Today we will see some methods of simulating spatial data

Why do we care about it?
- Sometimes we have to demonstrate our statistical method works by showing that the method exhibits satisfactory long-run behavior
- We may have to perform randomization tests or other hypothesis tests (we will see an example in a minute)
- Many times we lack of replication in spatial data sets
Example of usage of simulated spatial data set

- Jun, Knutti, and Nychka (2008, *JASA*)
- We have 20 climate models (each model is a gigantic system of PDEs)
- The assumption is that some of these climate models have correlated errors
- For a given time point, we have one output from each climate model
- If we estimate correlation between a certain pair of model errors, we cannot do inference on the statistical significance of it
- What do we do?
Our idea is that, we build a spatial (or spatial-temporal) model for each climate model.

Based on our spatial model, we do *independent* simulation many many times for each climate model.

If climate model errors are independent, the correlation that we get from the independent simulation should be “similar” to the actual correlation that we get from the actual climate model outputs.

Our spatial model and simulation from it gives us a ground for the test for the significance of the correlation.
From now on, we will assume the spatial field has a Gaussian distribution

This is due to the fact that all we know about the process is mean and the covariance structure

Gaussian distribution is a special distribution that is determined 100% by mean and the covariance

We can not do much if the process is not Gaussian
Conditional vs unconditional simulation

- Suppose our spatial domain is $D \in \mathbb{R}^2$ and we have observations on $s_1, \ldots, s_m$ for some $m$
- We consider simulating values of the spatial process (with the same mean and same covariance, and thus same distribution under Gaussianity) on the locations of $D$ other than $s_1, \ldots, s_m$ for some $m$
- *Conditional* simulation means that we respect the observation values. That is, our simulated values on the locations of observations should be the same as the actual observations
- *Unconditional* simulation means that we do not have such restriction
- Obviously unconditional simulation would be easier than the conditional simulation
Suppose we consider simulating $\mathbf{Y} \sim N(\mu, \Sigma)$

Note that since $\Sigma$ is symmetric and positive definite, we can write $\Sigma = \Sigma^{1/2}(\Sigma^{1/2})^T$

Then if $\mathbf{X} \sim N(0, I)$ ($I$ is an identity matrix with the same dimension as $\Sigma$), we can show that $\mu + \Sigma^{1/2}\mathbf{X}$ has the same distribution as $\mathbf{Y}$

Now the question is now to find $\Sigma^{1/2}$
**Unconditional simulation of Gaussian random fields**

1. **Cholesky decomposition**
   - As we discussed in the previous lecture, we can decompose \( \Sigma = U^T U \) for \( U \) being a upper triangular matrix and we can have all of the diagonals of \( U \) being positive (and they are unique)
   - You can use \( \Sigma^{1/2} = U^T \)

2. **Eigenvalue decomposition**
   - There is another way of decomposing \( \Sigma \)
   - That is, we can let \( \Sigma = P \Delta P^T \) where \( \Delta \) is a diagonal matrix with its diagonal values as eigenvalues of \( \Sigma \) (they should be all positive)
   - Also \( P \) should be orthonormal matrix \( (P^T P = I) \)
   - Since the diagonals of \( \Delta \) are positive, we can let \( \Sigma^{1/2} = P \Delta^{1/2} P^T \)
Suppose we want to simulate the random field $Z(s)$, $s \in \mathcal{D}$.

Suppose also that we have observations $Z(s_1), \cdots, Z(s_m)$.

We denote the simulated values of $Z$ as $S$.

A conditional simulation produces $n = m + k$ values such that

$$S(s) = [Z(s_1), \cdots, Z(s_m), S(s_{m+1}), \cdots, S(s_{m+k})]$$
Conditional simulation of Gaussian random fields

1. Sequential simulation for Gaussian random field
   - Note the fact from the multivariate Gaussian distribution that if
     \[
     \begin{bmatrix}
     Z(s_0) \\
     Z(s)
     \end{bmatrix}
     \sim
     \mathcal{N}
     \left(
     \begin{bmatrix}
     \mu_0 \\
     \mu
     \end{bmatrix},
     \begin{bmatrix}
     \sigma^2 & c^T \\
     c & \Sigma
     \end{bmatrix}
     \right)
     \]
     then, \( Z(s_0) | Z(s) \sim \mathcal{N}(\mu_0 + c^T \Sigma^{-1} (Z(s) - \mu), \sigma^2 - c^T \Sigma^{-1} c) \)
   - Using the above fact, we calculate the conditional distribution of \( S(s_{m+i}) \) given \( Z(s_1), \ldots, Z(s_m), S(s_{m+1}), \ldots, S(s_{m+i-1}) \) which is Gaussian

2. Conditioning a simulation by Kriging
   - Consider the decomposition \( Z(s) = p_{sk}(s; Z) + Z(s) - p_{sk}(s; Z) \)
   - Then we replace \( Z(s) - p_{sk}(s; Z) \) by \( S(s) - p_{sk}(s; S_m) \), where \( p_{sk}(s; S_m) \) denotes the simple kriging predictor at location \( s \) based on the values of the unconditional simulation at \( s_1, \ldots, s_m \)
   - Then we can show that the above quantity has the desired property
In statistics, we care about the property of estimators of parameters. Roughly speaking, as we have infinite number of data points, we want our estimators converge to the true values of the parameters. However, for Matérn covariance model, if our spatial domain is fixed and if we assume we get more dense observations, we cannot get such nice property for all of the parameters.
In particular, Zhang (2004, *JASA*) showed that we can only have such nice property for a combination of Matérn parameters.

- Use the parameterization that

\[ K(x) = \frac{\sigma^2 (x/\alpha)^\nu}{2^{\nu-1} \Gamma(\nu)} K_\nu(x/\alpha) \]

- Then only \( \sigma^2 \alpha^{-2\nu} \) can be estimated consistently.

- Now I will show two figures that illustrate some of the above points (pictures provided by Jonathan Stroud).
Figure 1
About Figure 1

- It shows the likelihood function for the range vs sill for the exponential (\(\nu=1/2\)) and Whittle models (\(\nu=1\)).
- The bottom row shows the corresponding posterior distributions using Berger, DeOlivera & Sanso’s objective prior.
- This is for a small dataset with \(n=100\) obs.
It shows the likelihood for the range vs sill in the exponential model.

The bottom row shows the likelihood function for range vs the transformed parameter $\psi = \frac{\text{sill}}{\text{range}}$.

This is for a large dataset with $n=262144$ observations.
A FEW REMARKS

- Note the long ridge in the likelihood function for range vs sill, especially as $n$ gets large.
- The curvature of this ridge is determined by the slope of the covariogram at the origin.
- This suggests the transformation $\psi = \frac{sill}{range^{2\nu}}$, as proposed by Zhang (2004).
- The transformation fixes the problem for large $n$, as range and $\psi$ are nearly orthogonal.