# Lab 1

Zhong Liu
zliu@stat.ubc.ca

Department of Statistics
University of British Columbia
Vancouver, BC Canada

# Outline

- Data sources.
- Variogram and Kriging.
- Linear model of co-regionalization and co-Kriging.
- Bayesian MCMC.

# Data source

- Measurements of various air pollutants from the AQS (Air Quality System). The AQS data is available at http://www.epa.gov/ttn/airs/airsaqs/detaildata/requestingaqsdata.htm.
- The Clean Air Status and Trends Network (CASTNET) data is available at http://www.epa.gov/castnet/
- The acid rain data is available at http://bqs.usgs.gov/acidrain/. Each of the site has the measurement of more than one chemical items. The data is multivariate.
- As an example, the AQS data from May 1995 to Sep 1995 is available in the folder. The file "AIRS.complete" store the hourly ozone measurements from 375 sites. In total, there are 2880 hours. The file "airsSites.txt" stores the sites information.

# *Preliminary data analysis*

- For the time series data, we can use time series plot, box plot, ACF (auto-correlation function), PACF (partial auto-correlation function) to see whether there is a temporal trend within the data. If you want to remove the temporal trend and auto-correlation, you can use appropriate time series model.
- How to deal with missing values. In the original AQS data, many sites (almost all the sites) have missing values. I use the 24-hour mean to fill them in. You can use your own creative methods to fill them in.
- Empirical study of the spatial correlation: plot of correlation versus the Euclidean distance.
- Exercise: After removing the temporal trend and correlation, make the same plot again to see how much spatial correlation left in the residuals. You can only use a subset of the sites.

# *Some basic concepts in spatial statistics*

- Suppose we have a real-valued random process $\{Z(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, which is observed at locations $\{\boldsymbol{s}_i : i = 1, ..., n\}$ over a geographic region $D \subset \Re^d$, $d$ being a positive integer.
- The random process $\{Z(\boldsymbol{s})\}$ is defined as second order stationarity if it satisfies for all $\boldsymbol{s}$

$$\begin{aligned} &\mathsf{E}(Z(\boldsymbol{s}+\boldsymbol{h}) - Z(\boldsymbol{s})) = 0, \\ &\mathsf{var}(Z(\boldsymbol{s}+\boldsymbol{h})) = \mathsf{Var}(Z(\boldsymbol{s})), \\ &\mathsf{cov}(Z(\boldsymbol{s}+\boldsymbol{h}), Z(\boldsymbol{s})) = C(\boldsymbol{h}), \end{aligned}$$

- Definition of stationary is similar to such definition in time series. That is, the correlation only depends on the distance between two locations.
- Isotropic and anisotropic. If $C(\boldsymbol{h}) = C(|\boldsymbol{h}|)$, then we call the process isotropic, otherwise anisotropic. An example is when the air pollution are affected by the wind because wind has its direction.

# *Some basic concepts in spatial statistics*

- Definition of variogram is the following.

$$\mathrm{Var}(Z(\boldsymbol{s}+\boldsymbol{h})-Z(\boldsymbol{s}))=2C(\boldsymbol{0})-2C(\boldsymbol{h})=2\gamma(\boldsymbol{h}),$$

  $2\gamma(\boldsymbol{h})$ being known as a variogram. The so-called semi-variogram refers to $\gamma(\boldsymbol{h})$.

- Semi-variogram is an increasing function.
- The empirical estimate of variogram function is

$$2\hat{\gamma}(\boldsymbol{h})=\frac{1}{|N(\boldsymbol{h})|}\sum_{N(\boldsymbol{h})}(Z(\boldsymbol{s}_i)-Z(\boldsymbol{s}_j))^2, \tag{1}$$

  where the sum is over $N(\boldsymbol{h})=\{(i,j):\boldsymbol{h}-\boldsymbol{\delta}\leq|\boldsymbol{s}_i-\boldsymbol{s}_j|\leq\boldsymbol{h}\}$ and $|N(\boldsymbol{h})|$ is the number of distinct elements in $N(\boldsymbol{h})$.

# Some basic concepts in spatial statistics

- An more robust estimate is

$$2\bar{\gamma}(\boldsymbol{h}) = \frac{\{\frac{1}{|N(\boldsymbol{h})|}\sum_{N(\boldsymbol{h})}|Z(\boldsymbol{s}_i) - Z(\boldsymbol{s}_j)|^{1/2}\}^4}{0.457 + \frac{0.494}{|N(\boldsymbol{h})|}}. \qquad (2)$$

- A quick way to empirically check the isotropic assumption is to plot the semi-variogram function in various directions. The function "variog" function can estimate semi-variogram in different directions.
- Nugget effect. The semi-variogram function $\gamma(\boldsymbol{h})$ should be 0 when $\boldsymbol{h} = \mathbf{0}$. However, empirically, it is not always the case or we do not have enough information at mesoscale.
- Definition of nugget effect. If $\gamma(\boldsymbol{h}) \to \tau > 0$, as $\boldsymbol{h} \to \mathbf{0}$, $\tau$ is the called nugget effect. One of the sources of the nugget effect is measurement error and usually measurements are assumed to be the ground truth $Z(\boldsymbol{s})$ plus some measurement error.
- In the R package "geoR", you can specify whether there is nugget effect or not in estimating the variogram function.

# Some basic concepts in spatial statistics

- Choices of covariance functions. If we have replicates, we may not need parametric covariance function. But if we only have one replicate, we need to specific the function perimetrically.
- The matrix constructed from the covariance function has to be positive definite and symmetric. Bochner's theorem (Stein 1999) states the relationship between spectral density and covariance function.
- Matern function:

$$C_{\boldsymbol{\theta}}(d) = \frac{\sigma}{2^{\nu-1}\Gamma(\nu)}(2\nu^{1/2}|d|/\rho)^{\nu}K_{\nu}(2\nu^{1/2}|d|/\rho).$$

Here $\sigma$, the *sill* parameter, represents the variance of random process $Z(\boldsymbol{s})$ while $\rho$, the *range* parameter, determines how fast the correlation decreases when distance $d$ increases and $\nu$, the *smoothing* parameter controls the smoothness of the covariance function. $K_{\nu}$ is the modified Bessel function of type III.

# Some basic concepts in spatial statistics

- Matern function is commonly used because of its flexibility. When $\nu = 0.5$, it becomes exponential function

  $$C_{\boldsymbol{\theta}}(d) = \sigma \exp(-|d|/\rho).$$

  The Gaussian function

  $$C_{\boldsymbol{\theta}}(d) = \sigma \exp(-|d|^2/\rho)$$

  .

- We can use maximum likelihood, restricted maximum likelihood, ordinary least square or weighed least square to fit the parametric form of the covariance function.

# Kriging

- Given measurements of a random field $\boldsymbol{Y} = (Y(\boldsymbol{s}_1), ..., Y(\boldsymbol{s}_n))^T$, the question is how to predict the random variable $Y$ at a new site $\boldsymbol{s}_0$.
- General equations of kriging. The interpolated value at $\boldsymbol{s}_0$ is

$$Y^*(\boldsymbol{s}_0) = \sum_{i=1}^{n} w_i Y(\boldsymbol{s}_i),$$

  $w_i$ being the weight of $Y(\boldsymbol{s}_i)$.

- Depending on different assumptions, we have different types of Kriging.
- Simple Kriging assumes the mean is known and there is no constraints on $\sum_{i=1}^n w_i$. Other Krigings have the constraints that $\sum_{i=1}^n w_i = 1$.
- Ordinary Kriging assumes the mean of the process is constant over space. That is, $E(\boldsymbol{Y}(\boldsymbol{s}0)) = E(\boldsymbol{Y}(\boldsymbol{s}_i)) = \mu$, $i = 1, \cdots, n$.

# Kriging

- Universal Kriging assumes mean of the process is function of the coordinates. That is, $E(\boldsymbol{Y}(\boldsymbol{s})) = f(\boldsymbol{s})$. The function $f$ usually is a polynomial function of the coordinates at $\boldsymbol{s}$.
- kriging with external drift assumes that the mean of the process is defined through some auxiliary variable. For example, we know that temperature is related to the ozone level. If we have the temperature information at each site, we can use that information to model the mean of ozone level at each site.
- How to implement Kriging?
  1. Estimate the variogram function.
  2. Minimizing the MSE (mean square error) of the Kriging estimator by plug-in the variogram function to estimate the weights $w_i$.
- Example, in ordinary Kriging we only need to minimize the variance of the Kriging estimator since it is an unbiased estimator. The variance is

$$\begin{aligned}\sigma_E^2 &= E\left[(Y^*(\boldsymbol{s}_0) - Y(\boldsymbol{s}_0))^2\right] \\ &= -\gamma(\boldsymbol{s}_0 - \boldsymbol{s}_0) - \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(\boldsymbol{s}_i - \boldsymbol{s}_j) + 2\sum_{i=1}^{n} w_i \gamma(\boldsymbol{s}_i - \boldsymbol{s}_0).\end{aligned}$$

# Kriging

- By plugging the variogram, Kriging underestimate the uncertainties of the interpolation. We can use Bayesian Kriging to improve the classical Kriging at the price of more computation time.
- Kriging is a "pure" spatial model, which does not incorporate any temporal correlation. The ozone data from the AQS system are hourly time series. As an exercise, you can apply Kriging to every hour (for example the first 24 hours) or you can remove the temporal trend first then apply Kriging to the residuals. Use your imaginations!!
- When you use Kriging, compare the interpolation results between ordinary and universal Kriging with different degree of the polynomial function $f(\boldsymbol{s})$.

# co-Kriging

- .
- Kriging is a "pure" spatial model, which does not incorporate any temporal correlation. The ozone data from the AQS system are hourly time series. As an exercise, you can apply Kriging to every hour (for example the first 24 hours) or you can remove the temporal trend first then apply Kriging to the residuals. Use your imaginations!!
- When you use Kriging, compare the interpolation results between ordinary and universal Kriging with different degree of the polynomial function $f(\boldsymbol{s})$.

# Basic introduction to Gibbs sampling and Metropolis-Hastings algorithm

- Both algorithms are used to sample from a distribution.
- In Bayesian framework, we often need to sample from the posterior distribution, which is rarely a standard one.
- Suppose we have posterior distribution for parameters $\alpha$ and $\beta$, that is $p(\alpha, \beta|D)$, where $D$ stands for the available data. The Gibbs sampling (Gelfand and Smith, 1990) is implemented in an iterative way. In each iteration $i$, we sample $\alpha$ and $\beta$ from its full conditional distribution, which is the conditional distribution of parameter given all the other parameters and data. At the beginning of the algorithm, we choose some arbitrary values for the parameters, that is, $\alpha^0$ and $\beta^0$. At iteration $i$, we have the following two steps:

  1. Sample $\alpha^i$ from its full conditional distribution $p(\alpha|\beta^{i-1}, D)$, $\beta^{i-1}$ being the sample of $\beta$ from the previous iteration.

  2. Sample $\beta^i$ from its full conditional distribution $p(\beta|\alpha^i, D)$, $\alpha^{i-1}$ being obtained from the previous step in this iteration.
- Under some conditions, the Markov Chains of $\alpha$ and $\beta$ will converge to their posterior distributions.
- Gibbs sampling is of benefit when the joint posterior distribution $p(\alpha, \beta|D)$ is too difficult to derive while each parameter's full conditional

# Basic introduction to Gibbs sampling and Metropolis-Hastings algorithm

- Metropolis-Hastings (Metropolis,etc 1953 and Hastings 1970) is used to sample from distributions which do not have closed forms.
- Suppose the posterior distribution for parameters $\alpha$ is $p(\alpha|D)$, which does not have a closed form. Same as Gibbs sampling algorithm, we choose some arbitrary values for the parameter, that is, $\alpha^0$. At each iteration $i$, we have the following two steps:

  1. Propose a new value $\alpha^*$ from its proposal distribution $q(\alpha^*|\alpha^{i-1})$.

  2. Accept the new value with acceptance probability defined as

$$a = \min\{1, \frac{p(\alpha^*|D)q(\alpha^{i-1}|\alpha^*)}{p(\alpha^{i-1}|D)q(\alpha^*|\alpha^{i-1})}\}.$$

- Under some conditions, the Markov Chains of $\alpha$ will converge to its posterior distributions.

# Basic introduction to Gibbs sampling and Metropolis-Hastings algorithm

- If the proposal distribution is symmetric, that is $q(\alpha^{i-1}|\alpha^*) = q(\alpha^{i-1}|\alpha^*)$. Then the acceptance probability is the ratio of posterior densities of $\alpha^*$ and $\alpha^{i-1}$.
- If the support of $\alpha$ is on real line and does not have specific restrictions, usually we use normal as proposal distribution. The tuning parameter is the variance of the normal distribution. The acceptance rate tends to be big if the variance is too small and tends to be small if the variance is too big. It is recommenced that the acceptance rate is about $25\% \sim 60\%$ to achieve good convergence. However, it is not a rigorous rule, there are many quantitative ways to check the convergence of the Markov chain.
- We often use Metropolis-Hastings within Gibbs sampling. In the two steps of Gibbs sampling, we have to use Metropolis-Hastings if the full conditional distribution $p(\alpha|\beta, D)$ does not have closed form.

# Toy examples...

- Example 1. Linear regression. Instead of using least squares, we can also use Gibbs sampling in a Bayesian framework. The linear model is

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \\ \boldsymbol{\epsilon} &\sim N(\mathbf{0}, \sigma^2 \boldsymbol{I}). \end{aligned}$$

Prior and full conditional distribution of $\boldsymbol{\beta}$ is

$$\begin{aligned} &\pi(\boldsymbol{\beta}) \sim N(\boldsymbol{\beta}_0, \boldsymbol{F}) \\ &p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{Y}) \sim N(\boldsymbol{B}\boldsymbol{b}, \boldsymbol{B}) \\ &\boldsymbol{B}^{-1} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\boldsymbol{X} + \boldsymbol{F}^{-1} \\ &\boldsymbol{b} = \boldsymbol{X}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}Z + \boldsymbol{F}^{-1}\boldsymbol{\beta}_0 \\ &\Sigma = \sigma^2 \boldsymbol{I}. \end{aligned}$$

# Toy examples...

- Prior and full conditional distribution of $\sigma^2$ is

$$\pi(\sigma^2) = \frac{1}{\Gamma(\alpha)\gamma^\alpha} \frac{1}{(\sigma^2)^{\alpha+1}} \exp^{-1/(\gamma\sigma^2)} \quad \alpha > 0 \gamma > 0$$

$$\pi(\sigma^2) \sim IG(\tilde{\alpha}, \tilde{\gamma})$$

$$\tilde{\alpha} = \alpha + (n)/2$$

$$\tilde{\gamma} = \left( \frac{1}{\gamma} + \frac{1}{2} (\boldsymbol{Y} - \boldsymbol{X\beta})^{\intercal} (\boldsymbol{Y} - \boldsymbol{X\beta}) \right)^{-1}$$

n is the sample size.

- The R code for this toy example is in file "gibbstoy.r".

## Toy examples...

- Example 2: Suppose $\boldsymbol{X}_1, \cdots, X_n \sim BVN(0, \Sigma)$. Each $X_i = (x_{i,1}, x_{i,2})^\mathsf{T}$ and var$(x_{i,1})$ = var$(x_{i,2}) = 1$, cov$(x_{i,1}, x_{i,2}) = \rho$. The question is to estimate the correlation coefficient $\rho$.
- The R code for this toy example is in file "mhtoy.r".
- Advantage of using Metropolis-Hastings over empirical correlation estimate: standard error of the estimate.
- MCMC diagnosis: assessment of the convergence.
    Make a time series plot of the MCMC samples.
    Histogram of the MCMC samples.
    Check the acceptance rate.
    Acceptance rate too high, decrease the proposal step size.
  Acceptance too low, increase the proposal step size.
    MCMC is easy to implement but sometime can be quite tricky.
  Every run may result different results!!
- Exercise: suppose $\boldsymbol{X}_1, \cdots, X_n \sim BVN(0, \Sigma)$. Each $X_i = (x_{i,1}, x_{i,2})^\mathsf{T}$ and var$(x_{i,1})$ = var$(x_{i,2}) = \sigma^2$, cor$(x_{i,1}, x_{i,2}) = \rho$. The question is to estimate the correlation coefficient $\rho$ and the variance parameter $\sigma^2$. You can use Metropolis-Hastings within Gibbs sampling.