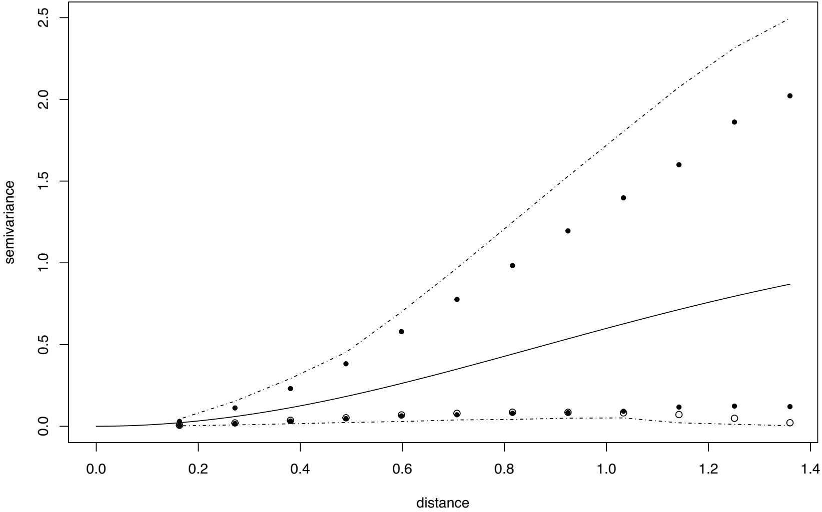
```
# Instead of plotting max/min envelope, plot the 5%
# and 95% quantiles from the simulations
gtmp=gsim
nbins=length(vgram$u)
vgsims=matrix(0,nrow=100,ncol=nbins)
for(i in 1:100) {
    gtmp$data=vgram.env$simulated.data[,i]
    vg=variog(gtmp)
    vgsims[i,]=vg$v
}
```

```
## variog: computing omnidirectional variogram
## variog: computing omnidirectional variogram
## variog: computing omnidirectional variogram
## variog: computing omnidirectional variogram
## variog: computing omnidirectional variogram
## variog: computing omnidirectional variogram
```

# Example

```
plot(vgram, envelope = vgram.env)
for(i in 1:nbins) points(vg$u[i],quantile(vgsims[,i],0.05)
for(i in 1:nbins) points(vg$u[i],quantile(vgsims[,i],0.95)

# Add in a line at the MLE estimate of the model
lines.variomodel(vg$u,cov.model="gaussian",
    cov.pars=vgram.ml$cov.pars,nugget=1e-10)
```

# Example



# Sample Properties of GP's

- path-wise properties depend on behaviour of $Z(x) = Z(x, \omega)$ when $\omega$ is fixed.

# Sample Properties of GP's

- path-wise properties depend on behaviour of $Z(x) = Z(x, \omega)$ when $\omega$ is fixed.
- process-wise properties given as usual specification of mean and covariance.

# Continuity

**Definition:** Suppose $Z(x)$ is a process with finite second moments. Then, $Z(\cdot)$ is *mean-square continuous* at $x_0$ if $E[|Z(x)|^2] < +\infty$ and

$$lim_{x \to x_0} E_\Omega \left[ |Z(x) - Z(x_0)|)^2 \right] = 0.$$

**Definition:** $Z(x)$ has almost surely continuous sample paths on $\chi$ provided

$$P\left( \omega : \forall x_0 \in \chi, z(x) \to z(x_0) \text{ as } x \to x_0 \right) = 1.$$

- this one is more abstract. It says that the only events that occur with probability 1 are those where the sample path of the process is continuous.

# Continuity

- Let $c_Y(\cdot)$ be the covariance function of a stationary process $Y(\cdot)$. Then,

$$\begin{aligned} E\left[(Y(x) - Y(x_0))^2\right] &= E\left[Y^2(x) - 2Y(x)Y(x_0) + Y^2(x_0)\right] \\ &= 2c_Y(0) - 2c_Y(x - x_0) \\ &= 2\left(c_Y(0) - c_Y(x - x_0)\right) \end{aligned}$$

So,

$$\begin{aligned} lim_{x \to x_0} E\left[(Y(x) - Y(x_0))^2\right] &= lim_{x \to x_0} c_Y(0) - c_Y(x - x_0) = 0 \\ &\Leftrightarrow lim_{x \to x_0} c_Y(x - x_0) = c_Y(0) \end{aligned}$$

i.e. the process is mean-square continuous $\forall x_0 \in \chi$ provided $c_Y(\cdot)$ is continuous at the origin. Or similarly, $R(h) \to 1$ as $h \to 0$ and $R(h)$ is continuous at origin.

# Continuity

- For GP's, mean-square continuity can be expressed as

$$E\left[|Z(x_i) - Z(x_j)|^2\right] \to 0 \text{ as } ||x_j - x_i|| \to 0.\dagger$$

† Theorem 3.4.1 in Adler: The Geometry of Random Fields (1981).

‡ Pg. 60 in Adler: The Geometry of Random Fields (1981).

# Continuity

- For GP's, mean-square continuity can be expressed as

$$E\left[|Z(x_i) - Z(x_j)|^2\right] \to 0 \text{ as } ||x_j - x_i|| \to 0.\dagger$$

- It has also been shown that the sample paths of a GP are almost surely continuous if $R(h)$ converges to 1 sufficiently slow:

$$1 - R(h) \leq \frac{c}{|log(||h||_2)|^{1+\epsilon}} \forall ||h||_2 < \delta$$

for some $0 < c < \infty$, some $\epsilon > 0$ and some $\delta < 1$ then $Z(\cdot)$ has a.s. continuous sample paths.

† Theorem 3.4.1 in Adler: The Geometry of Random Fields (1981).

‡ Pg. 60 in Adler: The Geometry of Random Fields (1981).

# Continuity

- For GP's, mean-square continuity can be expressed as
$$E\left[|Z(x_i) - Z(x_j)|^2\right] \to 0 \text{ as } ||x_j - x_i|| \to 0.\dagger$$
- It has also been shown that the sample paths of a GP are almost surely continuous if $R(h)$ converges to 1 sufficiently slow:
  - ‡ If $Z(\cdot)$ is a stationary GRF with correlation function $R$ satisfying

$$1 - R(h) \leq \frac{c}{|log(||h||_2)|^{1+\epsilon}} \forall ||h||_2 < \delta$$

for some $0 < c < \infty$, some $\epsilon > 0$ and some $\delta < 1$ then $Z(\cdot)$ has a.s. continuous sample paths.

† Theorem 3.4.1 in Adler: The Geometry of Random Fields (1981).

‡ Pg. 60 in Adler: The Geometry of Random Fields (1981).

# Continuity

- For GP's, mean-square continuity can be expressed as
$$E\left[|Z(x_i) - Z(x_j)|^2\right] \to 0 \text{ as } ||x_j - x_i|| \to 0.\dagger$$
- It has also been shown that the sample paths of a GP are almost surely continuous if $R(h)$ converges to 1 sufficiently slow:
  - ‡ If $Z(\cdot)$ is a stationary GRF with correlation function $R$ satisfying

$$1 - R(h) \leq \frac{c}{|log(||h||_2)|^{1+\epsilon}} \forall ||h||_2 < \delta$$

for some $0 < c < \infty$, some $\epsilon > 0$ and some $\delta < 1$ then $Z(\cdot)$ has a.s. continuous sample paths.

  - In other words, $(1 - R(h))|log(||h||_2)|^{1+\epsilon} \leq c$ is bounded.

† Theorem 3.4.1 in Adler: The Geometry of Random Fields (1981).

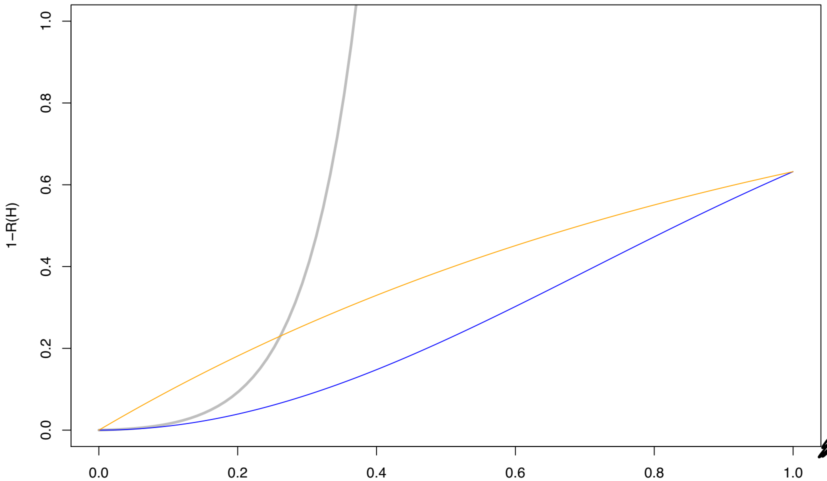
‡ Pg. 60 in Adler: The Geometry of Random Fields (1981).

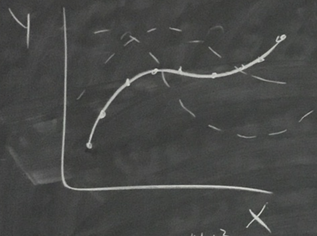
# Continuity

```
c=1
eps=4
h=seq(0,1,length=100)

# Bound from Adler, 1981
bound=c/( abs(log(h))^(1+eps) )

# Gaussian correlation with theta=1
Rh=exp(-h^2)

# Exponential correlation with theta=1
Rhe=exp(-abs(h))
```

# Continuity

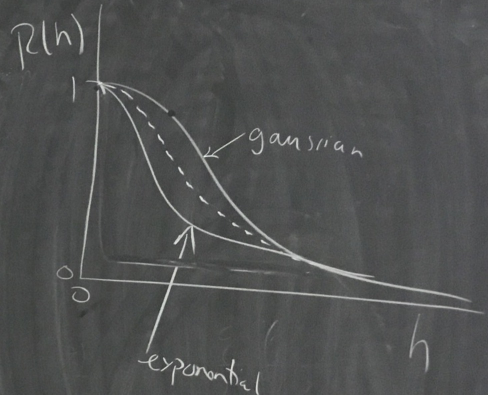
```
plot(h,bound,type='l',lwd=3,col="grey",ylim=c(0,1),xlab="h'
lines(h,1-Rh,col="blue")
lines(h,1-Rhe,col="orange")
```






$$R(h) = e^{-\frac{\|h\|^2}{\theta}} = \rho^{\|h\|^2}$$

$$\rightarrow \hat{\theta}$$



# Continuity

- For the Gaussian correlation model, $R(h) = exp(-\theta h^2)$, then $Z(\cdot)$ is a.s. continuous and a.s. infinitely differentiable.

# Continuity

- For the Gaussian correlation model, $R(h) = exp(-\theta h^2)$, then $Z(\cdot)$ is a.s. continuous and a.s. infinitely differentiable.
- For the Matern correlation model, $Z(\cdot)$ is a.s. continuous and $\lfloor \nu \rfloor$ times a.s. differentiable.

# Continuity

- For the Gaussian correlation model, $R(h) = exp(-\theta h^2)$, then $Z(\cdot)$ is a.s. continuous and a.s. infinitely differentiable.
- For the Matern correlation model, $Z(\cdot)$ is a.s. continuous and $\lfloor \nu \rfloor$ times a.s. differentiable.
- For the cubic correlation model, $Z(\cdot)$ is a.s. continuous and 2 times differentiable.

# Correlation/Covariance of Derivative Processes

- $E\left[\frac{\partial}{\partial x_i} Z(x)\right] = \frac{\partial}{\partial x_i} E\left[Z(x)\right]$

Morris, Mitchell and Ylvisaker: Bayesian design and analysis of computer experiments: use of derivatives in surface prediction (1993)

# Correlation/Covariance of Derivative Processes

- $E\left[\frac{\partial}{\partial x_i} Z(x)\right] = \frac{\partial}{\partial x_i} E\left[Z(x)\right]$
- $\text{Cov}\left(\frac{\partial}{\partial x_i} Z(x), Z(x')\right) = \frac{\partial}{\partial x_i} \text{Cov}\left(Z(x), Z(x')\right) = \frac{\partial}{\partial x_i} c(x - x').$

Morris, Mitchell and Ylvisaker: Bayesian design and analysis of computer experiments: use of derivatives in surface prediction (1993)

# Correlation/Covariance of Derivative Processes

- $E\left[\frac{\partial}{\partial x_i} Z(x)\right] = \frac{\partial}{\partial x_i} E\left[Z(x)\right]$
- $\mathrm{Cov}\left(\frac{\partial}{\partial x_i} Z(x), Z(x')\right) = \frac{\partial}{\partial x_i} \mathrm{Cov}\left(Z(x), Z(x')\right) = \frac{\partial}{\partial x_i} c(x - x')$.
- $\mathrm{Cov}\left(\frac{\partial}{\partial x_i} Z(x), \frac{\partial}{\partial x_j} Z(x')\right) = \frac{\partial^2}{\partial x_i \partial x_j} \mathrm{Cov}\left(Z(x), Z(x')\right) = \frac{\partial^2}{\partial x_i \partial x_j} c(x - x')$.

Morris, Mitchell and Ylvisaker: Bayesian design and analysis of computer experiments: use of derivatives in surface prediction (1993)

## Correlation/Covariance of Derivative Processes

- $E\left[\frac{\partial}{\partial x_i} Z(x)\right] = \frac{\partial}{\partial x_i} E\left[Z(x)\right]$
- $\mathrm{Cov}\left(\frac{\partial}{\partial x_i} Z(x), Z(x')\right) = \frac{\partial}{\partial x_i} \mathrm{Cov}\left(Z(x), Z(x')\right) = \frac{\partial}{\partial x_i} c(x - x')$.
- $\mathrm{Cov}\left(\frac{\partial}{\partial x_i} Z(x), \frac{\partial}{\partial x_j} Z(x')\right) = \frac{\partial^2}{\partial x_i \partial x_j} \mathrm{Cov}\left(Z(x), Z(x')\right) = \frac{\partial^2}{\partial x_i \partial x_j} c(x - x')$.
- Clearly we at least need $c(\cdot)$ to be twice differentiable to obtain the correlation/covariance function of the derivative process.

Morris, Mitchell and Ylvisaker: Bayesian design and analysis of computer experiments: use of derivatives in surface prediction (1993)

# Single-Path Inference

- Can we do inference form a single sample path to determine the $\Omega$-properties?

† Blum: Ergodic Theorems. Entry in Encyclopedia of Statistical Sciences, Vol 2, by S. Kotz and N.L. Johnson, eds., Wiley, NY pg. 541-545 (1982).

# Single-Path Inference

- Can we do inference form a single sample path to determine the Ω-properties?
- Requires that correlation between points decays to zero as the distance increases - *ergodic property*.

† Blum: Ergodic Theorems. Entry in Encyclopedia of Statistical Sciences, Vol 2, by S. Kotz and N.L. Johnson, eds., Wiley, NY pg. 541-545 (1982).

# Single-Path Inference

- Can we do inference form a single sample path to determine the $\Omega$-properties?
- Requires that correlation between points decays to zero as the distance increases - *ergodic property*.
  - if $Z(\cdot)$ is a stationary GP, then $Z(\cdot)$ is ergodic provided $C(h) \to 0$ as $||h||_2 \to \infty$.†

† Blum: Ergodic Theorems. Entry in Encyclopedia of Statistical Sciences, Vol 2, by S. Kotz and N.L. Johnson, eds., Wiley, NY pg. 541-545 (1982).

# Ergodicity

- Greek for "wandering"

† Cressie: *Statistics for Spatial Data* (1993)

# Ergodicity

- Greek for "wandering"
- Put simply, allows expectations over $\Omega$-space to be estimated by averages over the covariates (e.g. space, time) from a single realization.

† Cressie: *Statistics for Spatial Data* (1993)

# Ergodicity

- Greek for "wandering"
- Put simply, allows expectations over $\Omega$-space to be estimated by averages over the covariates (e.g. space, time) from a single realization.
- We'll describe a simple motivating case for the Ergodic Theorem†.

† Cressie: *Statistics for Spatial Data* (1993)

# Ergodicity

- Greek for "wandering"
- Put simply, allows expectations over $\Omega$-space to be estimated by averages over the covariates (e.g. space, time) from a single realization.
- We'll describe a simple motivating case for the Ergodic Theorem†.
- Let $T$ be a translation operator:

$$T(\{Z(t)\} : t = 1, 2, \ldots) = \{Y(t) : t = 1, 2, \ldots\}$$

where $Y(t) = Z(t + 1)$.

† Cressie: *Statistics for Spatial Data* (1993)

# Ergodicity

- Greek for "wandering"
- Put simply, allows expectations over $\Omega$-space to be estimated by averages over the covariates (e.g. space, time) from a single realization.
- We'll describe a simple motivating case for the Ergodic Theorem†.
- Let $T$ be a translation operator:

$$T(\{Z(t)\} : t = 1, 2, \ldots) = \{Y(t) : t = 1, 2, \ldots\}$$

  where $Y(t) = Z(t + 1)$.
- $T^k = T(T(\ldots))$

† Cressie: *Statistics for Spatial Data* (1993)

# Ergodicity

- Let $A$ is a measureable set (e.g. $A = \{z(\cdot) : z(1) \leq z_1, z(2) \leq z_2, \ldots, z(p) \leq z_p\}$) and define
$$T^{-1}(A) = \{b : T(b) \in A\}$$

# Ergodicity

- Let $A$ is a measureable set (e.g. $A = \{z(\cdot) : z(1) \leq z_1, z(2) \leq z_2, \ldots, z(p) \leq z_p\}$) and define
$$T^{-1}(A) = \{b : T(b) \in A\}$$
- Then $Z(\cdot)$ is stationary iff $P(T^{-1}(A)) = P(A)$ for every measurable set $A$. (e.g. just our usual translation-invariance in our GP models).

## Ergodicity

- Let $A$ is a measureable set (e.g. $A = \{z(\cdot) : z(1) \leq z_1, z(2) \leq z_2, \ldots, z(p) \leq z_p\}$) and define
$$T^{-1}(A) = \{b : T(b) \in A\}$$
- Then $Z(\cdot)$ is stationary iff $P(T^{-1}(A)) = P(A)$ for every measurable set $A$. (e.g. just our usual translation-invariance in our GP models).
- And $Z(\cdot)$ is *ergodic* if $T^{-1}(A) = A$ (i.e. $A$ is invariant with respect to time) implies $P(A) = 0$ or $P(A) = 1$.

# Ergodicity

- Let $A$ is a measureable set (e.g. $A = \{z(\cdot) : z(1) \leq z_1, z(2) \leq z_2, \ldots, z(p) \leq z_p\}$) and define
$$T^{-1}(A) = \{b : T(b) \in A\}$$
- Then $Z(\cdot)$ is stationary iff $P(T^{-1}(A)) = P(A)$ for every measurable set $A$. (e.g. just our usual translation-invariance in our GP models).
- And $Z(\cdot)$ is *ergodic* if $T^{-1}(A) = A$ (i.e. $A$ is invariant with respect to time) implies $P(A) = 0$ or $P(A) = 1$.
- This means than the trajectories generated by this process (almost) all belong to a single invariant set.

# Ergodicity

- Let $A$ is a measureable set (e.g. $A = \{z(\cdot) : z(1) \leq z_1, z(2) \leq z_2, \ldots, z(p) \leq z_p\}$) and define
$$T^{-1}(A) = \{b : T(b) \in A\}$$
- Then $Z(\cdot)$ is stationary iff $P(T^{-1}(A)) = P(A)$ for every measurable set $A$. (e.g. just our usual translation-invariance in our GP models).
- And $Z(\cdot)$ is *ergodic* if $T^{-1}(A) = A$ (i.e. $A$ is invariant with respect to time) implies $P(A) = 0$ or $P(A) = 1$.
- This means than the trajectories generated by this process (almost) all belong to a single invariant set.
- For our time-indexed process analogy here, this definition says that almost every realization of the time-series $\{Z(t), t = 1, 2, \ldots\}$ when successively translated completely fills the space of possible trajectories.

# Ergodicity

- Let $A$ is a measureable set (e.g. $A = \{z(\cdot) : z(1) \leq z_1, z(2) \leq z_2, \ldots, z(p) \leq z_p\}$) and define
$$T^{-1}(A) = \{b : T(b) \in A\}$$
- Then $Z(\cdot)$ is stationary iff $P(T^{-1}(A)) = P(A)$ for every measurable set $A$. (e.g. just our usual translation-invariance in our GP models).
- And $Z(\cdot)$ is *ergodic* if $T^{-1}(A) = A$ (i.e. $A$ is invariant with respect to time) implies $P(A) = 0$ or $P(A) = 1$.
- This means than the trajectories generated by this process (almost) all belong to a single invariant set.
- For our time-indexed process analogy here, this definition says that almost every realization of the time-series $\{Z(t), t = 1, 2, \ldots\}$ when successively translated completely fills the space of possible trajectories.
- i.e. the future of this ergodic process holds within it any type fo behaviour allowable under its probability measure $P$.

# Ergodicity

**Ergodic Theorem**†: Let $\{Z(t), t = 1, 2, \ldots\}$ be an ergodic time-series and suppose that the measureable function $g(\cdot)$ is integrable ($\int gdP$ exists, e.g. if $g = \mathcal{I}_A$ then $\int gdP = P(A)$). Then for almost every realization $\{z(t), t = 1, 2, \ldots\}$,

$$lim_{n\to\infty}\frac{1}{n}\sum_{j=0}^{n-1} g(T^j(\{z(t) : t = 1, 2, \ldots\})) = \int gdP.$$

† Birkhoff: *Proof of the Ergodic Theorem* (1931).

# Single-Path Inference

- Can we do inference form a single sample path to determine the $\Omega$-properties?

† Blum: *Ergodic Theorems.* Entry in Encyclopedia of Statistical Sciences, Vol 2, by S. Kotz and N.L. Johnson, eds., Wiley, NY pg. 541-545 (1982).

# Single-Path Inference

- Can we do inference form a single sample path to determine the $\Omega$-properties?
- Requires that correlation between points decays to zero as the distance increases - *ergodic property*.

† Blum: *Ergodic Theorems.* Entry in Encyclopedia of Statistical Sciences, Vol 2, by S. Kotz and N.L. Johnson, eds., Wiley, NY pg. 541-545 (1982).

# Single-Path Inference

- Can we do inference form a single sample path to determine the $\Omega$-properties?
- Requires that correlation between points decays to zero as the distance increases - *ergodic property*.
  - if $Z(\cdot)$ is a stationary GP, then $Z(\cdot)$ is ergodic provided $C(h) \to 0$ as $||h||_2 \to \infty$.†

† Blum: *Ergodic Theorems.* Entry in Encyclopedia of Statistical Sciences, Vol 2, by S. Kotz and N.L. Johnson, eds., Wiley, NY pg. 541-545 (1982).

# Single-Path Inference

- Can we do inference form a single sample path to determine the $\Omega$-properties?
- Requires that correlation between points decays to zero as the distance increases - *ergodic property*.
  - if $Z(\cdot)$ is a stationary GP, then $Z(\cdot)$ is ergodic provided $C(h) \to 0$ as $||h||_2 \to \infty$.†
- In practice: can we make inference of a GP observed over a fixed domain, e.g. $[0, 1]$? This is known as *infill asympototics*.

† Blum: *Ergodic Theorems.* Entry in Encyclopedia of Statistical Sciences, Vol 2, by S. Kotz and N.L. Johnson, eds., Wiley, NY pg. 541-545 (1982).

# Identifiability

- First, a small aside.

# Identifiability

- First, a small aside.
- The log-likelihood function of a GP for the observed vector $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} log(2\pi) - \frac{1}{2} log|\Sigma| - \frac{1}{2}\mathbf{y}^T \Sigma^{-1}\mathbf{y}.$$

# Identifiability

- First, a small aside.
- The log-likelihood function of a GP for the observed vector $\mathbf{y} = (y_1, \ldots, y_n)$ is

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2} log(2\pi) - \frac{1}{2} log|\Sigma| - \frac{1}{2}\mathbf{y}^T \Sigma^{-1}\mathbf{y}.$$

- Then the parameters $\boldsymbol{\theta}$ are identifiable if $\ell(\boldsymbol{\theta}_1) = \ell(\boldsymbol{\theta}_2)$ iff $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

## Single-Path Inference

- It turns out these results are very challenging to come by.

Ying: *Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process* (1991)

# Single-Path Inference

- It turns out these results are very challenging to come by.
- † For a 1-D GP with mean zero and exponential covariance $\sigma^2 exp(-\theta|h|)$ observed over the domain $[0, 1]$, then $(\theta, \sigma^2)$ and $\theta', \sigma^{2'})$ are **not** distinguishable with certainty from the sample path as long as $\theta\sigma^2 = \theta'\sigma^{2'}$.

Ying: *Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process* (1991)

# Single-Path Inference

- It turns out these results are very challenging to come by.
- † For a 1-D GP with mean zero and exponential covariance $\sigma^2 exp(-\theta|h|)$ observed over the domain $[0, 1]$, then $(\theta, \sigma^2)$ and $\theta', \sigma^{2'})$ are **not** distinguishable with certainty from the sample path as long as $\theta\sigma^2 = \theta'\sigma^{2'}$.
- So, only $\theta\sigma^2$ is identifiable, not the individual parameters.

Ying: *Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process* (1991)

# Single-Path Inference

- It turns out these results are very challenging to come by.
- † For a 1-D GP with mean zero and exponential covariance $\sigma^2 exp(-\theta|h|)$ observed over the domain $[0, 1]$, then $(\theta, \sigma^2)$ and $\theta', {\sigma^2}')$ are **not** distinguishable with certainty from the sample path as long as $\theta\sigma^2 = \theta' {\sigma^2}'$.
- So, only $\theta\sigma^2$ is identifiable, not the individual parameters.
- For dimension $\geq 2$, with a separable covariance, this is not the case (they are identifiable).

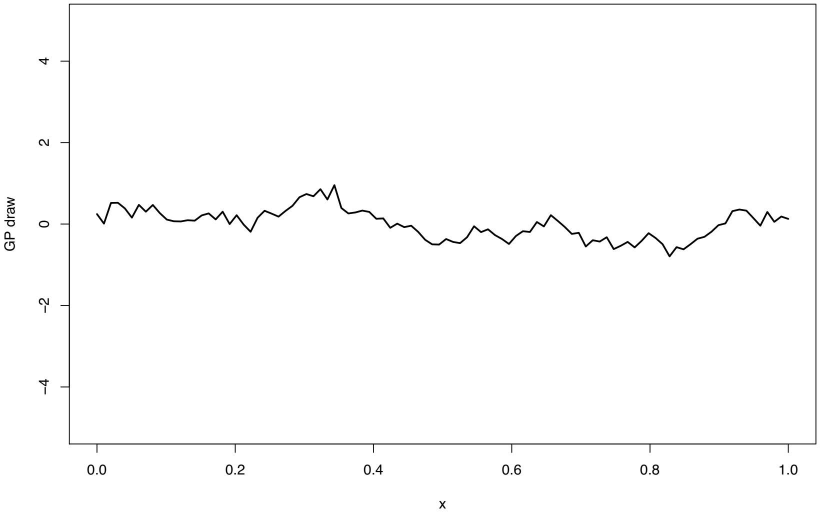
Ying: *Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process* (1991)

# Example

```
source("dace.sim.r")
n=100
design=matrix(seq(0,1,length=n),ncol=1)
l1=list(m1=abs(outer(design[,1],design[,1],"-")))
l.dez=list(l1=l1)
rho=0.2
alpha=1

s2s=10
rs=1/s2s
seed=sample(1:1e5,1)
set.seed(seed)
s2=1
rho=0.2  # -log(.2)*s2=1.6ish
se2=0
z0=sim.field(l.dez,rho,s2,se2=se2,alpha=alpha)
z0=z0-mean(z0)
```

# Example

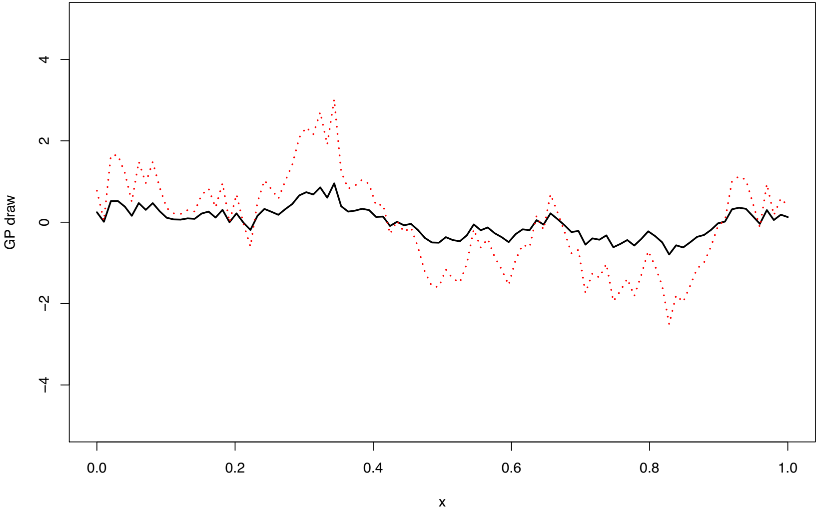
```
plot(design,z0,type='l',xlab="x",ylab="GP draw",
     lwd=2,ylim=c(-5,5))
```



# Example

```
set.seed(seed)
s2=1*s2s
rho=0.2
z1=sim.field(l.dez,rho,s2,se2=se2,alpha=alpha)
z1=z1-mean(z1)
```

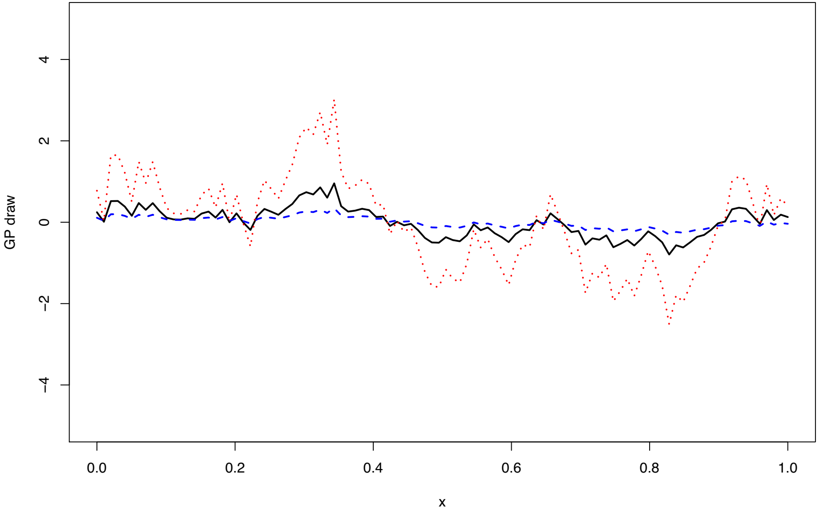
# Example



# Example

```
set.seed(seed)
s2=1
rho=0.2^rs
z2=sim.field(l.dez,rho,s2,se2=se2,alpha=alpha)
z2=z2-mean(z2)
```

# Example



# Example

```
set.seed(seed)
s2=1*s2s
rho=0.2^rs # -log(rho)*s2=1.6ish
z3=sim.field(l.dez,rho,s2,se2=se2,alpha=alpha)
z3=z3-mean(z3)
```

# Example