Applied Spatial Statistics: Lecture 2B Identifying a covariance function

Douglas Nychka,
National Center for Atmospheric Research

Supported by the National Science Foundation   *Boulder, Spring 2013*

Outline

- The additive model
- Spatial data $\rightarrow$ covariance function
  using a variogram
- Finding variograms in R

Estimating a curve or surface.

The additive statistical model:

Given n pairs of observations (x_i, y_i), $i = 1, \ldots, n$

$$y_i = g(x_i) + \epsilon_i$$

ϵ_i's are random errors – usually $\boldsymbol{e} \sim (0, \sigma^2 I)$
and g is an unknown smooth function.

The goal is to estimate a function g based on the observations

Develop a statistical model for g that suggests an estimate and also its uncertainty

Estimating the covariance function

The variogram is a clever way to tease out the covariance function from the data by looking at differences of observations.

Isotropic $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \rho\Phi(d)$ Also to make things easier $\Phi(0) = 1$ (Φ is a correlation function).
i.e. The covariance only depends on the distance of separation between points. (Isotropic)

The variogram function:
From the beginning of the Lecture.

$$E[\frac{(U_1 - U_2)^2}{2}] = (1/2)VAR(U_1) + (1/2)VAR(U_2) - 2COV(U_1, U2)$$

for two spatial locations

$$E[\frac{(g(\boldsymbol{x}_1) - g(\boldsymbol{x}_2))^2}{2} = \rho - \rho\Phi(d)$$

where $d = distance(\boldsymbol{x}_1, \boldsymbol{x}_2)$

The variogram

Sample variogram

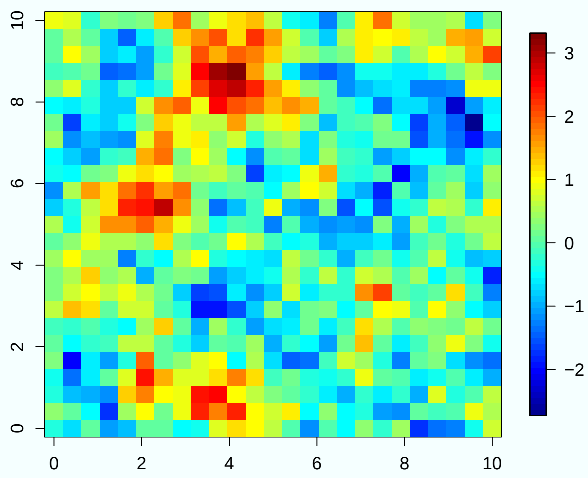
• Plot for a spatial data set or spatial field plot $\frac{(\boldsymbol{y}_i - \boldsymbol{y}_j)^2}{2}$ against the distance of separation. "On the average" this should be follow the theoretical curve $\rho - \rho\Phi(d)$.

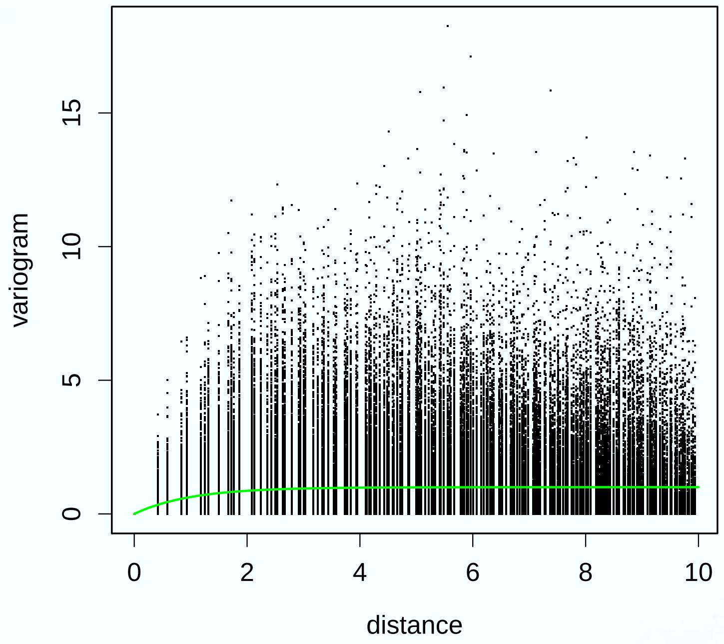
• If one can find $\rho - \rho\Phi(d)$ then one can transform back to the covariance function. (Subtract it from ρ !)

• Need to be careful about how the variogram behaves close to zero distance. The variogram estimates $\rho + \sigma^2$ right at zero – not just ρ

A variogram example

Sample field (standard exponential covariance) on a 25×25 grid.

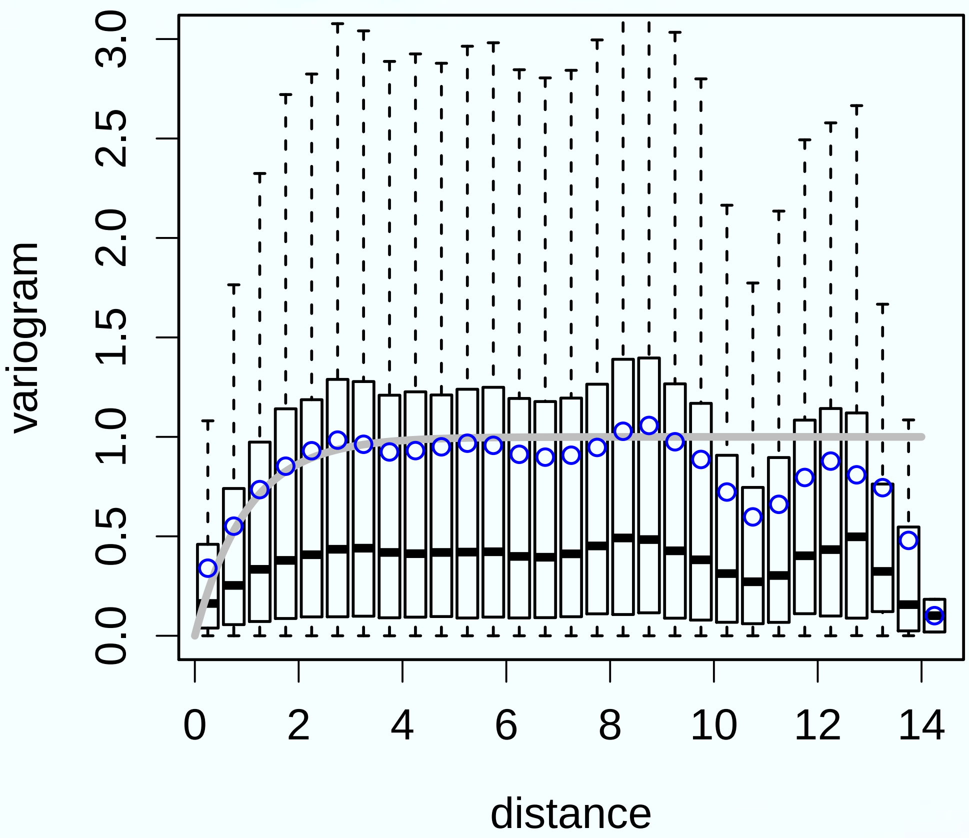




Houston, we have a problem.

Improving the variogram

Binning observations by distance ranges and adding the mean for each bin.



Computing in R

The advantage of the `vgram` function is that it finds the statistics over over the distance classes. But see the R code for this lecture for the basic computation of the variogram using two `for` loops. From the lecture 2 R script `x.grid` grid of x locations and `Z` the simulated field at those locations.

```
# use 30 equally spaced binds for the distances
 look<- vgram( x.grid, Z, N=30)
#plotting the variogram
 bplot.xy( look$d, look$vgram, breaks= look$breaks,
          outline=FALSE, ylim=c(0,4),
          xlab="distance", ylab="variogram")
 points( look$centers, look$stats[2,], col="blue")
# adding true covariance
 dgrid<- seq(0,10,,400)
 lines(dgrid, 1 - exp(-1*dgrid), col="green", lwd=2)
```

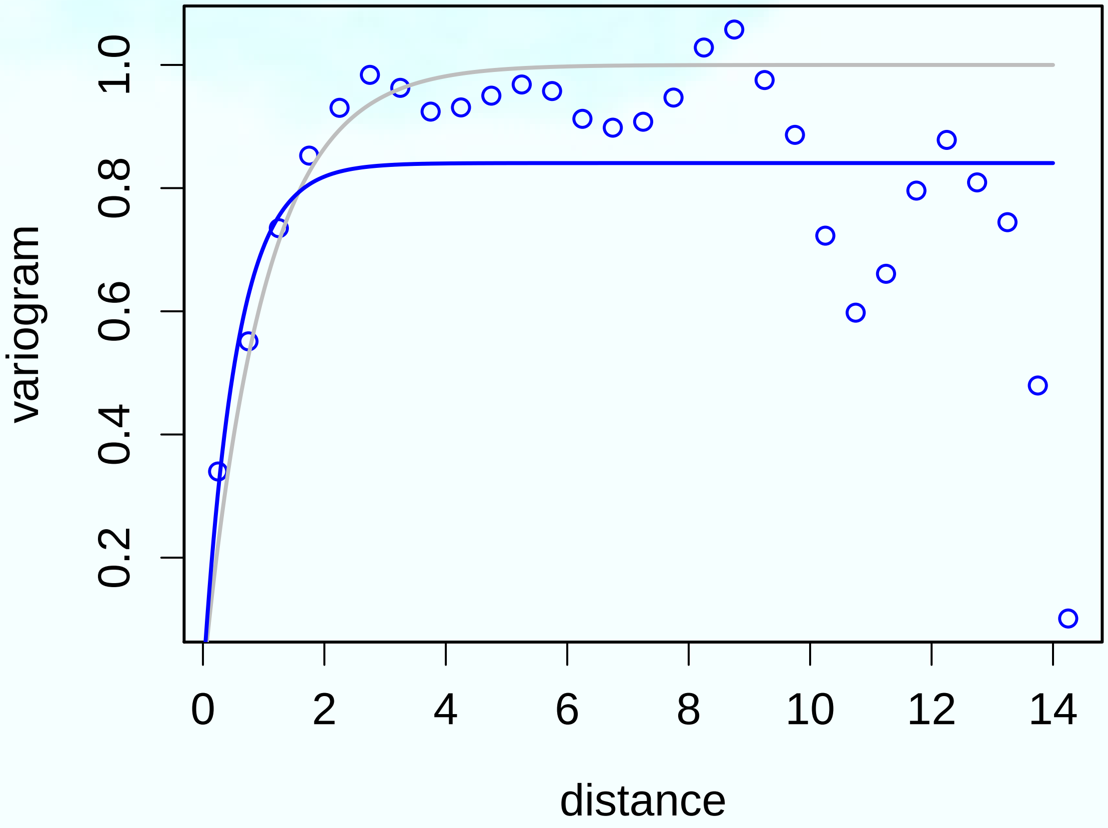
The fact that these are on a regular grid is not exploited by the `vgram` function see `vgram.matrix` to handle larger grids.

Estimating the variogram

Simplest way is to nonlinear least squares to fit a function to the bin means

```
 xd<- look$centers
 ym<- look$stats[2,]
 nls.fit<- nls( ym ~  rho*( 1- exp(-xd/theta)),
               start= list(  rho=1.5, theta=1.5),
               control=list( maxiter=200) )
#adding curve to plot
 pars<- coefficients(nls.fit)
 rho<-pars[1]
 theta<-pars[2]
 dgrid<- seq(0,14,,400)
 lines(dgrid, sigma^2 + rho*(1 - exp(-1*dgrid/theta)),
                                          col="blue", lwd=3)
```

Fitted variogram

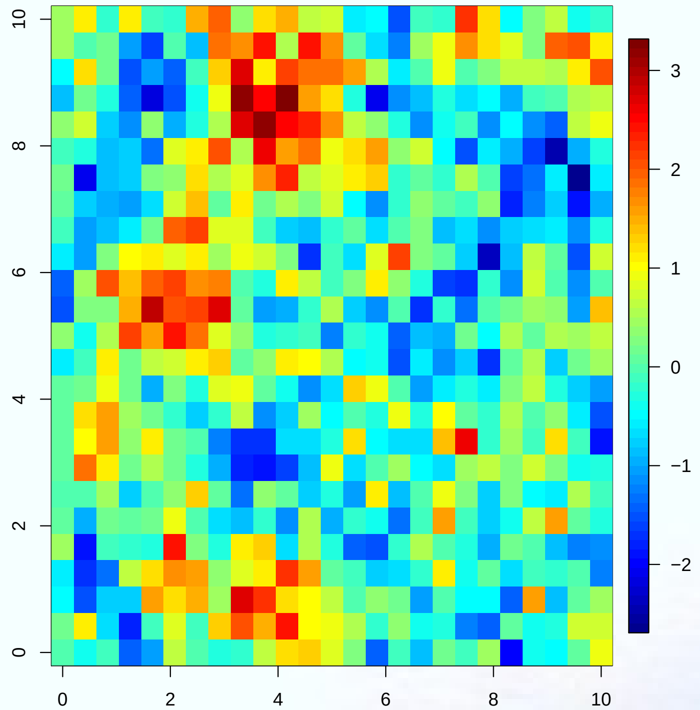


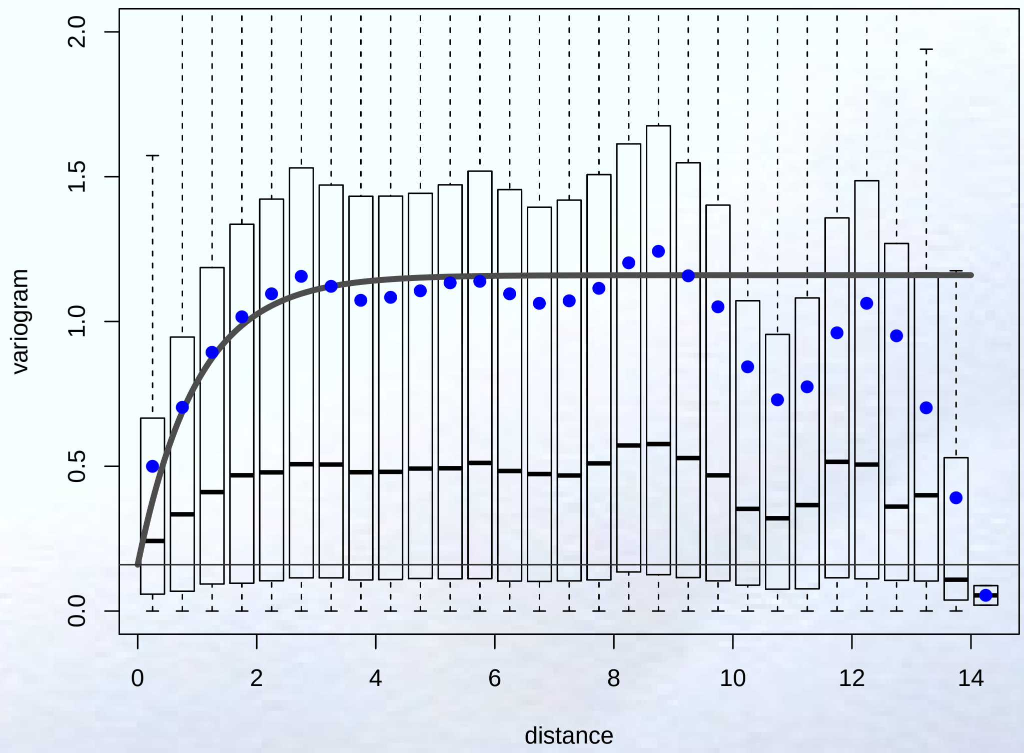
Identifying the nugget σ^2

Recall the additive model:

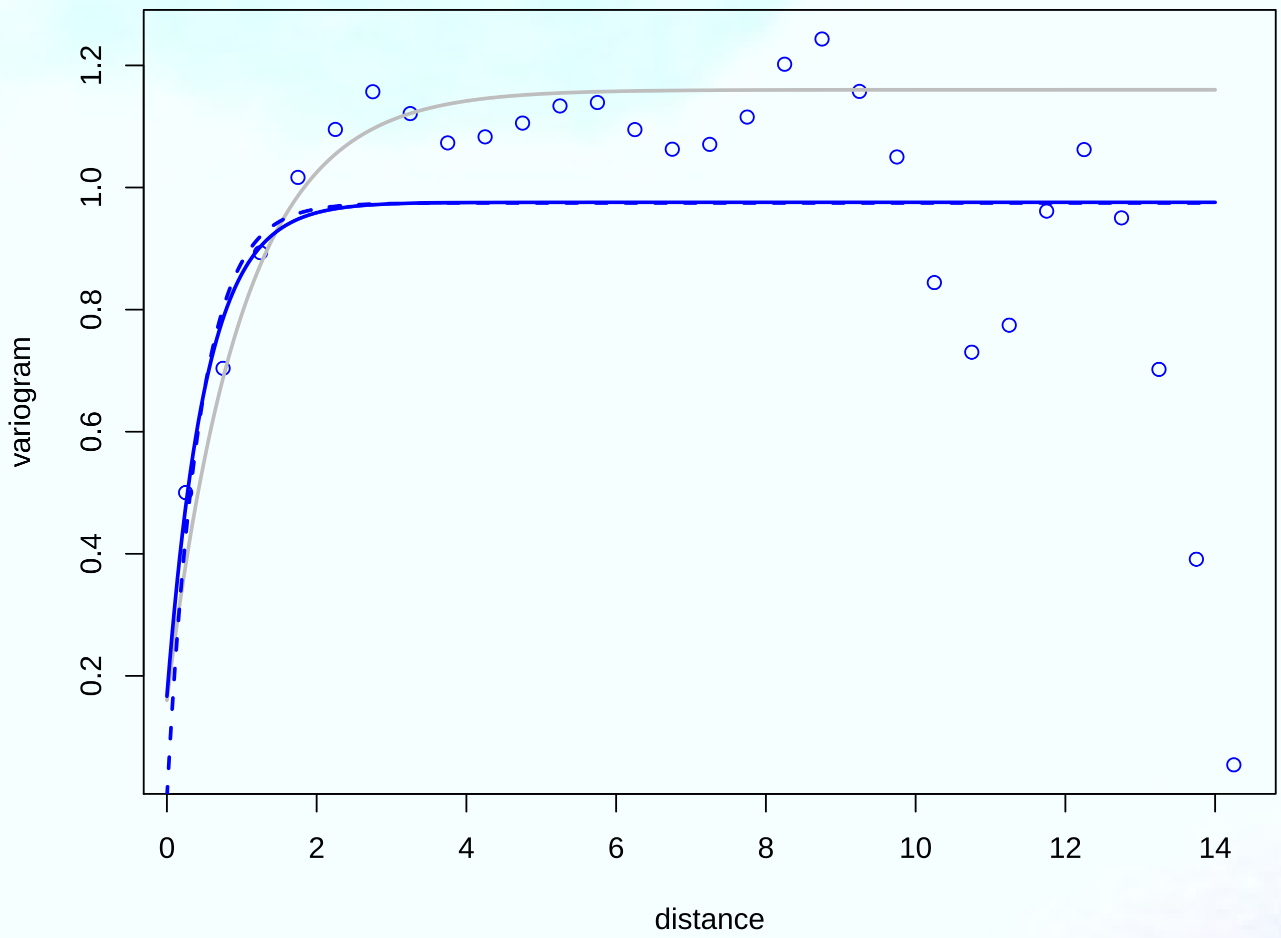
$$y_i = g(x_i) + \epsilon_i$$

Correlations among the observations due to the smooth field but the measurement error is uncorrelated. Adding measurement error to the example ($\sigma = .4, \sigma^2 = .16$)





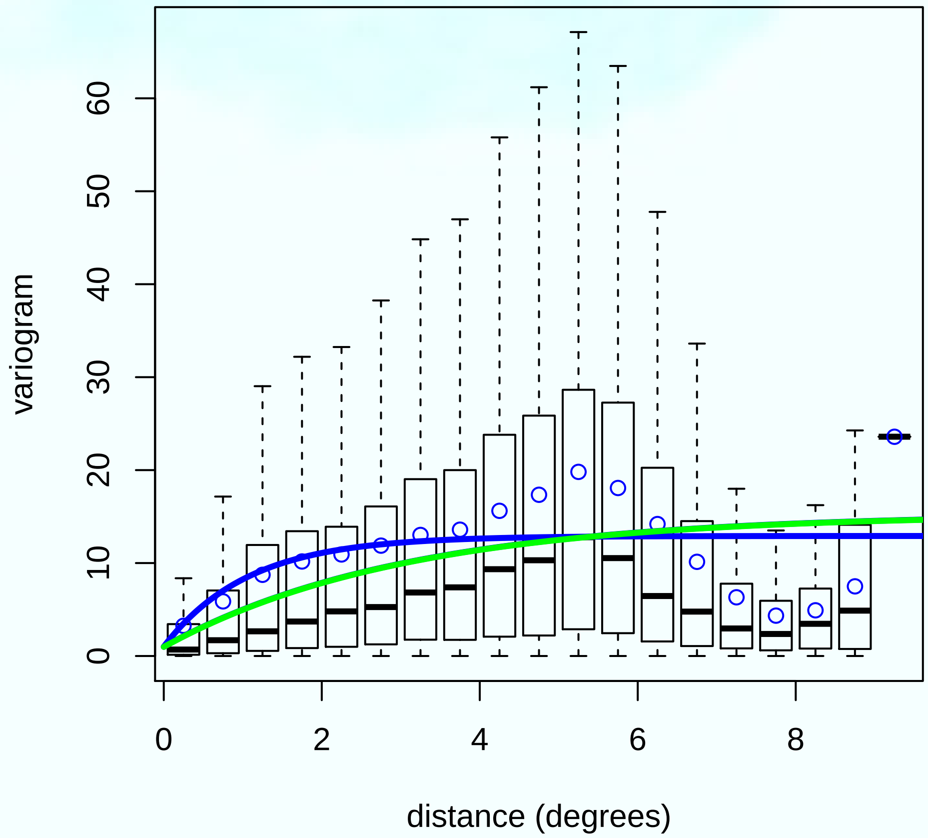
Fitting variogram with a nugget



Nonlinear regression fitting with a nugget:

```
nls.fit<- nls( ym ~   sig2 + rho*( 1- exp(-xd/theta)),
               start= list(  rho=1.5, theta=1.5, sig2= .1),
               control=list( maxiter=200) )
```

Back to the CO climate data



To fit the curve by maximum likelihood:

```
MLE.fit<- MLE.Matern( xHW1, yHW1, smoothness=.5, m=1)
MLE.fit$pars
      rho       theta       sigma
14.2873743   3.0516346   0.9893552
```

Where we are going

Through the variogram (and possibly other analyses) we can identify a covariance model for the random field g and the measurement error variance $sigma$.

A linear estimator

$$\hat{g}(\boldsymbol{x}) = \sum w_i \boldsymbol{y}_i$$

Expected squared error for this estimator has a closed from if we know the covariance function.

Kriging

Find the set of weights that minimized the expected squared error.