ST433/533 Applied Spatial Statistics

Lab activity for 10/21/2020

A. CLARIFICATION QUESTIONS

(1) Why is it that "the full conditional distributions of CAR models depending only on neighbors" means that it is necessarily a GMRF? If the initial spatial covariance function is assumed to be non-gaussian, is this still the case?

This is the definition of the Markov model. In time series, a Markov model means that given all the past, the present only depends on the recent past. In space, a Markov model means that given all the other regions, a region only depends on its neighbors. This definition does not rely on normality (see the autologistic example below).

(2) Are there inferences that the CAR model allows for that the SAR model does not allow for since the SAR model is a specific case of the CAR2 model? If so, what are those inferences and why?

The inferences that are available under these two models are basically the same. Although they are conceived from different assumptions, in the end they just give two slightly different covariance matrices, sort of like the difference between an exponential and squared exponential covariance function in geostatistics.

(3) How graphics help to make a decision about the models. I haven't seen thresholds or criteria for reject or accept ranges

I assume you're referring to the histograms of the sample distribution of Moran's I. The relevant threshold is p-value<0.05. These histograms are just to help us learn about sampling distributions.

(4) I have question about the definition of 'large Moran's I' and 'small Geary's C'. Can we treat I as small value when it is between 0 and 0.5 and treat C as large value when it is between 0.5 and 1?

Well, these statistics aren't like correlations so those rules of thumb don't apply. Probably there are rules of thumb about strong/medium/weak Moran's I, but I don't like these rules of thumb generally. IMO, it's just an oversimplification. The most important thing here is the p-value.

(5) Is there a distance-based model similar to the CAR?

Sure, you could set the weights based on distance. For example

$W_{ij} = 1$ if $d_{ij}<D$ for some threshold D

$W_{ij} = 1/d_{ij}$

$W_{ij} = \exp(-d_{ij})$.

are all doable. A drawback of the first choice is that you have to pick D, and a drawback of the last two choices is that the adjacency matrix will no longer be sparse (which will slow computation but this shouldn't stop you from fitting a good model).

(6) How do you plan to deal with covariates that are significant on the spatial model, but not significant or even opposite in the normal covariate model?

I assume by "normal covariate model" you mean non-spatial linear regression. This is one of many examples where your conclusion depends on the assumed model/estimator. Other examples are X1 is significant for the non-spatial linear model without X2 but not significant when X2 is included. Or you get different results based on removing outliers, or transforming the response, etc. In all of these cases it's best to qualify your results with the caveat that they are sensitive to the assumed model and discussing the other models you fit. A related question is how many models/estimators to try. You could try models for years and years, but obviously you don't need to report models that fit poorly. To make the story somewhat intelligible it's best to get down to 3-5 models. Wow, that was a long rant for a short question, sorry.

(7) In the example of Guassian CAR model, the result of function "S.CARleroux" contains the estimate of parameter nu2. But I cannot see anything related to nu in the slides. So I am confused what it is.

They call the nugget variance nu2 and the spatial variance tau2. Seems weird, but there you go.

(8) To find P-values for Moran's I, we are using N data set with predefined distribution. Therefore, the distribution 'X' from the Moran's I of those N data sets will always cluster around '0'. Is it possible to simply rely on a decent value of Moran's I for the response variable instead of comparing it with the distribution 'X'? Or am I not understanding something?

As in any frequentist analysis, to compute the p-value you assume the null hypothesis is true. In our case, the null hypothesis is I=0 for independence. The idea is that you sample from I under H_o to get an idea of how far from zero it might be just due to random chance under the H_o, and then if the I for the real data is way outside this distribution you reject H_o and conclude there is spatial dependence.

(9) Can you give an interpretation of adjacency matrix W with more detail? like what are the values meaning and why most of them are 0.

(Assuming binary adjancies for simplicity) If $W_{ij} = 0$ then regions i and j are not adjacent and if $W_{ij} = 1$ then they are adjacent. So if i = NC and j = SC then $W_{ij}=1$ and if i=NC and j=CA then $W_{ij}=0$. Most of the W_{ij} are zero because most pairs of regions are not adjacent, e.g., NC is adjacent to only 5 (?) of the 49 other states.

(10) When to use the CAR model with constant variance/varied variance? Does it make a difference when setting the prior for Bayesian analysis?

99.99% of CAR model applications do not have constant variance because the constant variance CAR model is really complicated to fit. If a non-constant variance is a concern, I would recommend fitting a geostatistical model with distance between centroids defining the covariance.

(11) In the CAR model, how can we intuitively understand the covariance matrix of the joint distribution of Z before the inverse is taken (M - rho*W) and how do we connect it to the variance in conditional

distribution of Zi (sigma^2/mi)? Why isn't the sigma^2 parameter found anywhere in the covariance matrix formula?

This is a good question, but I don't know of any way to make sense of the covariance matrix. The full conditional distribution and inverse covariance matrix are intuitive, but the covariance is not. This is the opposite of a geostat model where the covariance is intuitive but the full conditional distribution and inverse covariance matrix are not. You have to pick your battles I guess.

The covariance of a CAR should have a sigma. It should be $\text{sigma}^2(\text{M-rho*W})^{-1}$. Sorry if this was a typo.

(12) What is the best plan when the effective n-count is below 1000 for certain variables that you are trying to test?

I assume you mean the effective sample size from MCMC sampling. Some solutions are (i) run longer chains, (2) pick better initial values, (3) pick a simpler model or (4) pick tighter priors (less variance).

(13) What does the process for pooling information across counties look like?

Pooling happens naturally from the model fit. The spatial model (prior) for site i is a function of the mean of its neighbors, and this is combined with the data at site i (likelihood), so the estimate (posterior) uses data from both site i and its neighbors.

B. STUDENT DISCUSSION QUESTIONS

(1) What are other examples of areal data, other than political/government boundaries?

Ecotones (body of water, forest), pixels/voxels, crop fields, (physiographic? regions) coast/piedmont/mountains

(2) Which way of calculating adjacency do you prefer and why? If you think it depends on the data, how so?

Border adjacency (rook, queen). BR: me too, no tuning parameters like the number of neighbors or distance threshold to be considered adjacent.

(3) What is the advantage/disadvantage of using Geary's C over Moran's I?

C is better for local variation, I code was faster (has a matrix expression). Computing both is good.

(4) Is there a matrix form of Geary's C? If so, what is it? BR: The key is writing the numerator sum

$$\sum_{i=1}^{n}\sum_{j=1}^{n}(r_i - r_j)^2 W_{ij}.$$

as a matrix multiplication. Can you do it?

In R the function is

```
r           <- scale(x)
r_square <- r^2

numerator <- sum(t(r_square) %*% W) –
                  2* t(r) %*% W %*% r +
                 sum(W %*% r_square)
# or in two terms
numerator <- 2* t(r) %*% W %*% r + 2 * sum(W %*% r_square)
```

(5) Geary's C is ranging from 0 to 2 where 1 indicates no autocorrelation/no spatial dependence of the data, and 2 indicates perfect "negative" autocorrelation/dispersed. What does "negative" autocorrelation mean on spatial data?

BR: Territorial species, maybe others.

(6) A potential group discussion topic is: scenarios where areal data models are more appropriate than geostatistical models and scenarios in which the choice is ambiguous. Choosing an adjacency matrix is like choosing a covariance model. Would we prefer one matrix over the other? How does it affect computation? Is the best way to choose is to use a AIC / BIC / cross validation approach?

AIC works!

(7) In class Dr. Reich discusses how uses the center of mass of an area can be used to apply geospatial models to areal data. Does it ever make sense to aggregate (long,lat, Y) spatial data into areal data to do areal models?

If you only care about regional summaries, you could average high-resolution geostat data and analyze the averages with areal data. Or you collect lat/long data and aggregate for privacy.

(8) How can we handle the marginal counties of the CAR model or SAR model? Because they will not be the center of their neighbors.

BR: Regions on the edge will have few neighbors and thus higher variance. In some sense this feels right.

(9) The idea of borrowing strength got me thinking about an expansion of the CAR model. So far we've seen it applied when the unit of observation is the same as the unit of spatial grouping (ie county), but what if we have multiple observations per county that are not georeferenced (ie. counts of instances of cancer), each of which has a continuous predictor, quantity of perfluorooctanoic acid (a suspected carcinogen). We want to estimate the effect of this chemical in each county, so we can set up a hierarchical model with random slopes that are modeled as being drawn from a distribution with common hyperparameters -- this is the context that I've often seen for "borrowing strength" -- where counties with few observations have their estimates pulled towards the hypermean. Could we merge the CAR model and this hierarchical model by adding the Zi errors and their covariance structure to the hierarchical model with the goal of estimating the effect of the chemical by county but also accounting for spatial correlation?

BR: Yes. Say Y_{ij} is the j^{th} response in region i. Then $Y_{ij} = a_i + b_i X_{ij} + e_{ij}$ where $a = (a_1,\ldots,a_n)$ and $b = (b_1,\ldots,b_n)$ follow CAR model and e_{ij} are iid normal errors. In this model, a is the spatially-varying intercept and b is the spatially-varying slope. Find the right R package to implement this model might be tough though. The key phrase is “spatially-varying coefficients”.

C. BRIAN'S DISCUSSION QUESTIONS

(1) If you had to conduct a geostatistical analysis of a large spatial dataset this afternoon, which of the methods presented in class last week would you try first and why?

LatticeKrige is efficient on one core. NNGP is slow, but we like it. Predictive process was slowest.

(2) Define $Y_i = 1$ if county i has a hospital with ICU beds and $Y_i = 0$ otherwise. The covariates are X_{1i} = population of county i and X_{2i} = Percent African American of county i.

(a) Define a spatial model for these areal binary data.

Prob(Y_i=1) = p_i where logit(p_i) = $b_0 + b_1X_{1i}+b_2X_{2i} + Z_i$ and Z = (Z_1,…,Z_n) ~ CAR

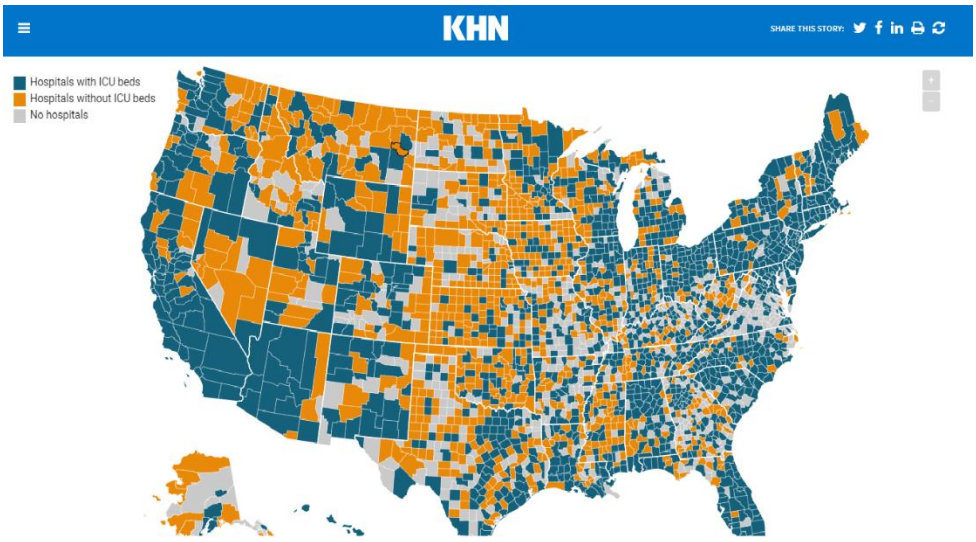
(maybe fit with the BayesCAR package)

(b) How would you determine if your model in (a) fits better than a non-spatial logistic regression?

AIC/BIC/DIC (if Bayes).

(c) If the goal is just to estimate the effect of X_{2i} on a Y_i, what's wrong with fitting a simple non-spatial logistic regression, `glm(Y~X2,family=binomial)`? Why is the spatial model preferred?

Why not run both? The standard errors from a non-spatial model are questionable if you ignore spatial correlation (they will probably be too small).



https://khn.org/news/as-coronavirus-spreads-widely-millions-of-older-americans-live-in-counties-with-no-icu-beds/

(3) The autologistic model is an extension of the CAR model to binary data. The full conditional distributions (ignoring covariates) are

Prob(Y_i=1|Y_j for all j≠i) = p_i where logit(p_i) = $\beta + \rho\sum_{j\sim i} Y_j$

and $\sum_{j\sim i} Y_j$ is the number of regions that neighbor region i and have Y_j=1. What is are the interpretations of the two parameters β and ρ?

The slope rho is the increase in log odds of Yi = 1 for every neighbor that is equal to 1.

(4) The CAR model was initially defined through full conditional distributions and then we found a MVN joint distribution that lead to these full conditional distributions. This worked out for the CAR model, but there are some full conditional distributions that do have a valid joint distribution. These are called incompatible distributions. If you are going to define a model via full conditionals, you have to verify the are compatible.

The following two conditional distributions are incompatible:

$$X|Y \sim N(2Y,1) \quad \text{and} \quad Y|X \sim \text{Normal}(-5X,2)$$

That is, there is no bivariate normal distribution f(X,Y) that gives these conditional distributions.

(a) In words, what do the conditional distributions of Y|X and X|Y tell you about the relationship between X and Y?

X|Y implies a positive relationship between X and Y, and Y|X implies a negative relationship.

(b) Why do the conditional distributions seem to be incompatible?

X and Y can't be simultaneously positively and negatively correlated.