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

Lecture - 9C
Spatial Autocorrelation Implications for Inference - Monte Carlo Stimulations

Alright. So, welcome back to lecture 9. Today we are going to sort of cover this lecture in the third part, where I am going to talk about what is called Monte Carlo Simulations.

Monte Carlo simulations are a popular tool for attaining numerical solutions for problems that might otherwise have analytical solutions. And, we are going to cover this topic on Monte Carlo simulations for specifically assessing statistical inference under spatial dependence.

So, that would mean that I am going to sort of specify or talk about a computational algorithm or an experimental design that can be implemented on a personal computer which can provide us you know the results that we have derived in the previous lecture for statistical inference under spatial autocorrelation or a given spatial dependence structure alright.

So, a little bit of introduction on Monte Carlo simulations. So, Monte Carlo algorithms or Monte Carlo experiments are computational algorithms that employ randomness and repeated sampling to arrive at numerical solutions. So, critically I am pointing out that Monte Carlo simulations will exploit stochasticity right. So, they will exploit randomness in terms of designing and experimenting and they will then also rely on repeated sampling right? So, you are going to be somehow sampling again and again from a parent distribution ok.

And, the example that we covered in the last lecture is of a circular city. We can establish whether or not the mean estimator and the confidence bounds, the 95 percent confidence bound are indeed consistent estimators in the presence of spatial dependence using simulations right? So, that is the sort of power here. And, these methods provide an alternative to the analytical framework that yields a deterministic solution right?

So, there you know we had a when we said that you know the mean of prices in a circular city are n plus 1 over 2, we did not say that they could be n plus 1 over 2 plus minus something right? So, we kind of had an analytical solution that the mean of prices for the circular city is n plus 1 over 2, where n is simply the number of homes in this city right? So, we are let us begin looking at the steps in Monte Carlo simulations.

Before we do that, you know I just want to sort of give you a very quick recap of this practice problem, that we have worked with where we introduced the circularity. And, we solved for the mean estimator, the variance estimator, the standard deviation estimator, and the 95 percent confidence bounds for home prices, you know spatially distributed on a circle with a circumference of K units ok.

So, here you know if you just a quick reminder we are working with first-order neighbors on a circle. So, we had you know entities P_1, P_2, P_3, and P_4 which are you know denoting both prices as well as location. So, the index on the price P is the location. So, location 1 location 2 location 3, and the corresponding price P_1, the corresponding price P_2, and P_3 respectively on these locations; all the way till location n right.

And, we are given a specific spatial dependence structure, we talked about this in detail. You can refer to the previous lecture for details of what this means.

And, for this particular scenario, we estimated you know population mean, what would be the best guess of a population mean? We said it will be a P bar, we came up with an analytical solution of a P bar that is n plus 1 over 2. We came up with an analytical solution of variance of P bar which was sigma squared over n times 1 plus 2 lambda. Again, all of this was done step by step in detail in the previous lecture.

And, then we had you know the estimator for a standard deviation of P bar, and then we had the confidence bounds right. So, now, today's lecture, in this lecture we are going to try and you know arrive at these estimators using simulations rather than mathematical derivation or analytical derivation. So, we are going to arrive at them numerically is what the agenda of this lecture is ok.

So, let us go over the steps in Monte Carlo simulations and through this also learn what are Monte Carlo simulations. So, step 1 says to start with a true model that employs a stochastic component to specify the given spatial dependence structure. So, first of all, we say start with a true model, a true model of what? Of the price process on this circular city right. So, the true model for P is right, for each P_i we need a model that helps me arrive at this value of P_i.

This model should employ a stochastic component right? So, there will be some kind of random number that will be used or a random variable that will be used to specify a P_i. So, we can sort of you know say that let us say it is u_i with you know PDF f of u right? So, when we say stochastic component means a random variable right and I am just you know specifying it to be u i. And, then it is supposed to be able to explain or encapsulate the given spatial dependence structure.

So, when I say the given spatial dependence structure, what I am talking about is that prices for first-order neighbors are correlated ok and the degree of correlation was specified with this parameter lambda right? So, we need to parameter lambda right? So, this is the variance-covariance matrix or the correlation P_i P_j structure, that was specified in the practice problem earlier ok.

So, let us try to do that ok. So, we are going to say let P_i equals i plus u_i. Now, i is simply the index which is specifying the location as well as giving it a deterministic value. So, i by itself is the systemic or deterministic component of price. What do we mean by deterministic component? What it means is that if I am standing at location 1, I know that the level of price that I am looking at will be closer to 1 plus a random variable right?

And, you know if I am standing at location 10, then I have an index of what level of the price I am looking at, if I am standing at location 100, at location 1000; I am giving a deterministic level to prices through this systemic component i. And, then to this systemic component, I am then adding this random component, a random or stochastic component of price right. And, now next the final condition for this you know stochastic component is that it should specify a given spatial dependence structure right?

So, we are going to write u_i as lambda over 2 epsilon i minus 1 plus lambda over 2 epsilon i plus 1 plus epsilon i such that epsilon i's are iid. So, independently and identically distributed with you know according to a normal distribution with mean 0 and variance 1. So, epsilon i's are you know iid random variables which are which behave according to a standard normal distribution ok.

And, now we have to check you know we should validate or check that this indeed you know specifies the given spatial dependence structure right? So, our model our true model is all these equations combined right? So, we have P_i equals i plus u_i, where u_i equals lambda over

2 times epsilon i minus 1 plus lambda over 2 epsilon i plus 1 plus epsilon i right? So, intuitively you can see that I am using these epsilon i's to sort of specify spatial dependence.

So, the error component or the stochastic component at location i depends on their component at location i minus 1. And, on you know error component at location i plus 1 right? So, I am bringing in the spatial dependence through the stochastic term. But, I need to now make sure or validate that this given structure is indeed enough or sufficient for me to ensure that it follows first-order correlation with the degree of correlation being given by parameter lambda.

To do that I am going to sort of say check that covariance of Pi and Pi plus 1 should equal lambda and the covariance of Pi and Pi minus 1 should also equal lambda right? So, let us do that ok. So, the covariance of Pi and Pi plus 1 I am going to expand this, I am going to say covariance. So, Pi equals i plus lambda by 2 epsilon i minus 1 plus lambda over 2 epsilon i plus 1 plus epsilon i comma. And, then Pi plus 1 is i plus 1 plus lambda over 2 epsilon i plus lambda over 2 epsilon i plus 2 plus epsilon i plus 1 ok.

Now, i and i plus 1 are deterministic components. So, if I look at a deterministic component and think about the covariance of this deterministic component with all the other you know deterministic and random components, they are going to be 0 right? The covariance of epsilon i minus 1 with i plus 1 is 0, epsilon i, it is going to be 0 because remember epsilon i's are iid right. So, they are independent, i minus 1's are independently distributed of is right, epsilon is right.

So, this covariance is going to be 0 with epsilon i plus 2 will again is going to be 0, and then epsilon i plus 1, which is again going to be 0 right? For so, there is no contribution to the covariance of P and P_i and P_i plus 1 from either the deterministic term i or this lagged spatially lagged you know epsilon i minus 1 term. Now, let us come to the epsilon i plus 1 term. So, I am going to look at epsilon i plus 1 as going to have no correlation or no covariance with the deterministic, you know they cannot co-vary because the deterministic component does not vary at all, right?

And, then i plus 1 with epsilon i the covariance will be 0. Again, because i plus 1 is independent of epsilon i. Second, i plus 1 with i plus 2 is again going to be 0 because they are iid; finally, you have i plus 1 and i plus 1. So, I found my first you know component of covariance that is indeed going to sort of contribute to the covariance of P_i and P_i plus 1. So, that is the covariance of lambda epsilon i lambda over 2 i plus 1 i plus 1.

Now, I am going to come to the last term which is epsilon i. So, epsilon i with the deterministic component has 0 contributions to the covariance. Epsilon i and epsilon i, yes we will have you know a covariance contribution right? So, we can write covariance lambda over 2 epsilon i comma epsilon i. And, then with epsilon i plus 2, I have 0 contributions because they are iid, and with epsilon i plus 1 again 0 contributions because they are iid.

Now, if we come back to this right-hand side, we have a covariance of lambda over 2 with epsilon i plus 1 and epsilon i plus 1. So, this here I can bring out this you know lambda over 2 which is a constant will just come out and I will have covariance epsilon i plus 1 comma epsilon i plus 1 which is nothing but variance epsilon i plus 1. And, similarly, you will have lambda over 2 variance epsilon i.

Now, looking at you know the fact that the variance of epsilon is simply you know 1 for all i. So, we have the covariance between P_i and P_i plus 1 will simply sum to lambda. Similarly, you can show, and you should at your time show that the covariance of P_i comma P_i minus 1 will also be equal to lambda. So, indeed what is going on now, what is happening now is that we have a true model.

We have a true model, a true model of home prices that exhibit the spatial dependence structure on a circular city right? So, this spatial dependence structure is the given spatial dependence structure right? We should be able to sort of you know follow or specify the structure that we are working with right? So, we are indeed able to work with. So, thumbs up to step 1 ok.

So, let us move on to step 2. Now, it says drawing random components of the size of the circular city is n ok. So, for example, for this example, I am simply going to say n equals 100. So, what I am saying is that there are 100 homes located in this circular city right? For the purpose of understanding, there are 100 homes you know located in the city. So, this is P_1, P_2, P_3, P_4, P_5, P_6, P_7, P_8, keep going till P_{100} and just before P_{100} we will have P_{99} ok.

It says to draw the random components of the size of the circular city; that means, I have to go back to the model that I specified earlier. So, I am going to do that, I am going to write down the model at the top of your slide here. So, I have P_i equals i plus u_i, where u_i equals epsilon i plus lambda over 2 epsilon i minus 1 lambda over 2 epsilon i plus 1 such that epsilon is are simply iid normal 0 comma 1 for every i ok.

So, what am I doing now? Ok. So, now, what I am doing is I am generating or constructing my own data. I am simulating a data set of prices of these 100 homes located in the given circular city. What that would mean is that I can just go to an Excel sheet right and start populating it as follows. So, I have my rows ok. The first row is i, that is the id, the home id; I have 1^{st}, 2^{nd}, 3^{rd}, 4^{th}, 5^{th}, 6^{th} keep going 99 100.

Then, I have you know epsilon i minus 1 epsilon i, and epsilon i plus 1. Each epsilon i's that is epsilon i minus 1 epsilon i and epsilon i plus 1, are simply normally distributed with 0 mean and variance 1. So, I can simply apply you know a random draw from a standard normal distribution in an Excel sheet which is a coded exercise, it's already canned in the Excel sheet. So, you can go ahead and do that.

But, you will what you are going to expect is you are going to have these you know 0.01 0.03 minus 0.9 you know and so on and so forth. So, you are going to have these random entities generated individually one by one. So, there is no relationship between epsilon i minus 1 epsilon i and epsilon i plus 1. You are simply drawing from a standard normal distribution 3 times right? The first time you know you are doing it for epsilon i minus 1.

The second time you are doing it for epsilon i and the third time you are going to do it for epsilon you know i plus 1 right? So, let us say I get minus 0.23 here as the first row, then I have 0.56 and I have 0.8 and so on and so forth till 100 values. Let us say here I have 0.9 0.8 0.7 and so on and so forth right? So, these are random entities being generated. Of course, when you will take the mean of any of these random entities, they will turn out to be 0 because you are drawing from that distribution right? very very close to 0 right? that is what we should expect.

Finally, once you have done that you have created data; so, basically you have 4 columns that you have generated one. So, the first one is a deterministic column. It is simply you know the value of the id itself, right? So, it is just deterministic you know, when I say i equals 1; I do not mean i is 1 and minus something. I mean i equals 1, it's deterministic right?

And, then I have three different normally distributed independent random variables being drawn which I denote to be epsilon i minus 1, epsilon i, and epsilon i plus 1 ok. Of course, when you know when i is 1 epsilon i minus 1 means 100; here I am looking at the 100^{th} home. And, epsilon i plus 1 means that I am looking at the second home, you know assuming I am going clockwise on the circle. So, using these I can then define P_i as i plus lambda over 2 right epsilon i minus 1 plus lambda over 2 epsilon i plus 1 plus epsilon i ok.

Now, to populate P_i the column of P_i, data in the column of P_i, I must also specify lambda ok. So, in the true model, I will go back to my true model and I am going to specify, I am going to say let lambda equals 0.4 for the Monte Carlo simulations ok. And, then once I have my you know the specification, I am basically saying lambda by 2 is 0.2 and I have data on these values. So, Pi is nothing but 1 minus 0.2 multiplied by minus 0.01 plus 0.2 multiplied by 0.9 plus minus 0.23 which will turn out to be some numerical value right?

So, you have created this data set of P is ok. So, you have created you have simulated. These are simulated data for $P_{i's}$. The beauty of these data is that they follow the same first-order spatial order correlation structure that the practice problem provided us in the previous lecture right? So, now, we have actual values to work with that behave according to the analytical problem that you were solving in the previous lecture.

Step 3: So, evaluate $P_{i's}$ along with mean value, its mean value, variance, standard deviation, and 5th and 95th percentiles. So, we have already in step 2, we collected this data on Pi. So, now, I have i which goes from 1, 2, 3, 4, 5, 6, all the way to 99 and 100. I will also have these values of $P_{i's}$ which I have created in step 2. So, I have constructed data on P_1, P_2, P_3, P_4, P_5, P_6 so on and so forth till P_{99} and P_{100}.

Once I have this data, I can calculate the mean, which is simple. So, I can just say the P bar equals summation i equals 1 to 100 P_i, remember this P bar is a function of lambda and is set as 0.4. It is also a function of n being set as 100 right? So, it is for 100 homes, you could be working with 1000 homes. You could set n equal to 1000, it is not going to be a very difficult exercise to do that, right you can see that right?

So, what I am saying is that my P_i itself is a function of setting lambda equals 0.4 and n equals 100 and this is what it will look like. Similarly, I have sorry it is not P_i, it is P bar. Similarly, I have a variance of P_i variance of you know P_i or P bar that will you know the variance of P_i will be summation i equals 1 to n which is 100, P_i minus P bar the whole square divided by n minus 1 which is 99 right.

Now, this I can call as sigma hat squared P. With that, we know the variance of the P bar will be just sigma hat squared P divided by n is 100 right. The standard deviation of the P bar will be sigma hat squared P by 100 square root. And the 95th confidence bound is nothing but that I order these data order P is in you know descending right, with in descending values.

So, basically, you have the highest value as the first topmost value and then it just keeps going down to the minimum value. So, descending from the max of Pi to the min of Pi. And, then you pick from the bottom you pick the 5th value, the 5th value from min Pi will be the 95th percentile of the prices. And, the 95th value from the bottom is the minima of P_i will be denoted as sorry.

So, I made a typing error here. So, the top one will be P5, the 5th percentile and this will be the 95th percentile. And, the confidence bound for P is will be P 5th to P 95th alright. So, this is a simulated mean, a simulated variance, a simulated standard deviation, and a simulated bound right; that is very important ok. You could, I mean this is for P_i; if you wanted the confidence bound for P bar, I mean you could sort of you know we will see how to do that now next.

So, step 4 is replicate steps 2 and 3 500 times. So, now, we are going to replicate the simulated prices and their averages ok. So, let us see. So, step 2 was that we first construct data on P_i, step 3 was we collect these statistics on Pi ok, and then finally, we create these replicas of P_i right? So, we basically have you know; so, you know the first time we did that we can say that was replica 1. For that replica, we got P bar m, right? we got the variance of P bar m equals 1 sorry.

So, at P bar m equals 1, the standard deviation of P bar for m equals 1 ok. So, we got these values. And, then you know we can create the 2nd replica. The 2nd replica will be we start the process all over again. We have our i's, we draw you know we create this random draw for epsilon i minus 1 for epsilon i, epsilon i plus 1, again construct data on P is and then get our P bar right. So, I am going to get a P bar for the 2nd replica, and for the 3rd replica.

Remember, in each replica n is fixed as at constant 100, and lambda is kept at 0.4. So, we do not change those parameters. All the P bars are simply because we are drawing from the normal distribution with 0 means and variance 1 for the second time, for the third time, for the fourth time and we keep doing it 500 times right. So, we are going to keep doing it 500 times and we are going to collect our data right.

So, we are going to store these P bars for till m equals 500; sorry about that right? So, we have these replicas of the P bar going from the 1^{st} replication, the 1^{st} replica to the 500 replica. The only very very important thing to keep in mind is that for all these replications right? So, throughout all these replications, we fix n equals 100 and lambda equals 0.4. So, this is step 4.

Let us go to step 5. Step 5 says to contrast the numerical estimators with their analytical counterparts. So, now, we have you know the numerical estimators in the sense that if you go back, we have P bars which we have 500 values of P bars right? So, we have an Excel column of m's that goes from 1 to 500. And, then we have P bars which you know go from which are different values of average prices through each iteration of simulated construction of mean.

Once I have 500 means, I can take a meta to mean or some kind of a P bar which is nothing but summation m equals 1 to 500; P bar m over 500 right? So, I can get my P double bar. Similarly, I can get my variance of P bar as nothing but summation m equals 1 to 500 P bar minus P double bar the whole squared divided by n minus 1 which is here, which will be m minus 1 here which is 500 minus 1 that is 499 right.

The standard deviation of the P bar is basically the square root of the variance of the P bar calculated just now. And, the 95^{th} and the 5^{th} percentile will nothing, but I will simply take these values that I have generated from 500 replicas. I am going to order them from the smallest value or to the largest value or the other way around from the largest to the smallest.

Let us work from the smallest value to the largest value right? The value that I am looking at which is at the 95^{th} percentile is the 25^{th} value starting from the smallest value, all the way up to the 25^{th} value that I find will give me P bar 5^{th} percentile, right? And, the 475^{th} value starting from the minima all the way up to the maxima 475^{th} value in this simulation will be my P bar 95^{th} ok.

Now, realize that these estimators, the P double bar, a variance of P bar, a standard deviation of the P bar, and the 5^{th} and the 95^{th} percentiles are nothing but you know they are numerical solutions of you know the mean estimator for you know prices that are spatially autocorrelated. And, all other statistics; the second moment as well as the 5^{th} percentile and the 95^{th} percentile right.

So, these percentiles are simulated confidence bounds for the P bar right. So, as a next step what we are going to do is we are going to contrast them, we can then contrast them with the analytical solutions that we spent a full lecture previously to arrive at right. For example, now I have my numerical solution, I have my numerical solution where P bar P double bar is nothing but summation m equals 1 to 500 P bar indexed with replication m divided by 500.

Against the analytical solution, where I calculated this P bar to be n plus 1 over 2, n being 100; you know I basically have 101 over 2. So, I can contrast whether or not they are closed. And if you conduct this exercise, you will find that they are going to be indeed very very close right and which makes sense. But, one thing that we have to understand is that for both these cases n is fixed at 100 and lambda at 0.4, ok.

Similarly, the variance in case you know the simulated solution variance of P bar is summation m equals 1 to 100 P bar m minus P double bar whole square divided by 499 right? And, in the other case you know it was P bar equals sigma squared by n 1 plus 2 lambda. Remember, sigma squared in our case has been you know fixed to we will have to figure out the variance of P which will be you know the variance of P_i from where P_i equals i plus u_i.

Remember, u_i is a function of 3 epsilon i right which is i plus epsilon i by you know 2 into lambda i minus 1 plus epsilon i plus 1 into lambda over 2 plus epsilon i. So, the variance of P_i which I am going to just quickly calculate here is going to be lambda squared over 4 plus lambda squared over 4 plus 1 which is nothing but 1 plus lambda square over 2. So, the variance of P_i is 1 plus lambda square over 2 lambda is 0.4 right? So, lambda square is going to be 0.0016 right? So, 1 plus 0.0016 divided by 2 which is nothing but 0.0008. So, this is 1.0008 right?

So, I am going to now change this to my analytical solution will be 1.0008. You guys should check my calculations divided by 101 plus twice of 0.4. So, in black we have an analytical solution, in green, we have a simulated solution, right? They should be very close right? If our simulations are correct, they are going to be very close right?

Then, you know I can similarly I can sort of do the same thing for you know the standard deviation of P bar. And, the standard deviation of you knows P bar which is nothing but the variance of the P bar from the analytical solution. And then finally, my confidence bounds are P 5^{th} P bar $5^{th,}$ and P bar 95^{th} that we calculated previously right. We calculated on the previous slide, right?

And, in the analytical case, this was given by. So, the confidence bounds just a second the confidence bounds, but given by n plus 1 over 2 minus 1.96 into sigma over root n 1 plus 2 lambda to the square root comma n plus 1 over 2 plus 1.96 sigma over root n 1 plus 2 lambda square root right. So, these were my confidence bounds and you know they should match the ones on the left-hand side in green ok.

Finally, we can check for the consistency of our estimators. And, what does it mean to check for consistency? That would simply mean that as n approaches infinity, practically it would mean as I increase the value of n, the size of the circular city or the number of homes and the density of the city; what I should find is that the simulated mean, variance, standard deviation, confidence bounds should you know should become closer and closer to their analytical counterparts ok.

So, we have gone through a full-blown you know simulation exercise to arrive at numerical solutions for estimators of mean-variance. So, the first moment, the second moment, the confidence interval; so, to conduct statistical inference, in case we did not want to go through all that math right. So, we can use computational algorithms, and computational simulations to get there.

So, Monte Carlo simulations are very important; you know they are used widely. So, this is a very important topic and I urge all of you to sort of go over it one more time. And, I hope you had fun learning Monte Carlo simulations. See you next time.