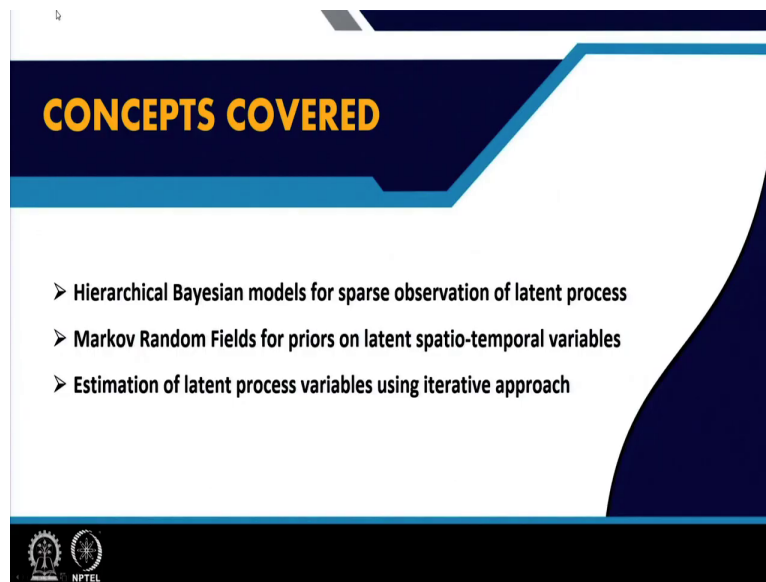


Machine Learning for Earth System Sciences
Prof. Adway Mitra
Department of Computer Science and Engineering
Centre of Excellence in Artificial Intelligence
Indian Institute of Technology, Kharagpur

Module - 03
Machine Learning for Discovering New Insights
Lecture - 22
Hierarchical Bayesian Models for Spatio-Temporal Processes

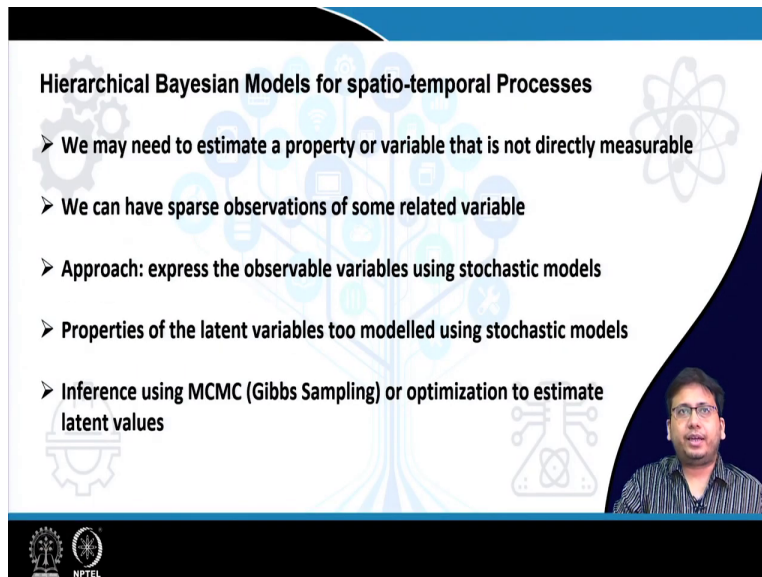
Hello everyone, welcome to Lecture 22 of this course on Machine Learning for Earth System Science. So, we today we are going to talk about Hierarchical Bayesian Models for Spatio-Temporal Processes, we are still in Module 3 where we are discussing Machine Learning for Discovering New Insights in the earth sciences.

(Refer Slide Time: 00:45)



So, the concepts which we are going to cover today are as follows; hierarchical Bayesian models for sparse observations of latent processes. Secondly, the Markov Random Fields for priors on latent spatio-temporal variables and third estimation of latent process variables using the iterative approach.

(Refer Slide Time: 01:06)



Hierarchical Bayesian Models for spatio-temporal Processes

- We may need to estimate a property or variable that is not directly measurable
- We can have sparse observations of some related variable
- Approach: express the observable variables using stochastic models
- Properties of the latent variables too modelled using stochastic models
- Inference using MCMC (Gibbs Sampling) or optimization to estimate latent values

The slide features a background with faint icons of a gear, a tree, and a network. A video inset in the bottom right corner shows a man with glasses speaking. The NPTEL logo is in the bottom left corner.

So, first of all, we may need to estimate the property or a variable that is not directly measurable. So, we have seen some examples of this in earlier lectures also, for example, in the previous lecture we talked about the heat wave, the binary variable. So, that is not directly measurable there can also be other.

So, that is of course, a conceptual variable that is like that does not exist, but there can be actual variables which exist, but they cannot be measured directly that is or even if they or like they can be we can measure some noisy versions of them, but not, but you like the noise components might be quite significant.

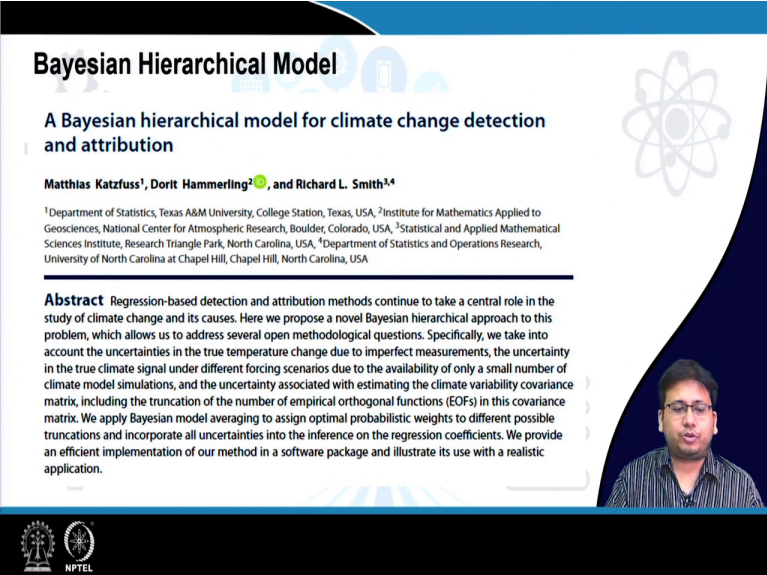
So, we have we may have some sparse observations of some related variable that is instead of measuring that thing that variable directly we can focus on some related variable for which we may have some sparse observations, some sparse observation may be either sparse in either space or in time.

If it is in space then we call it in-situ observations, where like something like we have suppose we are measuring global temperature, then we can have temperatures only at certain specific locations which may be distributed all over the world, but they are only specific locations. So,

from that we have to somehow from those sparse or in-situ observations they like we have to get the global picture somehow.

The general approach is that express the observable variables using some kind of stochastic models and the properties of the latent variables are also modeled using stochastic models. Now, the inference is done using some kind of MCMC or Gibbs sampling to optimize or and estimate the latent values. So, remember we had talked about in one of the lectures in Module 1 we had talked about the observation model, data model, the parameter model and so on. So, this is related to that same method.

(Refer Slide Time: 03:10)




Bayesian Hierarchical Model

A Bayesian hierarchical model for climate change detection and attribution

Matthias Katzfuss¹, Dorit Hammerling², and Richard L. Smith^{3,4}

¹Department of Statistics, Texas A&M University, College Station, Texas, USA, ²Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, Colorado, USA, ³Statistical and Applied Mathematical Sciences Institute, Research Triangle Park, North Carolina, USA, ⁴Department of Statistics and Operations Research, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

Abstract Regression-based detection and attribution methods continue to take a central role in the study of climate change and its causes. Here we propose a novel Bayesian hierarchical approach to this problem, which allows us to address several open methodological questions. Specifically, we take into account the uncertainties in the true temperature change due to imperfect measurements, the uncertainty in the true climate signal under different forcing scenarios due to the availability of only a small number of climate model simulations, and the uncertainty associated with estimating the climate variability covariance matrix, including the truncation of the number of empirical orthogonal functions (EOFs) in this covariance matrix. We apply Bayesian model averaging to assign optimal probabilistic weights to different possible truncations and incorporate all uncertainties into the inference on the regression coefficients. We provide an efficient implementation of our method in a software package and illustrate its use with a realistic application.



So, let us consider a couple of use cases of this. So, for the first paper which we will discuss today this is a Bayesian hierarchical model for climate change detection and attribution.

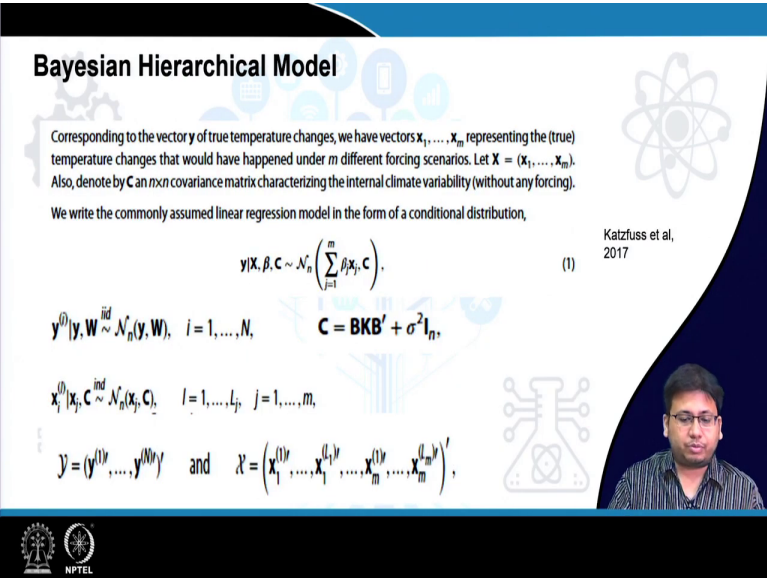
So, regression-based detection and attribution methods continue to take a central role in the study of climate change and its causes. Here we propose a novel Bayesian hierarchical solution to this problem, which allows us to address several open methodological questions. Specifically, we take into account the uncertainties in the true temperature change due to imperfect measurements, the uncertainty in the true climate signal under different forcing scenarios due to the availability of only a small number of climate model simulations and the uncertainty

associated with estimating the climate variability covariance matrix, including the truncation of the number of EOFs in the in this covariance matrix.

So, this covariance matrix which they are talking about this is related to the variability. So, like the when we are considering the that is we are considering the various or any particular variables related to the climate we have only some observations and in the other locations or time points we have to estimate it using the by using the covariance structure or by knowing the covariance relations with the other locations and times where the observations are present. However, these covariance structures are often not well known or cannot be well represented they may be known very approximately.

So, we apply Bayesian model averaging to assign optimal probabilistic weights to different possible truncations and incorporate all uncertainties into the inference of regression coefficients. We provide an efficient implementation of our method in a software package.

(Refer Slide Time: 05:12)



Bayesian Hierarchical Model

Corresponding to the vector y of true temperature changes, we have vectors x_1, \dots, x_m representing the (true) temperature changes that would have happened under m different forcing scenarios. Let $X = (x_1, \dots, x_m)$. Also, denote by C an $n \times n$ covariance matrix characterizing the internal climate variability (without any forcing).

We write the commonly assumed linear regression model in the form of a conditional distribution,

$$y|X, \beta, C \sim \mathcal{N}\left(\sum_{j=1}^m \beta_j x_j, C\right). \quad (1)$$

Katzfuss et al, 2017

$$y^{(i)}|y, W \stackrel{\text{ind}}{\sim} \mathcal{N}_n(y, W), \quad i = 1, \dots, N, \quad C = BKB' + \sigma^2 I_n,$$

$$x_j^{(l)}|x_j, C \stackrel{\text{ind}}{\sim} \mathcal{N}_n(x_j, C), \quad l = 1, \dots, L_j, \quad j = 1, \dots, m,$$

$$y = (y^{(1)'} , \dots, y^{(N)'})' \quad \text{and} \quad X' = \left(x_1^{(1)'} , \dots, x_1^{(L_1)'} , \dots, x_m^{(1)'} , \dots, x_m^{(L_m)'} \right)',$$

So, basically the idea is as follows, corresponding to the vector y of true temperature changes. So, let us say that in like in different places or different locations in the world the temperature change that will that is expected to take place in a given period of time that is like those values

are taken in a vector called y and then we have vectors x_1, x_2, \dots, x_m representing the temperature change that would have happened under the m different forcing scenarios.

So, like we have like let us say m like we do not know which like what exactly will happen in future, but let us say we have m simulations of that and then in like each a like in each of the cases we have some like in each of the scenarios also we may have several observations, because now all these observations or sorry all these simulations they have some they as uncertainty associated with it even in one particular scenario let us say x_1 we need to have several simulations or several runs of simulations.

And the like the that is what that in some sense captures the uncertainty of the simulations or the spread of the simulations because like even in a particular scenario we are never able to say for certainty what kind of change will take place. So, that is why we need an ensemble of values from one scenario and then on top of that we need an ensemble of scenarios.

So, let us say that we have m scenarios and a from each scenario we have several observations. Also, denote by C as $n \times n$ covariance matrix characterizing the internal climate change variable climate variability without any forcing. So, this C is the covariance which we have already talked about. Now so, there is this kind of a, so first of all let us say that the true temperature change y it is a like it follows a Gaussian distribution whose like whose mean is can be expressed as a linear combination of the different scenarios.

So, we can say that each scenario has certain probability. So, like we do not know those probabilities, but if we knew them then the like we could say that the actual change that will happen is some kind of linear the probability weighted linear combination of these individual scenarios and then of course, the covariance is there.

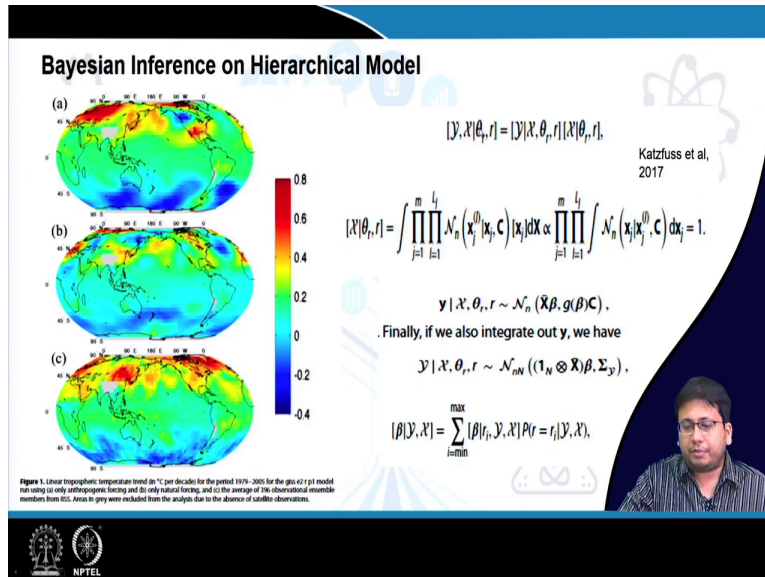
Now, if you consider the, so like. So, I already said that there are so many different observations, sorry yeah so many different runs of simulation. So, in like in any particular run of a simulation let us say that y_i is like the true value.

So, sorry y is the true value and y_i is what is observed and so it like we can say it follows this kind of a Gaussian distribution. So, this is like you can say this is the y_i is the measurement of y the true variability at all locations across the world. So, like for that we again we use another Gaussian distribution and then even for the x 's. So, like I already said that there are j sorry there are m scenarios and in each scenario there are certain number of runs or the simulation runs. So, let us say those runs are denoted by this by this variable l .

So, in the l^{th} run, the x_i is let us say i is a particular location then the temperature change at any given location i in the l^{th} run given all the like given the temperatures at all the temperature changes at all other locations that again follows a Gaussian distribution like for which we need this C as the spatial the covariance structure, because the covariance what it takes into account is the spatial relationship. So, i and j are two locations. So, this C in a sense in a sense incorporates the variance between the covariance between them.

So, let us. So, we have y which is a collection of these that of the actual y variables like this and we have x that is the different runs from the simulation runs from the different like according to the different scenarios. So, like we have l_1 number of x 's from the simulation 1, we have l_2 number of x 's from the scenario 2 and we have l_m number of x 's from the scenario m and so on ok.

(Refer Slide Time: 10:19)



So, basically we so we have the. So, both x and y are random variables in a sense. So, we write the joint distribution in terms of the different model parameters. So, we factorize the joint distributions as follows. So, first of all we want to have a like a distribution x . So, this is for the model runs. So, here we this joint distribution can be factorized in a like in a particular way, again over all the simulation all the scenarios and each run of the scenario. So, like assuming all the simulation runs to be independent we can like just express the each of the individual simulations as products of each other.

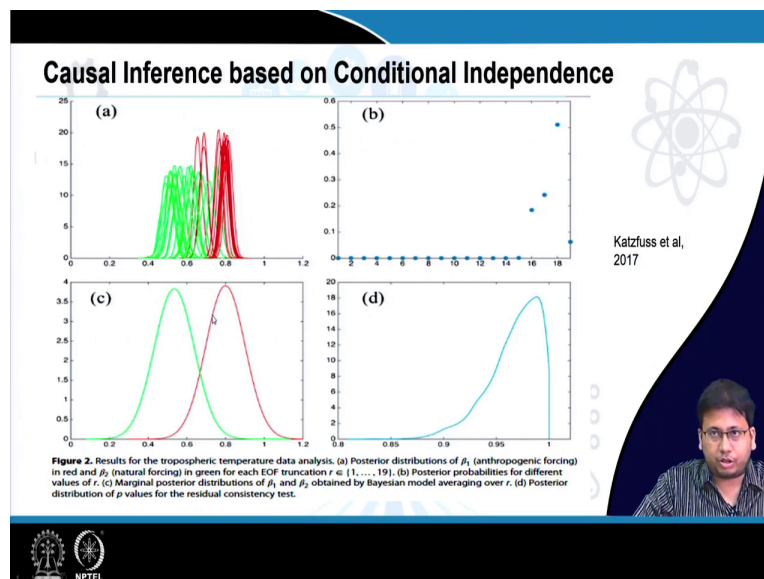
And then again within that we can do the simulation like we can express it in this way and similarly for y we can express. So, like we already know the relation between y and x so this is the relation. So, it can be expressed in another way also. So, now the thing is we have so many other variables like this C and so on like that is what we are finally, interested in is the joint distribution of y and x . So, so the C and all these things that is C itself we can write it in using this as a BKB' etcetera, where BK, σ these are all model parameters.

So, these all those model parameters are like they are compressed within this θ_r and r is like. So, like this for C we can consider. So, C is the covariance matrix, but we can like it is a it will have a certain rank, but we can go for a low rank approximation of r . So, in like whenever we are

considering any particular model we will consider only a certain number of or a only a certain number of principal components. So, that number let that number be r .

So, that the full instead of having the full covariance matrix which might be difficult to measure we can go for a like a truncated or as they say a low rank version of the covariance matrix. So, these are the like the truncated model parameters and so what also these β these are another set of model parameters. So, what we need to do is we like we need to like estimate these through the process of Gibbs sampling. So, like we do we again will not go into the details of the Gibbs sampling here, what we will instead do is that we will go for this Bayesian like inference.

(Refer Slide Time: 13:12)



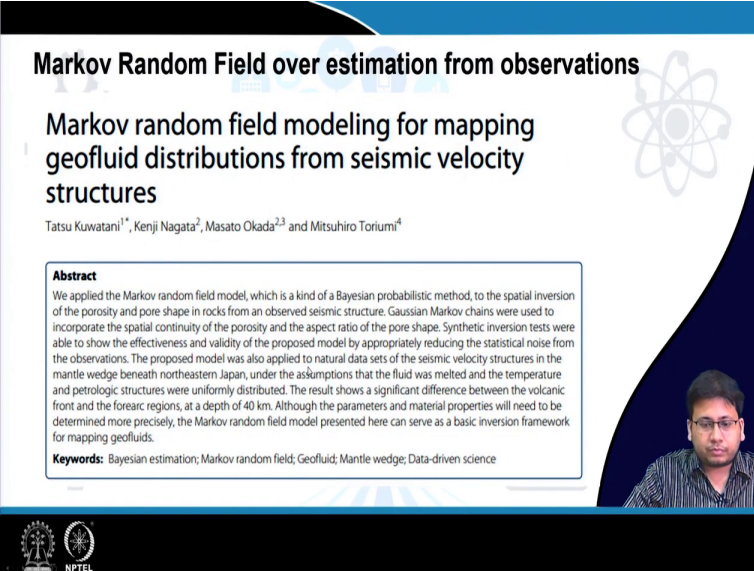
So, like once that. So, once we do the Bayesian inference then we have the results and here like this green and the red curves these indicates the yeah these are like if you see the posterior distributions β_1 the anthropogenic forcing in red and β_2 the natural forcings in green for each EOF truncation.

So, so like this is the different scenarios. So, like in the in one scenario where we do not have any human induced impacts. So, like here is what the that is the posterior distribution in that case looks like and in this I mean the posterior distribution on y and in this case we are talking about

the other scenario where we have the posterior distribution from the red that is the when we are considering the anthropogenic changes.

So, as you can see that there is a clear difference between the temperature in the absence of in the according to natural forcings which is this green curve and in the presence of anthropogenic or human induced forces this is the curve and so on.

(Refer Slide Time: 14:27)




Markov Random Field over estimation from observations

Markov random field modeling for mapping geofluid distributions from seismic velocity structures

Tatsu Kuwatani^{1*}, Kenji Nagata², Masato Okada^{2,3} and Mitsuhiro Toriumi⁴

Abstract
We applied the Markov random field model, which is a kind of a Bayesian probabilistic method, to the spatial inversion of the porosity and pore shape in rocks from an observed seismic structure. Gaussian Markov chains were used to incorporate the spatial continuity of the porosity and the aspect ratio of the pore shape. Synthetic inversion tests were able to show the effectiveness and validity of the proposed model by appropriately reducing the statistical noise from the observations. The proposed model was also applied to natural data sets of the seismic velocity structures in the mantle wedge beneath northeastern Japan, under the assumptions that the fluid was melted and the temperature and petrologic structures were uniformly distributed. The result shows a significant difference between the volcanic front and the forearc regions, at a depth of 40 km. Although the parameters and material properties will need to be determined more precisely, the Markov random field model presented here can serve as a basic inversion framework for mapping geofluids.

Keywords: Bayesian estimation; Markov random field; Geofluid; Mantle wedge; Data-driven science

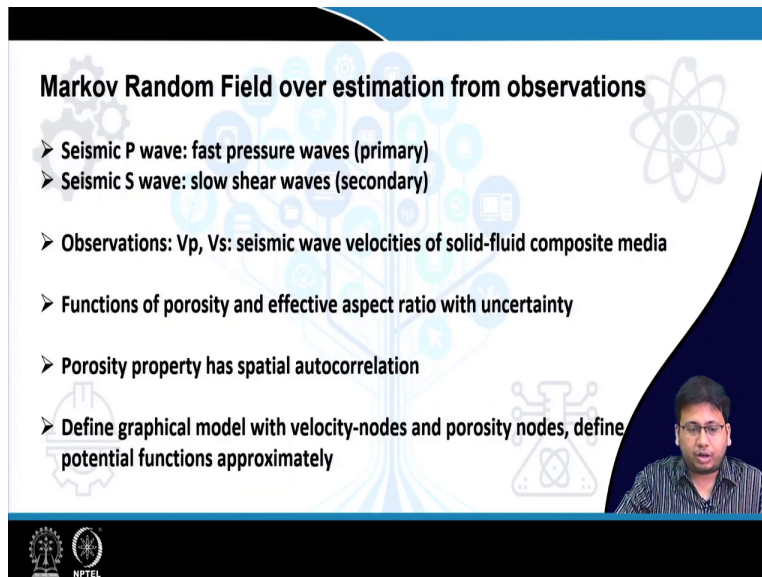


So, next we come to another similar case study where we use Markov random fields for modeling of for mapping of geofluid distributions from seismic velocity structures.

So, this like in this domain of seismology. So, here the we like in this paper we applied Markov random field model, which is a Bayesian probabilistic method to the spatial inversion of porosity and pore shape in rocks from an observed seismic structure. Gaussian Markov chains were used to incorporate the spatial continuity of the porosity and aspect ratio of the pore shape.

Synthetic inversion tests were able to show the effectiveness and validity of the proposed model by appropriately reducing the statistical noise from the observations. The proposed model was also applied to natural data sets of the seismic velocity structures in the observations then so on.

(Refer Slide Time: 15:30)



Markov Random Field over estimation from observations

- Seismic P wave: fast pressure waves (primary)
- Seismic S wave: slow shear waves (secondary)
- Observations: V_p , V_s : seismic wave velocities of solid-fluid composite media
- Functions of porosity and effective aspect ratio with uncertainty
- Porosity property has spatial autocorrelation
- Define graphical model with velocity-nodes and porosity nodes, define potential functions approximately

The slide features a blue header and footer. The footer contains the NPTEL logo and a small video inset of a man speaking. The background has a faint pattern of gears and a molecular structure.

So, like basically there are two kinds of seismic waves the P waves which are fast the pressure waves these are the primary seismic waves and then there are the S waves which are the slow shear waves these are they are the secondary waves.

So, we have observations of their velocities of both the P waves and the S waves at different places on the earth at certain depths and so on. Now, there are some properties of the earth like earth's mantle these are called the porosity and the effective aspect ratio. So, both of these factors they these properties they impact the velocities of the seismic waves.

Now, the task here is that based on the sparse observations of the seismic wave velocities we need to estimate the porosity and effective aspect ratio in different places in different points of the like earth's interior. The now the porosity property has spatial autocorrelation; that means, at two locations which are close to each other the special properties are likely to be very similar.

So, what we need to do is to define a graphical model with velocity nodes and porosity nodes and then define the potential functions approximately, I mean not only porosity nodes, but also the this aspect the effective aspect ratio nodes.

So, like some of the nodes will be corresponding to the observed variables and some will be corresponding to the latent or unobserved variables.

(Refer Slide Time: 17:11)

Markov Random Field over estimation from observations

$$V_p = f_p(\phi, \alpha),$$

$$V_s = f_s(\phi, \alpha).$$

$$p(V_p, V_s | \phi, \alpha) = \prod_{i=1}^N p(V_p^i | \phi^i, \alpha^i) \cdot p(V_s^i | \phi^i, \alpha^i), \quad (4)$$

where N is the total number of grid cells measured, and V_p, V_s, ϕ , and α indicate the respective set of variables V_p, V_s, ϕ , and α for the observed grid cells $i = 1, \dots, N$.
On the other hand, Bayes' theorem can be written as follows:

$$p(\phi, \alpha | V_p, V_s) = \frac{p(V_p, V_s | \phi, \alpha) \cdot p(\phi, \alpha)}{p(V_p, V_s)}. \quad (5)$$

$$V_p^i = f_p(\phi^i, \alpha^i) + \epsilon_p^i, \quad (2)$$

where ϵ_p^i is the observational noise for V_p^i for each spatial grid cell (i). If we assume a Gaussian noise with zero mean, Equation 2 can be rewritten in terms of the conditional probability as

$$p(V_p^i | \phi^i, \alpha^i) = \frac{1}{\sqrt{2\pi\sigma_p^2}} \exp\left(-\frac{(V_p^i - f_p(\phi^i, \alpha^i))^2}{2\sigma_p^2}\right),$$

$$p(\phi) = \frac{1}{Z_\phi} \exp\left\{-\frac{1}{2\sigma_\phi^2} \sum_{i \sim j} (\phi^i - \phi^j)^2\right\},$$

$$E(\phi, \alpha; \theta, V_p, V_s)$$

$$= \frac{1}{2\sigma_p^2} \sum_{i=1}^N (V_p^i - f_p(\phi^i, \alpha^i))^2$$

$$+ \frac{1}{2\sigma_s^2} \sum_{i=1}^N (V_s^i - f_s(\phi^i, \alpha^i))^2$$

$$+ \frac{1}{2\sigma_\phi^2} \sum_{i \sim j} (\phi^i - \phi^j)^2 + \frac{1}{2\sigma_\alpha^2} \sum_{i \sim j} (\alpha^i - \alpha^j)^2$$

$$+ \frac{N}{2} (\ln \sigma_p^2 + \ln \sigma_s^2) + \ln Z_\phi + \ln Z_\alpha + C,$$

Kuwatani et al, 2014

So, you if you remember the various graphical models which we had studied in one of the earlier lectures. So, there are like for Markov random fields we have some nodes corresponding to variables that are observed and we will also have some nodes corresponding to the variables that are not observed. And then there will be edges between different kinds of variables and each of those edges will have the edge potential functions also.

So, the like first of all the thing is we like we can express the V_p and V_s that is the p velocity and the s velocity as some functions of the two things we are interested in Φ and α , Φ is the porosity and α is the effective aspect ratio. So, now let us say that these functions f_p and f_s let us say that we already know those functions from our domain knowledge of seismology, but we do not know Φ and α at all locations.

So, now what we do is. So, like first of all we like we define some kind of a probability distribution on V_p and V_s at any location. So, let us say there are total N locations or grids where we will like we will do the do our study and let us also say that in each grid this the Φ values we have unique Φ and α values. Also let us further assume that the V_p and V_s the these velocities they are independent of each other so, that this joint distribution can be factorize as a product of their marginal distributions.

Then what we finally, have to do is or what we intend to do is to create the posterior distribution on this porosity and aspect ratio depending on the observations that we have, So, that is to be written in this way now. So, like our aim is to estimate this kind of a posterior distribution. So, for this purpose we define what is known as the energy of the Markov random field where this energy the like.

If you remember in a Markov random field the joint distribution is a product of the different click potential functions now instead like a like in a normal graph a click may in may just mean that individual nodes which are clicks of size 1 and edges which are clicks of size 2. So, for like. So, for that we need to define the.

So, the node potentials are there on top of that we put the edge potential functions and the edge potential. So, there are two kinds of edges one is the edge between 2 between the Φ variables or the α variables at different locations and then there are edges between the Φ variables and the V_p variables and V_s variables and so on.

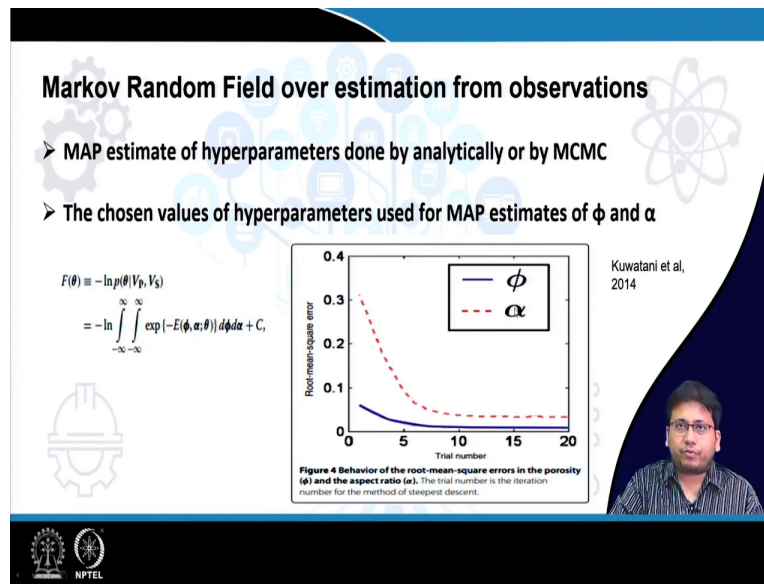
So, the in case of those the later kind of edges where like the like at any location the corresponding V_p is connected to the properties Φ and α . So, for such edges we have like the edge potential functions and then for the edges that are connecting to Φ variables we have this kind of a edge potential function. So, as you can see $\Phi^i - \Phi^j$. So, let us say that like it is an arrow between the Φ variables at location i and location j . So, between them we just simply have a the edge potential is simply the squared difference.

So, like if two of them are close to each other then like that should have a high probability those five values are close to each other that should have a high probability compared to the situation where the Φ values are broadly different. So, this is based on the spatial autocorrelation which we talked about. So, a similar property also holds for α and of course, the there is a variance component for both Φ and α . So, they are encoded like this.

So, accordingly we have this kind of a the what is called the energy function of the Markov random field which is nothing, but a representation of the joint probability distribution of all these random variables. So, these Z_Φ , Z_α these are known as the partition functions that are

necessary to make sure that it is indeed a joint like a valid probability distribution that is the sum of these values of the energy values for the different combinations they all add up to one in the proper way.

(Refer Slide Time: 22:01)



So, that now what we need to do is so, first of all there are these the variables Φ and α the latent variables and then there are the different hyper parameters these σ and so on, I mean also these f_p, f_s these functions they also they may have their own hyper parameters which have to be estimated. So, what is first done is that this the like all the parameters let us say we represent by θ . So, for the time being let us ignore the latent variables and let us try just try to estimate a posterior distribution on the hyper parameters based on the observations V_p and V_s .

So, the like we do this kind of marginalization where we integrate out the all possible values of Φ and α and then we get a posterior distribution of the parameters. Now, this kind of integration this cannot be done numerically because it is may that might be intractable. So, this kind of integration is again done through MCMC techniques.

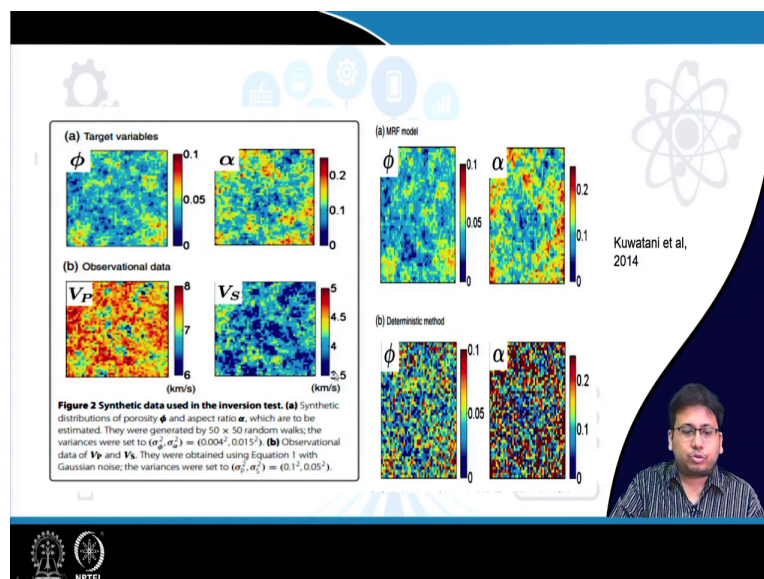
So, that is based on sampling. So, once we get this f_θ that is the posterior distribution of the parameters, then we choose the map estimate of the all the parameters that is the those the

parameters which maximize the value of this posterior distribution, that in a sense gives us the optimal values of the parameters.

Then using the optimal values of the parameters again we calculate the map estimates of Φ and α . So, note that Φ and α are not individual values, but they are like they are defined separately for each of the different locations that we are considering. So, that we will get n values of Φ and n values of α and like. So, like we need to do it in many trials and we can expect that like as we or many iterations and as we do the further and further iterations we can see the like convergence like.

So, unless the convergence happens we cannot expect to find the like any optimal values and like here the as their studies shows their experiments show that as they keep on doing more and more trials by this like using these optimization like they see the some kind of convergence of the root mean square error; that means. So, like in the like in the held out set which or rather on which they are validating they find that the using these approach the estimated value of Φ and α are coming close to the unknown values.

(Refer Slide Time: 24:37)



And. So, like. So, this is this might be a possible plot of the Φ and which the so, the observational data are of course, V_p and V_s . So, like the this primary wave it has higher velocities

up to 8 kilometers per second, while the secondary wave it is usually slower about 4 kilometers per second. So, this is let us say over a spatial region. So, as you can see this is graded data at different locations they the V_p and V_s vary.

So, using this and the corresponding target variables are Φ and α let us say this is their these are their spatial distributions. So, this is like this is the training data this is what is I mean this Φ and α is what is already known by some kind of measurement. So, now, what we need to do is to based on these observations of V_p and V_s we need to estimate the Φ and α . So, we can either go for some kind of deterministic method which was known to the seismology researchers earlier that is where they basically tried to like invert this function f_p and f_s in some ways.

And so, like this is the kind of estimates of Φ and α which they get using those methods and these are the estimates of Φ and α which is obtained by the MRF methods. So, as you can understand in the like you can compare these plots or these maps with the true the targets the or the true values and you can see that the estimated values are like very similar to what is the actual values, but if you compare these maps which are obtained by the old deterministic methods you can see that there is hardly any similarity.

(Refer Slide Time: 26:34)

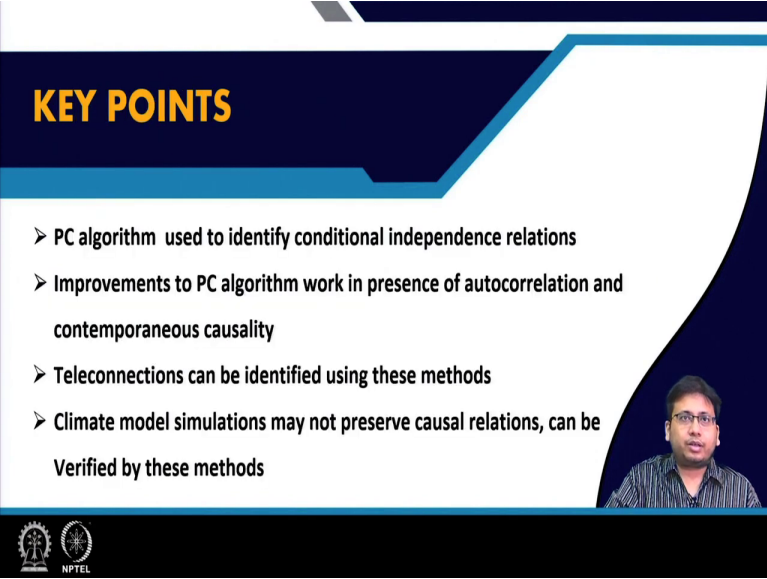
REFERENCES

- Katzfuss M, Hammerling D, Smith RL. A Bayesian hierarchical model for climate change detection and attribution. *Geophysical Research Letters*. 2017 Jun 16;44(11):5720-8.
- Kuwatani T, Nagata K, Okada M, Toriumi M. Markov random field modeling for mapping geofluid distributions from seismic velocity structures. *Earth, Planets and Space*. 2014 Dec;66(1):1-9.

The slide features a dark blue header with the word 'REFERENCES' in orange. Below the header, two references are listed with orange arrowheads. In the bottom right corner, there is a small video inset showing a man with glasses and a beard speaking. The bottom of the slide has a black bar containing the NPTEL logo on the left.

So, that shows that using this MRF model we are able to get a much better estimate of the maps of the unknown variables. So, these are the references of the two papers that we discussed today.

(Refer Slide Time: 26:45)



KEY POINTS

- PC algorithm used to identify conditional independence relations
- Improvements to PC algorithm work in presence of autocorrelation and contemporaneous causality
- Teleconnections can be identified using these methods
- Climate model simulations may not preserve causal relations, can be Verified by these methods

NPTEL

So, the key points to be taken away is like. So, there are the basically that is the key point these are the key points to be taken away which we discussed that Markov random fields are capable of like incorporating the spatial autocorrelations and this in turn helps us to identify the relation the or estimate the various unknown variables.

And secondly, like when we have observations of or multiple observations none of which are reliable of a particular quantity we can express the particular quantity as a linear combination of all those observations along with some uncertainty estimates and then the not only the linear coefficients for the for that relation as well as those coefficients they can all be estimated using some kind of a like using the inference algorithms like Gibbs sampling and optimization.

So, that ends us that brings us to the end of this lecture, in the further lectures we will see some more applications of machine learning to for a to discover new insights in earth sciences. So, till then bye.