ST433/533 Applied Spatial Statistics

Lab activity for 10/7/2020

A. CLARIFICATION QUESTIONS

(1) How we will use the MCMC for the AR1 model?

You could code it yourself in JAGS or BUGS, but there are many packages including `bsts`.

(2) You mentioned there were several multivariate stats classes at NCSU. Do you know the names? And could you recommend any one in particular for environmental/climate/ecology applications.

This is a good one: https://maityst537.wordpress.ncsu.edu/. It's taught at about the same level as this class. I doubt it covers these application areas specifically but these techniques are certainly relevant.

(3) Could you briefly explain how R packages are "created" or "governed"? It seems like certain academic groups or individuals write R packages, how do you go about making a whole new package? Why would they spend so much time doing that? Paper citations?

You can basically submit any code you'd like at https://cran.r-project.org/submit.html. There are some formatting requirements, but no rigorous testing or approval process. Putting your code in a package helps with citations for sure, but most people just want to contribute to science. What really impresses me are the extensive and super helpful comments on stack overflow, https://stackoverflow.com/questions/tagged/r. At least some nerdy corners of the internet remain cordial 😉.

(4) Can you explain more on the covariance function for different response types at different locations in slide 16 of multivariate?

It's probably best to first consider the separable case. LL' is the KxK covariance matrix of the K responses at a given site. If two sites are separated by distance d their covariance is reduced to L'L*rho(d), eventually going to zero and d increases.

(5) Can we simply assume nonstationary time series data to be stationary like we did to non-stationary spatial data?

I may have been too cavalier about this in the lectures. There is some risk of model misspecification effects if you ignore non-stationarity in spatial, temporal, or spatiotemporal settings. I don't know that the effects are more or less severe in the temporal case, but it is definitely easier to allow for non-stationarity in the temporal case because the data are ordered in one direction. For example, a non-stationary AR1 model is $Y_t = rho_t Y_t + e_t$ where rho_t can vary over time to capture nonstationarity.

(7) How to determine the coefficients / the autoregressive prior for slopes in DLM?

Usually they follow time series model like an AR1 model. The `bsts` package has defaults for priors, for what it's worth.

(8) Can you please write down the equation that you were describing at time 13:26 lecture 13 Spatiotemporal models? I can't visualize what you were saying.

The sample ACF of a times series of length n (in R code) is cor(Y[1:(n-1)],Y[2:n]).

(9) I didn't understand the plot in "Explore the covariance structure" in spatiotemporal lecture (13) . What were we plotting in this plot?

These are variograms computed separately by year.

(10) Same for "ACF at site 1" plot.

This is the sample autocorrelation function for the first location in the dataset.

(11) Just to clear my understanding: AR1 model is a lot similar to the first-order Marko chain model (for predicting 'State', not values).

Yes, it is a first-order Markov chain.

(12) R example: Spatial average mean of Ozone showing a downward trend from 2007 to 2008 & 2009 might be due to the great recession.

Maybe. Wow, there is always the silver lining 😉. How would you test for this? (making this a discussion question now…)

(13) When explaining the autocorrelation function in the lectures, where the examples given for data collected at one specific location or was it over many different locations?

Yes, a standard time series is just at one location.

(14) When we do include the autocorrelation function in the study of multiple locations, should we pray that the data is independent over each location?

Well, that would certainly make the analysis easier. Although it would make predictions at location without data impossible.

(15) Will the cross covariance always converge?

I'm not exactly sure what you mean here, but in a separable multivariate spatial model where the spatial correlation decays to zero as distance increases, the cross-covariance will also converge to zero.

(16) The spatial covariance matrix is n * n where n is the number of observations. What is the dimension for time series covariance matrix?

Yes, it is (# time step) x (# time steps). Although for simple models like an AR1 you don't need to write out the covariance matrix because the model has the form of a linear regression with Y_t as the response and Y_{t-1} as a covariate so you can use `lm`.

(17) Wouldn't MCMC have problems for AR_P models for P>1, since the the basic Markov chain only looks at the previous state as opposed to the P previous states? Or is the blackbox MCMC algorithm handling this on the fly?

I can see how this would be confusing. MCMC doesn't assume the data follow a Markov process, it samples the parameters by constructing a Markov process. So MCMC can be used for any degree of Markov model for the data and even data-models that are not Markov processes, like Bayesian Kriging.

(18) In factor analysis (and LMC), what do you need to know about the relationship between the latent factors and response prior to predicting how the response measures the combination of latent factors? Is this a model-free relationship (where the functional form is negligible) or do we need to know what the functional form is (where the factor is on the x and the response is on the y)?

The only required input is the number of factors, and I guess you don't need to know this value because it can be estimated by say fitting with different number of factors and comparing AIC/BIC/DIC/CV. There is some sensitivity to the way your order the responses in a Bayesian analysis because L is assumed to be lower-triangular (i.e., some elements are fixed at zero), so a sensitivity analysis to this is a good idea.

(19) In the spatiotemporal exploratory analysis example, how do the results affect the choice of spatiotemporal models?

We were looking for (i) which covariates to include in the mean, (ii) seeing if there is spatial and/or temporal correlation to guide our residual modeling, and (iii) examining the assumption of separability, again to determine if this is a reasonable assumption to impose in the model.

(20) In LMC, how do we get the loading matrix?

The loading matrix L is estimated from the data. You can use variograms, MLE, or Bayes for this.

(21) How many variables can a multivariate spatial analysis reasonably handle?

Off the shelf methods like spBayes can probably handle maybe 5 or 10, but more would be tricky. You can do more with LMC if you assume a small number of factors, but that doesn't always fit well.

B. STUDENT DISCUSSION QUESTIONS

(1) What are some examples of real-world data that would fit better with a spatiotemporal model as opposed to just a spatial model?

Climate and economic data models. Most ST model would also work as a spatial model, but if you are interested in time trends an ST model is preferred, e.g., climate change or understand effects of policies on unemployment.

(2) With a spatial autoregressive model, you can still use the Kriging equations to get optimal spatial predictions. When you do this, at what time point are the predictions made? Are they marginalized across all time points?

We set the time for prediction. Prediction way in the future is unreliable.

BF: If you're predicting Y(s,t) using the full training dataset, you are using training data from all times 1,…,t-1,t,t+1,….,n_t.

(3) In the lectures we distinguished the difference between discrete and continuous time. Couldn't continuous time be modeled as discrete? For example, data collected every second could be considered discrete for every 1 second is a time-point.

You can always discretize, but you might lose some information.

(4) What is the minimum discrete or continuous time steps to do a spatiotemporal analysis?

Depends on the context and the time scale. BR: Even with two times and a lot of spatial locations you could do a spatiotemporal analysis.

(5) One of the goals for multivariate analysis is to improve prediction accuracy. Is it the same as making some variates as predictors?

You could call X=Y1 and Y=Y2 and do a univariate spatial analysis. It's not the same, but could certainly be useful. If the goal is to study their covariance and make predictions for both, then use a MV model.

(6) Under what circumstances would a separable model be true/useful? What types of data might comply? Is it okay to assume separable models in all instances?

Separable needed when you don't have data at each timepoint for each location (google street view). Wouldn't make sense if… BR: if Y1 and Y2 clearly have different spatial correlation, e.g., precip (low correlation) and temp (high correlation).

(7) Would a multivariate analysis would have been helpful for the first midterm's analysis and if so, what other response(s) would be fitting? How would they relate to the additional covariates selected?

If there were multiple response, e.g., PM2.5 and Ozone or PM10 etc.

(8) What's the advantage to doing multivariate analysis as opposed to running separate geospatial models for each response variable? It seems like adding the additional response variable has the potential complicate matters

Helps with prediction, but only helps if the two responses have similar spatial correlation and are strongly correlated. If the goal is just prediction, there might be simpler methods than MV spatial.

(9) Can you do a multivariate spatiotemporal analysis? Any examples of when this would be applied or is this not realistic?

Yes! Pollution data for midterm 1 with PM, PM10, ozone over space and 10 years. How?

(10) Intuition suggests that multivariate spatiotemporal models are very common in the climate and weather sciences. What techniques are used in this case? Is it simply a combination between autoregressive models and LMC models?

Could assume separability. Your model sounds like a good place to start, i.e., Y_k(s,t) = rho*Y_k(s,t-1)+ error. There is likely no software to do this, but maybe it will be developed in the future.

(11) Since the normal Kriging equations apply to the spatiotemporal and multivariate analyses, does that mean that the big data techniques we are looking at would also apply to these other types of data?

Most can be applied with some careful tweaks. For example, predictive process could be MV by having a MV process at the knots.

C. BRIAN'S DISCUSSION QUESTIONS

(1) You are studying the change in the prevalence of forest fires in the Western US over time. You select 100 spatial locations (the same 100 sites each year) and record the number of days in each year that the site is in the plume of a forest fire. Let Y(s,t) \in {0,1,…,365} be the number of fire day at location s in year t.

(a) Write a spatiotemporal model for these data.

Y(s,t) = Z(s) + U(t) + e(s,t) where Z is spatial (matern, etc) and U is temporal (AR1, etc).

or

BR: Y(s,t) ~ Bin(363,p(s,t)) and logit(p(s,t)) = Z(s) + U(t).

or

BR: Y(s,t) ~ Bin(363,p(s,t)) and logit(p(s,t)) = Z(s,t) where Z(s,t) has a separable ST covariance.

or

BR: Y(s,t) ~ Bin(363,p(s,t)) and logit(p(s,t)) = Z1(s) + t*Z2(s) where Z1 (intercept) and Z2 (time trend) are a bivariate spatial process with separable MV covariance.

or

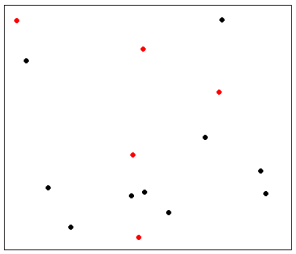
BR: Many more…be creative and have fun!

(b) Describe how you would test if the expected number of forest fires per year is increasing at a given location, s_0.

Test if U(t) is significantly different across years by looking at 95% intervals.

BR: In the last model from (a) you could test if Z2(s0)=0

(2) Say ozone is measured in the five red locations below and PM is measured at the 10 black locations below. Describe how you would test whether these two pollutants are correlated with each other.



One idea: Could fit an LMC model and test which the cross-covariance is zero using 95% intervals.

(3) In clarification question (12), a student questioned whether the decline in ozone during the time of the last recession could be attributed to the decline in economic activity. How might you go about testing this hypothesis? Write a model and/or analysis plan.

One idea: Fit a spatiotemporal model, and test if there is an **association** by including a bunch of other covariates and econ activity and test is the beta for econ activity is zero.