1 Gaussian Processes

Definition 1.1 A Gaussian process $\{x_i\}$ over sites i is defined by its mean function

$$E(x_i) = \mu_i$$

and its covariance function

$$c_{ij} = \mathrm{Cov}(x_i, x_j)$$

plus joint normality of the finite dimensional distributions.

Hence x restricted to the points labelled by $1, \ldots, n$ is $(x_1, \ldots, x_n)^T$ and it has a n-variate Gaussian distribution $N(\mu, \Sigma)$, where $\mu = (\mu_1, \ldots, \mu_n)^T$ and $\Sigma = (c_{ij})$. Note that the covariance function c_{ij} must be positive definite (ie. any covariance matrix created from a finite dimensional set of x_i's must be positive definite: $a^T \Sigma a > 0$, for any non-zero vector a).

Covariance functions The restriction that the function $\{c_{ij}\}$ be positive definite can make the search for valid covariance functions difficult. Most covariance functions model covariance between sites i and j as a function of distance between the two sites $d_{ij} = \mathrm{dist}(i, j)$, where $\mathrm{dist}(i, j)$ is typically Euclidean distance, or a simple modification of it. Hence $c_{ij} = C(d_{ij})$. It is standard to choose from a number of parameterized covariance functions, often called covariograms, listed below:

- Power family
  $$C(d|\theta, p) = \theta_1 \exp\{-|d/\theta_2|^p\},\ 0 < p \leq 2$$
  Two notable covariograms in this family are the *exponential* ($p = 1$) and the *Gaussian* ($p = 2$).
- Spherical
  $$C(d|\theta) = \begin{cases} \theta_1 \left[1 - \frac{2}{\pi}\left(\frac{d}{\theta_2}\sqrt{1 - \frac{d}{\theta_2}} + \sin^{-1}\frac{d}{\theta_2}\right)\right] & \text{for } d < \theta_2 \\ 0 & \text{for } d \geq \theta_2 \end{cases}$$
  For the spherical covariogram, if i and j are separated by a distance greater than θ_2, x_i and x_j are independent.
- Matérn
  $$C(d|\theta) = \theta_1 \frac{1}{2^{\theta_3 - 1}\Gamma(\theta_3)} \left(\frac{2\sqrt{\theta_3}d}{\theta_2}\right)^{\theta_3} \mathcal{K}_{\theta_3}\left(\frac{2\sqrt{\theta_3}d}{\theta_2}\right)$$
  where θ_2 is a scale parameter and θ_3 is a shape parameter, and $\mathcal{K}()_{\theta_\ni}$ is a modified Bessel function of the third kind of order θ_3 (Abramowitz and Stegun 1964, Chapter 9).

Why positive definite? Consider the power covariogram for large p. This makes the covariogram look like a step function $C(d) = I[0 \le d \le 1]$. So if sites 1,2,3 lie on a line with spacing $\frac{1}{2}$, then $\text{Cov}(x_1, x_2) = \text{Cov}(x_2, x_3) = 1 \Rightarrow x_1 = x_2 = x_3$, but $C(d = 1)$ requires that $\text{Cov}(x_1, x_3) = 0$, which is a contradiction. Such a difficulty can occur for any $p > 2$.

Example: Gaussian random walk Let $x = (x_0, x_1, x_2, x_3, ...)^T$ be a Gaussian process defined on the integers $\{0, 1, 2, 3, ...\}$ such that

$$x_0 \equiv 0, \quad x_i | x_{i-1} \sim N(x_{i-1}, 1), \quad \text{for } i = 1, 2, 3, \ldots$$

$x = (x_0, x_1, x_2, x_3, x_4)^T$ has density:

$$\begin{aligned}
\pi(x) &= (2\pi)^{-\frac{4}{2}} \exp\left\{-\tfrac{1}{2}\sum_{i=1}^{4}(x_i - x_{i-1})^2\right\}, \; x_0 = 0 \\
&= (2\pi)^{-\frac{4}{2}} \exp\left\{-\tfrac{1}{2} \begin{pmatrix} x_0 & x_1 & x_2 & x_3 & x_4 \end{pmatrix} \begin{pmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix}\right\}, \; x_0 = 0 \\
&= (2\pi)^{-\frac{4}{2}} \exp\left\{-\tfrac{1}{2} x^T W x\right\}, \; x_0 = 0
\end{aligned}$$

Alternatively, we can compute the means, variances and covariances of x using standard covariance formulas since $x_i = \sum_{j=1}^{i} z_j$ where the z_i's are iid $N(0, 1)$ random variables. Thus:

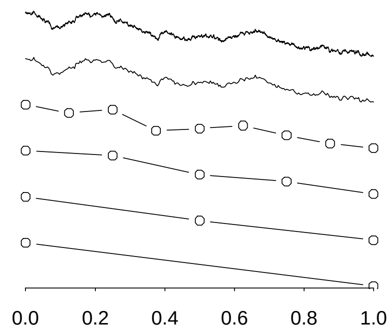
$$\begin{pmatrix} x_0 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 2 & 2 \\ 0 & 1 & 2 & 3 & 3 \\ 0 & 1 & 2 & 3 & 4 \end{pmatrix}\right) \sim N(\mathbf{0}, \Sigma).$$

Note that this distribution is degenerate since $x_0 = 0$. So the rank of Σ is only 4. One can check that W is a generalized-inverse of Σ.

The covariance function of this process is:

$$\text{Cov}(x_i, x_j) = \min(i, j) = i \wedge j$$

One can imagine defining a random walk on a grid with twice the density using increments with half the variance. The resulting covariogram on the integers would be exactly the same. The figure below shows a random walk on a successively finer grid.



The limiting process $\{x_t\}$ is called *Brownian motion*. It is defined over $t \geq 0$, it's continuous, and it is characterized by its mean function $Ex_t = 0$ and its covariance function $\text{Cov}(x_t, x_s) = t \wedge s$.

Note that the covariance function cannot be written as a function of distance here. In fact, $\text{Var}x_t = t$, so the process $\{x_t\}$ isn't even stationary. A process $\{x_t\}$ is said to be stationary if the distribution of a finite dimensional restriction of $\{x_t\}$ is unchanged if the process is translated over space (ie. $(x_{t_1}, \ldots, x_{t_n}) \stackrel{d}{=} (x_{t_1+s}, \ldots, x_{t_n+s})$ for any s in the case of Brownian motion). The random walk is an example of an *intrinsicly stationary* process – because the increments are all iid. Such processes can be characterized by their variograms, which give the variance of all pairwise differences. When the variance of any pair (x_i, x_j) depends only on the distance d_{ij} between sites i and j we can characterize the process $\{x\}$ with a variogram denoted by $2\gamma(d_{ij})$. For historical reasons $\gamma(d)$ is called the semi-variogram, and the variogram is defined to be $2\gamma(d)$. For the random walk

$$\text{Var}(x_i - x_j) = 2\gamma(d_{ij}) = |i - j|$$

In the example here, knowing the variogram doesn't completely specify the distribution of $\{x\}$. For example, the random walk process where $x_0 = 0$ has exactly the same variogram as does the process with $x_0 = 1$. However, as long as one of the x_i's (or more) are known, then the distribution for the remaining components of $\{x\}$ is completely specified.$\square$

The Variogram

Definition 1.2 *A Gaussian process $\{x_i\}$, is said to have a* variogram *if $Var(x_i - x_j)$ is a function of distance d_{ij} between sites i and j. We use $2\gamma(d_{ij})$ to denote the variogram.*

$$2\gamma(d_{ij}) = Var(x_i - x_j)$$

If a process $\{x_i\}$ has a covariogram, then the two functions are related by

$$\begin{aligned} \gamma(d) &= C(0) - C(d) \\ C(d) &= \gamma(\infty) - \gamma(h) \end{aligned}$$

If $\{x_i\}$ has a variogram, but the covariogram doesn't exist, we can compute a covariance function by conditioning on the event $x_0 = 0$.

$$\begin{aligned}
2\gamma(d_{ij}) = \text{Var}(x_i - x_j) &= \text{Var}(x_i - x_j | x_0 = 0) \\
&= \text{Var}((x_i - x_0) - (x_j - x_0) | x_0 = 0) \\
&= \text{Var}(x_i - x_0) + \text{Var}(x_j - x_0)) - 2\text{Cov}(x_i - x_0, x_j - x_0 | x_0 = 0) \\
&= 2\gamma(d_{i0}) + 2\gamma(d_{j0}) - 2\text{Cov}(x_i, x_j | x_0 = 0)
\end{aligned}$$

Thus one gets

$$\text{Cov}(x_i, x_j | x_0 = 0) = \gamma(d_{i0}) + \gamma(d_{j0}) - \gamma(d_{ij})$$

This is one way of obtaining Σ in the random walk example. In fact using $\Sigma_{ij} = c - \gamma(d_{ij})$ will give a valid covariance matrix provided c is sufficiently large.

One example of a Gaussian process which has a variogram, but not a covariogram is a generalized Brownian motion process:

- Power variogram:

  $$\gamma(d) \propto d^p, \ 0 < p \leq 2$$

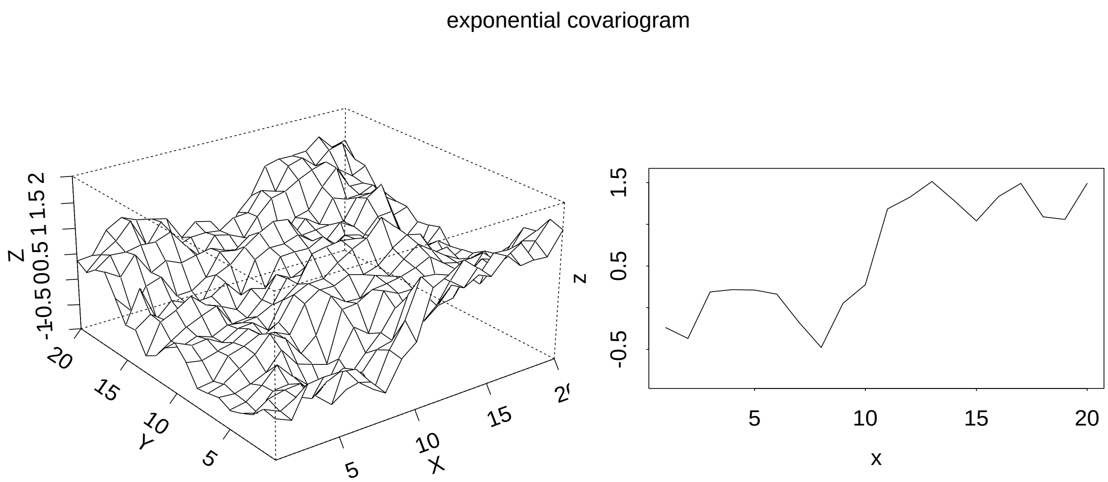
  For the case $p = 1$, this is general Brownian motion.

- 2-d Thin plate spline

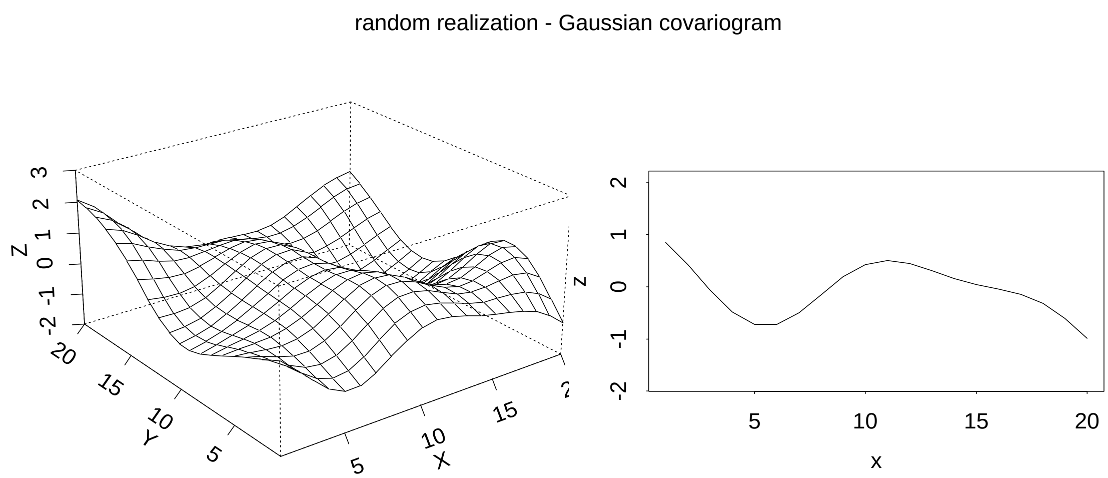
  $$\gamma(d) \propto d^2 \log d$$

Sample realizations:

$$C(d) \propto \exp\{-d/\sigma\}, \quad \text{here } \sigma = 25$$

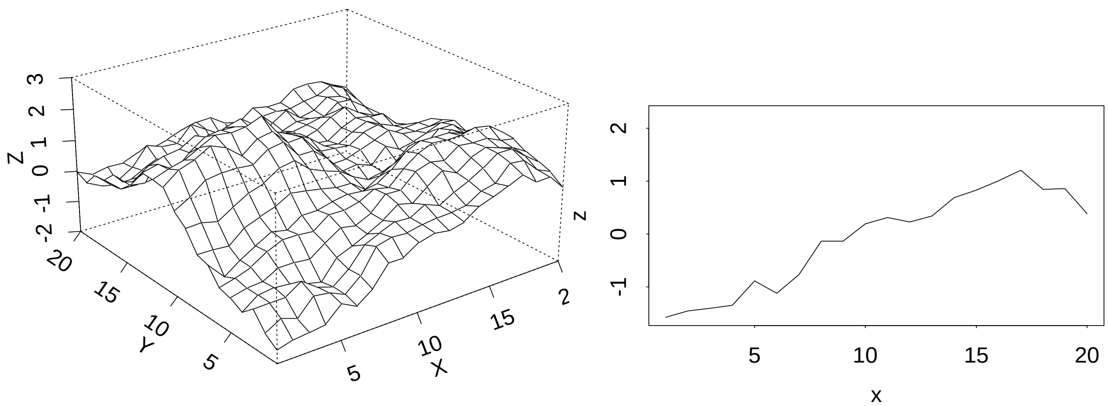


$$C(d) \propto \exp\{-(d/\sigma)^2\}, \quad \text{here } \sigma = 6$$



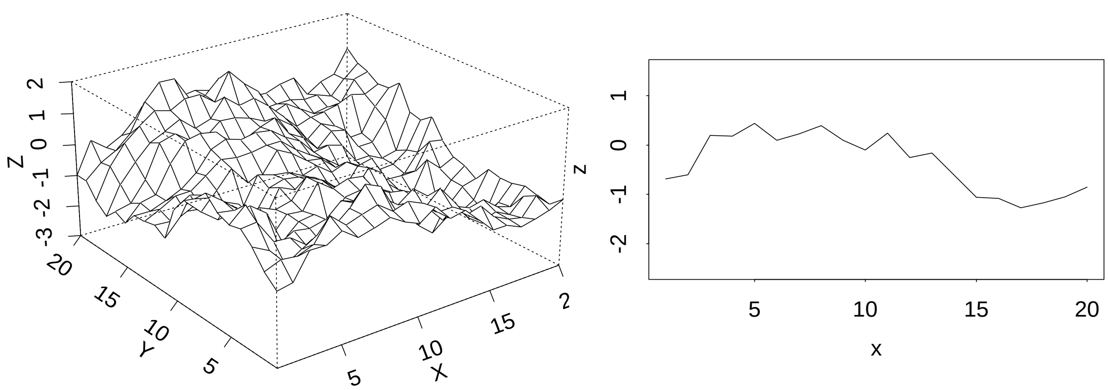
$$C(d) \propto \exp\{-(d/\sigma)^{1.5}\}, \quad \text{here } \sigma = 9$$

random realization - C(d) = exp(-d^1.5)



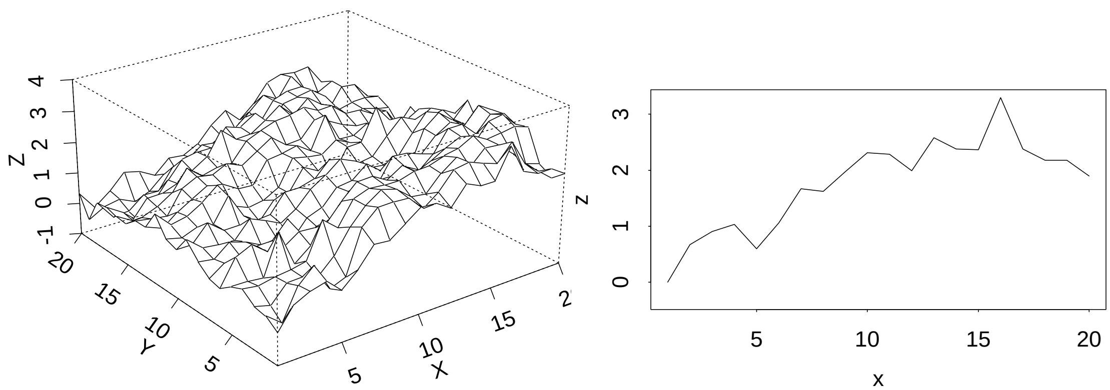
$$C(d) \propto 1 - \frac{2}{\pi}\left(\frac{d}{\sigma}\sqrt{1-\frac{d}{\sigma}} + \sin^{-1}\frac{d}{\sigma}\right) \text{ here } \sigma = 9$$

random realization - spherical



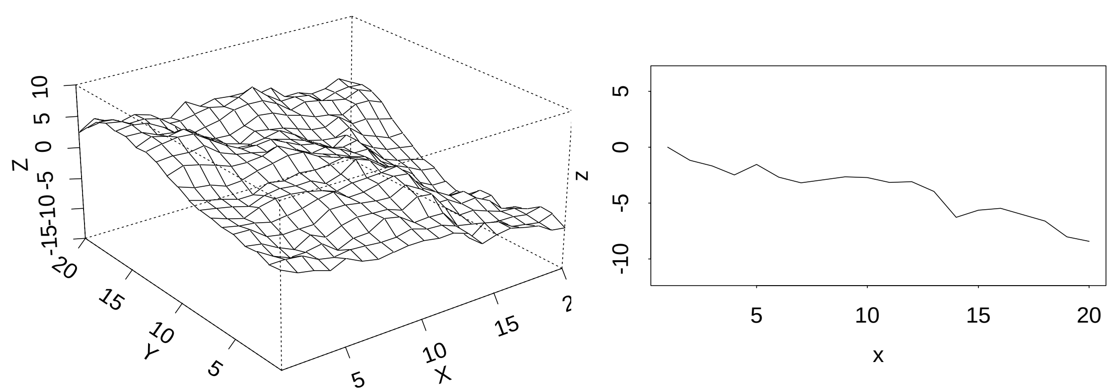
$$\gamma(d) \propto d/\sigma, \quad \text{here } \sigma = 5$$

random realization - Brownian motion p=1



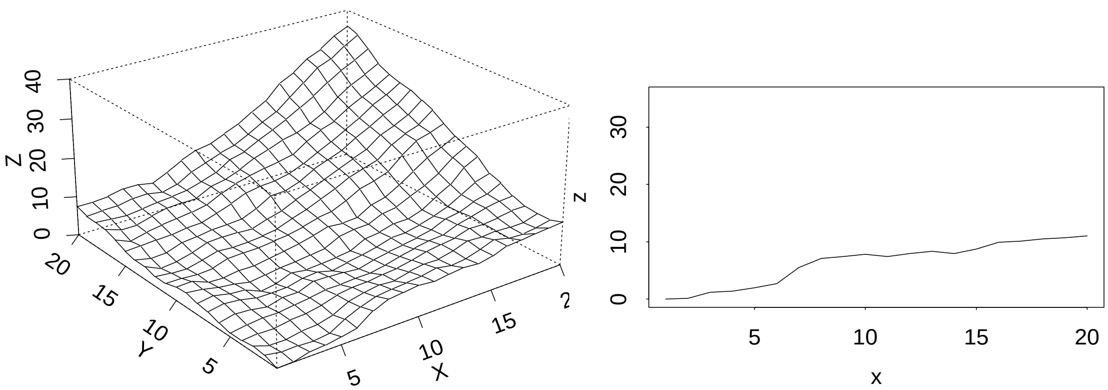
$$\gamma(d) \propto (d/\sigma)^{1.3}, \quad \text{here } \sigma = 5$$

random realization - Brownian motion p=1.3



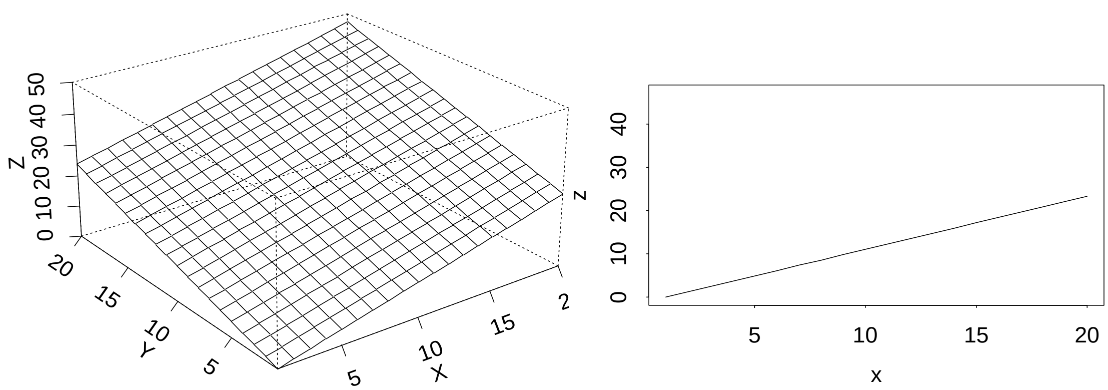
$$\gamma(d) \propto (d/\sigma)^{1.7}, \quad \text{here } \sigma = 5$$

random realization - Brownian motion p=1.7



$$\gamma(d) \propto (d/\sigma)^{2.0}, \quad \text{here } \sigma = 5$$

random realization - Brownian motion p=2.0



2 Interpolation

Given a set of points $\{1, 2, \ldots, n\}$ on a line or in a plane and their values $x_1, x_2, \ldots, x_n$ one can fit an interpolating surface just by choosing a covariogram or variogram model.

1. Assume the surface is a realization from a Gaussian field with a specified covariance function.

2. Compute the conditional mean of the process given the observed sites. This is the interpolating surface.

Here's the recipe:

- Suppose the process $\{x_1, \ldots, x_n\}$ is observed. Now suppose you wish to find the interpolating surface $\{x_{n+1}, \ldots, x_{n+m}\}$ at the sites $\{n+1, \ldots, n+m\}$.
- Assume

$$x = (x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m})^T \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$$

  where the elements of Σ are determined by the covariogram or variogram model.

- The conditional distribution of $(x_{n+1}, \ldots, x_{n+m})$ given $x_1, \ldots, x_n$ is:

$$(x_2|x_1) = ((x_{n+1}, \ldots, x_{n+m})|(x_1, \ldots, x_n)^T) \sim N\left(\mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\right)$$

  Note: if we re-express things in terms of precisions, then for

$$x = (x_1, \ldots, x_n, x_{n+1}, \ldots, x_{n+m})^T \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix}^{-1}\right)$$

  we have

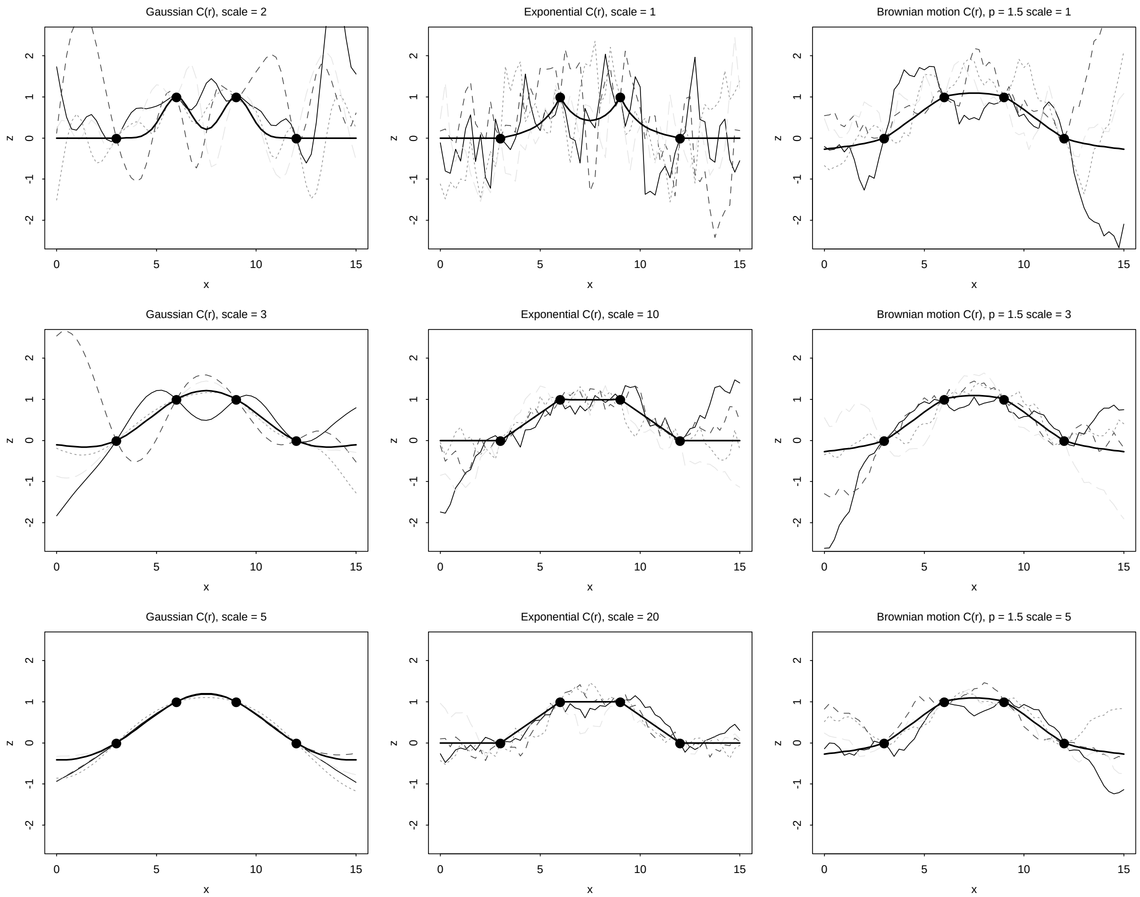
$$(x_2|x_1) = ((x_{n+1}, \ldots, x_{n+m})|(x_1, \ldots, x_n)^T) \sim N\left(\mu_2 + W_{22}^{-1}W_{21}(x_1 - \mu_1), W_{22}^{-1}\right).$$

- Define the interpolant to be the mean of x_2 given the observations x_1:

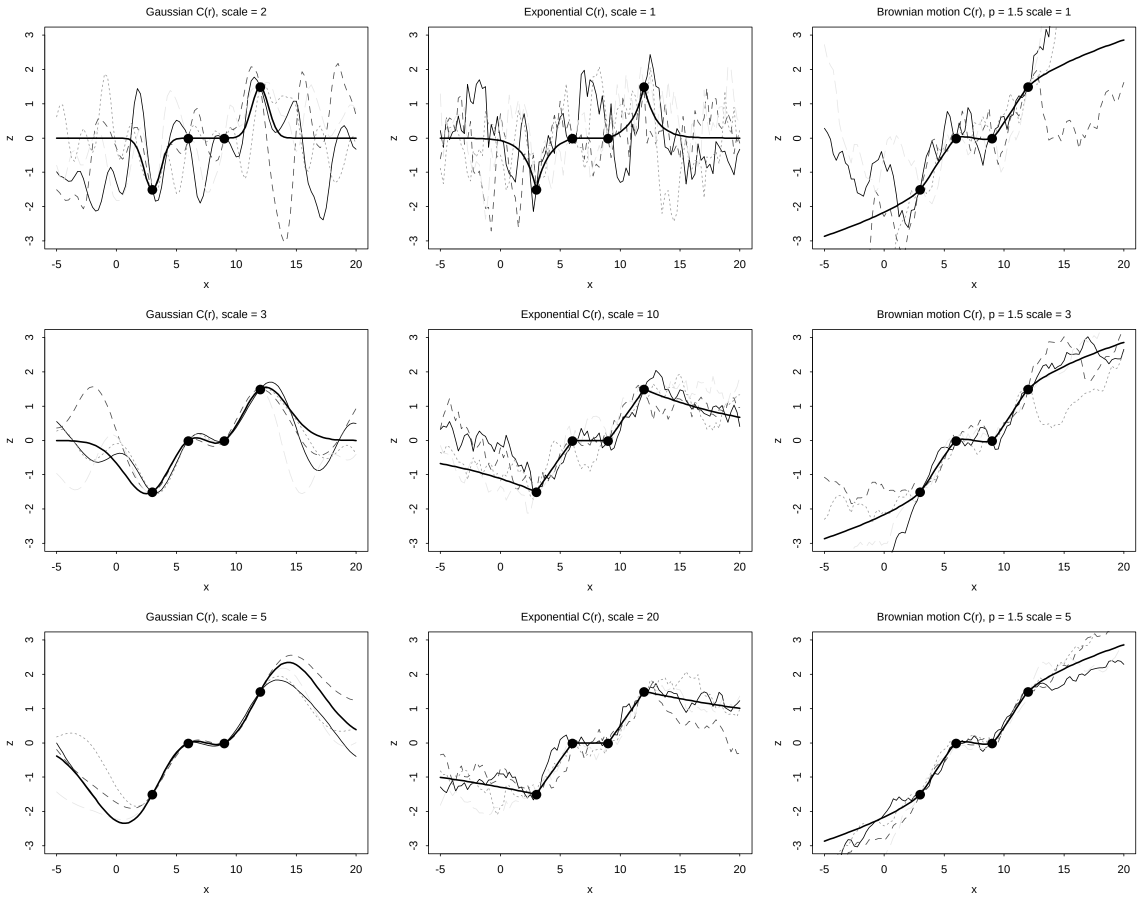
$$\mu_{22\cdot 1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1)$$

- If $\{x_i\}$ really did follow a Guassian process with the specified mean and covariance function, then the standard error of the enterpolant at x_j would be the jj component of $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$. We'll look at estimating parameters of the covariogram/variogram later. The conditional mean is the "optimal" interpolator under mean squared error loss.

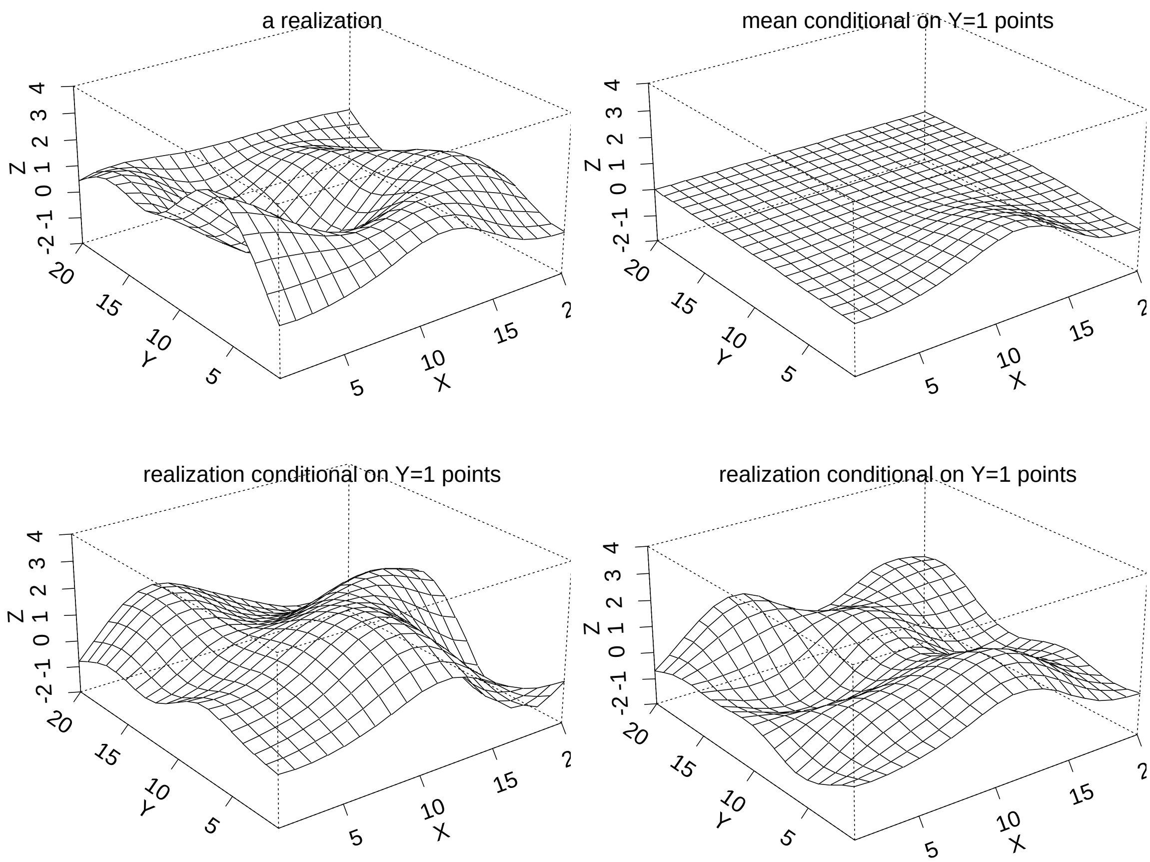
Some examples:

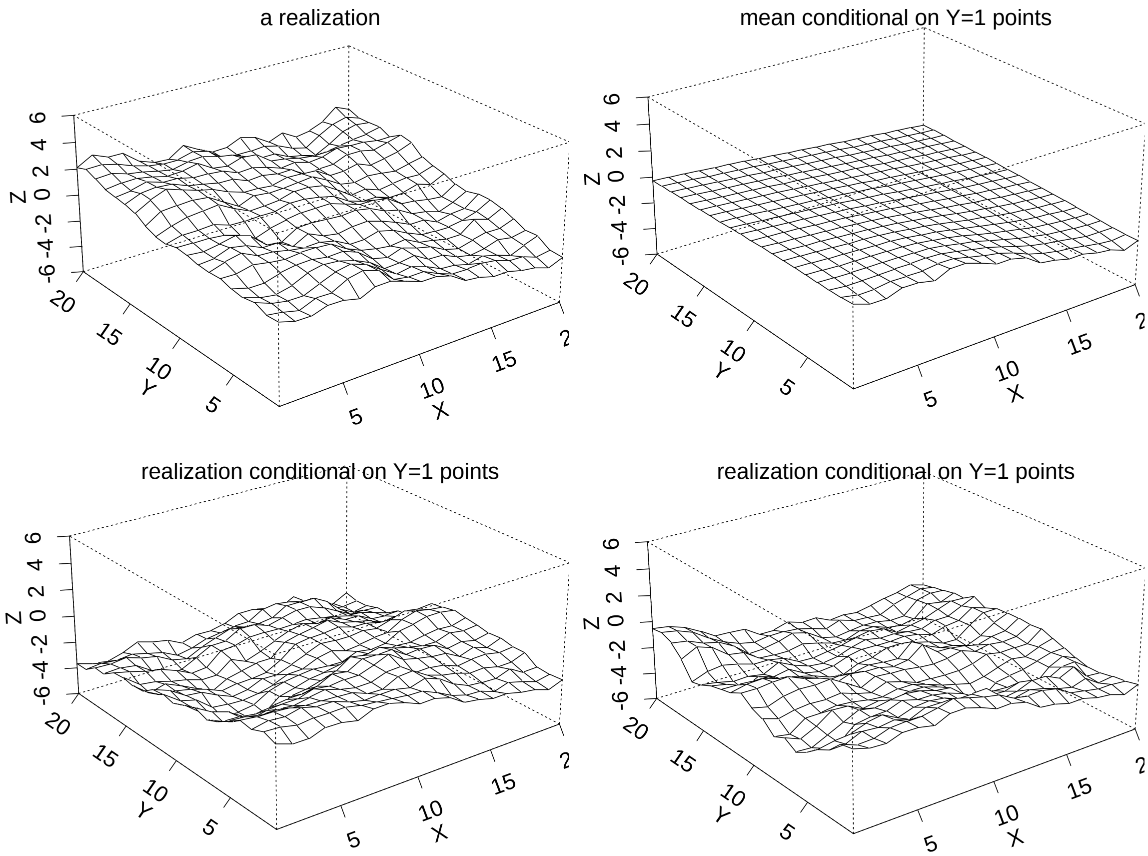


Gaussian C(r), scale = 2
Exponential C(r), scale = 1
Brownian motion C(r), p = 1.5 scale = 1
Gaussian C(r), scale = 3
Exponential C(r), scale = 10
Brownian motion C(r), p = 1.5 scale = 3
Gaussian C(r), scale = 5
Exponential C(r), scale = 20
Brownian motion C(r), p = 1.5 scale = 5
x
z



Gaussian C(r), scale = 2
Exponential C(r), scale = 1
Brownian motion C(r), p = 1.5 scale = 1
Gaussian C(r), scale = 3
Exponential C(r), scale = 10
Brownian motion C(r), p = 1.5 scale = 3
Gaussian C(r), scale = 5
Exponential C(r), scale = 20
Brownian motion C(r), p = 1.5 scale = 5
x
z

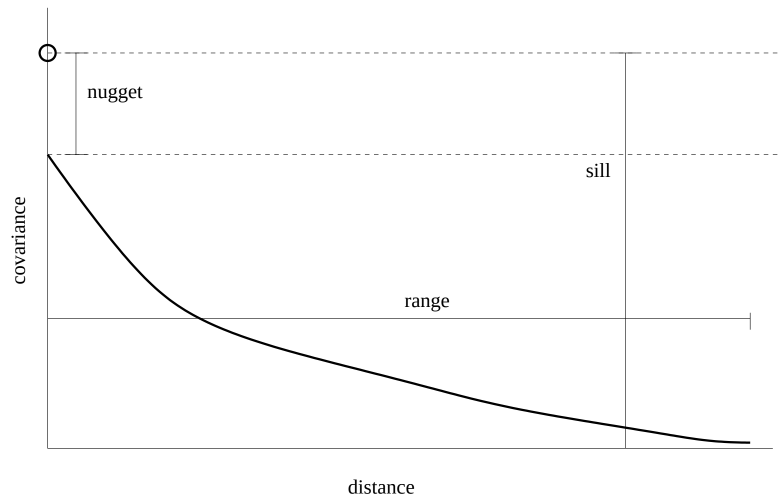

a realization
mean conditional on Y=1 points
realization conditional on Y=1 points
realization conditional on Y=1 points
X
Y
Z



3 A closer look at the covariogram

There are a number of features of a variogram/covariogram that are worth noting. It is also of use to understand how properties of the covariogram affect the resulting spatial process.

- **range**
- **scale**
- **nugget**



Examples:

- White noise process
- Effects of the "nugget" on the conditional distribution
- Effects of the range on the conditional distribution
- Effects of $\gamma'(0+)$ on the conditional distribution

4 Estimating the variogram

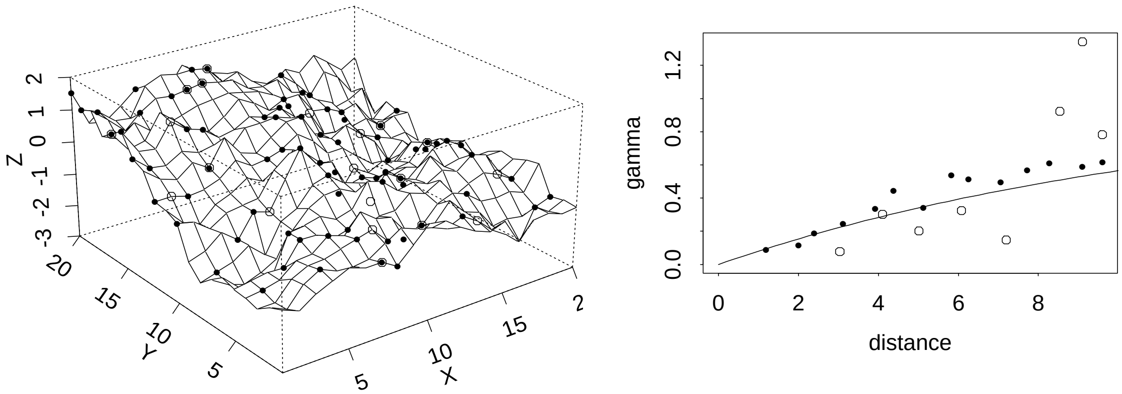
Given an observed set of points $x_1, \ldots, x_n$ at locations $1, \ldots, n$ one may wish to estimate various properties of the variogram governing their covariance. This can be done graphically using the empirical variogram.

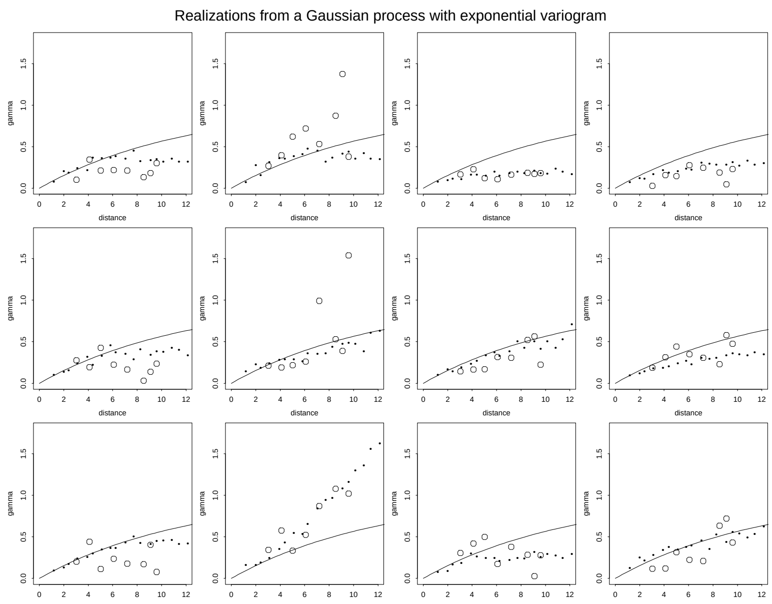
Definition 4.1 *The empirical variogram is determined by discretizing distance into a n_d bins and then estimating*

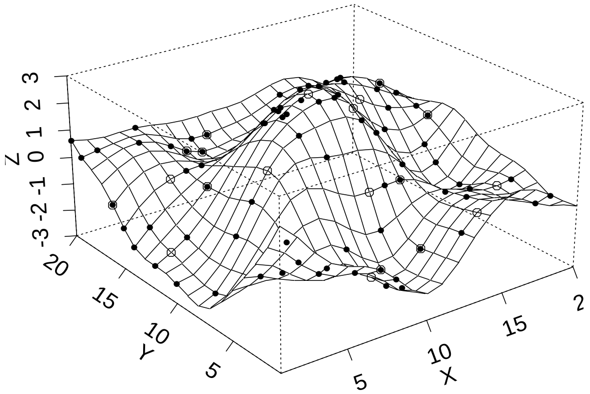
$$\hat{\gamma}(d) = \frac{1}{2N_d} \sum_{(i,j) \in d_\Delta} (x_i - x_j)^2$$

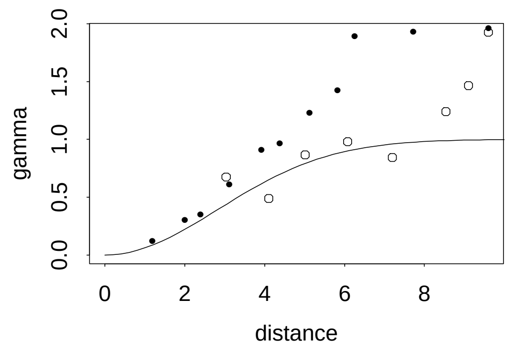
where d_Δ is the set of all pairs (i, j) such that the distance between i and j is within Δ of d and N_d is the number of pairs in d_Δ.

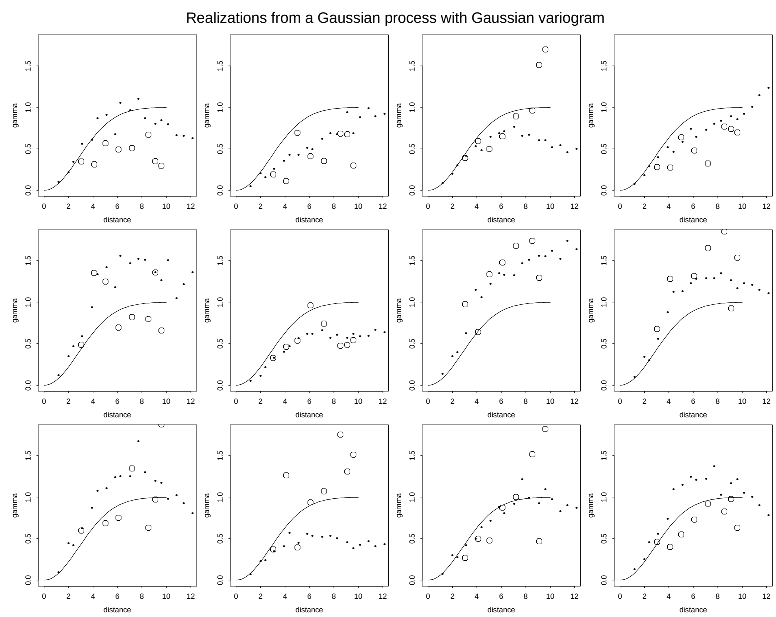
Note this estimate can be somewhat unreliable. What follows is a number of surfaces and their empirical variograms derrived from $x_1, \ldots, x_n$, which were sampled uniformly from the 400 points making up a 20×20 lattice.

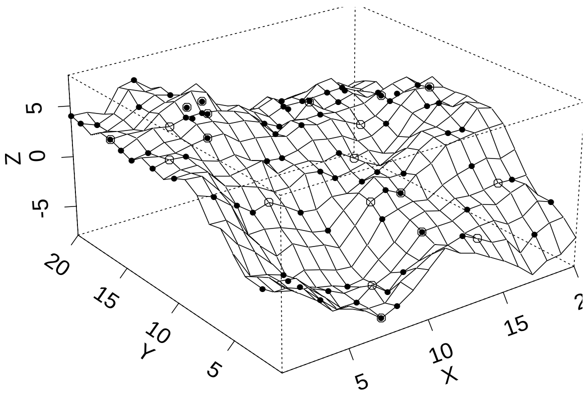

Z
X
Y
gamma
distance

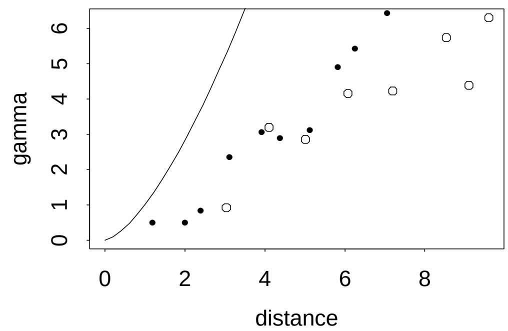

Realizations from a Gaussian process with exponential variogram
gamma
distance


Z
X
Y

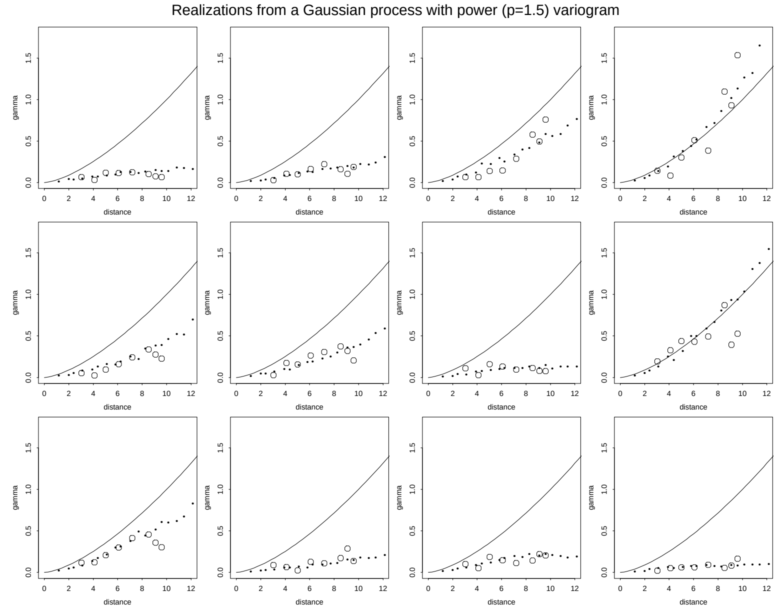

gamma
distance


Realizations from a Gaussian process with Gaussian variogram
gamma
distance





Realizations from a Gaussian process with power (p=1.5) variogram



- What features of $\gamma(d)$ are important to capture
- Fitting a variogram model to the empirical variogram

  1. Least squares: Choose θ so that

     $$\sum_k (\gamma(d_k|\theta) - \hat{\gamma}(d_k))^2$$

     is minimized.

  2. Weighted least squares: Choose θ so that

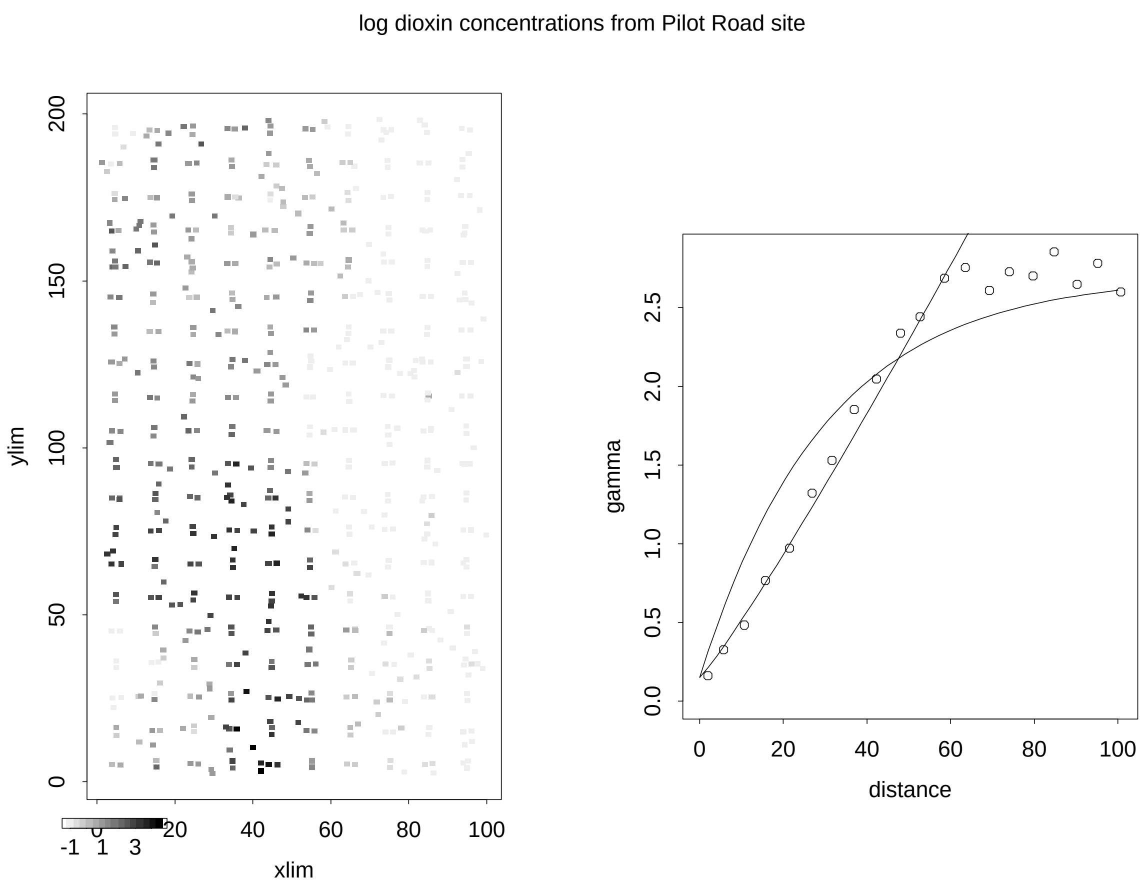
     $$\sum_k w_k(\gamma(d_k|\theta) - \hat{\gamma}(d_k))^2$$

     is minimized. Weights w_k may be chosen to be proportional to the number of pairs in each bin N_d. This will put more weight on pairs that are closer together.

3. Maximum likelihood: Suppose
$$y \sim N(\mu, \Sigma(\theta)),$$
then we can estimate μ and θ by the values $\hat{\mu}$ and $\hat{\theta}$ that maximize the likelihood:
$$L(\mu, \theta|y) \propto |\Sigma(\theta)|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(y-\mu)^T \Sigma(\theta)(y-\mu)\right\}$$
A variant of maximum likelihood is *restricted maximum likelihood* (REML) which uses a slightly modified version of the likelihood.
4. Bayesian estimation. Specify prior distributions for μ and θ and use the posterior mean or posterior mode to estimate θ.

- Anisotropy $d' = \Lambda R d$. Transform euclidean distance thru a rotation $R(\theta)$ and a stretching/shrinking of the principle axes via Λ. These additional parameters may be absorbed in θ for estimating the variogram parameters.

Example: Piazza Road Superfund site.



One can estimate the variogram parameters by eye, using ML or REML, or better yet, a Bayesian approach. Use the interpolation formulas of Section 2 to estimate the concentration at unobserved sites.

5 Modeling spatial data

$$y = X\beta + z + e$$

where $X\beta$ absorbs standard linear model terms, z absorbs the spatial trend, and e is a white noise term.

examples

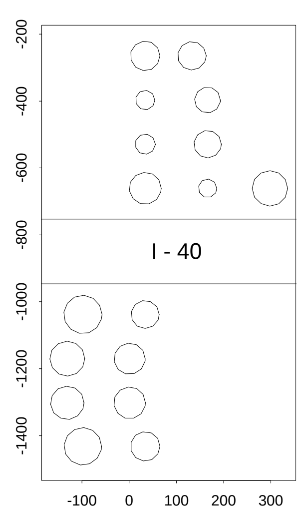
- Agricultural field trials.
- rainfall estimation
- environmental monitoring
- imaging

Observed data: $y = (y_1, \ldots, y_n)^T$
Unobserved spatial trend: $z = (z_1, \ldots, z_n)^T$
Covariates: X

Some equivalent formulations:

$$\begin{array}{rcl} y & \sim & N(X\beta + z, \sigma_y^2 I) \\ z & \sim & N(0, \Sigma(\theta)) \end{array} \qquad \begin{array}{rcl} y & = & X\beta + z + e \\ z & \sim & N(0, \Sigma(\theta)) \\ e & \sim & N(0, \sigma_e^2 I) \end{array} \qquad \begin{array}{rcl} y & \sim & N(X\beta, \sigma_y^2 I + \Sigma(\theta)) \end{array}$$

An example: Researchers want to know how carbon concentrations in a stream differ on two different sides of a culvert on I-40. Concentrations are measured at 9 upstream locations and 8 downstream locations and are given in the figure below. Is there evidence that the culvert is associated with differences in carbon concentration?



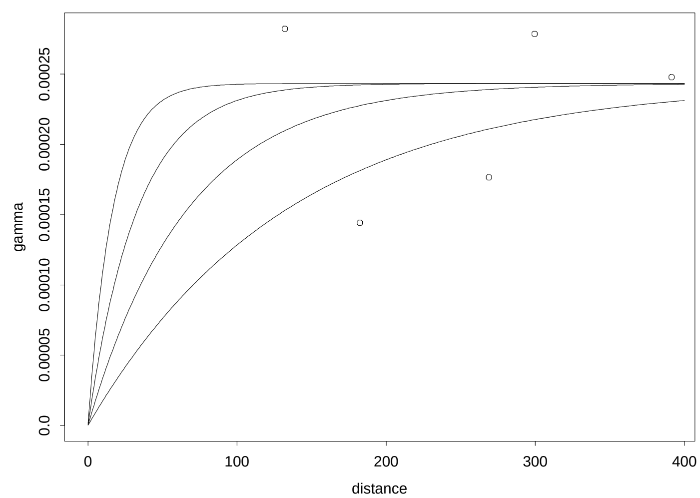
Here we can model the 17 measurements by

$$y \sim N(\mu + \alpha_i, \sigma_e^2 I + \Sigma(\theta))$$

where $\alpha_i,\ i = 1, 2$, denotes the upstream or downstream measurements. So the question can be put in statistical terms, does $\alpha_1 = \alpha_2$?

It turns out it depends on what covariance function is specified. If an exponential covriance function is used, then the inference depends on the specified range:

```
significance by range:
dist                 0    50    100    200    400
corr of nearest      0    .0     .1     .2     .5
p-value            .01   .01    .02    .06    .30
diff*10^2          .99   .99    .98    .92    .75
```



One could rely on maximum likelihood to specify the parameters of the exponential distribution. In fact it fits a covariance that is nearly all nugget effect (ie. no spatial dependence). However the uncertainty about the covariance parameters are unaccounted for.