Spatial Statistics and Spatial Econometrics

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

Lecture - 8B
Spatial Autocorrelation Implications for Inference - II

Alright. So, now, that we have understood you know when we are given an iid sequence, what are the different things that you know that we can characterize about it, you know we can characterize the sample mean, the sample variance, and there is a concept of degrees of freedom. And then, there is also statistical influence which is very interesting, you know which gave us an idea that Z bar could be an erroneous understanding of you know of mu.

And you may want to know rather than rely on this point estimate, go forward, one step forward, and evaluate a confidence interval which is for a given level of probability or that you are likely to see Z bar values in. So, in the previous lecture, we figured out, a 95% interval; confidence interval for the iid case for Z bar which was Z bar minus 1.96 sigma hat over root n and Z bar plus 1.96 sigma hat by root n. So, the confidence interval was in these bounds, right with the lower and the upper bound, right.

So, now, we generalize this iid case a little bit. How do we generalize it? We introduce spatial dependence to the data, right?

So, we introduce specifically a metric or a measure of spatial dependence which is spatial autocorrelation which we have you know articulated on this slide as the covariance of pairs Z_i and Z_j, for our given data set is given as sigma squared rho to the power i minus j in absolute terms, right. So, the absolute value of i minus j is an exponent, and i can go from 1 to n, right?

So, now, this gives us two distinct cases, one is when i is equal to j. That is when let us say we are talking about the covariance of Z_1 and Z_1. Now, the covariance of Z_1 and Z_1 is nothing, but just the variance of Z_1.

This definition says that it is going to be sigma squared rho 1 minus 1 which is sigma squared rho 0 which is just sigma squared, right? And you can evaluate for any I, you know case, where i is equal to j, right? So, I can in general I can say covariance Z_i, Z_i equals the variance of Z_i is going to be the sigma squared root of the power 0 which is sigma squared, for all i goes from 1 to n.

So, the first thing it tells us is that the variance of Z is the same which is sigma squared, right? So, I am still working with an identical variance you know a case. But what is happening is that you know, what happens is that as soon as I have i is not equal to j, right we have let us say we have covariance for example, 1 and 2, we have variance oh sorry it is sorry; it is going to be sigma squared rho absolute value of 1 minus 2 which is just 1. So, we have sigma squared rho.

If I look at covariance Z_1 and Z_3, I will have a sigma squared rho absolute value of 1 minus 3 which is sigma squared times rho squared. And then covariance Z_1 and Z_4 will simply be a sigma squared root cube. And you can keep going covariance of Z_1 and Z_n will be nothing but sigma squared rho n minus 1, right? So, now, what is happening is that I have a distance metric sitting on the exponent of rho, right?

So, typically, we say that rho is between 0 and 1, right? And it is an important condition. We will see at some point in the course later. And we come to the idea of stationarity and all that rho if it is higher than one it can have major consequences on what can be done with the data. Well, we will see that we cannot do much. But anyway, for now, we will just take rho to be between minus 1 and plus 1.

And the idea is that you know when I take any position in space, let us say position 1, Z_1 I can have you know when I move to Z_2 there is some dependence between Z_1 and Z_2 given by a multiple of rho. As I move further away from Z_1, the dependence you know goes down at an exponential rate by rho squared, rho cube, and to rho to the power n, ok.

So, when I am given my data set, right, so when I am given this data set, right; so, the data set let us say the data set can be ordered from Z_1 to Z_n. So, I am going to simply just put the location. So, this is a real space, right? So, location 1, location 2, location 3, location 4, keep going you have till location n, right? And the idea is at each of these locations you have data Z_1, data Z_2, Z_3, Z_4, and n. So, these are realizations of space.

The idea is that the value that I realize on Z1 has a dependence metric that specifies the spillover effect of location 2, the spillover effect of location 3, the spillover effect of location 4, and so on. Similarly, realize that I also have a covariance of Z_2 comma Z_1, which is also sigma squared rho 2 minus 1 which is equal to sigma squared rho. So, the idea is there is also a spillover in the opposite direction for all the values. Of course, as the distance, you know sort of increases the strength of the spillover declines. So, the strength of spatial dependence goes down.

Now, instead of this real space being visualized as you know, you know as some kind of in a spatial sense I could have also thought of it as a time you know dimension. So, in the case of time, you know you had time period 1, you had time period 2, time period 3, time period 4, and time period n. And there also we have this understanding of you know autocorrelation; temporal autocorrelation where you know the value at t_2 has a spillover from t_1. Similarly, t_3 has a spillover for t from t_1, I mean it also has a spillover for t_2 and so on.

The only difference between spatial autocorrelation and temporal autocorrelation is that in the case of time series data, the spillovers do not go back. So, the idea is because in time series, you know time is unidimensional and unidirectional, what happens is that value that is realized that t equals 4 you know cannot have any effect on what happened in t equals 1, because t equals 4 did not even happen then, right.

So, it necessarily will have only the effect will only be one way, that is the value that has happened, in the previous period will have may affect what has what is happening in the current period. But what is happening in the current period it could not have happened, could not have had any impact on what happened in the previous period just by the virtue that it happens on a later date, right? So, this unidirectionality is relaxed in the case of spatial dependence or spatial autocorrelation, right?

So, here the dependence unlike you know in one direction is increasing time periods in case of time series data also comes back. So, every entity has a spatial impact on every other entity with a strength which is given by rho to the exponent of the distance between those you know between those entities, ok. So, having understood this interpretation of this the spatial order correlation structure that we are working with, now let us go forward and you know evaluate different characteristics of this distribution, ok.

So, I am going to say note that Z_i's are still normally distributed with mean mu and variance sigma squared. However, now they or these let us say these entities are no longer independent. So, I still have an identical distribution, but they are not independent of each other, right?

So, what happens at Z_1 is impacted by what happens at every other location and whatever happens at every other location is impacted or is dependent to some degree on you know what happens at location 1. So, this is interesting. This is a very complex structure if you try and visualize it, start to visualize it, right?

And remember we are only working in one-dimensional space right now. If you, we go beyond one dimension and look at set two-dimensions or three-dimensions. You can imagine the math will really become complex and that is the point, right, you know, we can still work with those things in fairly simplistic you know, with fairly simplistic you know tools.

So, that is the point you know that we will sort of see pan out in future you know in future lectures of this course. So, we are working with entities that are no longer independent. And of course, we still care about the mean. So, what is mean, Do you know the sample representation of mean mu? So, I am asking the same question. I am asking the same question that I asked in the iid case which was to say that you know what is the best guess of mu.

Well, you know at the end of the day you know what information I have; I mean my information that I have is I have a sequence Z_1, Z_2 till Z_n. Of course, they are spatially dependent fine, but that is the information that I have. So, my best guess of mu is still going to be Z bar which is nothing but taking every value weight equally by 1 over n and just summing it, ok.

So, again even when you have spatial dependence, the best guess of mu or mu hat is equal to Z bar, right? So, the Z bar is indeed the best guess that data may be independent or spatially dependent. Now, what changes? Right, what changes you know is the variance of Z bar is going to change you know with respect to the iid case, ok. And to evaluate it you know, let us try and move forward, right?

So, let us see. So, we want to, we want to now we want to evaluate the variance of the Z bar because we know it is a random variable, it is composed of random variables. There is an error. I want to articulate that error and that error will come from the variance of the Z bar, and I want to get there, right? We want to evaluate the variance of the Z bar. So, before we do that, what I am going to do is, I am going to sort of you know try and go back to where we started, right?

So, we had articulated the variance of Z bar as the variance of 1 over n summation i equals 1 to n Z i and I said 1 over n squared and we had written this as summation i equals 1 to n, the variance of Z_i, the variance of Zi plus 2 covariances of Z_1, Z_2 plus 2 covariances of Z_1, Z_3, keep going plus 2 covariances of Z_n minus 1 Z_n, right. So, we need these pairs of covariances to be part of this variance definition.

In the case of iid, the covariance terms were all 0s. In the case of spatially autocorrelated data, they are all nonzero. So, this yellow you know the terms in the in within yellow curly brackets are no longer 0. In fact, we are given a definition of those and that is the covariance of Z_i and Z_j, right? So, we know the covariance of Z_1 and Z_2 is sigma squared rho. We know the covariance of Z_1 and Z_3 is sigma squared rho squared, right?

So, what we are going to do is we are going to introduce a variance-covariance matrix which is a very concise tool for encapsulating these covariance terms in you know at one go. So, the covariance of Z_i and Z_j, I am going to write in a variance-covariance matrix. So, I have n, I have n data points. So, I am going to have an n-by-n variance-covariance matrix each. So, I have n rows and n columns. The diagonal elements are going to be sigma squared.

So, each row is representing i's which is 1, 2, 3 all the way till n, and each column is represented by j's which is 1, 2, 3 all the way till n. So, I have n rows and n columns. What happens is that given my definition, so the definition that I have is that covariance Z_i, Z_j is equal to sigma squared rho, the absolute value of the difference between the locations i and j such that i goes from 1 to n and j goes from 1 to n. So, I am going to use this definition and fill in the variance-covariance matrix.

Quite clearly the diagonal elements are going to represent the variance terms when i equals j, right? So, i is 1 here and 1 here, i is 2 here, and it's 2 here, i is 3 here and it's 3 here, similarly i is n here and n here. So, all the diagonal elements are going to be just sigma squared, right? So, sigma squared, sigma squared, I am just going to use dot dot dot, for you to understand that all the diagonal elements are nothing but sigma squared.

Now, let us come back to the off-diagonal limits. The off-diagonal elements are; so, what is this off-diagonal element here? This represents covariance Z_1, Z_2, right. The covariance Z_1, and Z_2, which we have seen previously is equal to sigma squared times rho. What about this element here? It is sigma squared rho to the exponent absolute value of 1 minus 3. So, I have sigma squared rho squared. Similarly, I will keep going, I will have the sigma square root of the power n minus 1, ok.

What about this element here? This is the covariance of Z_2 comma Z_1. And this 2 will be nothing but sigma squared rho given my definition. And I will have Z_2 and Z_3, again sigma squared rho because the distance between 2 and 3 is still 1. So, sigma squared times rho and you keep going you have sigma squared rho n minus 2, right? For the third row which is third i, 3 n 1 sigma squared rho squared, 3 n 2 sigma squared rho, keep going sigma squared rho n minus 3, ok.

And then we keep coming down, ok. So, we have here n and 1, so sigma squared rho n minus 1, sigma squared rho n minus 2, sigma squared rho n minus 3, and keep going, ok. Similarly, you can fill in all of these values and the point that I am trying to make here is that the off-diagonal elements are nonzero. In fact, they are a representation of spatial dependence. By contrast, if you were to consider the case of iid, what would happen? Ok.

So, as an aside we are just going to quickly look at the case of iid, the case of iid. So, in the case of iid if I were to sort of you know get to this variance-covariance matrix for the iid case, I will still have the elements in the diagonal of this you know cells of these matrices which will be equal to sigma squared, sigma squared, sigma squared and again I have i's 1, 2, 3 all the way to n in the rows n j's as columns all the way till n.

But, what is different between the variance-covariance matrix in the case of spatial dependence or spatial autocorrelation as specified with the row parameter, is that while in that case in the case of spatial dependence diagonal elements are nonzero. In the iid case, all of these diagonal elements will be just 0s, so, right. So, we are full we have a much simpler case where everything is 0, right? One last thing I want to point out with the variance-covariance matrix is that it is symmetric, right?

So, you will see, if you see sigma squared rho here you see a mirror of it here, right? If you see a sigma square rho n minus 1, you see a mirror of it on you know a downward side of it as well, right? So, the left and right of the diagonal are symmetric. So, the variance-covariance matrix is always symmetric, right? So, it is sufficient to just write, so write down the once any one side of it, and the other side can be automatically filled in because we know that it is a symmetric tool.

So, with this understanding, now you know with this understanding of how to articulate the variance-covariance matrix, you know ultimately we aim to figure out the variance of the Z bar. And we want to fill in these values if you want to put in the variance terms and the covariance terms and so on and so forth, right? So, what I am doing is, I am summing everything, each cell of this variance-covariance matrix, I am summing it and I am multiplying it by 1 over n square, ok.

So, when you do that, what you are going to get is the variance of Z bar is equal to 1 over n squared, we can also take out sigma squared by the way, right? So, in the previous, you know on the previous page, I could have taken sigma squared out, I could have taken sigma squared out as a common multiplier from each cell of this matrix, right? So, I can just take sigma squared out and I will be left with ones in the diagonal elements and off-diagonal elements will be rho, rho squared, rho to the power n minus 1, and so on and so forth, right, ok.

So, here what we are going to do is we are going to, we are going to also take out sigma squared. So, I am going to have sigma squared over n squared and inside I have n plus. So, all the ones in the diagonal elements are being summed I get a 1 and n. So, this n is nothing but the sum of all the ones in the diagonal elements of the variance-covariance matrix plus twice of n minus 1 times rhos.

So, there is n minus 1 rhos, n minus 2 rho squared, plus n minus 3 rho cubes, and you keep going you have 3, rho to the power n minus 3s, we have 2 rho to the power n minus 2s and you have a 1 singular rho n minus 1, ok. So, remember we have taken this twice, this 2 multiple outs. So, you know all of these are occurring you know twice of times what you are writing here inside the square bracket, ok.

Now, this sum can be you know rewritten as the following. So, I am going to you know take out this sigma squared by n squared multiplied by n plus 2 sigma squared by n squared. Times, you have n times rho plus rho squared plus rho cubed, keep going plus rho to the power n minus 1, minus rho times rho plus 2 rho squared plus 3 rho cube plus n minus 3, rho to the power n minus 3 plus n minus 2, rho to the power n minus 2 plus n minus 1, rho to the power n minus 1, ok. Sorry about that; little fumble there, ok. So, now, it turns out that this first series is nothing but a geometric progression, ok.

So, rhos just increase geometrically as a sequence. So, we have rho, rho squared, rho cube, rho to the power 4, all the way to l to the power n minus 1. So, you have a sum of a geometric progression, right? So, it is a geometric series. And this one which is deducted from it, from this geometric series is an arithmetic-geometric series, right? So, it is a combination of an arithmetic progression and a geometric progression. And it turns out that we have, we can have these standardized you know formulae for summing a geometric progression as well as arithmetic geometric progression.

So, I am going to simply apply these. You can go and recall, I am just going to put a recall, you should go back and check how to write down the sum of a geometric progression. So, the sum of n elements that are you know characterized by a geometric progression. And you should sum, you know you should go and figure out what is the sum of n elements that are characterized by an arithmetic-geometric progression. I am not going to solve these, so you know provide these.

You can check it in any standard you know mathematics text. And you can use this to then write down the variance of the Z bar. You can simplify it as sigma squared 1 over n plus 2 over n rho 1 minus rho to the power n minus 1 divided by 1 minus rho minus twice 2 over n squared rho plus rho squared 1 minus rho n minus 1 over 1 minus rho, minus n rho to the power n, ok; n minus 1. And that is it; which can be then approximated as; let me write it here. So, that it is clearer. Let me write it below, sigma squared 1 plus rho over 1 minus rho times n, ok.

So, now, let us go back, and let us go to the next page, and then contrast between the iid case and the case where we have spatial dependence characterized by the rho parameter.

So, we have a spatial autocorrelation scenario or the case where covariance of Z_i and Z_j is given by sigma squared rho modulus i minus j. And the other case is we have spatial i equals 1 to n and j equals 1 to n, where if we have spatial independence, right, we have seen these two cases which are nothing but the iid case, right? We have seen this and we have just seen both the cases, right, so spatial dependence and spatial independence, so iid and spatial dependence.

Now, in the spatial autocorrelation case, my variance of Z bar, turns out to be sigma squared 1 plus rho over 1 minus rho n, ok. I am just going to write sigma square over n multiplied by 1 plus rho over 1 minus rho. And in the case of the spatial independence case, this variance of the Z bar was simply sigma squared by n. So, to contrast them, what I can do is, I can write a rewrite variance of Z bar in case of spatial autocorrelation as sigma squared over n prime, where n prime is nothing but n times 1 minus rho over 1 plus rho.

Clearly, 1 plus rho which is a denominator that will be higher than 1 minus rho, right, is going to be higher; that means, n minus will be less than n. Sorry, n prime will be less than n. So, what we are seeing here is that the case where you introduce dependence in the data, right, the variance of the Z bar, the way you articulate the error in the Z bar, right?

The error or the variance in this random variable Z bar is using you know a denominator on sigma square which is n bar which is, which is an equivalent of spatially independent values of data, right? So, this is equivalent spatially independent data, ok. So, what happens is because of dependence on data there is a similarity in data. And what happens is that on the net we have only n prime spatially independent values instead of the n values that we are looking at, because there is dependence on data.

It compresses or kind of reduces the data size to an equivalent n prime size which is obviously, less than n in terms of you know evaluating the variance of the Z bar. And what will happen is because the n prime is less than n, a variance of the Z bar will be higher when you have spatial dependence relative to when you did not have spatial dependence, and that is because that says there is going to be a higher error. There is going to be a different error in how good a guess is Z bar of the population mean right?

So, what we are saying is that statistical inference right?, so this will imply that statistical inference, the right, the statistical inference was the bound of Z bar values, the bound of Z bar, the confidence interval of Z bar will now be different, right? So, now, you know will be different than in the iid case, right?

More specifically, you know in the iid case the 95 percent confidence interval for the Z bar was given as Z bar minus 1.96 sigma hat over root n comma sigma bar or Z bar plus 1.96 times sigma hat over root n and you have a lower upper bracket round bracket. In case when you have spatial dependence, such that covariance Z_i, Z_j is equal to sigma squared rho i minus j, for i 1 to n, j 1 to n. And I am writing it again and again because I am talking about this specific example. This is not the most general answer. This is for this example so that we understand the consequences one step at a time.

So, spatial dependence which is given as covariance Z_i and Z_j, you know it gives me you know where the 95 percent confidence interval for Z bar will be given as the lower bound Z bar which is the same as the iid case minus 1.96 sigma hat over root n prime, ok. This n prime is lower. So, I have a bigger bound on the, I have you know a smaller lower bound than in the case of iid. And a higher upper bound than the iid case, right?

So, I am trying to draw your attention to the point that you have an n prime in the denominator which is different from n, and the n prime is less than n, right? So, where n prime is strictly less than n, ok.

So, now you know I am going to end this lecture with a small exercise. So, I am going to say class exercise, and then we will end this lecture. We will come back and review this in the next lecture. You know because it is been a little bit of an involved material today. So, say you have a sequence of data, sequence of data Z_1, Z_2 to Z_n, where n is given as 10, ok n is given as 10. And there is a spatial autocorrelation correlation with the same structure as our example in this class, but with rho given as 0.26, ok.

We have to evaluate an equivalent of independent observations for the given case, right, the given spatial dependence case that is I am asking you to evaluate n prime and consequently the new 95 percent confidence bounds for Z bar. And I want you to sort of go over this exercise and then we will evaluate it, we will solve it in the next class.

Thank you for your attention.