Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Center of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 01
Spatio - Temporal Statistics
Lecture - 04
Gaussian Process Regression and Inverse Problems

Hello everyone, welcome to lecture 4 of the 1st module of our course on Machine Learning for Earth System Science. So, like this 1st module is about Spatio-Temporal Statistics and we are in the 4th lecture of it. In the previous lecture we considered the geo-statistical equation and how it can be expressed as some kind of a stochastic model for spatio-temporal processes.

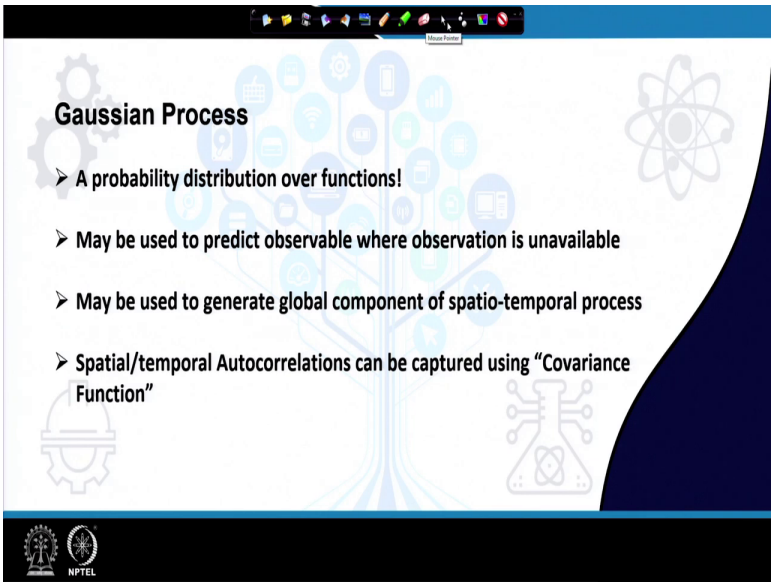
And we came to the conclusion that to incorporate the spatial and temporal auto correlations between different variables we need to use, we need to sample them from some kind of joint distributions rather than sampling them individually from individual locations or time points.

(Refer Slide Time: 01:15)

So, one like we at the end of the lecture we mention that one way to achieve that is through Gaussian process. So, what is this Gaussian process all about? So, in today's lecture we will first

consider Gaussian process regression, then we will see how the spatio-temporal processes as we considered in the last lecture can be modeled using the Gaussian process and finally, how the inverse problem can be solved by using sampling.

(Refer Slide Time: 01:39)



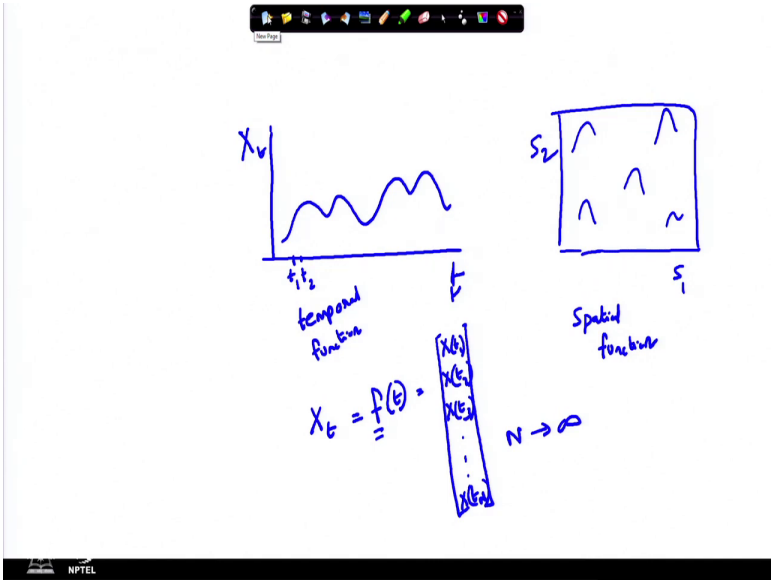
So, first of all let us talk about the Gaussian process. So, this Gaussian process is nothing but a probability distribution over functions. Now that may sound strange because we like we have considered various kind of probability distributions so far. Like for example, the Bernoulli distribution it is a probability distribution over like we can say over only two things the may be the heads and tails of a coin. Or if you consider the Gaussian distribution that is a probability distribution over real numbers where every real number has a probability attached to it.

Similarly, there are other kinds of probability distributions. There might be some probability distributions which are over categorical or like categorical quantities and things like that. Now this Gaussian process is such a probability distribution it is defined over functions, that is, entire functions have a certain probability under the Gaussian process that may sound strange, but actually it is not so strange. We will try to explain that in a bit.

So, now this so what is this Gaussian process used for? This Gaussian process, it may be used to predict the observable at places where the observation is unavailable. What do I mean by that?

So, suppose you are thus so I already mentioned that Gaussian process is a distribution over function. Now what function? So, since we are dealing with spatial and temporal variables we are we will focus only on spatial and temporal functions only.

(Refer Slide Time: 03:16)



So, like if you consider, say, the spatial function like this kind of a spatial like let us consider this kind of a temporal function over t ok. So, this like you can call this as let us say X_t or let us say we have a spatial location like some spatial field. Let us say it is indexed by the latitude S_1 and longitude S_2 and at different parts like we have like different values of this spatial process.

So, the idea here is that we so this kind of a curve it can be called considered as a temporal function. This can be considered as some kind of a spatial function. So, this function is nothing in both cases, is nothing but a collection of values over an infinite number of points. They may be in the in this case, those points are time points in the right hand case in case of the spatial function those points are the spatial points ok.

So, like when we are talking the like this when we are talking about this X_t we like we can call it as some kind of $f(t)$, but in reality it is nothing but a very long vector.

$$X_t = f(t) = \begin{bmatrix} X(t_1) \\ X(t_2) \\ X(t_3) \\ \vdots \\ X(t_N) \end{bmatrix}, N \to \infty$$

So, note the that at every time point at every single time point I have noted an observation. It is not discrete like t_1, t_2 and so on. So, basically it is like the points at which this function is evaluated is uncountably infinite. So, the we can look upon this temporal function f as a vector of the observation X, but that vector is like it is like it has it is infinitely long vector ok.

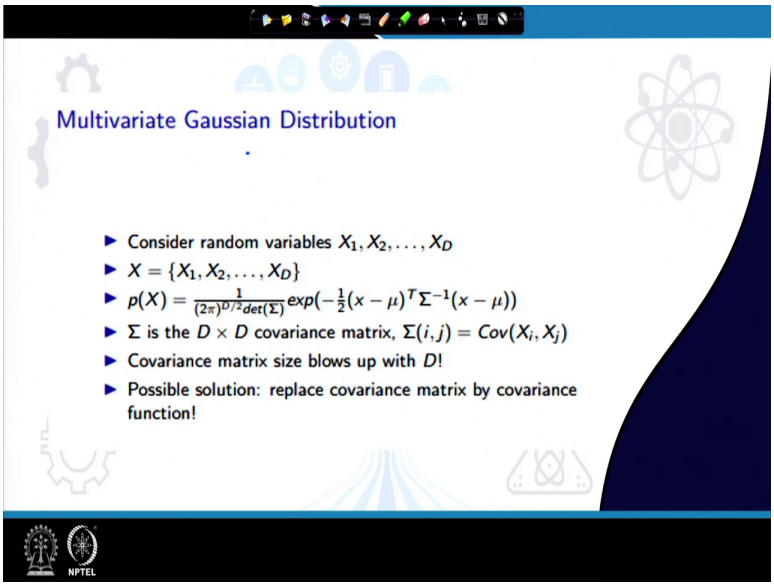
So, now in this like from this point of view you can consider the Gaussian process as some kind of a distribution over the functions. Now suppose you have estimated this kind of a function at different time points; that means what? That means, if you have any new time point you should be able to or you may be able to estimate the value of the observation at that time point. Once you I mean once you have learnt or once you have been able to parameterize that function somehow ok.

So, that is the most important application of Gaussian process where it is basically used for interpolation, that is to predict the value of the observable where the observations is not available. It can also be used to generate the global component of the spatio temporal process. So, in the in lecture 3, we mentioned that this global component needs to cannot be sampled at individual time points or locations, but they has to be sampled from some kind of a joint distribution.

Now, this Gaussian process by being a distribution over the functions its actually what its actually being like something like a joint distribution over multiple time points or multiple locations ok. So, the spatial and temporal auto correlations can be captured using the covariance function. Now this Gaussian process it has a covariance function which we will like we which we will discuss once we come to the formulation of the Gaussian process.

And that the task of that covariance function is to capture the spatial and temporal autocorrelations between the different locations and different time points, that is to say when we whenever we are sampling any observation from a Gaussian process like we if the Gaussian process is basically giving us the values at different time points and at or at different locations, however, those values we will be following some kind of like some relation between the let it will basically it will be it will follow the or it will respect the covariance between those locations or the time points which might be known from already from the data ok.

(Refer Slide Time: 07:53)



Now, coming to the actual formulation of this kind of a Gaussian process, before the Gaussian process, let us first consider the Gaussian distribution. So, the simple Gaussian, the univariate Gaussian distribution we all know that is the standard normal distribution where we have a mean μ and a variance σ.

Now, the same thing can be considered in multiple dimensions also in high dimensions also. So, instead of having a single variable let us consider a collection of D variables like this X_1, X_2, … , X_D ok. And now this all these random variables they can be taken together to form a vector of random variables which we can call as X ok.

So, this vector of random variable is nothing but I mean this X itself can be considered as a vector random variable which is nothing but a vector of individual scalar random variables. Now what we want to do is we want to define the probability distribution on this X, that is the input will be a vector D dimensional vector X and corresponding to that we will get a value of the PDF. So, this is what that PDF looks like.

And so like you might see similarities between the this PDF and the PDF of a standard normal distribution. So, in this case just like the standard normal distribution has its parameters μ and σ this one also has μ and Σ, but the difference is that here this μ the mean is not a real number, but it is a vector, it is a d-dimensional vector same dimension as the input X.

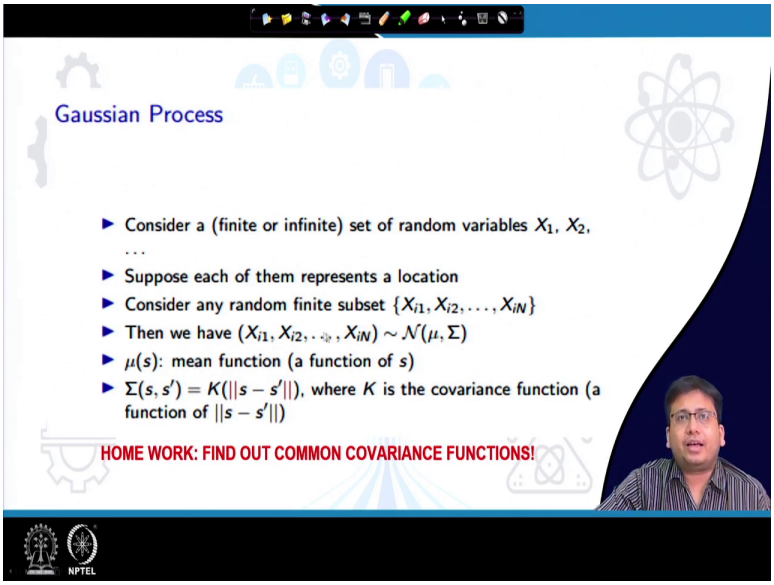
And the instead of having a σ as a I mean a variance, we have a full matrix or $D \times D$ matrix which we call as the covariance matrix. So, what is that matrix what does it consist of? So, if you consider any element let us say the i, j^{th} element of this covariance matrix Σ, it basically contains the covariance between the random variables X_i and X_j ok. So, remember that X is the collection of these D random variables.

So, I take up any two of them. Let us say X_i and X_j I measure the covariance between them and that covariance is like will be present in $\Sigma(i, j)$ of this covariance matrix and you can understand that this is going to be a symmetric matrix because if $\Sigma(i, j) = \Sigma(j, i)$because $Cov(X_i, X_j) = Cov(X_j, X_i)$ right.

And like not only is this covariance matrix it not only is it symmetric, but you can actually show that it is positive semi definite ok. And because of that property you can always define its inverse in this way, it is invertible matrix and you can also and it is determinant will also be non zero. So, this, the problem is that this covariance matrix its size blows up with D that is instead of considering a small number of variables like this, if you consider a very large number of variables, then, this covariance matrix like it will become $D \times D$ that is its size is going to be bigger and bigger. As a result, carrying out these kind of inverses or the determinants is going to become more and more difficult. So, the solution in that case is to replace this kind of a covariance matrix by what is called as a covariance function.

Now, what is the covariance function? So, the covariance function is that instead of having to store the covariance of the X at different locations I am sorry between different variables what if we can somehow express that difference like that covariance using some kind of a function. So, that will happen only is if this like that is we can say that it can happen easily if this the covariance is stationary.

(Refer Slide Time: 12:18)



So, let us explain that in a bit. So, the idea here is that like instead of considering like a finite set of D random variables like this. Let us consider a finite or infinite set of random variables X_1, X_2, .. ok. So, so basically we are relaxing this to the situation where $D \to \infty$. So, we still have the variable X, I mean the vector random variable X, but it has become an infinite vector.

So, theoretically we can still use this kind of a multivariate Gaussian distribution on X, but the problem here is that like we can because like the number of I mean the these the number of these variables has become so large, we can now no longer do the determinate calculate the determinants and inverses efficiently.

Now, let us also consider that these D different variables that we are considering they actually are specific to locations. So, you can say that there is a geophysical variable X which we are

measuring at an infinite number of locations ok, maybe that is temperature or something like that. Now, if you consider any finite random subset of these locations $X_{i1}, X_{i2}, \ldots, X_{iN}$.

So, I mean N is any let us say any integer. So, we what we are doing is we are focusing on only a certain number of these or a finite number of these locations and what we are trying to do is that now we are trying to define some kind of distribution on them. Now, in the like Gaussian process is a model which states that if you consider any random subset of these of such variables we will get the like they are joint distribution we will still be the Gaussian distribution with a mean μ and a covariance matrix Σ.

So, note that the μ in this case will be an N-dimensional vector and the covariance matrix will be $N \times N$ matrix. Also note that I am calling these as $X_{i1}, X_{i2}, \ldots, X_{iN}$. So, you may be wondering what these $i's$ are. So, this is to emphasize that we are choosing any set of N points that is X_{i1} that is if I say $X_1, X_2, \ldots, X_N$ it may mean only the first N points, but by writing $i1, i2, .., iN$, I am emphasizing it that it can be any N points.

X_{i1} maybe 6, X_{i2} may be 13, X_{iN} maybe say let us say 555 and so on. So, like if you consider any subset of N variables out of all these collection of like infinite variables you in like whether you are considering 1 variable or 2 variables or 17 variables, whether they are the 1st variables or the say the 4th and 9th variables or whatever they will all they are joint distribution will always be the Gaussian distribution.

And the now the μ the mean of that Gaussian distribution is the so is the mean function that is it is such that whenever you are considering any particular point in space the its mean is to be expressed as a function of that location. So, that if that location is expressed in terms of latitude and longitude you should be able to calculate some function of that latitude and longitude and get that value μ.

If that is difficult you can simply like consider it as some kind of uniform or constant thing, that is a as a you can consider it as a spatially stationary process, which means that μ is equal at all the locations. The more interesting part is the covariance sigma. So, this Σ this has to be an $N \times N$ covariance matrix. Now every element of this rather than actually storing the all the

elements of the covariance matrix, I will express every element of the covariance matrix as some kind of a function.

And if whenever we are considering the covariance between any two variables who which are like correspond to let us say locations s and s' we will say that the covariance they are covariance follows some kind is a like is a some kind of a function K of the distance between those two locations s and s'.

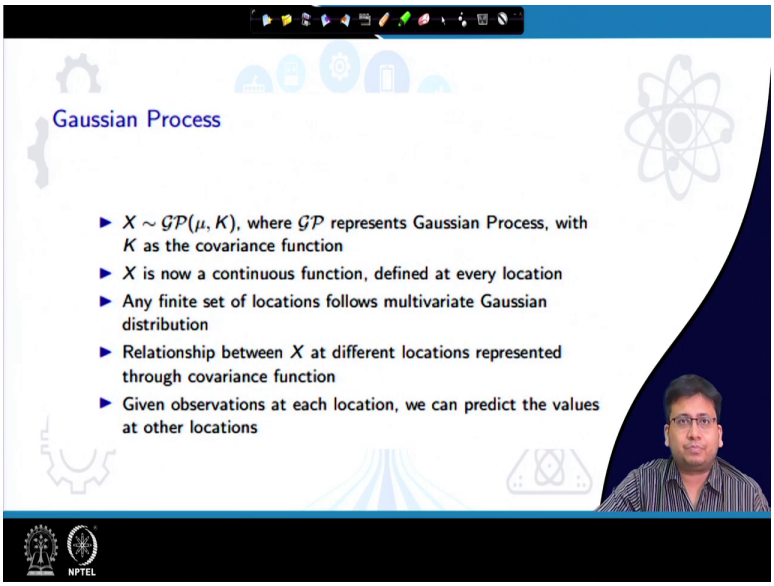
So, you may remember in lecture 2 we considered the concept of is covariance stationarity, which means that the covariance between any two variable or any two locations is I mean the covariance of a particular variable at two different locations is nothing but the function of the distance between those two locations.

So, that is what is happening in this case. So, $\Sigma(s,\ s')$ is basically the covariance between the value a I mean the variable X at location s and location s', but that covariance is being expressed as a function of only $(s\ -\ s')$ that is the distance between the two locations not of any other thing.

$$\Sigma(s,\ s') = K(||(s\ -\ s')||\}$$

So, now the question might arise that what kind of function is K like what function we may use to define this kind of a covariance function? So, there is a set of standard functions for this. So, I leave it as an as a homework for you to find out about common covariance functions ok.

(Refer Slide Time: 18:23)



So, that brings us to the definition of a Gaussian process. So, when we say that X follows a Gaussian process with mean function μ and covariance function K. This is what I mean it means that X is a like is a potentially an infinite dimensional or like an collection of an infinite number of variables such that if you consider any subset of them any random subset of them they follow the Gaussian distribution with parameters which you can obtain from this mean function and the covariance function ok.

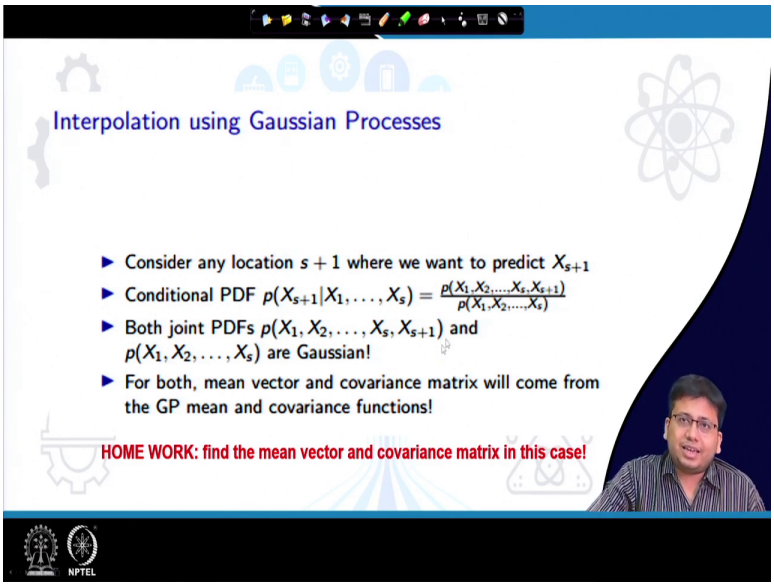
Any finite subset of locations they follow the multivariate Gaussian distributions. Now of course, this need not be locations this can be time also that is in that case it like we in if there are locations then X you can say it is a it follows a spatial Gaussian process. If it is a time points then we can say that it follows a temporal Gaussian process yes and so on.

So, the now the relation between X at different locations or different time is represented through this kind of a covariance function and so ok. So, what is achieved by this Gaussian process? The gauss the what is achieved is as follows: Once you have parameterized this μ and K, now you can like sample the value at any particular point.

So, like as you can understand it will follow the Gaussian distribution only with some parameter μ and Σ which like of course, you can calculate using this the mean function and the covariance function.

So, even if you do not have the observations once you have estimated the mean function and the covariance function then you will be able to generate the value or at any like any location or any time where you may not have had the these observations initially ok.

(Refer Slide Time: 20:31)



So, that is how this Gaussian process is used for the purpose of interpolation. So, to be more specific let us say we have the observations at s locations X_1, X_2, … , X_s. And I want to make the prediction at location $s + 1$ which is the $X(s + 1)$. So, what I will do is I will generate or I will create a conditional probability distribution for a $X(s + 1)$. So, which can obviously, be written in this way like using the standard definition. Now if you consider both the numerator and the denominator both are effectively joint distributions over random variables.

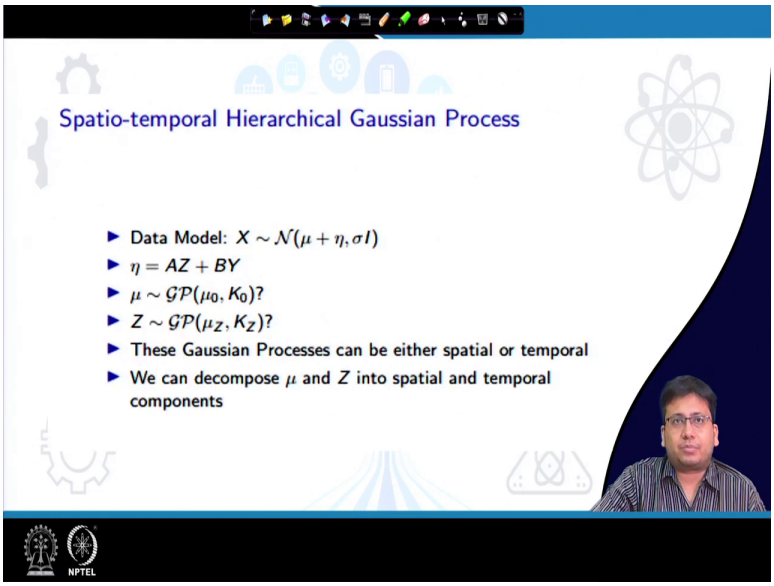
The denominator is a joint distribution over s random variables the numerator is over $s + 1$ random variables. Now because it follows a Gaussian process so any collection of random variables will jointly follow the Gaussian distribution. So, the numerator will also be Gaussian and the denominator will also be Gaussian ok.

So, if the PDF of $X(s+1)$ conditioned on the observation is actually follows this kind of a distribution where like its PDF is like the is some kind of a Gaussian PDF divided by another Gaussian PDF and you can actually work out the algebra and find that after doing this division the resulting PDF is still a Gaussian with the of course, the parameters of that Gaussian will change. I mean the mean vector covariance matrix of this new Gaussian will change compare, but they can, but it will still be a Gaussian distribution.

I mean this the numerator here is a Gaussian distribution, the denominator is a Gaussian distribution and the result the quotient is also going to be another Gaussian distribution. So, what will be the mean and the covariance matrix of this distribution that is I leave it as a exercise for you to find out.

You can like you basically have to write down the PDFs of both of these things do a lot of algebra and you can work it out like. In case it is difficult there are always various resources you can look up to like Christ Bishops book on pattern recognition for machine learning is a very good reference to follow in this case ok.

(Refer Slide Time: 22:57)



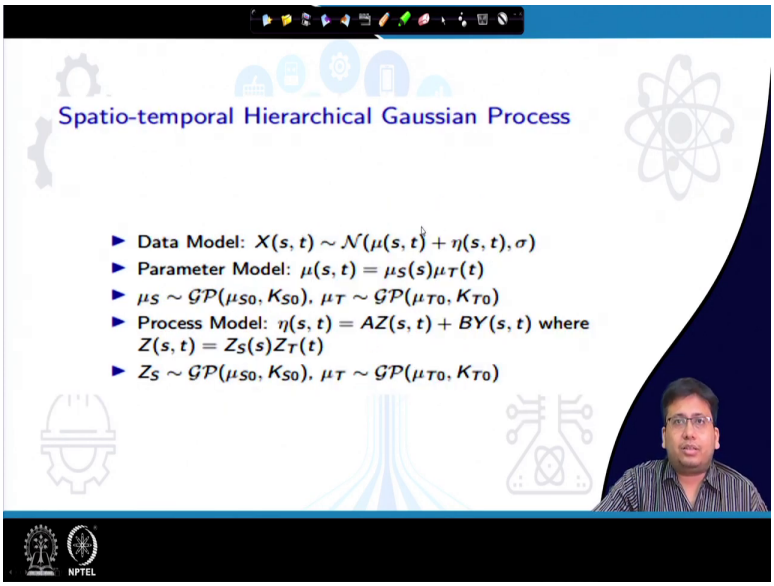
So, now we use the concept of Gaussian process in the spatio-temporal hierarchical process model that we considered in the previous lecture. So, the data model is always there. The Gaussian process is to be used on the process model.

So, if you remember the η is like which is which we called as the global component that can be expressed as $AZ + BY$ where Z is the latent variable at different locations. Y is the set of covariates at the different locations and A and B are the transformation matrices. That is they basically they combine the latent variable values at different locations and pass them on to each location ok.

So, the this μ and these Z these are variables to be defined over all the locations. μ is something like the static component the local component of every location and Z is the latent variable which is also defined over all the locations.

Now I can like in this version I can sample both μ and Z from Gaussian processes. So, what does it help us? It helps us in preserving the spatio-temporal relations between the different sorry not temporal I mean the spatial covariance or correlation between the different locations ok. And these Gaussian processes can be either spatial or temporal and this μ and Z we can decompose them into the spatial and the temporal components somewhat like this.

(Refer Slide Time: 24:38)

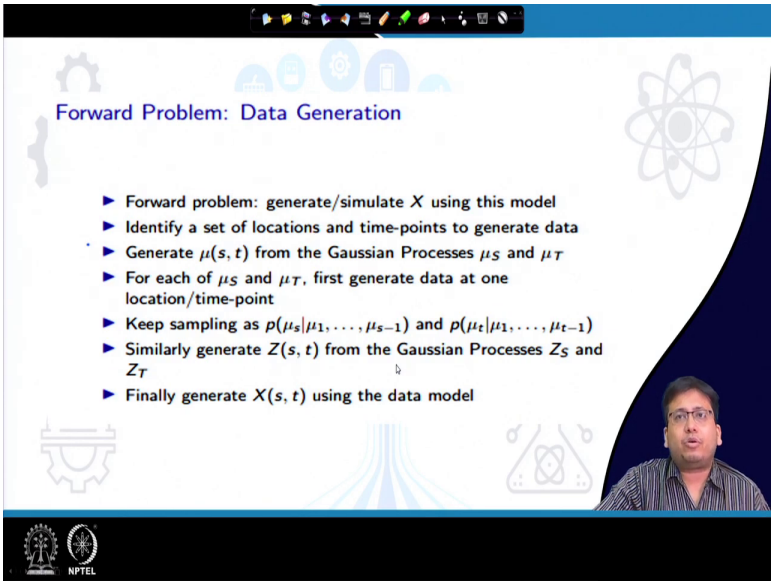


So, let us say that I my model is spatio-temporal that is I am trying to build the model for $X(s,\ t)$. So, it follows the Gaussian distribute at a any location s and time t it follows the Gaussian distribution with mean $\mu(s,t) + \eta(s,t)$and now the $\mu(s,t)$ let us say we have a spatial distribution I mean a spatial component $\mu(s)$ and a temporal component $\mu(t)$ and $\mu(s,t)$ is nothing but a product of these two things.

Now, $\mu(s)$ may individually follow a Gaussian process and $\mu(t)$ may also individually follow a Gaussian process. So, this is the spatial Gaussian process which is defined over locations and this is a temporal Gaussian process which is defined over time points. When I multiply them what I get is a spatio-temporal Gaussian process which is like I mean I get a spatio-temporal variable which is defined for all locations as well as all time points.

Similarly, when we are considering the process model the basically the η we know that η is to be expressed as $AZ + BY$. Now this Z and I mean Y is of course, the exogenous variable which is not a random variable. Now Z is the random variable which is spatio-temporal. So, we can decompose it as thus with as spatial component as well as a temporal component and the spatial component again it follows a Gaussian process. The spatial Gaussian process the temporal component again follows a temporal Gaussian process ok.

(Refer Slide Time: 26:16)

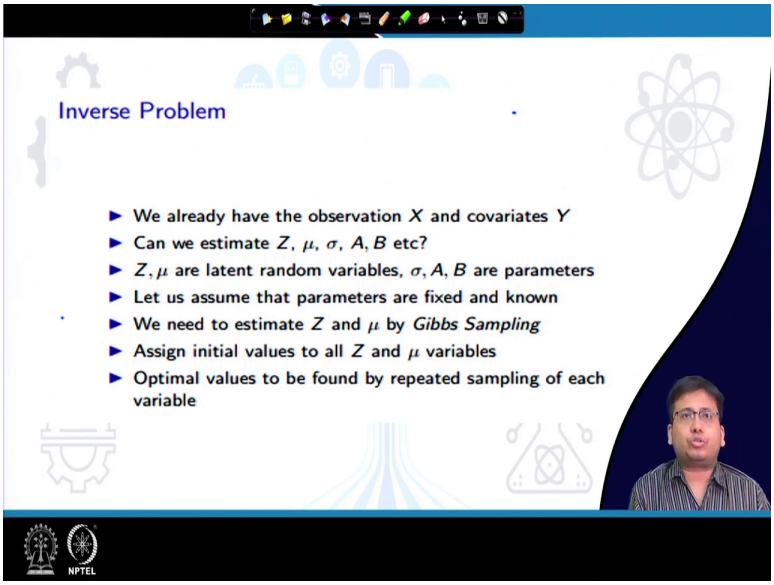


So, now that brings us to the forward problem of data generation which we have said that it helps us to create or sample new data or maybe data for the future. So, once we have developed this model by specifying all these parameters of the I mean all the hyper parameters of the Gaussian process and so on by choosing all the all these covariance function, the mean function etcetera. I mean choice of these is all lies on the in the hands of the model designer.

Once we have done that we simply identify a set of locations and time points to generate the data and then we just keep on sampling $\mu(s)$ and $\mu(t)$ based on the values which have already been generated. Once you have the $\mu(s)$ and $\mu(t)$ then you similarly generate a $Z(s,t)$ at all locations at all time points in the same way from the Gaussian process $Z(s)$ and$Z(t)$.

So, once again you go on sampling $Z(s1)$ then $Z(s2)$, $Z(s3)$ and so on and similarly you also go on generating $Z(t1)$, $Z(t2)$ etcetera and then multiply them together to get $Z(s,t)$ and finally, once you have got the $Z's$ then you use them to generate the actual values $X(s,t)$ using the data model this is the data model right. So, basically you like you just perturb it with the noise and generate that data.

(Refer Slide Time: 27:46)



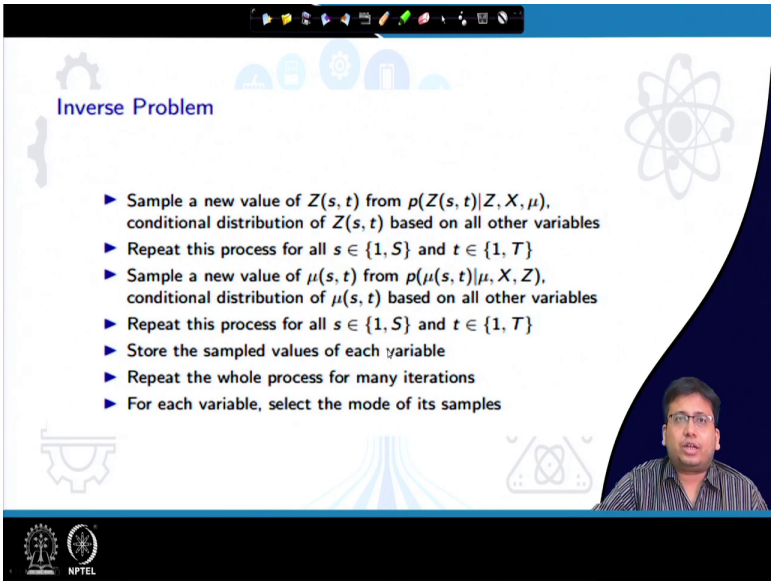
Now, the inverse problem what is the inverse problem that is more complicated? So, the I mean, but that is in a sense more natural also. So, the as you can see the model which we defined here

has so many parameters as and where will all these parameters come from? Because, what you actually have is the data the observations and the covariates also. So, the task when we are trying to build a model the task is to estimate all these other variables.

So, there are two sets, the random variables which are Z as well as the various parameters. So, we can estimate the random variables by a process of called the Gibbs sampling.

So, the process is as follows you assign initial values to all the variables and the optimal values are found by repeatedly sampling for of each of the variables keeping the other as constant. That is you may so Z is a collection of variables at different locations and different time points. So, you can keep focus on one location sampled values keeping all the $Z's$ other $Z's$as constant, then you sample the value of Z at a different location again keeping all the other of them constant and so on.

(Refer Slide Time: 29:01)



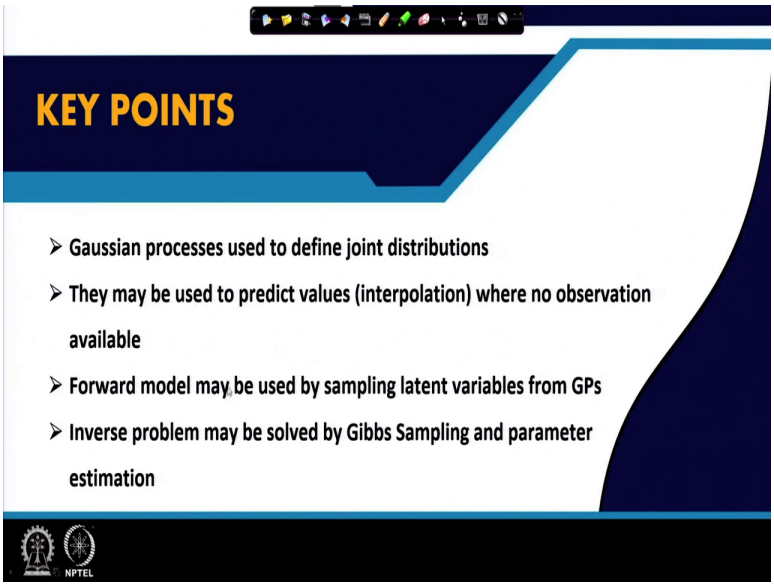
So, that is the process of Gibbs sampling. Now I in this lecture I will not go into the details of the of how exactly the Gibbs sampling is done, but basically you like you repeat this process for all locations as well as for all time points like just go on sampling new values from these kinds of conditional distributions and once you have done this, once you have sampled a large enough number of these values for every location and every time point for all the different random

variables that are latent random variables that are involved here; you store those sampled variables and then and repeat this whole process for many iterations. So, that for each random variable you get a large number of samples and so those using those samples you can get something like a posterior distribution over them and then you can like select the mode of those samples to get the final values ok.

(Refer Slide Time: 30:03)

So, that brings us to the end of this lecture.

(Refer Slide Time: 30:07)



So, with the key points to take away is that the Gaussian processes are used to define joint distributions they may be used to predict values or do the interpolation. The forward model may be used for sampling latent variables and it generate and creates new data while the inverse problem to estimate the model parameters and the latent variables may be solved using Gibbs sampling and other parameter estimation techniques. So, we will continue further details in the following lecture. So, that is it for today. Bye.