STAT 8025

Lecture 5: Estimation and Prediction (II)

Dr. Emily Lei Kang

Division of Statistics & Data Science
Department of Mathematical Sciences
University of Cincinnati

UNIVERSITY OF Cincinnati

Suppose we are interested in a spatial process $\{Y(\mathrm{s}) : \mathrm{s} \in \mathcal{D}\}$. We have data $\mathrm{Y} = (Y(\mathrm{s}_1), \ldots, Y(\mathrm{s}_n))'$ and would like to fit a model and predict $Y(\mathrm{s}_0)$.

- Strategies:
  1. Variogram:
     - Make assumptions (e.g., intrinsic stationarity, weak stationarity)
     - Estimating variogram
     - Kriging (Spatial BLUP)
  2. Maximum likelihood:
     - Make assumptions: GP
     - MLE for parameters
     - Prediction using conditional distribution from multivariate normal distribution
  3. Bayesian inference
     - Make assumptions: GP, priors
     - MCMC for estimation and prediction

Gaussian Process

- Assume $\{Y(\mathrm{s}) : \mathrm{s} \in \mathcal{D}\}$ is a Gaussian Process (GP) with mean function $\mu(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta}$ and a Matérn covariance function.
- We can write the model as:

$$Y(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + \delta(\mathrm{s})$$

where $\delta(\cdot)$ is a GP with mean zero and covariance

$$C(|\mathrm{s} - \mathrm{u}|) = Cov(\delta(\mathrm{s}), \delta(\mathrm{u})) = \sigma^2\rho(|\mathrm{s} - \mathrm{u}|; \phi, \nu) + \tau^2 I(\mathrm{s} = \mathrm{u})$$

  - τ^2 nugget
  - ϕ range
  - ν smoothness
  - $\rho(\cdot)$ Matérn correlation function

Suppose we observe $Y(\cdot)$ at $\mathrm{s}_1, \ldots, \mathrm{s}_n$. Define $\mathrm{Y} = (Y(\mathrm{s}_1), \ldots, Y(\mathrm{s}_n))'$. We have

$$\mathrm{Y} \sim \mathcal{N}(\mathrm{X}\beta, \Sigma(\Theta))$$

- X is $n \times p$ with the i-th row $\mathrm{X}(\mathrm{s}_i)'$
- $\Sigma(\Theta)$ is $n \times n$ with the (ij)-th element $C(|\mathrm{s}_i - \mathrm{s}_j|)$
- Covariance parameters $\Theta = \{\sigma^2, \tau^2, \phi, \nu\}$

Maximum Likelihood Estimation

We can use MLE to estimate $\boldsymbol{\beta}$ and Θ.

- The log-likelihood is:

$$l(\boldsymbol{\beta}, \Theta) = -\frac{1}{2}\log|\Sigma(\Theta)| - \frac{1}{2}(Y - X\boldsymbol{\beta})'\Sigma(\Theta)^{-1}(Y - X\boldsymbol{\beta}) + \text{const.}$$

- Given Θ, we know that the solution of $\boldsymbol{\beta}$ is the GLS estimator:

$$\hat{\boldsymbol{\beta}}(\Theta) = [X'\Sigma(\Theta)^{-1}X]^{-1}X'\Sigma(\Theta)^{-1}Y$$

- So we can profile $\boldsymbol{\beta}$ out and get the profile loglikelihood:

$$l(\boldsymbol{\beta}, \Theta) = -\frac{1}{2}\log|\Sigma(\Theta)| - \frac{1}{2}(Y - X\hat{\boldsymbol{\beta}}(\Theta))'\Sigma(\Theta)^{-1}(Y - X\hat{\boldsymbol{\beta}}(\Theta)) + \text{const.}$$

- MLE for Θ can be found by optimization routines. Good initial values can help (e.g., from fitting the empirical semivariogram)
- Restricted MLE (REML) can be used to reduce bias in variance parameter estimator
- For $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}(\hat{\Theta}) = [\mathrm{X}'\Sigma(\hat{\Theta})^{-1}\mathrm{X}]^{-1}\mathrm{X}'\Sigma(\hat{\Theta})^{-1}\mathrm{Y}$$

  - Confidence interval for elements in $\boldsymbol{\beta}$?
    Note
    $$Cov(\hat{\boldsymbol{\beta}}(\Theta)) = [\mathrm{X}'\Sigma(\Theta)^{-1}\mathrm{X}]^{-1}$$
    Use the plug-ins
- Asymptotics for MLE of Θ?

- The spatial locations $s_1, \ldots, s_n$ are fixed. We just increase the replicates
  - We will have $Y_i \sim \mathcal{N}(X\beta, \Sigma(\Theta))$ for $i = 1, \ldots, N$ independently.
  - Under the usual regularity conditions, we have
    - Consistency: $\hat{\theta} \to \theta$ in probability as $N \to \infty$
    - Asymptotically normal: $\sqrt{n}(\hat{\theta} - \theta) \to^d \mathcal{N}(0, \frac{1}{I(\theta)})$ where $I(\theta) = -E_\theta l''$ is the Fisher information.
- Infill asymptotics: Only ONE replication and we fix the spatial domain, but $n \to \infty$ i.e., increasing sampling density
- Increasing domain asymptotics: Only ONE replication and we fix the sampling density, but the extent of the spatial domain increases

Reading assignment:
Zhang, H. (2004) Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics, JASA, 99, 250-261.
Tang W, Zhang L, Banerjee S. (2021) On identifiability and consistency of the nugget in Gaussian spatial process models, J R Stat Soc Series B., 83, 1044-1070.

- We need to be very careful in claims of estimation accuracy

Prediction

- Assuming $Y(\cdot) \sim \mathcal{GP}(\mu(\cdot), C(\cdot,\cdot))$, we want to predict $Y(\mathrm{s}_0)$ given data $\mathrm{Y} = (Y(\mathrm{s}_1), \ldots, Y(\mathrm{s}_n))'$.
  - Think about the joint distribution of $(Y(\mathrm{s}_0), \mathrm{Y}')'$:

$$\begin{pmatrix} Y(\mathrm{s}_0) \\ \mathrm{Y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu(\mathrm{s}_0) \\ \boldsymbol{\mu} \end{pmatrix}, \begin{pmatrix} C(\mathrm{s}_0, \mathrm{s}_0) & \mathrm{c}(\mathrm{s}_0)' \\ \mathrm{c}(\mathrm{s}_0) & \Sigma \end{pmatrix} \right)$$

  - So we have the conditional distribution:

$$Y(\mathrm{s}_0)|\mathrm{Y} \sim \mathcal{N}(\mu_{pred}, \sigma^2_{pred})$$

  with
    - Predictor: $\mu_{pred} = \mu(\mathrm{s}_0) + \mathrm{c}(\mathrm{s}_0)'\Sigma^{-1}(\mathrm{Y} - \boldsymbol{\mu})$
    - Prediction variance: $\sigma^2_{pred} = C(\mathrm{s}_0, \mathrm{s}_0) - \mathrm{c}(\mathrm{s}_0)'\Sigma^{-1}\mathrm{c}(\mathrm{s}_0)$

Assume $\mu(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta}$ and covariance function $C(\cdot,\cdot) = \sigma^2\rho(\cdot,\cdot;\phi,\nu)$. We assume ϕ and ν are known, but $\boldsymbol{\beta}_{p\times 1}$ and σ^2 are unknown.

You need to show

$$Y(\mathrm{s}_0)|\mathrm{Y} \sim t(\hat{Y}(\mathrm{s}_0), \hat{\sigma}^2 c^*, n-p)$$

where $\hat{Y}(\mathrm{s}_0) = \mathrm{X}(\mathrm{s}_0)'\hat{\boldsymbol{\beta}} + \mathrm{r}(\mathrm{s}_0)'\mathrm{R}^{-1}(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})$

$\hat{\sigma}^2 = \frac{1}{n-p}(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})'\mathrm{R}^{-1}(\mathrm{Y} - \mathrm{X}\hat{\boldsymbol{\beta}})$

$c^* = \rho(\mathrm{s}_0,\mathrm{s}_0) - \mathrm{r}(\mathrm{s}_0)'\mathrm{R}^{-1}\mathrm{r}(\mathrm{s}_0) + (\mathrm{X}(\mathrm{s}_0) - \mathrm{X}'\mathrm{R}^{-1}\mathrm{r}(\mathrm{s}_0))'(\mathrm{X}'\mathrm{R}^{-1}\mathrm{X})^{-1}(\mathrm{X}(\mathrm{s}_0) - \mathrm{X}'\mathrm{R}^{-1}\mathrm{r}(\mathrm{s}_0))$

$\hat{\boldsymbol{\beta}} = (\mathrm{X}'\mathrm{R}^{-1}\mathrm{X})^{-1}\mathrm{X}'\mathrm{R}^{-1}\mathrm{Y}$

Here, R is the $n \times n$ correlation matrix for Y, and $r(\mathrm{s}_0) = (\rho(\mathrm{s}_0,\mathrm{s}_1),\ldots,\rho(\mathrm{s}_0,\mathrm{s}_n))'$.

These are Equations (2.3) and (2.4) from Gu, M. and Berger, J. O. (2016) Parallel partial Gaussian process emulation for computer models with massive output, Annals of Applied Statistics, 10, 1317-1347.

Bayesian Analysis

- Courses offered at 6000 and 8000-level
- In this course I will just give a VERY BRIEF introduction on
  - Bayesian modeling
  - Gibbs sampling and Metropolis-Hastings sampling
- Analysis in the context of geostatistics

- Frequentist vs Bayesian: two interpretations
  - Frequentist: frequency with which an event happens in repeated identical trials
  - Bayesian interpretation: a numerical representation of our belief according to the Bayes' Rule

$$P(A|Obs) = \frac{P(prior)P(A|Prior)}{P(Obs)}$$

- Think about how you explain confidence intervals in Frequentist's interpretation

$$p(\theta|y) = \frac{p(\theta)p(y|\theta)}{p(y)}$$

- Brad Efron, "Using Bayes rule doesn't make one a Bayesian. Always using Bayes rule does"
- With Bayesian analysis, Bayes rule is used to carry out all inference
- In the paradigm of Bayesian analysis:
  - Prior: $p(\theta)$ our belief about the plausible values of θ before seeing data
  - Likelihood: $p(y|\theta)$ how data depends on the unknown parameters θ
  - Marginal distribution: $p(y)$ usually not of interest, just a normalizing constant
  - Posterior: $p(\theta|y)$ our updated belief about θ after seeing the data

Posterior Distributions

- Bayesian inference relies on the posterior distribution $p(\theta|y)$
  - We often need to calculate posterior mean $E(\theta|y) = \int \theta p(\theta|y) d\theta$, posterior mode, variance, or a credible interval
  - We may need to do integration. Often the integrals cannot be worked out in closed form and we need to resort to numerical methods, such as Monte Carlo integration.
    - Draw $\theta^{(1)}, \ldots, \theta^{(M)}$ from $p(\theta|y)$

$$\frac{1}{M}\sum_{i=1}^{M} g(\theta^{(i)}) \to E(g(\theta)|y) = \int g(\theta)p(\theta|y)d\theta$$

Sampling from Posterior Distributions

- The ability to sample from the posterior distribution is essential for Bayesian inference
- Direct sampling from the posterior is rarely possible.
- Two widely used mechanisms to sample from the posterior distributions are:
  - Gibbs sampling
  - Metropolis-Hastings algorithm

Metropolis-Hastings

Suppose that we have a Markov chain with $X^{(t)} = x$. The Metropolis-Hastings algorithm is as follows:

1. Propose a move to x^* with $q(x^*|x)$
2. Calculate the ratio
$$r = \frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)}$$
3. Accept the proposed move $X^{(t+1)} = x^*$ with probablity
$$\alpha = \min\{1, r\}$$
otherwise, remain at x, i.e., $X^{(t+1)} = X^{(t)} = x$.

*$p(\cdot)$ is the distribution we would like to sample from or we can use a function whose value can be calculated and is proportional to $p(\cdot)$.

- Markov chains converge to the posterior distribution regardless of where they start
- It may take a while to converge, so we usually discard the beginning values of the chain, called burn-in. For higher dimensional problems, converging can be slower (an active research area)
- For the proposal distribution in M-H, there are theoretical arguments indicating that the optimal acceptance rate is 44% for one dimension, and has a limit of 23.4% as the dimension goes to infinity
- In practice, people run a short chain and see whether we need to adjust the proposal distribution, say, increasing/decreasing σ^2 the variance of the normal proposal
  - Look for a fat hairy caterpillar in the trace plot

Gibbs Sampling

- We split $\boldsymbol{\theta}$ into blocks and sample each block separately
- It simplifies sampling from a complicated (high-dimensional) joint distribution by breaking it down into simpler (low-dimensional) problems
  - Often many of the blocks can be sampled easily
  - M-H can also be used when we sample the blocks

Suppose we want to sample from $\boldsymbol{\theta} = \{\theta_1, \theta_2, \ldots, \theta_k\}$. At the $(t+1)$th iteration:

1. Draw $\theta_1^{(t+1)}$ from
$$p(\theta_1 | \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)})$$
2. Draw $\theta_2^{(t+1)}$ from
$$p(\theta_2 | \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)})$$
3. $\cdots$

The distribution $p(\theta_1 | \theta_2, \ldots, \theta_k)$ is called the full conditional distribution of θ_1

- I only describe the two most widely known methods, M-H and Gibbs
- If you are not in Stats graduate program, STAT 6043 Applied Bayesian Analysis is a good course to learn Bayesian analysis at a less-technical level
- If you are a PhD students in Stats,
  - you need to know how to draw sample from posterior distribution: obtaining full conditional, writing (your own) code for M-H and Gibbs (and other methods)
  - you may even work on a related topic for your dissertation: Computational approaches for sampling is an important and very active research area, especially for large data and high-dimensional setting

Bayesian Methods

- Advantages:
  - We can incorporate prior knowledge into the model
  - It is very flexible and natural to build complicated models: Hierarchical modeling
  - All uncertainties are taken into account, including those related parameter estimates
- Caveats:
  - How to set priors
  - It may be computationally more expensive: running MCMC

Widely Used Priors in Spatial Regression

$$Y(\mathrm{s}) = \mathrm{X}(\mathrm{s})'\boldsymbol{\beta} + \delta(\mathrm{s})$$

where $\delta(\mathrm{s})$ is a $\mathcal{GP}$ with mean zero and a Matérn covariance function with parameters $\Theta = \{\sigma^2, \tau^2, \phi, \nu\}$

- $\boldsymbol{\beta} \sim \mathcal{N}(0, 10000\mathrm{I})$, i.e., large variance, not necessarily 10000
- Range parameter ϕ: $log(\phi) \sim N(a, b)$ choosing a and b to ensure ϕ smaller than the maximum distance
- Smoothness parameter ν: $log(\nu) \sim N(0, 1)$. But in some studies, people also choose and fix ν instead of estimating it.
- Define the sill $\delta = \sigma^2 + \tau^2$, and let $log(\delta) \sim \mathrm{InvGamma}(0.1, 0.1)$
- Define $r = \frac{\tau^2}{\delta}$: $logit(r) \sim N(0, 1)$ or $r \sim Unif(0, 1)$

Inference

- We will sample from the posterior distribution $p(\boldsymbol{\beta}, \Theta | \mathrm{Y})$

$$(\boldsymbol{\beta}^{(1)}, \Theta^{(1)}), \cdots, (\boldsymbol{\beta}^{(N)}, \Theta^{(N)})$$

- For parameter estimation, we can summarize:

	Posterior Mean / Median	Credible interval (2.5%, 97.5%)
β_0		
$\vdots$		
ν		
ϕ		
σ^2		
τ^2		

Prediction

- We can sample from posterior distribution $p(Y(\mathrm{s}_0|\mathrm{Y})$ directly from MCMC, easy step in Gibbs for GP
- Note

$$\begin{aligned} p(Y(\mathrm{s}_0)|\mathrm{Y}) &= \int p(Y(\mathrm{s}_0), \boldsymbol{\beta}, \Theta|\mathrm{Y}) d\boldsymbol{\beta} d\Theta \\ &= \int p(Y(\mathrm{s}_0)|\boldsymbol{\beta}, \Theta, \mathrm{Y}) p(\boldsymbol{\beta}, \Theta|\mathrm{Y}) d\boldsymbol{\beta} d\Theta \end{aligned}$$

So for $\boldsymbol{\beta}^{(i)}, \Theta^{(i)}$, $i = 1, \ldots, N$, drawn from the posterior distribution $p(\boldsymbol{\beta}, \Theta|\mathrm{Y})$,

$$\frac{1}{N} \sum_{i=1}^{N} p(Y(\mathrm{s}_0)|\boldsymbol{\beta}^{(i)}, \Theta^{(i)}, \mathrm{Y}) \rightarrow p(Y(\mathrm{s}_0)|\mathrm{Y})$$

- In practice, composition sampling is used to draw from $p(Y(\mathrm{s}_0)|\mathrm{Y})$

$$Y(\mathrm{s}_0)^{(i)} \sim p(Y(\mathrm{s}_0)|\boldsymbol{\beta}^{(i)}, \Theta^{(i)}, \mathrm{Y}) \text{ for } i = 1, \ldots, N$$

Spatial Design

- In many scenarios we don't get to select where to take measurements, e.g., instruments on ISS
- For some monitoring networks, we can design and choose a design based on the network's purpose
  - Measure at some critical points
  - Enable prediction at unmeasured responses
  - Estimate process parameters
  - Enable future forecasts
  - Address societal concerns
- Reading: Zidek, J. and Zimmerman, D. (2009) Monitoring network designs. in Handbook of Spatial Statistics (ed. A. Gelfand, P. Diggle, M. Fuentes and P. Guttorp), 131-148.
- Preferential sampling: Diggle, P. J., Menezes, R. and Su, T.-I. (2010), Geostatistical inference under preferential sampling. Journal of the Royal Statistical Society: Series C (Applied Statistics), 59: 191-232.

Comparison

- In practice, we may need to compare and/or choose among: different mean/covariance, different methods
- We usually need to carry out simulation/numerical studies to compare a proposed method with the state of the art
- AIC/BIC for MLE and DIC for Bayesian methods
- Cross validation: Comparing $\hat{Y}$ with Y
- Missing at random and/or Missing in a block
- MSE or RMSE

$$MSE = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \hat{Y}_i)^2, \; RMSE = \sqrt{MSE}$$

- MAD

$$MAD = \frac{1}{m}\sum_{i=1}^{m}|Y_i - \hat{Y}_i|$$

- Coverage probability of 95% prediction intervals
- Average length of 95% credible intervals (ALCI)
- Continuous ranked probability score (CRPS)

Summary

- MLE
- Bayesian analysis
- Design and model comparison

Preview:

- Implementation in R
- Analysis of large spatial data