ST433/533 Applied Spatial Statistics

Lab activity for 9/23/2020

A. CLARIFICATION QUESTIONS

(1) This is just a general random question. Can you do a survey to see how long people spent on the exam? I feel like I spent a lot of time on it gathering my covariates and making sure they worked and am curious if other people also struggled and spent a lot of time on this exam or not.

A brilliant question for today's ice breaker! But the exams are not all turned in, so don't get technical.

(2) I was wondering what the timeline is for Midterm 2. I see that it's due 10/9 so I was wondering when we would be getting it and how long we will have to complete it since 10/9 is less than 2.5 weeks away

I'll send a doodle poll to form teams later this week. I'll give you the exam soon after. I hope you can send me PDF by 10/9 for the presentations over zoom the following Wednesday.

(3) Can you talk about how we have been estimating Z in the standard spatial regression? Are the exponential and matern models a part of calculating the spatial processes in Z?

We build a model for Z and estimate the parameters in its covariance (range, sill, kappa, etc). Kriging then gives use the prediction for E(Y) = Z+Xb, so if you want to estimate Z you can subtract Xb from the Kriging prediction.

(4) Can you clarify telling whether a specific covariate is significant for spatial GLMs.

You can declare it is significant if its 95% interval excludes zero.

(5) It seems like we only learned about MLE to learn that Bayesian analysis is way better. How common are applications where MLE is genuinely the best choice/how common is each in real world usage?

I wouldn't say Bayes is way better. It's slower and requires priors, but more flexible and better at quantifying uncertainty. Both are acceptable and my experience is that the usage varies by field, e.g., Bayesian very common in ecology while MLE is very common in ag. It's definitely good to be comfortable with both approaches.

(6) spLM() does not return the parameter estimate for the beta but spGLM() does. Why is that? spLM() only returns the estimates for the theta (sigma^2, tau^2 and phi).

spLM does return beta but it's in a separate slot than the covariance parameters. I'm not sure why they formatted things differently.

(7) Maybe you have mentioned it but I missed it. Why is there no nugget term to estimate in the Bayesian spatial logistic regression example? Did you check and see that no nugget term is needed for the model?

Yes, this was briefly mentioned in the videos. Given the spatial term Z, the responses are independent across space following a Bernoulli, binomial or Poisson distribution. This independent variation is a nugget effect. But you can also add a Gaussian nugget to Z if you feel it's needed.

(8) Also, the negative odd ratio (i.e. -p) can be interpreted as the predictor variable is p times less likely to cause any change in the dependent variable?

Yes!

(9) Does the MLE classification methods implemented on many remote sensing programs use any of the approaches we have seen to now?

I don't know, sorry. My guess is that for large datasets they use some approximations like the ones we'll cover soon.

(10) So the only reason we cannot use the classical MLE technique would be that the calculation of the joint likelihood is tricky? Would there be some cases where its not too bad to calculate, and would be preferred because it would be faster than Bayes?

Well, yes, for Gaussian data you could marginalize over Z and use MLE. This is the only case that comes to mind, but it's a really important case!

(11) What's the marginal (over Z) correlation of Y?; Why GLM cannot marginalize out the Zi?; In slide 21, why is Cov{expit(Zi),expit(Zj)} called "intractable quantities"?

In our model, if we knew Z, then the Y's would be independent. Of course, we don't know Z and the marginal covariance of Y averages over possibilities for Z. The requires an integral over the joint PDF of Z and it's really hard to solve analytically.

(12) It is mentioned that non-Gaussian spatial data is better suited for logistic regression, but from what I can tell, Gaussian covariance functions are still being used even in logistic regression, why is that?

Right, so the link function relates a Gaussian Z to the non-Gaussian Y. This is done because we understand multivariate normal distributions far better than we understand multivariate binary distributions, but it's surely not the only option.

(13) This is applicable to non-spatial GLMs too: In GLMs, do we have to assume (and verify) that the errors have the same distribution as the response? Or is there no assumption of how the errors are distributed?

The concept of ``error'' is a little tricky for non-Gaussian data. The response minus the mean isn't super useful for say binary data because this could be negative. For non-Gaussian data I would think of a model for the response distribution and not think in terms of errors.

(14) Could you give a rough idea of how you would set up a Bayesian GLM model in JAGS?

The website for my Bayes class has many examples, including Bayes GLMs and spatial models, https://bayessm.wordpress.ncsu.edu/

(15) How do we decide when to use GLM instead of MLE? What is the criteria?

A GLM is a model, MLE is an estimation algorithm. For example, non-spatial GLMs are typically fit using MLE.

B. STUDENT DISCUSSION QUESTIONS

(1) How do we ensure that a potential model is likely to have good convergence when the run time is significant?

Convergence is more likely with a good prior and simple model. Run with a subset of data or a short chain to get an idea how long convergence will take.

(2) Should the transformation be monotonic? If multiple transformations are feasible, do they lead to similar results?

Any non-linear transformation will change the result. BR: generally monotonic transformation and link function are good so you can get back to the original scale.

(3) In the slides, it is stated that slight deviations in normality are acceptable but extreme ones are not, is there some sort of normality test that can help quantify this?

Yes, a quantile-quantile plot of the sample v normal quantiles and there is a test associated with this plot (KS).

(4) What is the best way to address the assumptions of a Spatial GLM model?

Chi-square goodness-of-fit for the proposed distribution (Poisson). Visual inspections of data for isotropy etc.

(5) Give examples for when you would use regular logistic regression vs. spatial logistic regression. Repeat for Poisson regression.

Reg logitstic: election data (D v R) ~ education level or income (w/ and w/o spatial random effects)

Poisson: Number of clouds ~ mean precip or temp (w/ and w/o spatial random effects)

(6) When doing Bayesian spatial logistic regression. When not use a simple way like KNN, LDA, and QDA. What is the advantage of using Bayesian spatial logistic regression rather than the methods mentioned?

Account for uncertainty in parameters and select priors, and do testing/intervals for the posterior.

(7) What would be done in the case the response cannot be modeled by traditional distributions (Normal, Gamma, Binomial, Poisson, etc)?

BR: This is tough and I don't have a simple answer. There are many distributions and you don't have to be perfect, so usually you could find a decent working model. There are nonparametric models, but they are not trivial, e.g., https://arxiv.org/abs/2006.15640.

(8) Since we now have a Zi for every Yi plus the other parameters in the model, we have more parameters than observations. Does this mean that using a Bayesian method is the only way to get uncertainty quantification (via the posteriors of each parameter) in spatial GLMs?

BR: Yes, there are more parameters than observations, but because we are assuming the Z are correlated across space there is hope. Generally, for high-dimensional problems you need to make assumptions. There are frequentist methods that can deal with large number of parameters too, such as Gaussian spatial regression, so I don’t want to over-generalize. But I personally find Bayes to be a good solution for this in general.

C. BRIAN’S DISCUSSION QUESTIONS

(1) Say that X is available at 100 locations and Y is available at 200 locations. The analysis plan is to first use Kriging to predict X at the 200 Y-locations, and then fit a spatial linear regression using the 200 X/Y pairs. Treating the imputed X as a known covariate is an example of the ``error in covariates’’ or ``measurement error’’ problem.

(a) What problems might this cause?

Bias for beta (BR: beta is biased toward zero actually)

(b) What are some remedies for these problems?

Instrumental variables? Multiple imputation?

(2) Say we have binary spatial data, i.e., the responses Y(s) are either 0 or 1.

(a) Why can’t we simply model the mean as $E[Y(s)] = b_0 + b_1X(s)$?

The mean could then be out of range (0,1), predictions could be nonsense (not zero or one) too.

(b) Why can’t we simply apply the variogram as for Gaussian data?

BR: You could use a variogram or modified variogram to look for some idea about the spatial dependence, but since we don’t know the formula for the covariance function you can compare empirical and true variogram.

(c) Why can’t we simply apply the Kriging equations?

It requires the spatial covariance, which we can’t compute.

(3) Say the response at location s is Y(s) = 1 if the person who lives at location s gets the flu in 2020 and Y(s) = 0 if they do not get the flu. Let X(s) be the age of the person who lives at location s.

(a)0 Write a non-spa0tial generalized 0linear regression model for these data.

Y(s) ~ indep Bern(p(s)) where logit(p(s)) = b0 + b1*X(s)

(b) Write a generalized spatial linear regression model for these data.

Y(s) ~ indep Bern(p(s)) where logit(p(s)) = b0 + b1*X(s) + Z(s) where cov(Z(s),Z(t)) = sig2*exp(-d(s,t)/phi)

(c) Given a layman’s interpretation of the parameter that controls the effect of age on the response.

The log odds of getting the flu increase by b1 for a unit increase in age, or the odds increase by exp(b1).

(4) Researchers survey n locations, $s_1,\dots,s_n$. At each location they sample 100 trees and record the number (0,1,2,…,100) that have the emerald ash borer (a kind of disease that kills trees).

(a) Write a generalized spatial linear regression model for these data.

Y(s) ~ indep Poisson(p(s)) where log(p(s)) = b0 + b1*X(s) (if n is large and p is small)

OR

Y(s) ~ indep Binomial(n,p(s)) where logit(p(s)) = b0 + b1*X(s)

Both with Z as in 3b.

(b) Describe how to use the model to make a map of the distribution of the emerald ash borer.

BR: You could plot the posterior mean of p(s) of space, and if needed threshold it as say 0.05.

(c) For a new location s_0, how would you estimate the probability that a sample of 100 trees at s_0 would have at least one occurrence of the emerald ash borer?

BR: (1) Draw a sample of Z and b using MCMC, (2) Draw Y(s_0) given this draw for Z and b, repeat many times and compute the proportion of draws with Y(s_0)=0.

(5) Say the response at location s, Y(s), is the number of microbial species found in a soil sample. The relevant covariates are X_1(s) = annual average temperature at s and X_2(s) = average annual precipitation at s. The goal of the study is to test whether these two climate variables affect the response.

(a) Write a spatial generalized linear model for this problem.

Y(s) ~ indep Poisson(p(s)) where log(p(s)) = b0 + b1*X1(s) + b2*X2(s) + Z(s) with Z as in 3b.

(b) Describe how you would carry out the relevant tests.

If the 95% intervals of b1 and/or b2 excludes zero then we can say there is a significant effect.