What’s for today

- How to estimate/fit parametric covariance (or variogram) models
There are several ways of estimating parametric models for the empirical variograms. In this course, we will discuss two ways:

1. Least squares method
2. Maximum likelihood estimation

We will first discuss the method of maximum likelihood estimation.
Maximum likelihood estimation (MLE)

This is a general overview for those who do not remember the details of the method!

- Suppose a random variable $X$ has a density (pdf) $f(x, \theta)$ such that $P(X \leq x) = \int_{-\infty}^{x} f(y, \theta) dy$

- If we observe $n$ data points, $X_1, \cdots, X_n$, a random sample of size $n$ from the distribution of $X$

- we consider the likelihood $L(\theta) = \prod_{i=1}^{n} f(x_i, \theta)$ and $\hat{\theta}$ that maximizes

  $L(\theta)$ is the maximum likelihood estimator of $\theta$

- An example for $X \sim N(\mu, 1)$ with $x_1, \ldots, x_n$
MLE for spatial data

- If we observe \( Z = \{Z(s_1), \cdots, Z(s_n)\}^T \) and if we assume \( Z(s) \sim N(\mu 1, \Sigma(\theta)) \)

- we have likelihood:

\[
L(\theta) = \frac{1}{(2\pi)^{n/2} \det(\Sigma(\theta))^{1/2}} \exp\{-(Z - \mu 1)^T \Sigma^{-1}(\theta)(Z - \mu 1)\}
\]

- In modeling \( \Sigma(\theta) \), we may use one of the covariance models (e.g. Matérn, exponential, spherical, etc)
Suppose I assume \( x \sim \mathcal{N}(\mu, \Sigma(\theta)) \) for \( x \) in data 1 and I model \( \Sigma(\theta) \) using the following covariance model: 
\[
K(x) = \alpha \exp(-x/\beta) + \epsilon 1_{(x=0)}
\]
Recall we have 500 observations in data 1
Our loglikelihood is,
\[
\log l(\theta) \propto -\ln(\det(\Sigma(\theta))) - (x - \mu 1)^T \Sigma^{-1}(\theta)(x - \mu 1)
\]
where
\[
x = (x_1, \cdots, x_{500})^T
\]
and
\[
\Sigma(\theta)_{i,j} = \alpha \exp(-d_{i,j}/\beta) + \epsilon 1_{i=j}
\]
\((d_{i,j} \text{ is the distance for } x_i \text{ and } x_j)\)
Therefore, loglikelihood is a function of \( \mu, \alpha, \beta \) and \( \epsilon \)
Practical issues

- In practice, we may maximize $L(\theta)$ by numerical optimization
  - You may maximize analytically
  - However, if you use Matérn covariance function, getting derivatives with respect to the smoothness parameter $\nu$ is tricky
- In R, there are two functions ($nlm$ and $optim$) for numerical optimization
- Note that in numerical optimization, we should transform the parameters in the loglikelihood functions!
If I use `nlm` in R, I get something like this:

```
iteration = 24
Parameter:
[1] 1.1832631 8.2902927 -3.5134316 0.2625171
Function Value
[1] 78.78118
Gradient:
Relative gradient close to zero.
Current iterate is probably solution.
```

Note that `nlm` minimizes the given function.
Now we turn our attention to the Least Squares method
This is a quick overview of LS method

Suppose we have two vectors of data, $X = (X_1, \cdots, X_n)^T$ and $Y = (Y_1, \cdots, Y_n)^T$

Now we think about the regression problem (linear regression) of $Y$ on $X$

Our model is $Y = aX + b + e$, where $a$ and $b$ are parameters to be estimated and we assume $e \sim N(0, \sigma^2 I)$

Then we find $a$ and $b$ that minimize the sum of squares

$$\sum_{i=1}^{n}(Y_i - a - b \cdot X_i)^2$$

This is called the Least Squares (LS) method
Estimating variogram using LS

- Suppose $2\gamma(h; \theta)$ is the true variogram and we have empirical variogram values at $n$ distance lag: $\hat{\gamma}(h_i), i = 1, \cdots, n$
- We consider a model $\hat{\gamma}(h) = \gamma(h; \theta) + e(h)$ and we assume $e$ is mean zero.
- The tricky part is the covariance structure of $e$
  - Since we have $n$ “observed” variograms at distance lags $h_i$, $i = 1, \cdots, n$, we need to consider a covariance matrix of $e$ with dimension $n \times n$.
  - Are $e(h_i)$ and $e(h_j)$ for $i \neq j$ independent?
  - How can we calculate $\text{Cov}\{e(h_i), e(h_j)\}$?
- If we denote the covariance matrix of $e$ by $R(\theta)$, we need to find $\theta$ that minimizes

$$\{\hat{\gamma}(h) - \gamma(h; \theta)\}^T R(\theta)^{-1} \{\hat{\gamma}(h) - \gamma(h; \theta)\}$$
Simplest thing to do: OLS

- OLS stands for ordinary least squares
- Here we pretend that $e(h_i)$’s are independent and have a constant variance

- We let $R(\theta) = \phi I$
- Then the problem reduces to finding $\theta$ that minimizes

$$\{\hat{\gamma}(h) - \gamma(h; \theta)\}^T \{\hat{\gamma}(h) - \gamma(h; \theta)\}$$
Next simplest thing to do: WLS

- If we do not like the independent and constant variance assumption, we can do a little more sophisticated thing.
- Here we still pretend that $e(h_i)$’s are independent but we replace the diagonal of $R(\theta)$ by

$$\text{Var}\{\hat{\gamma}(h_i)\} \approx 2\frac{\gamma(h_i)^2}{|N(h_i)|}$$

- $|N(h_i)|$ denotes the number of observations in the bin that corresponds to the lag $h_i$.
- Some papers in the literature show that the estimators from OLS and WLS are more or less equally efficient (or equally bad).
WLS is still not good since it ignores the off diagonals of $R(\theta)$ while there should be clearly dependence between $e(h_i)$ and $e(h_j)$, $i \neq j$

We can work out the expressions for the covariance matrix of $e$, $R(\theta)$

If we denote $T_{ij} = Z(s_i) - Z(s_j)$ and $h_{ij} = |s_i - s_j|$, then we can show that

$$\text{Cov}(T_{ij}^2, T_{kl}^2) = 2\{\gamma(h_{ij}, \theta) + \gamma(h_{jk}, \theta) - \gamma(h_{jl}, \theta) - \gamma(h_{ik}, \theta)\}$$

While clearly GLS should perform better than OLS or WLS, there is a great computational difficulty

In practice, people commonly use OLS or WLS
You may use the function `variofit` and other related functions in the package `geoR`.

They let you do OLS or WLS and you can control the size of the bins.

These functions may not work as you expect though.

We will discuss this in more details once you try them later.
Homework 2 (due Sep 18 12:30 CST by email)

1. Variogram:
   - Using the data set (data 1) from the course webpage, plot variograms (try variogram cloud, lines that give binned averages, boxplots of variograms, directional variograms, etc).

2. Optimization problem:
   - Download the data file (data 2) from the course webpage.
   - Fit a function \( y = f(x) = a \log(x) + bx^2 + \epsilon(x) \) to the data when \( a \) and \( b \) are parameters.
   - Try least squares and assume \( \epsilon(x) \sim N(0, \sigma^2) \) for a constant \( \sigma \).
   - Specifically use “nlm” function in R and get numerical estimates of \( a \) and \( b \).
   - Try at least 5 different starting points (initial values).
   - Report your convergence result (parameter estimates, number of iteration taken) for each starting points.
   - Now try the same with the constraints that \( a > 0, b < 0 \). Report the results in the same way.