ST433/533 Applied Spatial Statistics

Lab activity for 9/2/2020

A. CLARIFICATION QUESTIONS

(1) I don't fully understand the idea behind the kappa in Matern model. The difference between matern and exponential is that the Matern is smoother. Looking at the formula, I don't understand what in the formula makes it smoother?

Unfortunately, the formula is too complicated to help with intuition. There is an even more complicated proof that shows nu/kappa determines the number of derivatives that exist before you get to white noise. Realizations from the model are more informative.

(2) In for example AIC , AIC = −2 log{L(β^, θ^)} + 2k ; 2k penelizes models with larger numbers of parameters. In −2 log{ MLE parameter estimates } part of the equation, what is the −2 log doing for the estimate here, why chose −2 log { }?

This is well beyond the scope of this class, but FYI, AIC is derived in information theory as the distance between the fitted and true models, and the -2 arises from the measure of distance between these models. Similarly, BIC approximates the Bayesian probability of the model, and the -2 just comes from this approximation. As side note, for Gaussian data the log likelihood has a -SSE/2 in it, so by taking -2*log_like that term becomes SSE.

(3) A little more explanation on why you prefer BIC over AIC?

It gives smaller models since the penalty term for the number of parameters is larger than for AIC. This is just my preference, and AIC is just as popular as BIC, so don't let me bias you.

(4) In lecture 6b, while describing "Summarize regression coefficients" you mentioned that you do not trust the standard error, Z-score, and the P values, why is that?

I think this was the SE, Z and p of a least squares regression that doesn't account for spatial correlation. The simulation at the end of this document will hopefully convince you not to trust these values. If the model does account for spatial correlation, I trust it.

(5) Why we assume the covariance model to be "exponential" in the beginning?

Exponential is just one simple example of a covariance function. We have to start somewhere, right?

(6) I have a question about the labs not the lectures. Is the plan to have labs every Wednesday or are you still planning to change them up to be on Mondays some weeks?

Yes, the syllabus hard codes in Wednesday for discussion/lab.

(7) If Σ is singular, can we use generalized inverse?

Yes, but you have to be careful. GeoR doesn't do this though so it will crash is Sigma if singular.

(8) Under the anisotropy example, I notice that for the isotropic MLE estimation, this shows in the output:

```
  anisotropy parameters:
  (fixed) anisotropy angle = 0 ( 0 degrees )
  (fixed) anisotropy ratio = 1,
```

and for the anisotropic MLE estimation, the output shows:

```
  anisotropy parameters:
  (estimated) anisotropy angle = 0.6926 ( 40 degrees )
  (estimated) anisotropy ratio = 11.46
```

What do "angle" and "ratio" mean? Can you also briefly go through the example during lab meeting please? Thank you.

The anisotropy model rotates the coordinates by "angle" radians, so s* = R(angle)s. Then the correlation is exp(-d/rho) where

d^2 = ratio*(distance in the first coordinate of s*)^2 + (distance in the second coordinate of s*)

The isotropic model has angle 0, so no rotation, and ratio 1, so the same in all directions.

(9) A large dataset must not be uncommon in spatial problems, so how do we deal with the computational complexity? In the video you said that for 1000s of observations, it can take very long to estimate the \Sigma matrix or construct the likelihood. What alternative ways can be used? In such a case if there is non-stationary correlation, splitting the dataset will have the added benefit of reducing computational complexity by running things in parallel?

Yes, you will tackle this problem for your second midterm. There are now many options, so I anticipate that this will be fun!

B. STUDENT DISCUSSION QUESTIONS

(1) How do we set the values of the initial_rho, initial_sigma^2 and initial_tau^2 in the beginning?

First use a variogram, or maybe some background knowledge.

(2) How to settle the interval range for unknown parameters like "tau_2 and sigma_2 in [0,3]" in slides 14 since the maximum likelihood method could be time-consuming and initial values aren't so good?

Variogram, or use an optimization algorithm that doesn't need bounds.

(3) On the figure showing the numerical optimization using iterations going "uphill", the MLE was different than the true sigma2 and tau2 that were used to build the model. Why is that?

MLE isn't going to get the true value, it's just an estimate. If we had a bigger dataset maybe it would get closer.

(4) At what point is a more complex model worth it as opposed to a simpler model with a slightly higher BIC?

It depends on which variables are considered important enough to exclude (maybe from other studies?), and maybe keep in mind that BIC often selects small models.

(5) What other information criteria have you used in the past, and could they be applicable to our spatial analysis?

Mallow's Cp (BR, not sure if it can be used for spatial data?), cross-validation (spatial CV uses Kriging).

(6) Depending on your AIC/BIC value, one can find which model is "better." Given that the likelihood function is part of AIC/BIC calculations, if you obtained the initial likelihood function from a different (non-Gaussian) parameter distribution, could you compare that AIC/BIC value to a model that used a Gaussian likelihood function (i.e. Y~Normal(mu,sig^2))?

Yes, you can compare different parametric models for the data. BR: You can't compare AIC for Y and AIC for log(Y).

(7) It seems that both forward and backward selection chooses a local optimal covariate in each step. I wonder if this always gives us a global optimal selection of covariates.

BR: There is no guarantee they will find the global optimal solution, in fact, with many covariates they almost certainly won't.

(8) When fitting a spatial model, we assume the data are gaussian, but what if they're log Normal instead? Can we use a log Normal distribution as our likelihood?

BR: Yes! Or anything else you think is appropriate. We will discuss this more in the generalized linear model section.

(9) Discuss why standard errors are unreliable with smaller datasets

Some methods are based asymptotic arguments. Also, the SE for beta depends on theta, and for small datasets the theta estimates could be off.

(10) In what situations would ignoring uncertainty of covariance, be harmful when using the "plug in" method for standard error?

Always! Unless uncertainty is small as for large datasets.

C. BRIAN’S DISCUSSION QUESTIONS

(1) What is the effect of selecting poor initial values for a maximum likelihood analysis?

As long as the initial values aren’t too crazy you should get the same answer, but it might take more iterations if the initial values aren’t close to the MLE.

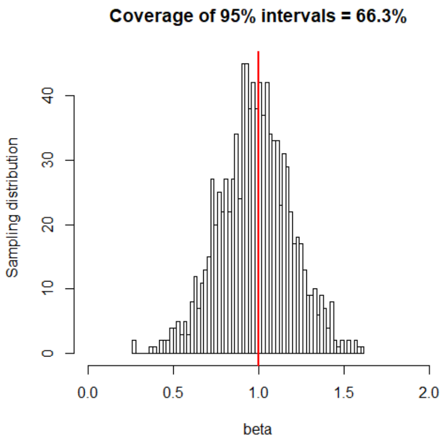
(2) What are the pros and cons of variogram versus maximum likelihood analysis?

Variogram is fast and easy to compute, MLE is more precise and rigorous.

(3) This question revolves around a simulation study. In a simulation study, we generate many datasets from a model where we know the true values of the parameters, apply a statistical method to each dataset, and compare the estimates/test/intervals produced by the method to the true parameter values. This used to understand when the method performs well, and to compare methods. In this example, we evaluate the performance of non-spatial regression applied to spatial data.

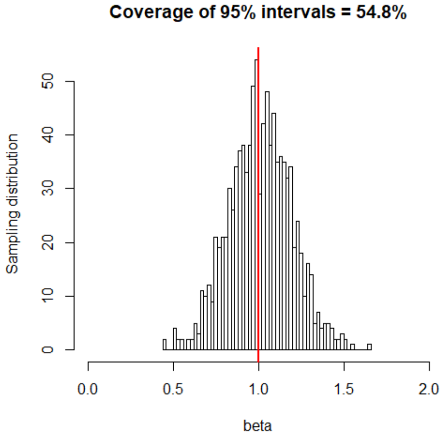
We simulate n=100 observations with s on the unit square, so s_1,s_2 are in (0,1) The data are Y = X*beta + Z + E where X and Z are Gaussian processes with exponential correlation and E is the nugget effect. We can vary n, the true value of beta, the spatial range of X, and the spatial covariance parameters in Z and E. For a set of true values, we simulate 1000 datasets, apply non-spatial least squares to each dataset, and then record the proportion of the datasets for which the 95% interval of beta includes the true value.

(a) The code below takes the spatial range of X and Z to be 0.2, the nugget variance to be zero and true value of the partial sill and beta to be one. Below is the histogram of the 1000 least squares estimates of beta and the coverage. Summarize the method’s performance.

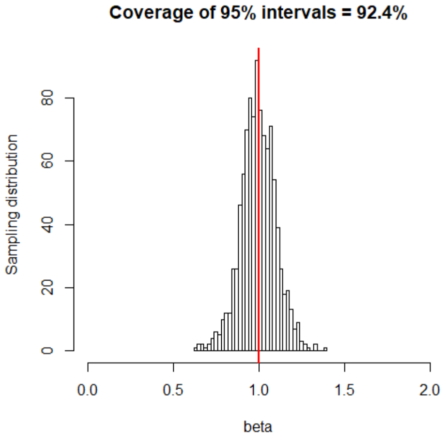


Undercoverage, but it’s unbiased!

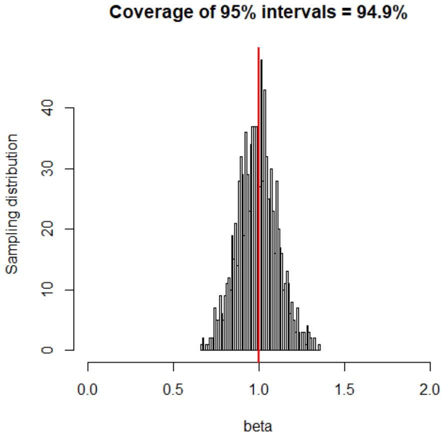
(b) How do you expect the results to change when the sample size increases from 100 to 200?



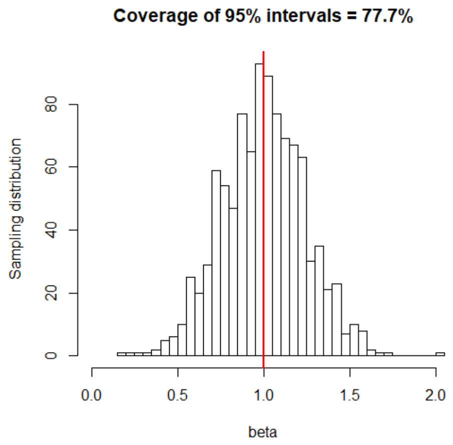
(c) How do you expect the results to change when the range of X changes from 0.2 to 0.02?



(d) How do you expect the results to change when the range of Z changes from 0.2 to 0.02?



(e) How do you expect the results to change when the nugget SD increases from 0 to 1?



(f) What parameter value do you think would be the most problematic for the non-spatial regression?

(g) Summarize the performance of non-spatial regression for spatial data. When does it perform well? When does it perform poorly? Why?

Least squares is OK when the covariate or residuals have low spatial correlation. In all other cases we tried coverage is too low, despite the fact that the estimate is unbiased.

```
library(mvtnorm)
set.seed(919)

# True values of parameter
n    <- 100
rhox <- 0.2  # Spatial range for the covariate
rhoz <- 0.2  # Spatial range for the errors
tau  <- 0.0  # Nugget SD
sig  <- 1.0  # Partial sill SD
beta <- 1    # True value of beta

s    <- cbind(runif(n),runif(n))
d    <- as.matrix(dist(s))
Sx   <- exp(-d/rhox)
Sz   <- exp(-d/rhoz)

nsims    <- 1000 # Number of datasets to generate

# Space to store the output
beta_hat <- beta_lo <- beta_hi <- rep(0,nsims)

for(sim in 1:nsims){

  # Generate fake data
  X   <- as.vector(rmvnorm(1,rep(0,n),Sx))
  Z   <- as.vector(rmvnorm(1,rep(0,n),Sz))
  Y   <- X*beta + sig*Z + rnorm(n,0,tau)

  # Conduct a non-spatial least squares analysis
  fit <- lm(Y~X)
  ols <- summary(fit)$coef

  beta_hat[sim] <- ols[2,1]
  beta_lo[sim]  <- ols[2,1]-1.96*ols[2,2]
  beta_hi[sim]  <- ols[2,1]+1.96*ols[2,2]
}

print(ols) # Check the form of the output for one (the last) dataset

                    Estimate Std. Error   t value     Pr(>|t|)
      (Intercept) -0.2265563 0.08675047 -2.611586 1.042877e-02
      X            1.2190238 0.09393041 12.977946 5.236577e-23

# Compute the coverage of the 95% interval
cov <- round(100*mean((beta_lo<beta) & (beta<beta_hi)),1)


# This produces the plot above
hist(beta_hat,breaks=50,xlim=c(0,2),
     xlab="beta",ylab="Sampling distribution",
     main=paste0("Coverage of 95% intervals = ",cov,"%"))
abline(v=beta,col=2,lwd=2)
```