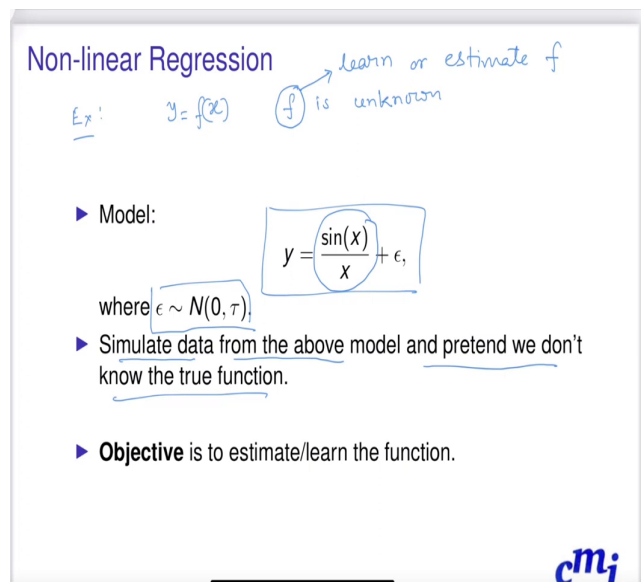


Predictive Analytics - Regression and Classification
Prof. Sourish Das
Department of Mathematics
Indian Institute of Technology, Madras

Lecture - 49
Gaussian Process Regression

Hello all, welcome back to the Predictive Analytics Regression and Classification course.
This is lecture 15 part A.

(Refer Slide Time: 00:25)



Non-linear Regression




Ex: $y = f(x)$ f is unknown → learn or estimate f

► Model: $y = \frac{\sin(x)}{x} + \epsilon$

where $\epsilon \sim N(0, \tau)$

► Simulate data from the above model and pretend we don't know the true function.

► Objective is to estimate/learn the function.

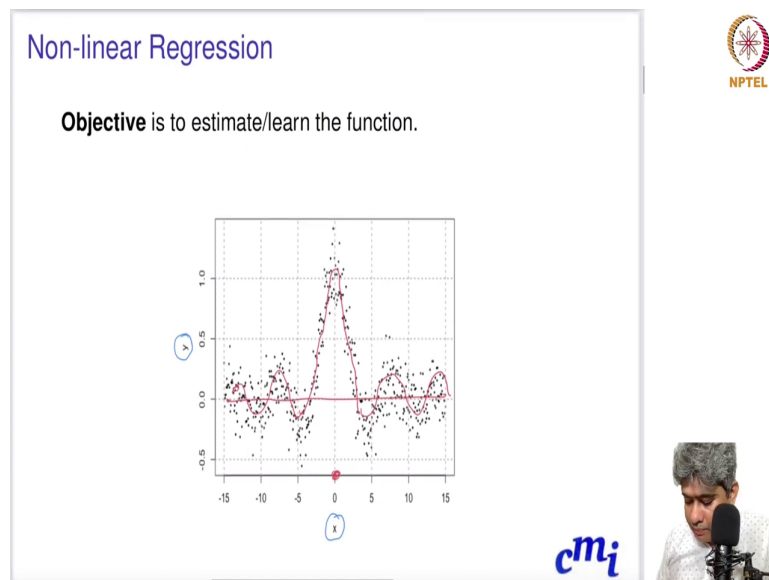


So, in this lecture in this discussion we will start non-linear regression. So, suppose let me start with an example, ok. Let me start with an example that suppose the there is only one y variable and one x variable and the true relationship between y and x is sort of f is unknown ok f is unknown.

And suppose true good relationship is this function which is generating the data this model which is generating the data where epsilon follow some normal swiipe noise or normal 0 tau and y equal to sin x by x is the actual true model, but we do not know that this is the model. So, what we are going to do? We are going to simulate some data from the above model and pretend that we do not know the true function as if.

And obviously, our target is to learn or estimate this function, ok learn or estimate the unknown function f that is our target.

(Refer Slide Time: 01:48)



Now, if we simulate from this function and what we will have? We will and we pretend as if we do not know what is the true relationship between x and y then what we have is only the data the x and the y that I s all. And if we plot them this is the kind of data that you are seeing.

Now, what happens is typically this kind of data can see sometimes in physics or in some you know biology also that you have most of your data which is hovering around 0 and then there is a point somewhere suddenly there is a signal. So, maybe its just going like this and then suddenly there is a signal and signal bust and came down as you deviate from the signal point and then again its hovering around 0. So, this kind of things actually happens.

Now, question is how do you estimate this function? Ok. It is clearly there is no trend there is some you know seasonality looks like, but its difficult just seeing the data, but can we estimate the function here.

(Refer Slide Time: 03:04)

Non-linear Regression Basis Functions


► Consider i^{th} record

$$y_i = f(x_i) + \epsilon_i$$

represents $f(x)$ as



$$f(x) = \sum_{j=1}^K \beta_j \phi_j(x) = \phi \beta$$

we say ϕ is a basis system for $f(x)$.



$$\phi = \begin{bmatrix} \phi_1(x) \\ \phi_2(x) \\ \vdots \\ \phi_k(x) \end{bmatrix}$$

$$\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}$$

So, let us try to formalize the mathematical construction of this through something called basis function ok something called basis function, ok. So, let me just take a blue color, ok. So,

this is the basis function expansion. So, we will talk about it lets talk about it. So, suppose the i th record is y_i is some function of x_i plus ϵ_i .

Now, f of x we can write it as some linear combinations of $\beta_j \phi_j(x)$ where ϕ is known as the basis system for f of x . We can write it as ϕ times β . So, ϕ here we can write it as $\phi_1(x), \phi_2(x) \dots \phi_k(x)$ and β is $\beta_1, \beta_2 \dots \beta_k$, ok.

And then we just you know if you just take the dot product what we get is essentially f of x . So, this is typically how we assume that ok you just approximate with the ϕ if you do not know the f you approximate with ϕ times β .

(Refer Slide Time: 04:42)

The slide is titled "Representing Functions with Basis Functions". It includes the NPTEL logo in the top right corner and a small video inset of a speaker in the bottom right corner. The main content of the slide is as follows:

Ex:

- Terms for curvature in linear regression *Polynomial fit*

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 x_i^2 + \dots + \epsilon_i$$

implies

$$f(x) = \beta_1 + \beta_2 x + \beta_3 x^2 + \dots$$

The slide also features the "cmj" logo in the bottom right corner.

Now, can we have some examples? Yes; obviously, this example we have seen before y equal to $\beta_1 + \beta_2 x_i + \beta_3 x_i^2 \dots + \epsilon_i$. So, this is like polynomial

functions bunch of this is typically called polynomial functions f of x is my this basis expansion kind of thing ok polynomial basis.

(Refer Slide Time: 05:16)

Fourier Basis


Ex: Consider Chennai Temperature Data modeling Exercise in previous lectures.



- ▶ sine cosine functions of increasing frequencies

$$y_i = \beta_1 + \beta_2 \sin(\omega x) + \beta_3 \cos(\omega x) + \beta_4 \sin(2\omega x) + \beta_5 \cos(2\omega x) \dots + \epsilon_i$$

- ▶ constant $\omega = 2\pi/P$ defines the period P of oscillation of the first sine/cosine pair.
- ▶ $\phi = \{1, \sin(\omega x), \cos(\omega x), \sin(2\omega x), \cos(2\omega x) \dots\}$
- ▶ $\beta^T = \{\beta_1, \beta_2, \beta_3, \dots\}$

$y = \phi\beta + \epsilon$



Then there could be Fourier basis that we have tried this model also previously you remember the Chennai temperature data right where we fit $\sin \omega x + \cos \omega x + \sin 2\omega x + \cos 2\omega x$ you have to choose the ω properly we chose $\omega = 2\pi/P$, P was our the you know sequence the size of the season. And we can write it as ϕ as our basis and β as our coefficient.




So, we can effectively remember that we wrote our Chennai temperature data model as $y = \phi\beta + \epsilon$. So, example here is consider Chennai temperature data modeling exercise in previous lectures, ok.

(Refer Slide Time: 06:36)

Other Basis

- ▶ **Exponential Basis** $\phi = \{1, e^{\lambda_1 x}, e^{\lambda_2 x}, \dots\}$ ✓
- ▶ **Gaussian Basis**
 $\phi = \{1, \exp(-\lambda(x - c_1)^2), \exp(-\lambda(x - c_2)^2), \dots\}$ ✓
- ▶ **Basis corresponds to Spline Regression**
$$y = \beta_0 + \sum_{k=1}^K \beta_k (x - \xi_k)_+^D + \dots + \epsilon$$

 $\phi = \{1, (x - \xi_1)_+^D, (x - \xi_2)_+^D, \dots\}$



What are the other basis functions that we can consider? We can consider something called exponential basis, we can consider Gaussian basis, we can basis correspond to spline regression these are the basis which corresponds to spline regression. So, there are different kind of basis are there which you can use for modeling your data.

(Refer Slide Time: 06:59)




Functional Estimation/Learning

▶ We are writing the function with its basis expansion

$$y = \phi\beta + \epsilon$$

▶ Lets assume basis ϕ is fully known

▶ Problem is β is unknown - hence we estimate β .



Now, how we estimate? So, remember that my phi is completely known the basis is complete with the basis function the basis expand basis the set of basis function is completely known ok what is unknown is the beta are unknown the coefficients are unknown. So, essentially my basis expansions you can think of as a feature engineering these are engineered features, but we are doing it in such a way essentially that we have already discussed and then we are putting it into the you know we are trying to estimate the beta here, ok.

(Refer Slide Time: 07:53)

Bayesian method

▶ Model:




$$y = f(x) + \epsilon$$

▶ $\epsilon \sim N(0, \sigma^2 I) \Rightarrow y \sim N(f(x), \sigma^2 I)$ $k \rightarrow \infty$

$$f(x) = \phi \beta = \sum_{k=1}^{\infty} \phi_k(x) \beta_k$$

▶ β is unknown and want to estimate
Assuming β 's are uncorrelated random variable and $\phi_k(x)$ are known deterministic real-valued functions.

▶ Then due to **Kosambi-Karhunen-Loeve** theorem, we can say that $f(x)$ is a stochastic process.



So, far as so, good now I am going to expand it to an extent and to the you know. So, far we whatever we were doing you remember that we were expanding it to the k-many functions we were always expanding it to the k-many functions. If you look into here also we are saying dot dot dot, but we were always saying it is k-many basis functions. So, in how many basis how many features we will have that we are putting it as sort of a finite parameter.

(Refer Slide Time: 08:38)

represents $f(x)$ as

$$y_i = f(x_i) + \epsilon_i$$

parameter

$$f(x) = \sum_{j=1}^K \beta_j \phi_j(x) = \phi \beta$$

we say ϕ is a basis system for $f(x)$.

K is a # basis / feature chosen by user or data scientist

NPTEL

cmu

Representing Functions with Basis Functions

So, k turns out to be a parameter some sort of parameter. So, k is a number of features is number of basis or features chosen by chosen by user or data scientist data scientist. So, sometimes these choices are bit ad hoc. So, we you do not want to choose this as ad hoc you want somewhere to you know data to decide which k to choose. So, what mathematicians have done that can we go beyond k can we k push to infinity and its lets data decide where to capital K here goes to infinity and can we come up with a function can we estimate this function?

So, this is a very interesting idea. So, what they are saying that y is the function of x plus epsilon. Epsilon for a normal 0 sigma squared i this will lead to y as a normal of f of x comma sigma square a . But f of x is ϕ of β where this guy is goes up to infinity the summation goes up to infinity and it converges to ϕ of β . So, that means, at every points it

converges, every point of x if this summation this summation does not diverge to infinity it definitely converges to ϕ of β or the f of x . So, that is the idea.


So, and in that ϕ of x is completely known what is unknown is β , β is unknown and we want to estimate the β . Now, assuming that β is an uncorrelated random variable and ϕ of x are known deterministic real valued function. If you make these assumptions, ok then there is a theorem by Kosambi-Karhunen-Loeve theorem we can say that f of x is a stochastic process what is it mean? f of x is a stochastic process.

The word stochastic is a German for probabilistic, ok. So, that means, essentially for each value of x you will get a probability distribution proper probability distribution that is typically called properly defined distribution you will get as a value of x and that is called probabilistic process or in German stochastic process.

(Refer Slide Time: 11:44)

Gaussian Process Prior

Rasmussen's Book



- ▶ As $f(x)$ is a stochastic process if we assume $\beta \sim N(0, \sigma^2 I)$ then $f(x) = \phi\beta$ follow Gaussian process.
- ▶ Since $f(x)$ is unknown function; therefore induced process on $f(x)$ is known as '**Gaussian Process Prior**'.


Prior on β :

$$p(\beta) \propto \exp\left(-\frac{1}{2\sigma^2}\beta^T\beta\right)$$

Induced Prior on $f = \phi\beta$:

$$p(f) \propto \exp\left(-\frac{1}{2\sigma^2}\beta^T\phi^TK^{-1}\phi\beta\right)$$

cm



Now, what f of x is a stochastic process and if we assume these betas follow normal 0 sigma square I then turns out f of x phi beta follow Gaussian process, ok. So, and this typically called Gaussian process prior. So, this comes under the Bayesian method I am not going into the detail of how these things happens and all, but I am just telling you these are called Gaussian process prior. If you want to know more about it, you can see it in Rasmussen's book, Rasmusen's book, ok.

Since f of x is unknown function therefore, induced process of f x is f of x is known Gaussian process prior. So, if you press if you induce p of beta a prior on the beta that will induce a prior on the f and that also a Gaussian process. So, that is how you induce a function a prior distribution on the unknown function.

(Refer Slide Time: 13:04)

Gaussian Process Prior

- ▶ The prior mean and covariance of $f(x)$ are given by


$$\mathbf{E}[f(x)] = \phi(x)\mathbf{E}[\beta] = \phi\beta_0$$


$$\begin{aligned} \text{cov}[f(x)] &= \mathbf{E}[f(x).f(x')^T] = \phi(x).\mathbf{E}[\beta.\beta^T]\phi(x')^T \\ &= \sigma^2\phi(x).\phi(x')^T = \mathbf{K}(x, x') \end{aligned}$$

$$f(x) \sim \mathcal{N}_n(\phi(x)\beta_0, \mathbf{K}(x, x'))$$

n: sample size

$$y(x) \sim \mathcal{N}_n(\phi(x)\beta_0, \mathbf{K}(x, x') + \sigma^2\mathbf{I})$$





And turns out that there are lot of mathematics goes into I am not going to the detail derivation of those things that f of x turns out to be a multivariate normal with $\phi(x)$ of β naught as mean and some covariance function $K(x, x)$ transpose and y of x , this is the most crucial part that y of x turns out to be multivariate normal of n -dimensional, n is the sample size, small n is the sample size, sample size.

So, small n is the sample size and ϕ of x , small n is. So, it is y of x is going to be coming from a multivariate normal with a dimension of sample size with mean as $\phi(x)$ of β naught and k covariance is $K(x, x)$ transpose plus $\sigma^2 I$.


(Refer Slide Time: 14:07)


Gaussian Process Regression

- ▶ The estimated value of y for a given x_* is the mean (expected) value of the functions sampled from the posterior at that value of x_* .
- ▶ Suppose $\mu(x) = \phi(x)\beta_0$, then expected value of the estimate at a given x_* is given by

$$\hat{f}(x_*) = \mathbf{E}(f(x_*)|x, Y)$$

$$= \mu(x) + \mathbf{K}(x_*, x) \cdot \underbrace{[\mathbf{K}(x, x) + \sigma^2 \mathbf{I}]^{-1}}_{\text{Matrix of order } n} \cdot (y - \mu(x))$$
- ▶ The time complexity of the matrix inversion is $\mathcal{O}(n^3)$





So, this is an interesting thing. So, then what happens that how do you. So, this is my likelihood function, ok.

(Refer Slide Time: 14:26)

$f(x) \sim N_n(\phi(x)\beta_0, \mathbf{K}(x, x')), \epsilon \sim N_p(0, \sigma^2 \mathbf{I})$
n: sample size *likelihood model.*

$y(x) \sim N(\phi(x)\beta_0, \mathbf{K}(x, x') + \sigma^2 \mathbf{I})$ *cm_i*

Gaussian Process Regression

- ▶ The estimated value of y for a given x_* is the mean (expected) value of the functions sampled from the posterior at that value of x_* .
- ▶ Suppose $\mu(x) = \phi(x)\beta_0$, then expected value of the estimate at a given x_* is given by —

So, this is my likelihood function let me write it down sort of likelihood model. So, this is actually our likelihood model likely hood model, ok. And then the estimated value of y given a x star, a new point x star this is the expected value of function sampled from the posterior at the value of x star.

Do not worry about the posterior prior and all these things if you are not familiar with the Bayesian methods, but what it tells us that you can estimate the function f hat if at new point test point x star as a function of $\mu(x)$ plus $\mathbf{K}(x, x^*) \mathbf{K}(x, x)$ and $\mathbf{K}(x, x) \sigma^2 \mathbf{I}^{-1} y$ minus $\mu(x)$.

Now, this is a matrix of order n . Now, if you so; that means, in the solution this is the final solution that is the final solution. And as you run the run this solution what happens that the time complexity of the matrix inversion is order of n cube. So, that means, if your sample size

increases the implementation of this solution is extremely difficult. So, this is almost impossible to implement. So, if for a very large data set, ok.

(Refer Slide Time: 15:49)

Likelihood Method: Gaussian Process Prior Model

▶ Data model:




$$y(x) \sim \mathbf{N}_n\left(\phi(x)\beta_0, \mathbf{K}_{\alpha, \rho}(x, x') + \sigma^2 \mathbf{I}\right)$$

▶ Static or Hyperparameters: $\theta = \{\beta_0, \alpha, \rho, \sigma^2\}$

▶ Likelihood function:

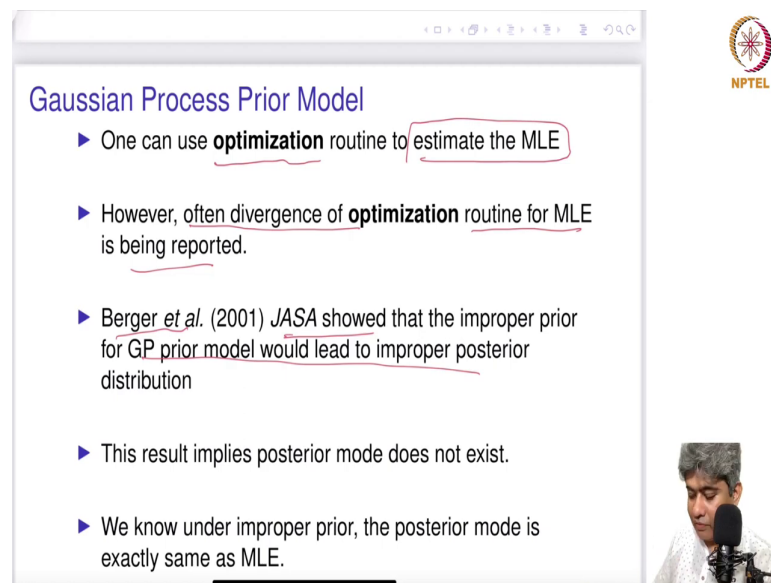
$$f(\beta|y, \phi, \sigma^2) \propto (\sigma^2)^{-p/2} \exp\left(-\frac{1}{2\sigma^2}(y-f)^T[\mathbf{K} + \sigma^2 \mathbf{I}]^{-1}(y-f)\right)$$

▶ Negative Log-likelihood function:

$$l(\beta) \propto \frac{1}{2\sigma^2}(y - \phi\beta)^T[\mathbf{K} + \sigma^2 \mathbf{I}]^{-1}(y - \phi\beta)$$


So, this is our data model and this is some hyper parameters that typically we set and this is the likelihood is a multivariate Gaussian distributions and this is negative log likelihood.

(Refer Slide Time: 16:11)



Gaussian Process Prior Model

- ▶ One can use optimization routine to estimate the MLE
- ▶ However, often divergence of optimization routine for MLE is being reported.
- ▶ Berger *et al.* (2001) *JASA* showed that the improper prior for GP prior model would lead to improper posterior distribution
- ▶ This result implies posterior mode does not exist.
- ▶ We know under improper prior, the posterior mode is exactly same as MLE.

And as you know if you have a negative log likelihood. you can pass it through or similarly you can write the negative log posterior distribution do not worry about it and you can run the negative log likelihood through optimization routine to estimate the maximum likelihood estimates.

So, however, often divergence of optimization routine for MLE is being reported because Berger et al in *JASA* showed that the you know the GP prior model if you do not put some prior on or penalty on the parameter space then it will have this problem so, that can be avoided.

(Refer Slide Time: 16:45)

16 of 21

► Perhaps this is a Bayesian interpretation of why **optimization** routine for MLE faces convergence issues.



cm

What prior to choose? And Why?

► Popular choice

$$\beta_0 \sim N_p(0, K), \quad K \text{ is large}$$
$$(\alpha, \rho, \sigma) \sim \text{InvGamma}(\epsilon, \epsilon), \quad 0 < \epsilon < 1.$$

► Robust Choice

$$\beta_0 \sim N_p(0, \tau),$$
$$\tau \sim \text{Half - Cauchy}(0, 1),$$
$$(\alpha, \rho, \sigma) \sim \text{InvGamma}(\epsilon, \epsilon), \quad 0 < \epsilon < 1.$$


(Refer Slide Time: 16:47)

What prior to choose? And Why?




► Popular choice

$$\beta_0 \sim N_p(0, K), \quad K \text{ is large}$$
$$(\alpha, \rho, \sigma) \sim \text{InvGamma}(\epsilon, \epsilon), \quad 0 < \epsilon < 1.$$

► Robust Choice

$$\beta_0 \sim N_p(0, \tau),$$
$$\tau \sim \text{Half - Cauchy}(0, 1),$$
$$(\alpha, \rho, \sigma) \sim \text{InvGamma}(\epsilon, \epsilon), \quad 0 < \epsilon < 1.$$

Ref: Gelman et al. (2006), Carvalho (2008), Dautta and Ghosh (2012) studied the robustness of this prior in regular regression model, but not in fda






So, there are some prior on the like you know is being like inverse gamma on the sigma alpha rho and sigma some robust choice of prior has been also been given. We will do we will understand when we will do it in hands on.

(Refer Slide Time: 17:03)

Experiment with GP Regression

We do not know the true f_{θ} .

- ▶ Model:
$$y = \frac{\sin(x)}{x} + \epsilon,$$
- where $\epsilon \sim N(0, \tau)$.
- ▶ Simulate data from the above model and pretend we don't know the true function.
- ▶ **Objective** is to estimate/learn the function. (y, x)



And so, here is the experiment that we are doing at the beginning, ok. So, y is a $\sin x$ by x and then when normal 0 τ square from this model we are going to simulate the data from the above model and pretend that we do not know the true function we will simulate from this model and we will pretend we do not know this model we do not know we do not know the true function, ok.

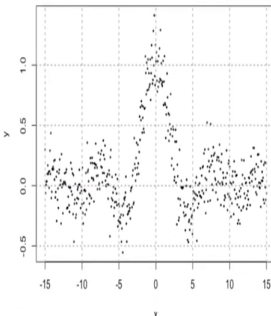
We do not know the true function our objective is to estimate or learn the function from just y and x .


(Refer Slide Time: 17:50)


18 of 21

Experiment with GP Regression

Objective is to estimate/learn the function.

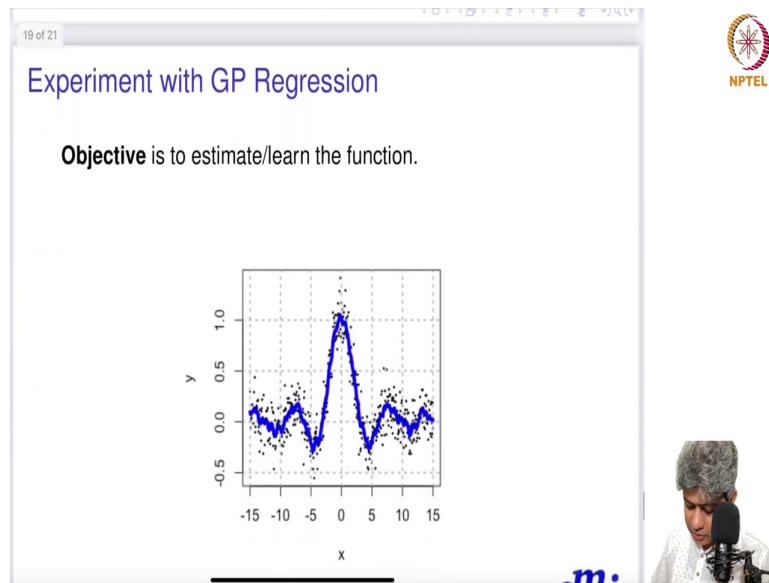






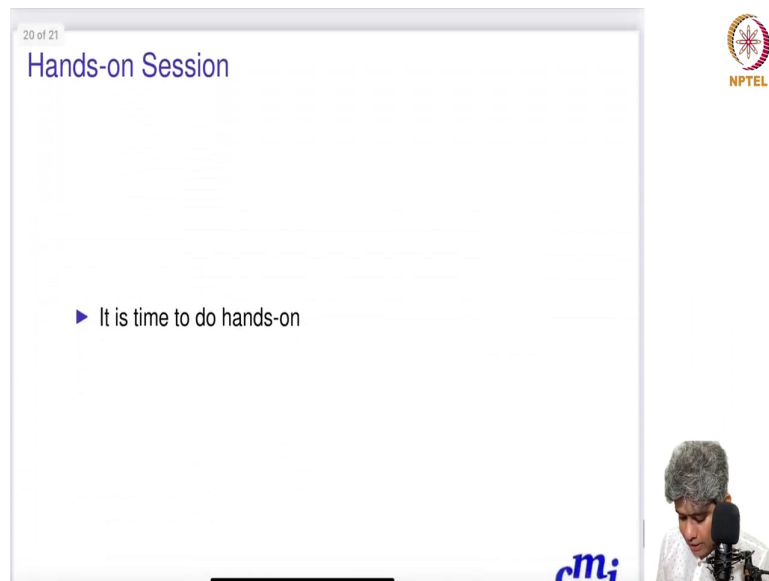
And what we found that this is the data this was the data that was simulated.

(Refer Slide Time: 17:57)



And the estimated function using GP regression was this that is going through and we trust us we this we have I have not given any information, but it just learn on itself that and almost you know picked up the true function as it should be.

(Refer Slide Time: 18:20)



20 of 21

Hands-on Session

- It is time to do hands-on

NPTEL

cmj

So, in the next video what I am going to do, we will going to do the hands on and in the hands on we will implement this thing and we will see that how nicely it can fit unknown completely unknown function as long as the function is somewhat smooth there is no major break or anything it will work pretty good. So, I hope you enjoyed this video. So, please watch the next video which will be hands on of implementation of Gaussian process regression in with R.

Thank you very much, see you in the next video.