Spatial Statistics and Spatial Econometrics

Prof. Gaurav Arora
Department of Social Sciences and Humanities
Indraprastha Institute of Information Technology, Delhi

Lecture - 9A
Spatial Autocorrelation Implications for Inference - II

Hello everyone, welcome back to the 9th lecture on Spatial Statistics and Spatial Econometrics.

So, in this lecture, we are going to do a quick recap of a very important topic that we started you know last time which was the implications or the consequences of Spatial Autocorrelation for statistical estimation of mean and its confidence bands right. So, very quickly since it is a very important topic you know what we started with was an iid sequence of value Z_1 till Z_n.

So, we have a sequence of data points from Z_1 till Z_n and they are independently and identically distributed sequences of random variables from a Gaussian distribution having unknown population mean mu and variance sigma squared ok. And we said that the minimum variance unbiased estimator of mu was nothing but Z bar, so we said that Z bar is our best guess of what you know the population mean would be.

And then we also said that look once you construct this Z bar it is a function of these random entities Z_i's and hence Z bar is itself a random variable. So, if the Z bar is a random variable then it must have a possibility of error that is to say that as a statistician I cannot say that you know I have known the Z bar with 100 percent certainty.

There is going to be some error and some uncertainty around what this Z bar value can take. And on your screens, you can see that if the Z bar is distributed normally because it is a function of the standard of normally distributed random variables Z_i’s, so Z bar is also normally distributed. Its population mean is mu right that is the expectation of Z bar which is what we are using to you know get a sense of a best guess of mu itself.

And its variance is sigma squared by n. So, in the previous lecture, we established that if you have iid random variables which are themselves you know normally distributed with variance sigma squared then the sample mean has variance sigma squared by n. So, it has a smaller variance than each random variable by itself.

Then we can write out the 95 percent confidence interval for the Z bar that is you know we are 95 sure that the Z bar will lie between this interval that is Z bar minus 1.96 sigma by root n and Z bar plus 1.96 times sigma by root n right. So, what we are saying is that we have a sequence of values Z_1 to Z_n you know distributed across a real number line.

Here is a real number line right and given these values we have some distribution normal distribution that is a bell curve which each Z_i follows. So, we have Z_i's we have Z_1 say Z_2, Z_3, Z_4, Z_5 so on and so forth ok. So, the Z bar is nothing but the mean of these values, so it is going to be you know around the central tendency of these values. Now, the Z bar can take you to know values which are in this range of Z bar minus 1.96 sigma by root n and plus on the right-hand side of Z bar 1.96 sigma by root n right.

So, what we say is that the Z bar will lie in this middle area with a probability of 95 percent ok or 0.95. This is the case for the iid distributed sequence of data. So, this is for iid very important.

And then we sort of relax this assumption of iid and we say all these Z is are not independent of each other, they are somehow spatially dependent. That is to say that if I am able to observe if I realize Z_1 is the Z value of Z at location 1, then I can know something about Z_2 right because Z_1 and Z_2 are dependent on each other and more specifically they are spatially dependent on each other right.

So, they are co varying together. So, covariance is an indicator or a measure of a linear relationship. So, we have this structure of spatial dependence covariance Z_i comma Z_j is equal to sigma naught squared times rho to the power of modulus i minus j. So, this modulus absolute value of modulus of or the modulus of i minus j is nothing but a metric of distance a measure of distance between these locations i and j.

Remember covariance is the way the measure to which or the degree to which Z_i and Z_j move together which is what covariance is telling me is a function or is a virtue of their respective locations and not the values by themselves right? So, the covariance structure only depends on the locations between and the difference of these locations i and j right.

So, this is a very specific, rather restrictive form of spatial dependent structure, but a very good point to start with ok. And also we are saying that look these are positively correlated that is rho is between 0 and 1, if rho were a negative number then these numbers you know these entities would be spatially negatively correlated ok.

So, if I see a very high realization at a particular location on the real number line it's quite likely that you know the neighboring values are also likely to be higher values rather than you know values on the lower side right. So, what we are saying is if we want to sort of visualize the problem we again have our real number line and we have our you know Z_i's distributed across this real number line you know we can just say Z at location 1, Z at location 2, Z at location 3, at location 4 keep going till Z at location n ok.

So, once I have these you know these values the spatial dependence structure is telling me that once I get to know Z_3 there is some dependence of Z_3 with Z_4. So, whatever value I realize that location 3 will have a dependent spillover on Z_4 and also a dependence spillover on Z_2 ok.

Because 2 and 3 and and 3 and 4 are equidistance equidistant that is the unit of distance between these two is 1. The spillover degree of spillover is going to be exactly the same that is what the covariance structure is telling me right or the spatial dependence structure is telling me right?

The deep spillover of Z_3 on Z_5 and Z_1 will be slightly smaller because now we will have rho you know raised to their to the power 2 right? So, rho now has a power 2 rho is between 0 and 1. So, rho squared is certainly smaller than you know rho ok. So, the spillover the degree of spillover that is retained from locations that are farther away from Z_3 is going to become smaller and smaller, and rather it is going to be exponentially smaller because the distance is on the exponent of rho right alright.

So, that is, what I believe is well understood. So, with this correlation structure, I have just written down some of the covariances between different locations as examples covariance of Z_1 and Z_1 which is nothing but the variance of Z_1 again something that we studied in the last lecture right?

And then the covariance of Z_1 and Z_2 is rho sigma naught squared rho to the power 1 that is 1 minus 2 absolute value right and then you have covariance Z_1 and Z_3 like I said when 3 is 2 positions away from 1 the degree of spillover, is going to be of the order of rho squared because you know the exponent is simply the absolute value of 1 minus 3 and so on, right.

So, this is how my structure is given. By the way, this also means there is one more piece of information here that not only 3 is going to have a spillover on 1; 1 is also going to have a spillover on 3 and just like that 1 is also going to have a spillover on 2; 2 is also going to have a spillover on 1 and so on and so forth right?

So, in space when we do spatial analysis every location impacts what is happening at every other location, and on the other hand, every location is also impacted by what is occurring at all other locations in the domain of our study right? How would these you-know impacts be defined, what will be their measures, and how strong they will all depend on the spatial dependence structure as specified mathematically?

So, these devices' spatial dependent structures, this is a device covariance offers the covariance operator offers a device to specify the spatial dependence structure ok, alright.

Just a little note before we sort of move on to the conclusion of this exercise that this spatial autocorrelation structure is quite similar to the autoregressive time series or serial correlation structure of first order which is called the AR1 structure, right?

So, I mean this is not very surprising if I were to take these indices 1 to n instead of space if I were to take them as time right. So, if I take time equals 1, 2, 3, 4 keep going till N or I could use the notation t it's just notation right? all I am saying is I am realizing values at each time period t right?

So, then I have a timeline which goes from 1, 2, 3, 4 all the way till N and I am able to specify random variables at each location that are which will have realization Z_1, Z_2, Z_3 and so on and so forth. The autoregressive structure with the with of order 1 which is called AR1 basically specifies that every value that we observe at a given time period has a spillover from the previous period like.

So, for example, habit persistence is right. So, if we look at the credit card spending of an average you know credit card user then it is quite likely that the amount of spending that we observe today a lot of it would be explained by what they were doing in the previous month or in that month in the previous year right.

The kind of consumption pattern that we would observe for someone in a cafeteria will a lot of it will be explained by you know what they were consuming in the previous visit to the cafeteria. So, the idea of the AR1 structure is that there is this one-period lag spillover that happens every time period as we sort of as time moves on, right?

So, we have a similar thing going on with spatial. However, in the case of spatial as we have seen that the spillovers can also go in the other direction which is not the case with time series right? So, what happens at time period 2 cannot impact what happens at time period 1 because time period 1 came before the time period, right?

So, the spillover impacts are unidirectional they can only happen in a direction that is a fundamental difference between time series analysis and spatial analysis. Another fundamental difference between spatial data and time series data is that when we talk of forecasting right? so if this were a time series forecasting exercise, then a typical query would be what is the value of Z at time period 1?

So, you know for a central bank for finance machinery of any country or a state the idea is can you forecast what would be the GDP growth in the subsequent time period, right? So, I have observed the GDP of India from you know right from independence to let us say 2021 or 2022 and I want to now predict what will be the GDP growth rate for 2023. So, I have a time series of data and I want to predict what will happen in the next step right?

In the case of spatial analysis, we do not necessarily have to hop on one unit right. Spatial forecasting is as good as to be done at a point like let us say 3 by 2. Any location on the real number line is a valid forecasting point; obviously, including N plus 1 right? So, that is the difference between spatial data and time series data, right?

If the time series started at Z_1, Z_0 we cannot go in retrospect and say what was the GDP of India in 1945 using the time series from let us say 1948 or 1950 right? It is not possible to go back, but in the case of space you know that is a valid query right that is the difference between spatial analysis and time series analysis ok.

So, we showed last time that under spatial correlation now what changes is the variance of the Z bar. So, the Z bar remains the consistent estimator of mu, but the variance of the Z bar is no longer sigma squared by n.

Rather it is a value that is sigma squared by n times 1 plus rho over 1 minus rho. And in the previous lecture, we also said that we could set this value n 1 minus rho over 1 plus rho equals some value n prime. So, we could sort of re-specify this variable into a more concise n prime or n tilde variable and then sort of try and contrast between the variance of the Z bar in the iid case and in the case when we have spatially dependent data, right?

So, we said what we found was this n prime which is the new effective strength of data if we were to represent the variance of the Z bar as it appears in the iid case, then this n prime will be effectively smaller than n. And I said in the previous lecture that what this means that because of spatial dependence in data, the effective or net information available to conduct our analysis is smaller than you know the unique data points or unique points where you have realizations.

Because all these points are to a degree dependent on each other, so they effectively contain an information set that is smaller than the quanta that you see in terms of the number of realizations that is n right? This is very interesting, but also quite intuitive ok.

So, we learned that there is bias in the variance of the Z bar if we ignore spatial independence, and indeed the data that you know exhibits that kind of structure. Of course, if there is bias and variance in the Z bar there is also bias in the standard deviation of the Z bar you know which is nothing but the square root of the variance of the Z bar, right?

So, if there is bias in the variance of the Z bar then there is bias in the standard deviation of the Z bar. And can we quantify this bias? Of course, we can quantify this bias. Well, this bias will be nothing but the true value of you know variance of Z bar is 1 plus rho sigma squared times 1 plus rho over sigma n times 1 minus rho minus what you would have thought if you took the data to be iid and they were spatially dependent.

So, if you ignore spatial dependence this value here is the quantum of bias that you are going to have to work with right. And in the other case in the case of standard deviation well it will be the square root of sigma squared 1 plus rho over n times 1 minus rho minus the square root of sigma squared by n right. So, this is the bias in the standard deviation of you know offset bar.

This then leads to the bias in the 95 percent confidence interval right and what we said last time is because n prime is smaller than n the confidence interval the 95 percent interval where in which I believe Z bar or the mean will lie for this given population this confidence interval is larger right.

So, to be able to attain you know confidence to the extent of 95 percent out of 100 times the Z bar will lie in this range that range will become slightly larger right? So, there is a little bit more error in the data and the reason is that you have an effectively smaller set of information because of all these dependent values.

And it will turn out that as the value of rho increases that is spatial dependence increases, n prime which is the effective size of data set independent data points will become smaller and smaller and the variance of Z bar will become larger and larger and hence the confidence interval will also become larger and larger alright.

So, you guys can sort of you know you guys should do some kind of a thought exercise on this you know before you move forward. A 2-minute thought exercise will help you understand this concept better ok alright. So, let us move forward.

So, let us try and do an example or an exercise to understand the concept of the dependence on the role of this bias of invariance of Z bar and the confidence interval of Z bar if that is a sample mean you know for a given sequence of data, size of the data set and a given value of rho. So, let us read the question. Consider a sequence of 10 data values. So, now, I have n equals 10 which is the apparent quantum of information, I have values observed or realized at 10 locations in space.

Z_1 till Z_{10} indexed by their location on a real number line ok. All these Z_i's it is not an apostrophe sorry that all these Z_i's exhibit a spatial dependent structure which you have already seen before and have an understanding of from earlier and now we are given rho is about 0.3 ok. Now rho could go between 0 and 1 if they are positively spatially correlated, right?

So, we still have a positive spatial correlation and we have rho equals 0.26 and we are asked to evaluate an equivalent number of observations that would exhibit spatial independence given the 10 spatially dependent observations having rho equals 0.26 ok. And this spatial dependent structure which is covariance Z_i Z_j is equal to sigma squared times rho to the power absolute value of i minus j right. And i and j are simply locations of data index in this on this real number line and i and j can only go from 1 and 10; 1 to 10 ok.

And the second query is to evaluate the spatial autocorrelation impacts on 95 percent confidence bounds for Z bar, ok. So, we have a pretty good understanding of these things now, so let us look at the first question. The first question is basically asking you what is n prime, what is the effect you know amount or quanta of data that would be independent spatially independent would bring to the table you know that these 10 values would bring you to know all together at once.

So, we have we know that for question 1, n prime is nothing but n times 1 minus rho over 1 plus rho ok. So, I have n as 10, 1 minus 0.26 over 1 plus 0.26 which if you solve you will find that it is equal to 6 ok. So, there is an equivalent of 6 independent observations that this data set you know provides as far as the information that it brings to the table to estimate the unknown population mean mu ok.

Of course, we know that the Z bar will still be summation you know i equals 1 to 10 Z_i divided by 10, but the variance of the Z bar will be sigma squared over you know n prime which is 6. So, we have sigma squared over 6 ok and that implies the standard deviation of Z bar; the standard deviation of Z bar is equal to sigma over square root of 6 that is n prime right? and this value will sort of or you know.

So, this value is what is going to be used to now construct the 95 percent confidence bounds for the Z bar. So, the 95 percent confidence bounds for Z bar are given as Z bar minus 1.96 into sigma over square root of 6 comma Z bar plus 1.96 into sigma over square root of 6 ok. What would be the confidence bounds for the iid case? Ok.

So, for the iid case, we will have Z bar minus 1.96 sigma over root 10 times Z bar plus 1.96 times sigma over root 10. And if you were to sort of if you want to contrast the iid case with the spatially dependent data set that we have here, then you can rewrite the confidence bounds in a more convenient form. You can say Z bar minus 1.96 times 1.3 into sigma over root 10, comma Z bar plus 1.96 times 1.3 sigma over root 10. This is 1.3 here. I suppose you will understand this is 1.3 ok alright.

So, what this really means is that the confidence bands are 30 percent longer on each side from the Z bar if you were to transition from the iid case to the spatially dependent case as given in this question, and remember this is because of the value of rho. If the value of rho increases if it were to increase you would have seen an even greater you know extension of the confidence bounds.

So, just to visualize it very quickly what we are saying is that we are given this data on the real number line. Data are all Gaussian in nature, right? So, they are all following this Gaussian distribution. In one case they are spatially dependent another case there is spatially independent, right? So, I have different data points Z_1, Z_2, Z_3, Z_4, Z_5, and $Z_{6,}$ and keep going all the way till Z_n, here it's 10.

So, n equals 10 right what we are saying is that you know the consistent estimator of the Z bar will be the same both for the iid case and also for a spatially dependent case right? So, we are saying that they will be the same for both cases right? What will differ though is the variance of the Z bar and by extension, the standard deviation of the Z bar and the confidence bounds of the Z bar. So, for the iid case, I would have to go 1.96 times the standard deviation.

So, Z bar minus 1.96 times sigma over root n and on the left an equal distance on the right that is 1.96 Z bar plus 1.96 sigma over root n. So, my confidence bounds for the iid case are in black ink which is the 95 percent confidence interval for Z bar ok, and the hashed area under the bell curve ok. And if I had instead I had you know I had spatially dependent data what happens is that my bounds sort of shift rightward on the right-hand side by a factor of 1.3. So, it is 30 percent higher, right?

So, what I do is I extend the right-hand side bound by 30 percent. So, you know I am going to do it approximately. I am going to extend it by about this much. So, I am going to take it forward and I am going to take it till here. I am going to say this here Z bar plus 1.96 times 1.3 sigma over root n ok, which is 10; so n is just 10. So, I hope you will be clear on that and I will try to go about the same, this figure is not to scale. So, pardon me for that, but I suppose it makes the point ok.

So, now the area under the bell curve that I am representing using the blue ink is the 95 percent confidence interval or confidence bound for you know that there is a 95 percent probability that the Z bar will lie between these two values in blue ink markers on the real number line and the area that you are looking at will now start to sort of provide me a 95 percent probability under the spatially dependent structure. So, that is about it.

So, now, we are going to sort of move on to a two-dimensional case and study you know spatial dependence and its consequences on mean estimator as a next step ok, alright?