ST433/533 Applied Spatial Statistics

Lab activity for 8/26/2020

A. CLARIFICATION QUESTIONS

(1) On slide 8 of Geostatistical estimation – Part 1, in the exponential covariance plot, how did we determine that the range is 0.2?

I just picked these parameter values to illustrate the shape of the curve.

(2) In Lecture05a_Estimation1, at time 10:07, (slide #7), to get the standard deviations for the denominator of corr function. How did we find the sqrt of { (sigma)^2 + T^2 } and end up with (sigma)^2 + T^2 ?

Since both Y_i and Y_j have SD sqrt(S^2+T^2), the product of their SDs is S^2+T^2.

(3) Is there any way to determine the ideal or minimum density sample to say that we choose the right model? Since if we have not enough samples or dispersed or not well spatial distributed, our model can be wrong.

We will discuss this in the spatial design lectures in a few weeks. Obviously if the effective range is X then you need pairs of observations within X to have a chance of estimating the correlation parameters. As a rule of thumb, I would say that less than 50 observations is a small spatial dataset and usually to get really good estimates of the covariance parameters you need 100+ observations.

(4) Are there any other ways to estimate the parameters instead of guessing them from the variogram? Maybe more accurately?

Definitely. We are just using this to gain intuition about the variogram and the role of different parameters. We will use maximum likelihood analysis starting next week to get final estimates.

(5) Are there other ways to find the mean value rather than linear regression?

Definitely. Right now we are using linear regression as an initial estimate to graphically explore residual covariance. But in the final production run (next week) we will use a spatial model.

(6) Why set d_l to be the twice of effective range and L so that the number of pairs in every bin is 30 at least?

It is nice to set the maximum distance to be well beyond the effective range to verity that the variogram actually plateaus. And the number of observations in each bin being at least 30 is to ensure the variogram isn't too noisy.

(7) Can you explain the 4 sections of the variogram again, please?

I'm not sure what you mean by "sections", but the key features are the y-intercept is the nugget variance, the x-value where the curve plateaus is the effective range, and the height of the plateau is the total variance, sigma2+tau2.

(8) In lecture 5b, in your ozone code, the total number of observations (vg$n summed) is over 220,000. Is this how big the data set is? The first few lines of code show it's only 1255 observations (I think). Why is this?

Yes, there are ~1000 observations. But the variogram counts each **pair** of observations.

(9) Why in the data subset section of code did you "multiply by 100 so the data are easier to visualize"? How does this make the data easier to visualize? Would you do this in a real situation or is this just for class benefit?

This is not needed and doesn't affect the results. It just changes the scale so that the axis labels don't have a bunch of leading zeros (this annoyed me when I first made the plots).

(10) What does an anisotropic variogram look like? Would it be drastically different from the isotropic variogram?

You would plot the empirical variogram for pairs of points in different directions given multiple curves. For example, you'd have an E-W, SW-NE, S-N and NW-SE variograms. If the all look the same this is evidence of isotropy and vice versa.

(11) I didn't understand how we tell from the plot whether the correlation is smooth. Also, does a more blurred graph indicate a nugget effect?

Plots of data with a large nugget effect look like random numbers/white noise. If there is no nugget than the surface is smooth.

(12) This is a technical question, but I was wondering how you got the equation for the variogram, i.e how did we get Var(Y_i - Y_j)^2 = Var(Y_i) + Var(Y_j) - COV (Y_i, Y_j)

This derivation uses several facts from mathematical statistics, such as V(X-Y) = V(X) + V(Y) – 2COV(X,Y).

(13) In cases where sigma^2 = 0, wouldn't the value of phi not matter at all?

Yes!

(14) Why do you choose p(d) = 0.05 and get the result that d is 3 * phi? Is 0.05 just by convention or for some other reasons?

Yes, just convention.

(15) Why is the nugget term equal to zero after the distance is not zero anymore?

The nugget is by definition independent across observations, and so the covariance of the nugget is zero when the distance is zero.

B. STUDENT DISCUSSION QUESTIONS

(1) What is the best approach for choosing maximum distance and number of bins for the variogram? Is it a trial and error process or is there telltale signs that would indicate the optimum values?

Starting with 30-100 bins, reduce until it's not noisy…so trial and error is part of it.

(2) In reality, you never know what the true variogram is or looks like. So, after you plot the data on the variogram and draw different exponential lines through the points, you look for the one with the smallest SSE to get an idea what the parameters may be (as shown in the last slide in Geostatistical estimation - Part I). However, in real data, there will be a lot of noise just like the ozone example in the lecture. How can you be sure the fit with the smallest SSE will be an ideal fit? Are there any other ways to check?

Yes, we will use MLE eventually.

(3) How does the scale and spatial resolution affect the models' implementation and outputs?

Should initially zoom in and fit/build the models based on this.

(4) Provide and discuss an example of a real life dataset that would not need a nugget in the model.

None!!! Some may have small measurement error, but all have some error.

If the height of plants is measured without error and two plants in the same location have the same height (same soil, same sunlight, etc)

Maybe if a covariate explains it all?

Systematic data, like government assigned IDs?

Climate model output is not random, so it's not clear you need a nugget.

Ocean SST is smooth over space and easy to measure. Maybe there is some error, but it is negligible and so assuming there is no nugget is a good modelling step.

(5) Since the Matern incorporates exponential and squared exponential models in special cases, why don't we just use the Matern every time?

You could use always the matern, but if the exponential fits well then this is preferred because the model has fewer parameters and is therefore it's easier to fit, explain, etc.

(6a) Would it be possible that more than one plateaus occur in a variogram plot? For example, the semivariance reaches a plateau first and then increases again, would it likely to reaches another plateau when the distance increases? (6b) To address the hump in the semi-variogram --> Can we cluster the locations based on some relevant features (instead of crudely dividing the data into six groups) and estimate different variograms for each cluster? (6c) What should you do if you see an empirical variogram with an irregular shape such as the shape seen in lecture 5. (6d) For lecture 5b, the variogram

plot does not obey exponential or Matern trend. Should we remove these strange data? Or should we do some transformation of the data?

This suggest that the exponential model doesn't fit well, so you could try a different correlation function (BR: though I don't know of one with two humps).

Transform the data, e.g., use log(Y).

Look for a missing covariate, especially spatial covariates like long^3.

Break the domain into subregions, divide and conquer.

C. BRIAN'S DISCUSSION QUESTIONS

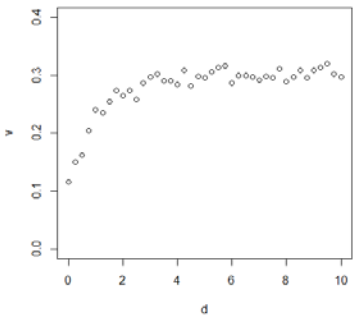
(1) Give a real-life situation where you might expect the true spatial covariance to be

(a) Matern with large smoothness, say nu=10: sea surface temperature data

(b) Matern with small smoothness, say nu=0.5: number of cats per household, precip or crime data at large scales.

(2) From this empirical variogram, make a guess at the value of each parameter in the model

$$Y_i = Z_i + E_i \text{ where } cov(Z_i,Z_j) = sigma2*exp(-d_{ij}/phi) \text{ and } E_i \sim N(0,tau2)$$



(a) Nugget variance, tau2 = 0.1

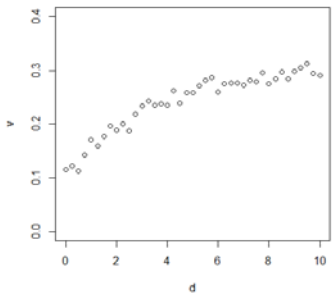
(b) Partial sill variance, sigma2 = 0.2

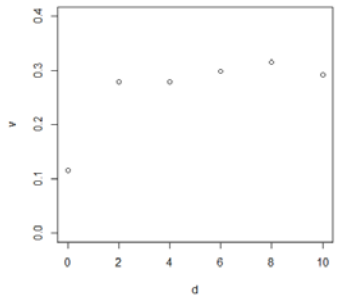
(c) Sill variance, sigma2+tau2 = 0.3

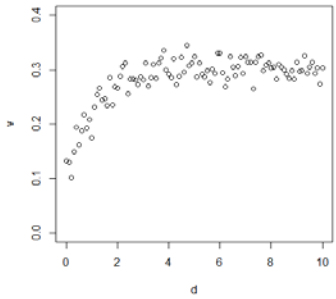
(d) Spatial range, phi = effective range = 3-4, phi = 1

(3) For each empirical variogram plot below, would you (a) use more or less bins and/or (b) increase or decrease the maximum distance (i.e., the range of the x-axis)?

(i) increase max dist? (ii) more bins (iii) fewer bins







(4) In layman's terms, why do you expect the variogram to increase with d?

Thinks that are closer together are more similar and vice versa

(5) In layman's terms, what does it mean if the variogram is the same for all d?

Data are random across space.