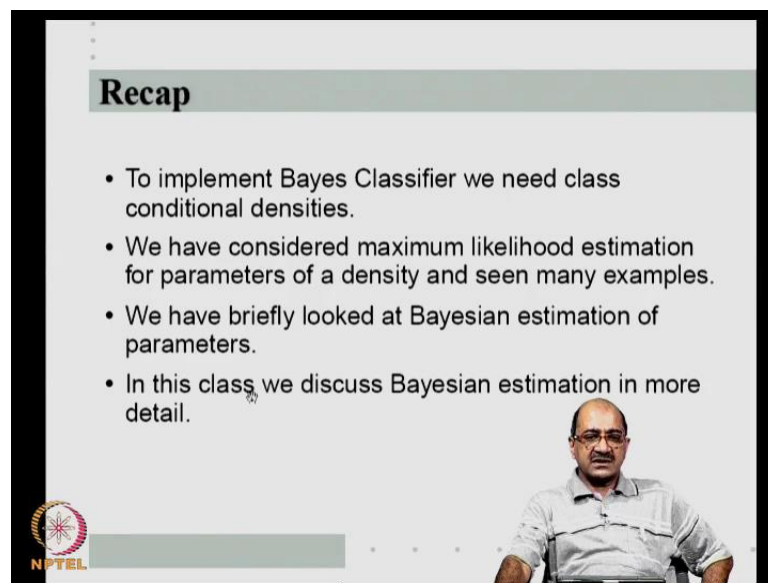


Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 7
Bayesian Estimation of Parameters of Density Functions, MAP Estimates

Hello, welcome to the next talk, in this Pattern Recognition course. Very briefly let us recall what you have been doing. We been considering ways to implement the Bayes classifiers, specifically we need to estimate the class conditional densities, for implementing the Bayes classifiers. So, we will be looking at various techniques for estimating class conditional densities, given iid samples from a density we want to estimate densities.

(Refer Slide Time: 00:46)



Recap

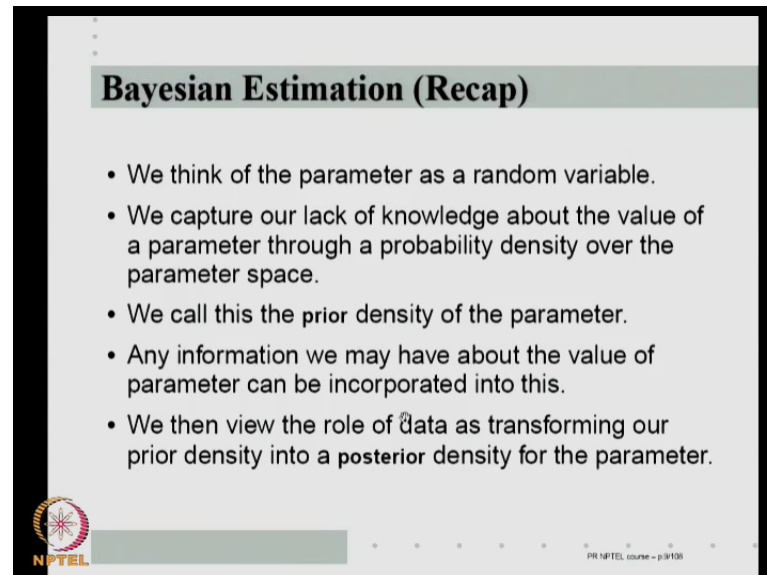
- To implement Bayes Classifier we need class conditional densities.
- We have considered maximum likelihood estimation for parameters of a density and seen many examples.
- We have briefly looked at Bayesian estimation of parameters.
- In this class we discuss Bayesian estimation in more detail.

NPTEL

Currently we are considering the parametric way of estimate density; that means, we assume the density is known but for values of some parameters; and we were looking at techniques for estimating the parameters. We have considered earlier the maximum likelihood estimation for parameters of a density. We seen many examples of how it is done and we seen in that many cases, we can actually obtain closed form solutions, say it is a very efficient technique. Then in the last class we briefly looked at another parametric way of estimating densities, what is called the Bayesian estimation of parameters. So, this class we will consider Bayesian estimation in more detail, we will


look at a few examples. So, both this class and part of next class we will be looking at Bayesian estimation.

(Refer Slide Time: 01:45)



Bayesian Estimation (Recap)


- We think of the parameter as a random variable.
- We capture our lack of knowledge about the value of a parameter through a probability density over the parameter space.
- We call this the **prior density** of the parameter.
- Any information we may have about the value of parameter can be incorporated into this.
- We then view the role of data as transforming our prior density into a **posterior density** for the parameter.

 PR NPTEL course - p18108

We discussed Bayesian estimation last class but let us briefly recall the basic idea of Bayesian estimation. In the maximum likelihood case, the parameters are unknown but are assumed constant. In the Bayesian estimation we think of the parameter itself as a random variable, what does that give us? It allows us to capture our knowledge or lack of knowledge, see we do not know the parameters that is what our lack of knowledge is, but it is may not be complete, completely unknown. At least for example, we know what space the parameter belongs to but we are certain about the actual value of the parameter.


So, this lack of knowledge or whatever partial knowledge we have, about the value of a parameter is captured through a probability density or the parameter space. As I said last class, we call this the prior density of the parameter, it is prior to seeing any data. So, before we do any experiments, our, whatever ideas we have about the parameter, are captured in this prior density. Any information we have about the value of the parameter can be incorporated into the prior density. Then we look at the estimation process, slightly differently the data itself is now used to transform our prior density into what is called a posterior density of the parameter, posterior is posterior to seeing the data. So, the data essentially transforms our prior density to a posterior density using Bayes theorem.

(Refer Slide Time: 03:21)



Bayesian Parameter Estimation

- As earlier, let θ be the parameter and let \mathcal{D} be the data
- Recall that
$$\mathcal{D} = \{x_1, \dots, x_n\}$$
is the set of iid data and each x_i has density $f(x_i | \theta)$ (which is the assumed model).
- Let $f(\theta)$ be the prior density of the parameter and let $f(\theta | \mathcal{D})$ be the posterior density.




This is the basic idea of the Bayesian estimation, we will see examples this class. So, to get our notation as earlier, unless we have some other symbol for the parameter, for generic parameters we use the symbol θ and \mathcal{D} is the data that we have. So, \mathcal{D} is simply x_1, x_2, \dots, x_n where, x_i 's are iid realization for the density, that is we have an underlying density model $f(x)$ given θ , which is known but, for the value values of θ .

So, each of these x_i 's have the same density $f(x_i)$ given θ , and they are independent because, this is the data we have and this f is the assumed probabilistic model for the density and we want the value of the θ . So, because we decided that the parameter itself is to be treated as a random variable, let $f(\theta)$ denote the prior density of the parameter and let $f(\theta | \mathcal{D})$ be the posterior density. Just one caution about notation just to keep notation simple we use f for densities of all random variables, when I want density of θ , I write it as $f(\theta)$ and when I want density of x , I am writing $f(x)$.

So, in that sense the symbol f is not specifying any specific function, normally in the in the probability literature. The random variable x is the density reverse to is put as a subscript on the density. So, for example, when I talking about the density of x , I should say $f_x(x)$ or $f(x)$, when I am talking about the density of θ , I should say $f_\theta(\theta)$ or $f(\theta)$ at an argument value θ and so on. It just clutters up the notation a little bit more. So, just using this unspecified f , as a density function for all random

variables that we are considering, the meaning will be clear from context and all of you should be a little cautious about this notation.


(Refer Slide Time: 05:21)



• Now, using Bayes theorem we get

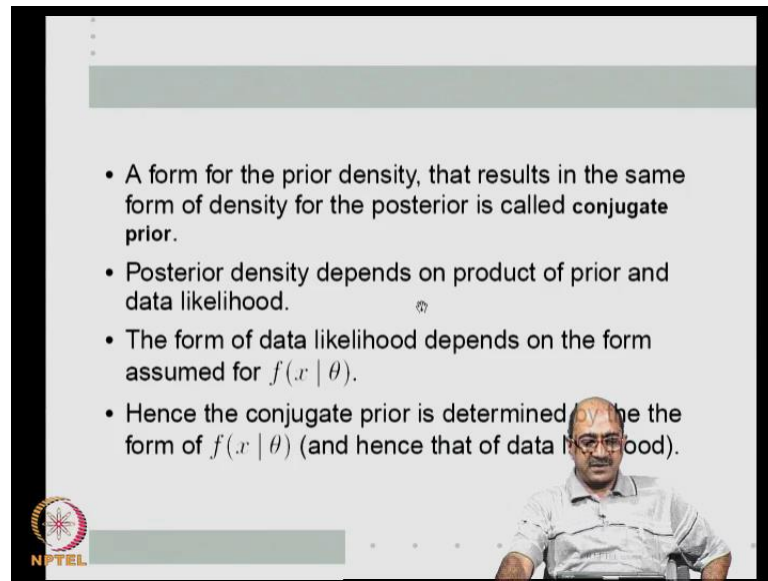
$$f(\theta | \mathcal{D}) = \frac{f(\mathcal{D} | \theta)f(\theta)}{\int f(\mathcal{D} | \theta)f(\theta) d\theta}$$

where $f(\mathcal{D} | \theta) = \prod_i f(x_i | \theta)$ is the data likelihood that we considered earlier.



So, once again $f(\theta)$ is the prior density of the parameter and $f(\theta | \mathcal{D})$ is the posterior density. So, the Bayes theorem tells us, how we can link the prior density to the posterior density. So, the posterior density $f(\theta | \mathcal{D})$ is given by $f(\mathcal{D} | \theta)f(\theta)$ into $f(\theta)$ by a normalizing constant where, $f(\mathcal{D} | \theta)$ because, \mathcal{D} consists of $n \times i$ each of them are iid, it is simply product of the individual marginal density values. So, $f(x)$ say given θ which is nothing but, the data likelihood that we have already seen.

(Refer Slide Time: 05:50)




- A form for the prior density, that results in the same form of density for the posterior is called **conjugate prior**.
- Posterior density depends on product of prior and data likelihood.
- The form of data likelihood depends on the form assumed for $f(x | \theta)$.
- Hence the conjugate prior is determined by the form of $f(x | \theta)$ (and hence that of data likelihood).


We also talk briefly about what are called conjugate priors last class, a form for the prior density that results, in the same form for the posterior density is called a conjugate prior. So, the prior for example, is normal, then we want posterior also to be normal, then the prior is conjugate. So, for a particular problem, if I choose the right form for the prior density, then the posterior may also become a density of the same class.

Such a prior is called a conjugate prior, see posterior density of course, depends on the product of the prior and the data likelihood. And the form of data likelihood depends on the form of the assumed $f(x)$ given θ the density that we are estimating. So, ultimately what will be a conjugate prior is determined by the form of $f(x)$ given θ that we use and hence that of the data likelihood, we will see examples later on.

(Refer Slide Time: 06:44)




- When we use conjugate prior, the prior and posterior would belong to the same class of densities.
- Hence calculating posterior would be like updating parameter values.



The reason why we want to use conjugate prior, is that the prior and posterior would belong to same class of densities and hence calculating the posterior would be like updating parameter values.


(Refer Slide Time: 06:59)



- Now, using Bayes theorem we get

$$f(\theta | \mathcal{D}) = \frac{f(\mathcal{D} | \theta)f(\theta)}{\int f(\mathcal{D} | \theta)f(\theta) d\theta}$$

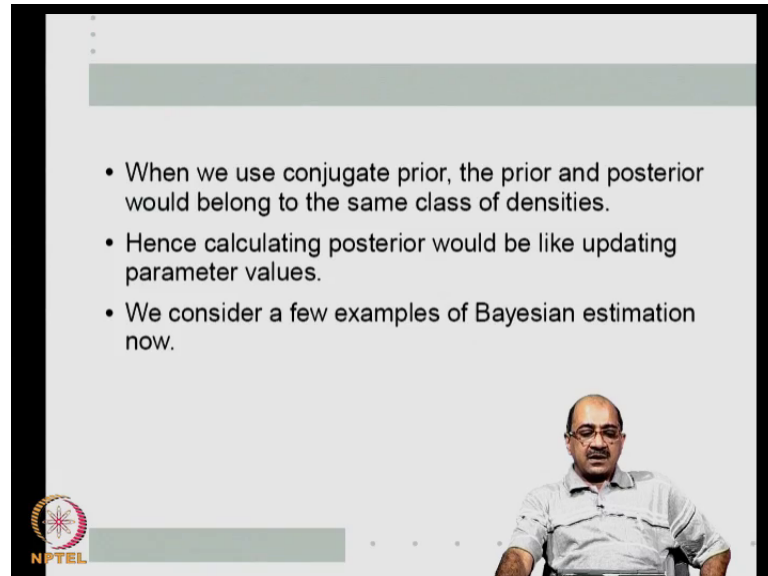
where $f(\mathcal{D} | \theta) = \prod_i f(x_i | \theta)$ is the data likelihood that we considered earlier.



What we mean is? In this expression if I know $f(\theta)$, let us say $f(\theta)$ is normal with some mean and variance. And for this problem the $f(\mathcal{D} | \theta)$ is of the form that when multiplied by $f(\theta)$ what I get is another normal density, then $f(\theta | \mathcal{D})$ also

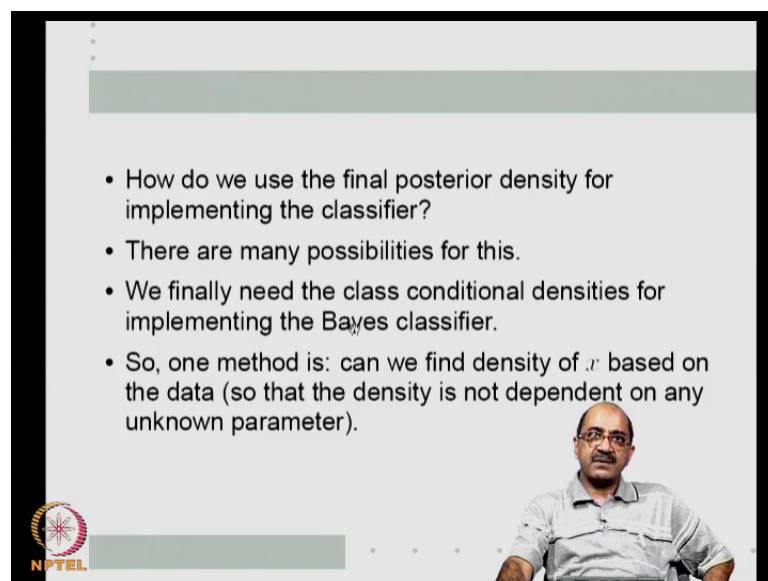
will be a normal. So, to transform $f(\theta)$ to $f(\theta)$ given D I have to only know what will be the new mean and variance.

(Refer Slide Time: 07:31)



So, in that sense because if the prior and posterior belong to the same family of densities, transforming prior to posteriors is just a parameter update step. Hence calculating the posterior would be like updating parameters values, we are going to consider examples.

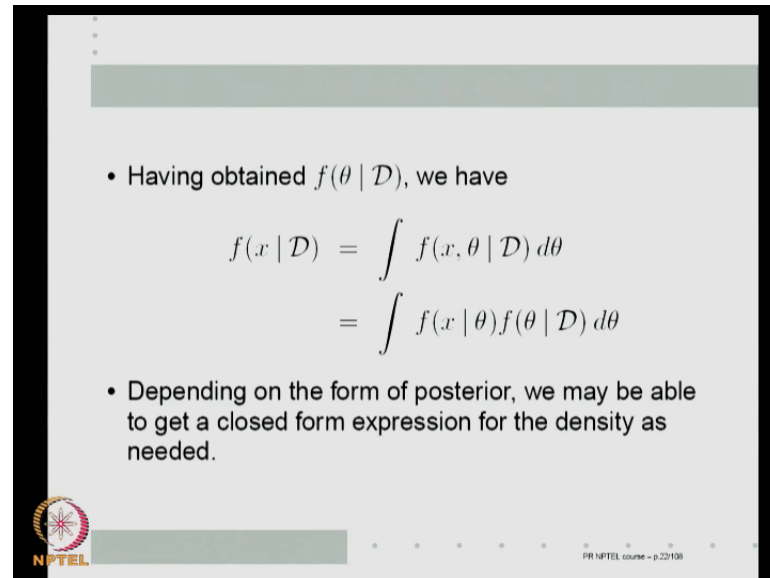
(Refer Slide Time: 07:40)



One final point, how do we use the final posterior density for implementing the classifiers, this is also what we discussed last time is worth recalling again. As I said last

time there are many possibilities for this. Ultimately, we need class conditional density for implementing Bayes classifier that is what we are doing. So, one method is given the posterior density, can we find a density for x based on the data.

(Refer Slide Time: 08:05)



- Having obtained $f(\theta | \mathcal{D})$, we have
$$f(x | \mathcal{D}) = \int f(x, \theta | \mathcal{D}) d\theta$$
$$= \int f(x | \theta) f(\theta | \mathcal{D}) d\theta$$
- Depending on the form of posterior, we may be able to get a closed form expression for the density as needed.

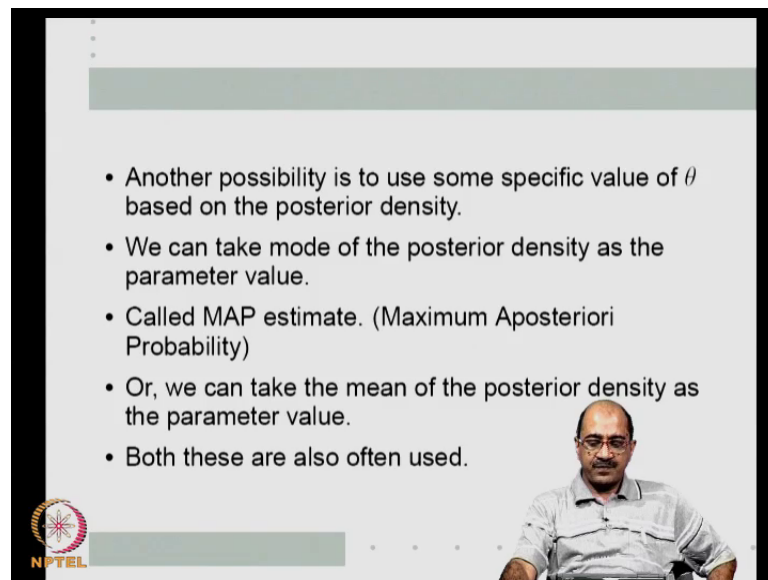
NPTEL

PR NPTEL course - p.22108

So, the density is not dependent on any unknown parameter, last class we seen how it can be done, if I know $f(\theta | \mathcal{D})$, we can write a density $f(x)$ given data that is density of x conditioned on the data. As as the marginal of the joint x comma θ conditioned on \mathcal{D} , by integrating with respect to θ . Now, this marginal can be split into $f(x | \theta)$ and $f(\theta | \mathcal{D})$ and given θ , x density of x does not depend on the data.

So, this integral becomes product of $f(x | \theta)$ $f(\theta | \mathcal{D})$ $d\theta$. $f(x | \theta)$ given θ is the assumed density model, $f(\theta | \mathcal{D})$ is the posterior that we have estimated. So, we can use this integral to find $f(x | \mathcal{D})$, which we can then use at the class conditional density. So, depending on the form of the posterior, we may be able to get a closed form expression for this density and then we can use that at the class conditional density. We will see examples of how to calculate this class, there are also other possibilities.

(Refer Slide Time: 08:58)

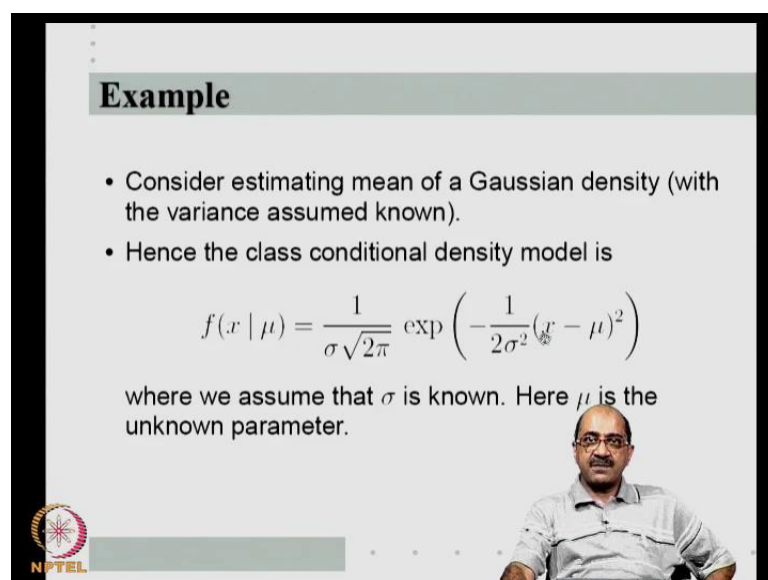


A slide with a light gray background and a dark gray header. The header contains a small green bar. The main content area has a list of five bullet points. In the bottom right corner, there is a small video inset showing a man with glasses and a mustache, wearing a light blue shirt, sitting at a desk. In the bottom left corner, there is a circular logo with a star and the text 'NPTEL' below it.

- Another possibility is to use some specific value of θ based on the posterior density.
- We can take mode of the posterior density as the parameter value.
- Called MAP estimate. (Maximum Aposteriori Probability)
- Or, we can take the mean of the posterior density as the parameter value.
- Both these are also often used.

We can use one particular value of theta based on the posterior. We can for example, take mode that the value at which the density has the highest value, when we do that is called a map estimate as I said last time maximum a posteriori probability estimate or we can take the mean of the posterior density, both these are also often used and we will see that in examples in this class, with that introduction let us move on to examples.

(Refer Slide Time: 09:23)



A slide with a light gray background and a dark gray header. The header contains the word 'Example' in bold. The main content area has two bullet points, followed by a mathematical equation, and then a line of text. In the bottom right corner, there is a small video inset showing a man with glasses and a mustache, wearing a light blue shirt, sitting at a desk. In the bottom left corner, there is a circular logo with a star and the text 'NPTEL' below it.

Example

- Consider estimating mean of a Gaussian density (with the variance assumed known).
- Hence the class conditional density model is


$$f(x | \mu) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x - \mu)^2\right)$$

where we assume that σ is known. Here μ is the unknown parameter.

So, first consider very, very simple example. We want to estimate the mean of a Gaussian density but, this time we assume even variance is known. So, only mean is

unknown we have iid samples from a Gaussian density, whose variance is known and the only unknown parameter is the mean. So, with that assumption what will be my assumed probability model of x given θ now becomes θ is the μ the mean of the Gaussian density is a one dimensional example. So, $f(x)$ given μ , this is the standard Gaussian density model. Once again I emphasize in this model, we are assuming σ is known and μ is the unknown parameter which we want to estimate.

(Refer Slide Time: 10:36)



- The likelihood is now given by

$$f(\mathcal{D} | \mu) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right)$$

- As a function of μ this has an exponential of a quadratic in μ .
- Hence, If the prior is normal (which has an exponential of a quadratic in μ) the product would once again be a normal density.
- Thus, the conjugate prior here is normal density.

PR NPTEL course - p.29/101

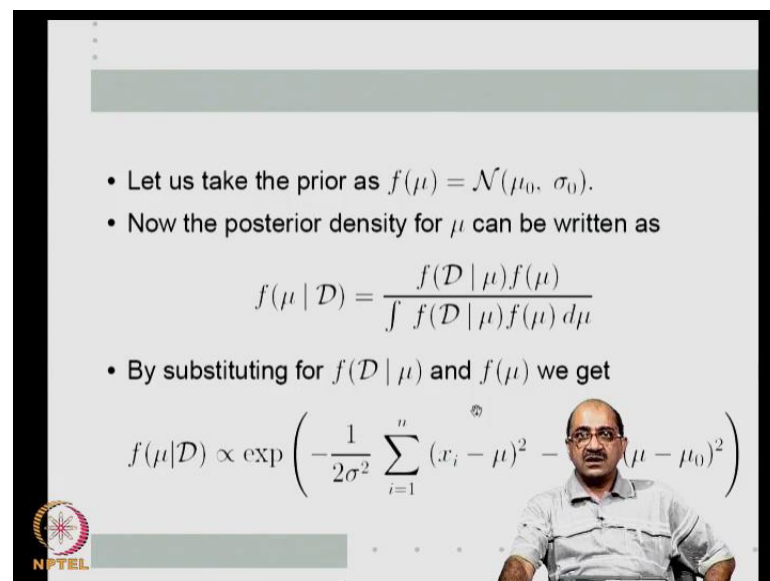
We have of course, as usual x_1 to x_n iid data from this density for estimation. So, what is the likelihood? Likelihood is simply product of $f(x_i)$ given μ . So, that becomes a just put x as x_i in this expression and then take n fold product, if I take n fold product I will get this term to the power n because this does not depend on x and here when I take the product all of them get added up inside the exponent right. So, my likelihood become this first term to the power n exponential some over i , x_i minus μ whole square.

Now, we are asking if this is the form of likelihood, what prior would be conjugate, for that this likelihood when viewed as a function of μ what is its main characteristic, is as a function of μ it has an exponential of a quadratic in μ . What is inside the exponent is some quadratic function μ , remember that x_i 's are given data they are known they are not the variable the variable is μ here.

So, inside exponential what I have is a quadratic in μ . So, as a function of μ is an exponential of a quadratic in μ . What is the density is it exponential quadratic in μ , a

normal density right. So, if the prior is normal, a normal density is exponential of quadratic right. Then if the prior density over μ is also normal, then what happens is I am multiplying one exponential, which has a quadratic in μ with another exponential there is a quadratic in μ because, inside the exponents thing added up I once again get an expression which is exponential of a quadratic in μ . And that means, my product would also be a normal density again, which means for this problem.

(Refer Slide Time: 13:08)

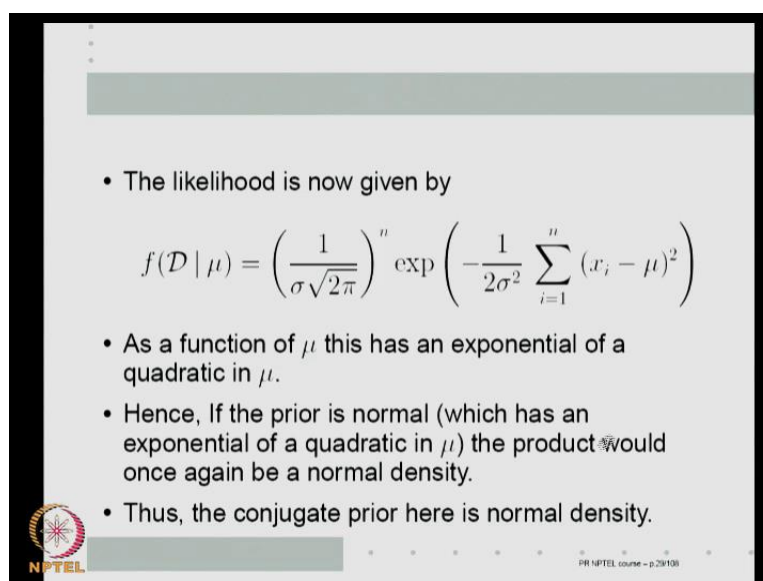


- Let us take the prior as $f(\mu) = \mathcal{N}(\mu_0, \sigma_0)$.
- Now the posterior density for μ can be written as
$$f(\mu | \mathcal{D}) = \frac{f(\mathcal{D} | \mu)f(\mu)}{\int f(\mathcal{D} | \mu)f(\mu) d\mu}$$
- By substituting for $f(\mathcal{D} | \mu)$ and $f(\mu)$ we get
$$f(\mu | \mathcal{D}) \propto \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

That is the problem of estimating mean of a one dimensional Gaussian with variance known the conjugate prior is a normal density. So, let us take the prior, prior is a density on μ remember we are thinking μ itself as a random variable. So, prior is a density on the variable μ . So, we are assuming it to be normal, we will use the script N to denote normal the notation is $\mathcal{N}(\mu_0, \sigma_0)$ is a normal density, whose mean is μ_0 and whose variance is σ_0 . So, we are thinking that the prior density is normal, we are not thinking, we are taking the prior density to be normal with mean μ_0 and variance σ_0 .

Now, we can calculate the posterior, the posterior of μ given \mathcal{D} is $f(\mathcal{D} | \mu)$ this is my likelihood, into $f(\mu)$ that is my prior, which is assumed to be normal with mean μ_0 and variance σ_0 , variance σ_0 square.

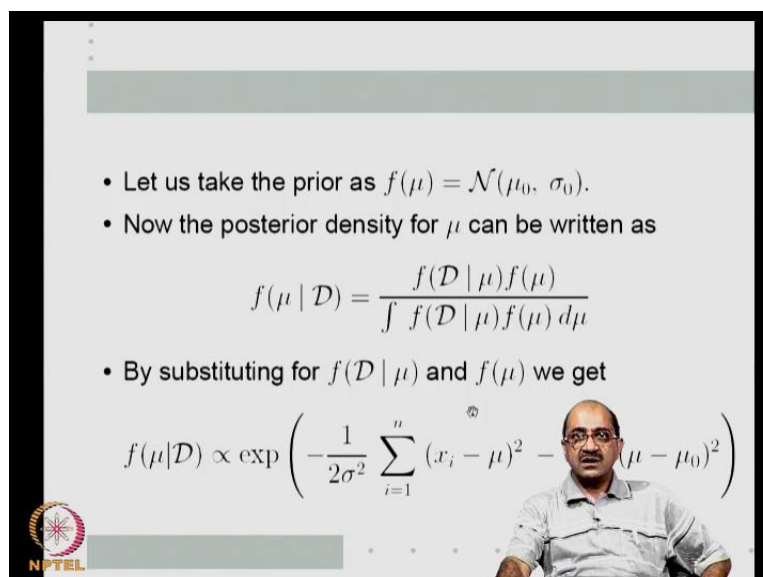
(Refer Slide Time: 13:16)

A presentation slide with a light gray background and a dark border. It contains a bulleted list and a mathematical equation. The NPTEL logo is in the bottom left corner, and the text 'PR NPTEL course - p.29/101' is in the bottom right corner.


- The likelihood is now given by
$$f(\mathcal{D} | \mu) = \left(\frac{1}{\sigma \sqrt{2\pi}} \right)^n \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 \right)$$
- As a function of μ this has an exponential of a quadratic in μ .
- Hence, If the prior is normal (which has an exponential of a quadratic in μ) the product would once again be a normal density.
- Thus, the conjugate prior here is normal density.

This is of course, a normalizing constant, so this is how I can calculate the posterior, so I have to just substitute. So, if I substitute $f(\mathcal{D} | \mu)$ we have from the previous slide right, that is $f(\mathcal{D} | \mu)$. So, I am skipping all the unnecessary constants, we only want the expression that depend on μ .

(Refer Slide Time: 12:12)

A presentation slide with a light gray background and a dark border. It contains a bulleted list, a mathematical equation, and a photograph of a man. The NPTEL logo is in the bottom left corner, and the text 'PR NPTEL course - p.29/101' is in the bottom right corner.

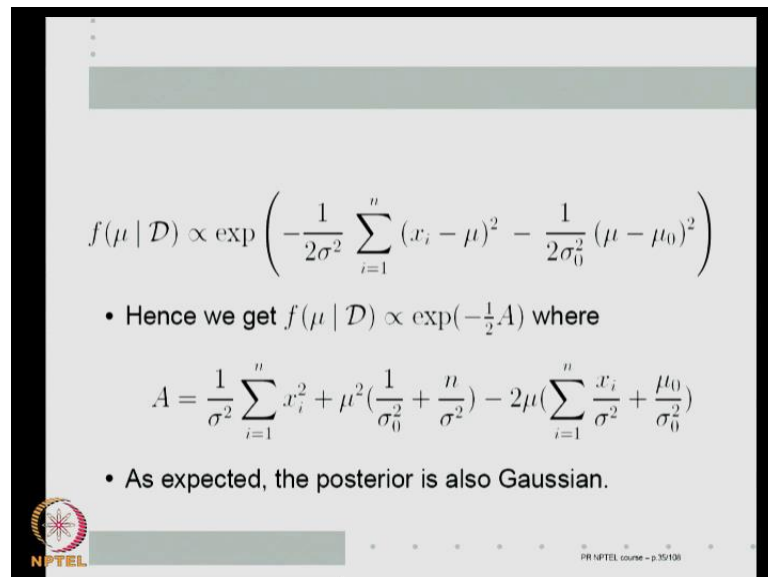
- Let us take the prior as $f(\mu) = \mathcal{N}(\mu_0, \sigma_0)$.
- Now the posterior density for μ can be written as
$$f(\mu | \mathcal{D}) = \frac{f(\mathcal{D} | \mu)f(\mu)}{\int f(\mathcal{D} | \mu)f(\mu) d\mu}$$
- By substituting for $f(\mathcal{D} | \mu)$ and $f(\mu)$ we get
$$f(\mu | \mathcal{D}) \propto \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

A small photograph of a man with glasses and a light blue shirt, sitting and looking towards the camera.

So, that is why I put a proportionality sign that is a unspecified constant. So, the posterior depends on the product of these two and the first I am $f(\mathcal{D} | \mu)$ is this exponential minus 1 by 2 sigma square sum of $x_i - \mu$ whole square and $f(\mu)$ is a normal

with mean μ_0 and variance σ_0^2 . So, that will add $\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2$, $\mu - \mu_0$ whole square. This is the only part of this product that depends on μ . The bottom the denominator in this expression is an integral with μ as the dummy variable is not dependent on μ , it's just some normalizing constant.

(Refer Slide Time: 14:21)



$$f(\mu | \mathcal{D}) \propto \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

- Hence we get $f(\mu | \mathcal{D}) \propto \exp(-\frac{1}{2}A)$ where

$$A = \frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 + \mu^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\sum_{i=1}^n \frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)$$

- As expected, the posterior is also Gaussian.

NPTEL PR NPTEL course - p.35/138

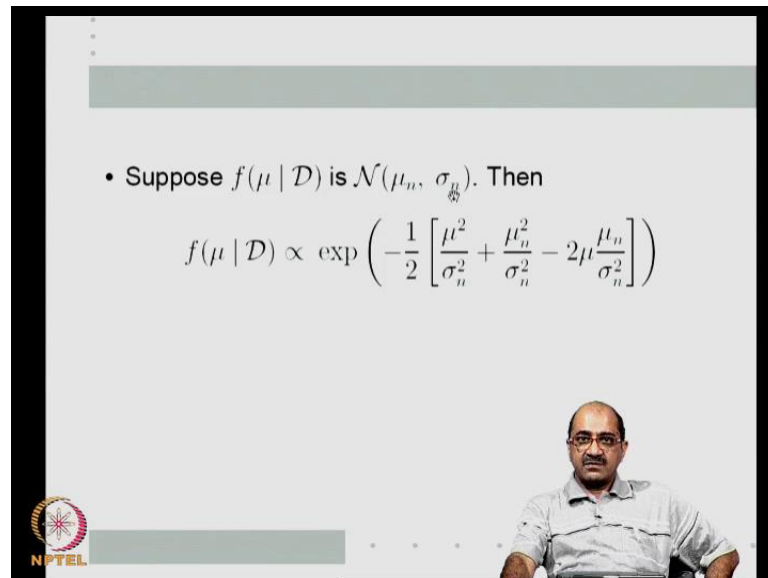
Similarly, both these densities have some constant that do not depend on μ . So, all those are subsumed in a constant that comes here. So, that is why I put a proportionality sign instead of equal to. So, the posterior is proportional to this expression. So, let us look a little more in this expression. So, posterior is proportional to $-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2$.

So, if I expand this and write it in a quadratic, I can think that this is proportional to $\exp(-\frac{1}{2}A)$, let us take the minus half factor out into some expression A, what is the expression A, what will come out of this, $\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 + \mu^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\sum_{i=1}^n \frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)$. So, one term is $\frac{1}{\sigma^2} \sum_{i=1}^n x_i^2$, then the μ^2 term comes out, I got a sum over i is equal to 1 to n of a constant 1 that will give me n .

So, I will get $\frac{n}{\sigma^2} \mu^2$, another μ^2 term will come from here and that will have factor $\frac{1}{\sigma_0^2}$. So, I will get μ^2 into $\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$, what is the left over terms minus 2μ term,

minus 2 mu here has coefficient summation x i by sigma square, the half is already being taken here right we are only looking at A. So, the 2 mu term from here, I will get summation x i by sigma square, the 2 mu term here will give me mu 0 by sigma 0 square that is the 2 mu term. So, finally, the posterior density is proportional to exponential minus half A, where A is a quadratic in mu given by this.

(Refer Slide Time: 16:23)



• Suppose $f(\mu | \mathcal{D})$ is $\mathcal{N}(\mu_n, \sigma_n^2)$. Then


$$f(\mu | \mathcal{D}) \propto \exp \left(-\frac{1}{2} \left[\frac{\mu^2}{\sigma_n^2} + \frac{\mu_n^2}{\sigma_n^2} - 2\mu \frac{\mu_n}{\sigma_n^2} \right] \right)$$

The slide also features an NPTEL logo in the bottom left corner.

So, what does this tell us because $f(\mu)$ given \mathcal{D} is proportional to exponential of quadratic in μ , as expected the posterior is also a Gaussian density. Now, I have to figure out for this Gaussian density. What is the mean and variance? Then I have completely characterized the posterior density. Because, I know that the posterior density is Gaussian, let us assume that posterior density is Gaussian with mean μ_n and variance σ_n^2 .

Then the form over the posterior density would be exponential, forgetting about the constants will be exponential minus half, μ minus μ_1 whole square by σ_n^2 . That is μ^2 by σ_n^2 , $\sigma_n^2 \mu_n^2$ by σ_n^2 minus 2 μ , μ_1 by σ_n . So, this will be the quadratic inside the exponent if the posterior is normal with mean μ_n and variance σ_n^2 .

(Refer Slide Time: 17:09)



$$f(\mu | \mathcal{D}) \propto \exp \left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2 - \frac{1}{2\sigma_0^2} (\mu - \mu_0)^2 \right)$$

- Hence we get $f(\mu | \mathcal{D}) \propto \exp(-\frac{1}{2}A)$ where


$$A = \frac{1}{\sigma^2} \sum_{i=1}^n x_i^2 + \mu^2 \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right) - 2\mu \left(\sum_{i=1}^n \frac{x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2} \right)$$

- As expected, the posterior is also Gaussian.

PR NPTEL course - p.35/136

We know what is the actual quadratic in the posterior, this is the actual quadratic in the posterior. So, if I want to find what is the μ_n and σ_n I have to just compare the terms right between that quadratic and this quadratic.

(Refer Slide Time: 17:12)




- Suppose $f(\mu | \mathcal{D})$ is $\mathcal{N}(\mu_n, \sigma_n)$. Then

$$f(\mu | \mathcal{D}) \propto \exp \left(-\frac{1}{2} \left[\frac{\mu^2}{\sigma_n^2} + \frac{\mu_n^2}{\sigma_n^2} - 2\mu \frac{\mu_n}{\sigma_n^2} \right] \right)$$

- Now, comparing with the earlier expression, we get

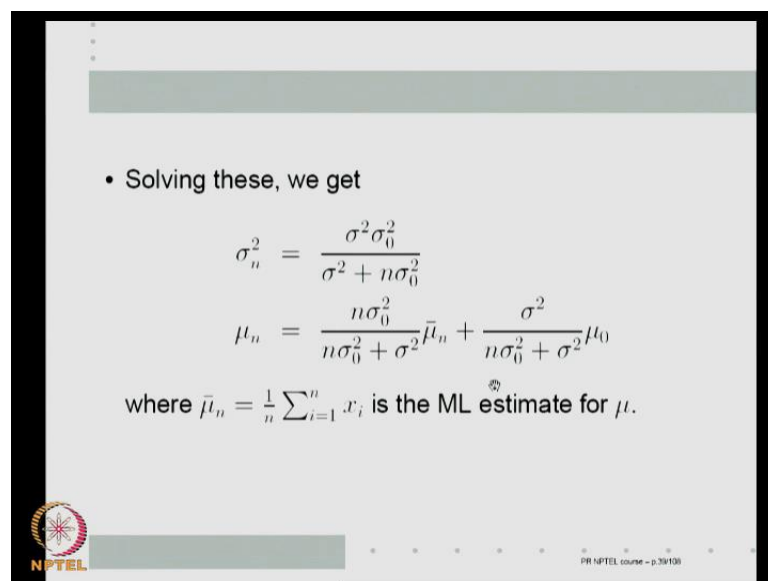
$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i + \frac{\mu_0}{\sigma_0^2}$$


So, comparing terms we get the following. So, first what is the coefficient of μ square here 1 by σ square n . (Refer Slide Time: 14:21) What is the coefficient of μ square here, 1 by σ_0 square plus n by σ square right. (Refer Slide Time: 17:12) So, I get 1 by σ square n is 1 by σ_0 square plus n by σ square.

Similarly, if I take the coefficient of 2μ that is μ_n by σ^2_n . (Refer Slide Time: 14:21) What is the coefficient of 2μ , we got earlier is 1 by σ^2_0 right. (Refer Slide Time: 17:12) So, by comparing coefficients we get these equations, where the unknowns are σ^2_n and μ_n right, those are that is what we want to find. We want to find what is the posterior density, we can easily solve them. For example, from this expression I get 1 by σ^2_n is σ^2_0 plus $n\sigma^2_0$ by σ^2_0 square, now you invert it we get σ^2_n .

(Refer Slide Time: 18:23)




• Solving these, we get

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n\sigma_0^2}$$


$$\mu_n = \frac{n\sigma_0^2}{n\sigma_0^2 + \sigma^2} \bar{\mu}_n + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2} \mu_0$$

where $\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i$ is the ML estimate for μ .

 PPT NPTEL course - p.39/101

So, that is σ^2_0 plus $n\sigma^2_0$ by σ^2_0 square.

(Refer Slide Time: 18:33)




- Suppose $f(\mu | \mathcal{D})$ is $\mathcal{N}(\mu_n, \sigma_n)$. Then

$$f(\mu | \mathcal{D}) \propto \exp \left(-\frac{1}{2} \left[\frac{\mu^2}{\sigma_n^2} + \frac{\mu_n^2}{\sigma_n^2} - 2\mu \frac{\mu_n}{\sigma_n^2} \right] \right)$$


- Now, comparing with the earlier expression, we get

$$\frac{1}{\sigma_n^2} = \frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}$$

$$\frac{\mu_n}{\sigma_n^2} = \frac{1}{\sigma^2} \sum_{i=1}^n x_i + \frac{\mu_0}{\sigma_0^2}$$


Similarly, now substituting that substituting that sigma square n in this expression, I can calculate mu n and that turns out to be this expression.

(Refer Slide Time: 18:39)




- Solving these, we get

$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n \sigma_0^2}$$

$$\mu_n = \frac{n \sigma_0^2}{n \sigma_0^2 + \sigma^2} \bar{\mu}_n + \frac{\sigma^2}{n \sigma_0^2 + \sigma^2} \mu_0$$

where $\bar{\mu}_n = \frac{1}{n} \sum_{i=1}^n x_i$ is the ML estimate for μ .

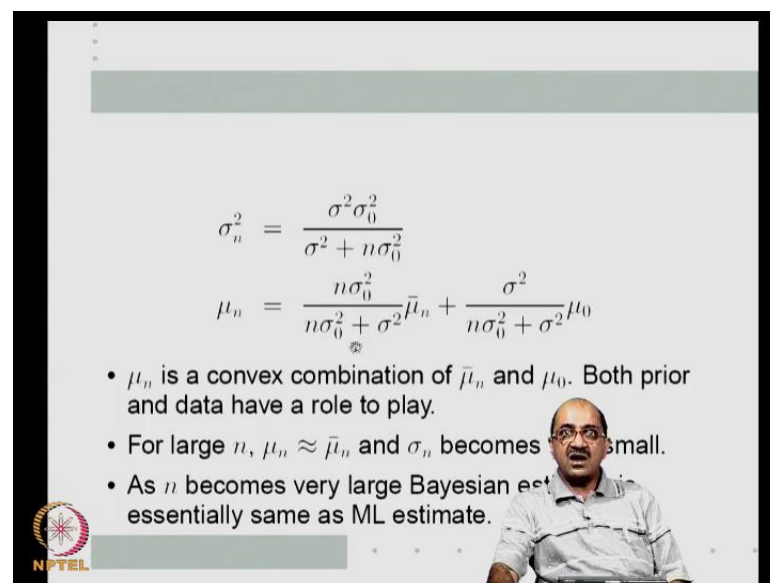
- The μ_n and σ_n completely specify the posterior density (after we have seen n examples)



This expression is very interesting. So, I am write that expression as n sigma 0 square by n sigma 0 square plus sigma square into mu n bar plus sigma square by n sigma 0 square plus sigma square into mu n into mu 0 where mu n bar, is just a symbol for this 1 by n sum i is equal to 1 to n x i, I just taken this expression out.

The reason for giving a symbol for this is this is the ML estimate right. So, the final posterior density is Gaussian with mean μ_n and variance σ_n^2 . And these completely specify the posterior density because, posterior density is Gaussian mean and variance completely specify it just a word about, why we chose this. So, we think of μ_n and σ_n^2 is the posterior density after we have seen n examples, in that sense μ_0 σ_0^2 are the density when we seen no examples that is the prior right. So, because μ_0 σ_0^2 is for prior μ_n σ_n^2 is the are the parameters of the posterior after we see n example.

(Refer Slide Time: 19:57)



$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n \sigma_0^2}$$

$$\mu_n = \frac{n \sigma_0^2}{n \sigma_0^2 + \sigma^2} \bar{\mu}_n + \frac{\sigma^2}{n \sigma_0^2 + \sigma^2} \mu_0$$

- μ_n is a convex combination of $\bar{\mu}_n$ and μ_0 . Both prior and data have a role to play.
- For large n , $\mu_n \approx \bar{\mu}_n$ and σ_n becomes small.
- As n becomes very large Bayesian estimate is essentially same as ML estimate.

Now, let us take a look at this μ_n and the expressions for μ_n square and σ_n^2 to see some interesting structure there. So, this is what we derived, as our final posterior density. Posterior density is Gaussian with mean μ_n and variance σ_n^2 which are given by this, in this μ_0 σ_0^2 are what we have chosen those are the prior densities that we have started with n is the number of examples.

σ^2 or σ_n^2 is the variance of the Gaussian of from which data is come which is assumed known. We are we are estimating the mean of a Gaussian whose variance is known. So, σ^2 is known, and $\bar{\mu}_n$ is simple an expression for $\frac{1}{n} \sum x_i$ which happens to be the ML estimate for μ in, in the same problem. So, thus μ_n is a convex combination I hope you can see that if it is some some constant into $\bar{\mu}_n$ plus some other constant into μ_0 both the constants are positive and they


add up to 1. So, μ_n is a convex combination of $\bar{\mu}_n$ and μ_0 . $\bar{\mu}_n$ is nothing, sample mean is the ML estimate which is the sample mean. So, if we did not do anything else all this Bayesian thing we could have simply taken sample means as the sample mean as the estimate.

But, we are not doing that we are taking as convex combination of the sample mean and the mean of the prior density, mean of the prior density is what we guessed originally without seeing any data as to what what is the most probable or what the expected mean expected value for the unknown mean is. So, our final estimate the the mean of the posterior density is a convex combination of μ_0 the mean of a prior density and $\bar{\mu}_n$ which is the sample mean right.


So, the final estimate has both the prior and the data play a role in the ML we are simply taking $\bar{\mu}_n$, as the as the final estimate. Here we are letting our initial beliefs about the values of the unknown unknown μ also to effect the final estimate right. That is how this became a convex combination of the sample mean $\bar{\mu}_n$ and the prior μ_0 . Second thing to notice is as n becomes large right once n becomes large $n \sigma_0^2$ square no matter what is the value of σ_0^2 square is will certainly dominate σ_n^2 square as as it keeps growing large and large. So, this this term becomes 1 and this term becomes 0.

So, as n becomes large the mean of the posterior simply become the sample mean. So, if I have large data my estimate is same as the maximum likelihood estimate, as it should be because maximum likelihood estimate is consistent, as n tends to infinity, it will give me the right value. So, I should go there any way and that is also given by my posterior densities σ_n goes to 0 as n tends to infinity. So, as n becomes large, σ_n becomes very small, μ_n becomes same as sample mean that means, the posterior density becomes more or less a dirac delta at the sample mean.

(Refer Slide Time: 23:42)


$$\sigma_n^2 = \frac{\sigma^2 \sigma_0^2}{\sigma^2 + n \sigma_0^2}$$
$$\mu_n = \frac{n \sigma_0^2}{n \sigma_0^2 + \sigma^2} \bar{\mu}_n + \frac{\sigma^2}{n \sigma_0^2 + \sigma^2} \mu_0$$

- 'Large n ' means $n \sigma_0^2 \gg \sigma^2$.
- We can say: μ_0 is our initial guess on μ . σ_0 determines the level of uncertainty in this guess.



So, that is my Bayesian estimate also right. So, as Bayesian becomes very large Bayesian estimate is essentially same as ML estimate as we expect. ML estimate is consistent as n tends to infinity it gives me the true value. So, Bayesian estimate should also give us the same thing. So, as n tends to infinity it becomes same as the ML estimate but, at any reasonable n the actual μ_n is a convex combination of $\bar{\mu}_n$ and μ_0 that means, both prior and the data have a role to play.

Actually i we can see it even little more by asking you know how large is large for n . So, these are our σ_n^2 and μ_n . So, what does large n mean essentially μ_n becomes approximately equal to the sample mean, when $n \sigma_0^2$ is much larger than σ^2 . How large n should be for this to happen depends on σ_0 . If σ_0 is small, n has to be very large to achieve this, if σ_0 itself is large then n does not have to be too large to achieve this. What does that mean? What does our prior say? Our prior is normal with μ_0 σ_0 .

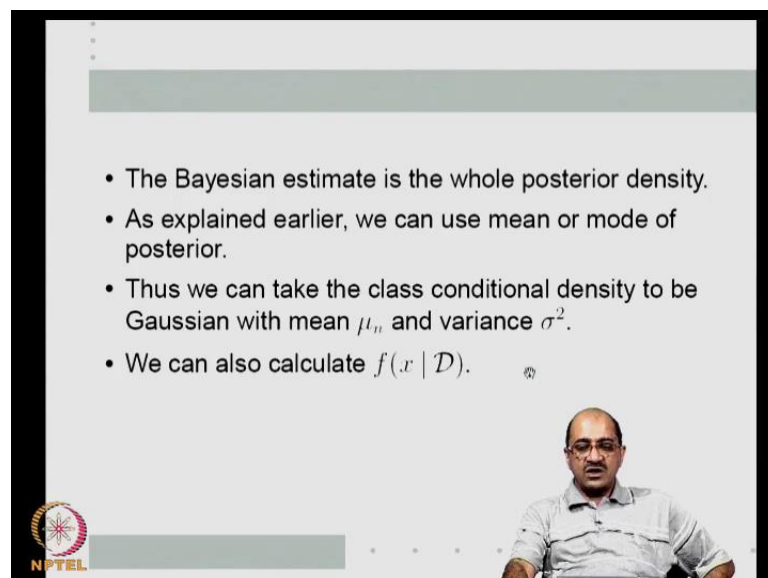
So, what we can think is initially we think the most probable or the or the expected value of the unknown parameter μ is μ_0 but, of course,, we are not sure and how 1 should is by determining by σ_0 . If my prior density has very small σ_0 that means, I have lot of faith in my initial guess μ_0 and on the other hand if σ_0 is very large; that means, I do not have much faith in my initial guess μ_0 . I can think of μ_0 as the initial guess and σ_0 as the label of uncertainty in this initial guess take a large σ_0

0 in my prior density means I am not too sure of my guess μ_0 right. That is why the the Gaussian will be very, very widely spread. If σ_0 is small I have lot of confidence.

So, if I have lot of if if μ_n becomes same as μ_0 , then the convex combination will always be same. If μ if anything μ_0 and sample mean happen to be the same then, obviously, this whole thing will be sample mean but, on the other hand if sample mean is much different from μ_0 . Then I would not believe it unless n is very large right, that is what small σ_0 would mean, on the other hand if σ_0 is large initially itself I do not have much faith in my guess.

Then moderate n would be enough for me to believing only the sample mean, by any case if the samples are very few then I would not let just sample mean take me as I have some control, by what prior I choose of course, μ_0 σ_0 is what we choose. So, that depends on whatever initial knowledge, we have about the about the area in the parameter space where, we think the true value of μ lies what its utility is that if the first few data that we got, by our misfortune happen to be out layers that immediately does not take away our guess too far our estimate too far.

(Refer Slide Time: 26:09)



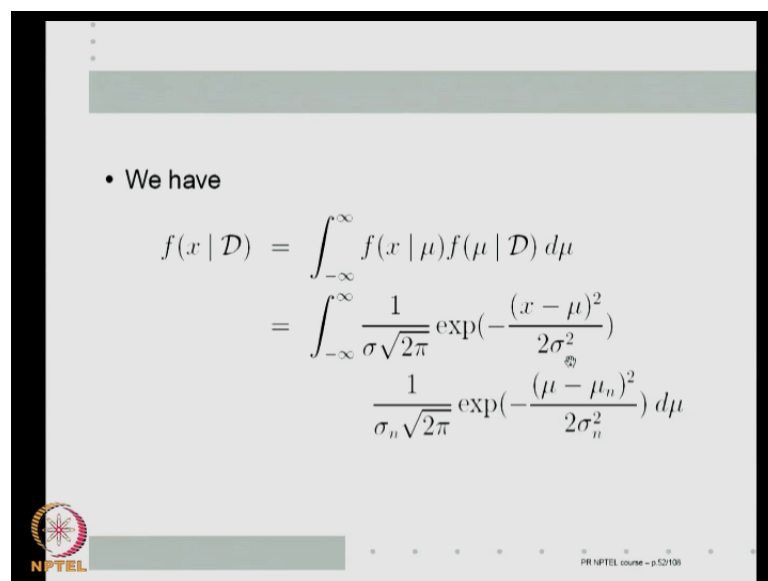
- The Bayesian estimate is the whole posterior density.
- As explained earlier, we can use mean or mode of posterior.
- Thus we can take the class conditional density to be Gaussian with mean μ_n and variance σ^2 .
- We can also calculate $f(x | \mathcal{D})$.

Now, of course, so, far (()) we just told you, we just calculated that the posterior density is Gaussian with mean μ_n and variance σ^2/n and we seen how to calculate μ_n and σ^2/n . So, the Bayesian estimate the whole posterior density. So,

what should we take as the final estimate, as we explained earlier today, we can use mean or more of the posterior density because the posterior density is Gaussian, it is mean and more both are equal to it is mean and that is μ_n .

So, which means I can simply take μ_n to be the value of the unknown mean that means, my class conditional density I am originally assuming it to be normal with mean μ which is unknown and variance σ^2 which is known. Now, my estimate could be simply mean or more of the posterior that means, I can simply take my class conditional density to be a Gaussian with mean μ_n and variance σ^2 this is one thing I can do. I can simply take the mean or more of the posterior density as the parameter value which happens to be μ_n and hence I will say my final class conditional density is Gaussian with mean μ_n and variance σ^2 because I started with a model where σ^2 is known.

(Refer Slide Time: 27:32)



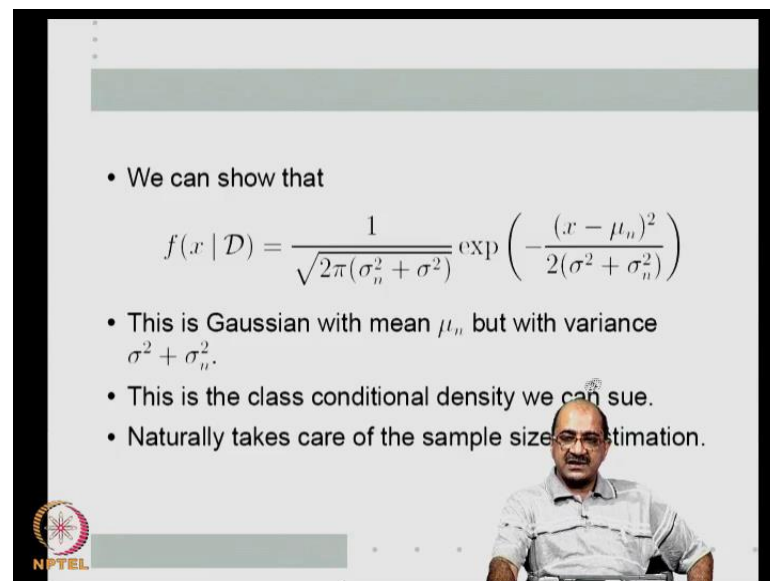
• We have

$$\begin{aligned}
 f(x | \mathcal{D}) &= \int_{-\infty}^{\infty} f(x | \mu) f(\mu | \mathcal{D}) d\mu \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{(x - \mu)^2}{2\sigma^2}\right) \\
 &\quad \frac{1}{\sigma_n \sqrt{2\pi}} \exp\left(-\frac{(\mu - \mu_n)^2}{2\sigma_n^2}\right) d\mu
 \end{aligned}$$

On the other hand as I said we can actually calculate, a density model for the data, for for x based only on the data right. So, in this particular problem it happens to be easy to calculate easy in a figurative sense we are not actually calculating the algebra involved is quite complicated as it would have taken at least four five slides for me to show you the algebra. But, let us just at least write the expression. So, $f(x)$ given \mathcal{D} is integral $f(x)$ given μ $f(\mu)$ given \mathcal{D} $d\mu$ right.

This is very integrated with respect to μ . So, only x and d will remain what is $f(x)$ given μ this is my assuming model, $\frac{1}{\sigma \sqrt{2\pi}} \exp\left(-\frac{x - \mu}{\sigma}\right)$ whole square by $2\sigma^2$, that is the first term multiply by second term $f(\mu)$ given D , μ given D is normal with mean μ_n and variance σ_n . So, that is $\frac{1}{\sigma_n \sqrt{2\pi}} \exp\left(-\frac{\mu - \mu_n}{\sigma_n}\right)$ whole square by $2\sigma_n^2$ right integrated with respect to μ . So, this is my expression for $f(x)$ given D . Of course, this looks a complicated integral, what I have is a quadratic in μ here right.

(Refer Slide Time: 28:59)



- We can show that

$$f(x | D) = \frac{1}{\sqrt{2\pi(\sigma_n^2 + \sigma^2)}} \exp\left(-\frac{(x - \mu_n)^2}{2(\sigma^2 + \sigma_n^2)}\right)$$

- This is Gaussian with mean μ_n but with variance $\sigma^2 + \sigma_n^2$.
- This is the class conditional density we can use.
- Naturally takes care of the sample size estimation.

X is x survives the integration. So, there is a quadratic in μ here quadratic in μ here μ_n and σ_n are given. So, ultimately I get some constant we have to exponential quadratic in μ . So, you can at least see that they will be some simplifications possible, by using the standard Gaussian integral, after lot of length the algebra one can actually crunch this integral. And show that this turns out to be this right, once I do all the algebra right that expression turns out to be this. What is this? This says $f(x)$ given D is Gaussian with mean μ_n and variance $\sigma^2 + \sigma_n^2$ right. So, for example, instead of just taking μ_n as which is the mean or mode of the posterior density as my estimate. I can take my class conditional density estimate to be this that means, I will use a class conditional density, which is still Gaussian with mean μ_n as earlier but, now variance is $\sigma^2 + \sigma_n^2$. If I take the mean or mode of the class conditional density as my estimate, I would have taken my mean and mode of the posterior density as my estimate. Then final class conditional density would have

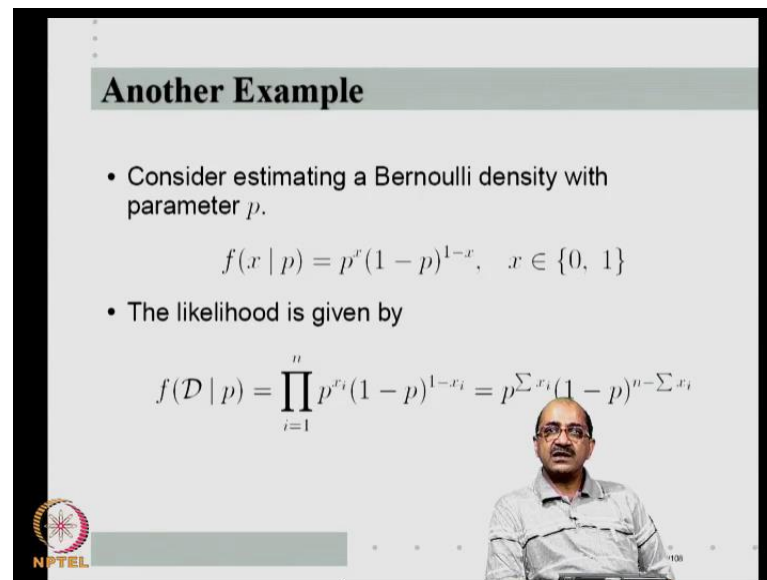
been Gaussian with mean μ_n and variance σ^2 which is originally assumed or no but, what our calculation $f(x)$ given D shows us. That the right class conditional density to use is Gaussian with mean μ_n but, with variance $\sigma^2 + \sigma_n^2$. The idea is μ_n is only estimate of the unknown mean and hence σ_n^2 is our initial our at that level uncertainty in our estimate.

So, it is better to take our class conditional density after seeing n samples, as Gaussian with μ_n and variance $\sigma^2 + \sigma_n^2$. Once again as n becomes larger and larger, this σ_n^2 becomes 0 and this becomes the ML estimate which is consistent right. So, seamlessly as the data increases ultimately it becomes Gaussian with sample mean as the as the value of the unknown parameter and variance to be σ^2 . So, which means if I use this class conditional density, naturally we take care of the sample size problems.

If you remember at the end of ML, we said the main problem with the ML is while it is consistent at small sample, sample size our estimate may be may not be too good. Now, here by calculating $f(x)$ given D we realize that if when I am taking μ_n as my estimate. I do not simply plug μ_n into the into the density model we assumed, which is Gaussian with mean μ_1 known variance σ^2 . I increase the variance of the class conditional density model I used to, to account for my current level of uncertainty in my estimate. So, this is another another very interesting thing about Bayesian estimation it it does take care of the sample size.

So, this one example of course,, as you can see when I derived the ML estimate, we actually taken a problem where both mean and variance of the Gaussian are known, we taken our first simple example as the unknown density is Gaussian and we directly estimated both unknown mean here. Just estimating the mean itself involves considerable algebra but, essentially it allows us to take prior beliefs into account, ensure that in at all sample sizes we do not we do not get, let us take too easily by out layers and so on. That is what we are getting gaining by Bayesian estimation, for the increase in complexity right. We will see one more simple example right.

(Refer Slide Time: 32:24)



Another Example

- Consider estimating a Bernoulli density with parameter p .

$$f(x | p) = p^x (1 - p)^{1-x}, \quad x \in \{0, 1\}$$

- The likelihood is given by


$$f(\mathcal{D} | p) = \prod_{i=1}^n p^{x_i} (1 - p)^{1-x_i} = p^{\sum x_i} (1 - p)^{n - \sum x_i}$$

NPTEL


Consider estimating a Bernoulli density, this is the second example we took even in our even we did ML maximum likelihood. So, like that we here also let us consider estimating a Bernoulli density with parameter p . So, what is the Bernoulli density with parameter p ? P is $(())$ with parameters. So, the densities of f of x given p is p power $(())$ for 1 minus $(())$ only 0 and 1 value. So, by density I mean is actually a mass function right. It has only two values $f(0)$ given p , $f(1)$ given p . $f(0)$ given p is $1 - p$ $f(1)$ given p , p is a Bernoulli. So, it takes value 1 with parameter with probability p and 0 with probability $1 - p$.

So, what is the likelihood? Likelihood \mathcal{D} is the product of this over x_i that is i is equal 1 to n , $f(x_i)$. So, what is $f(x_i)$? P to the power x_i $1 - p$ to the power $1 - x_i$. So, if I do this product I get p to the power of some x_i into $1 - p$ to the power of $n - \sum x_i$. That is my likelihood. So, the likelihood is of the form p to the power something into $1 - p$ to the power something. Now, I have to multiply this likelihood with some prior and the product should once again be the same form as the prior. So, if I choose prior also to be p to the power something into $1 - p$ to the power something, then the product will once again be p to the power something into $1 - p$ to the power something.

(Refer Slide Time: 33:59)




• Hence the conjugate prior should have the form

$$f(p) \propto p^a (1-p)^b, \quad p \in [0, 1]$$


So, the prior that we need right should be some density which is proportional to p to the power of a into $1 - p$ to the power b and this should be a density because p is a continuous value it can take any value between 0 and 1. So, it should be a well defined density over the space $0, 1$ and should have a form which say that the density is proportion density is some constant times p to the power of something into $1 - p$ to the power f something.

(Refer Slide Time: 34:25)




Another Example

- Consider estimating a Bernoulli density with parameter p .

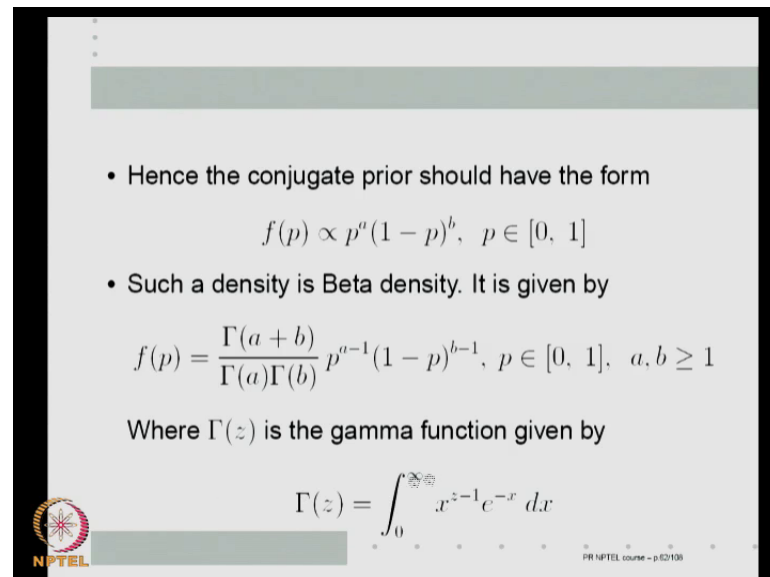
$$f(x | p) = p^x (1-p)^{1-x}, \quad x \in \{0, 1\}$$

- The likelihood is given by

$$f(D | p) = \prod_{i=1}^n p^{x_i} (1-p)^{1-x_i} = p^{\sum x_i} (1-p)^{n - \sum x_i}$$


The reason why this this should be the right prior is if that was the prior and then i multiply with the likelihood then what I get is p to the power of something into 1 minus p to the power of something. So, which means the posterior will once again be in the same class of the prior.

(Refer Slide Time: 34:33)



• Hence the conjugate prior should have the form


$$f(p) \propto p^a (1-p)^b, \quad p \in [0, 1]$$

• Such a density is Beta density. It is given by

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}, \quad p \in [0, 1], \quad a, b \geq 1$$

Where $\Gamma(z)$ is the gamma function given by


$$\Gamma(z) = \int_0^{\infty} x^{z-1} e^{-x} dx$$

 PR NPTEL course - p 62/108

So, the question is what are the densities that have this form? An important density which had this form is, what is called the beta density, it has two parameters a and b and it is given by f p is gamma of a plus b by gamma of a into gamma of b into p to the power of a minus 1, 1 minus p to the power of b minus 1 the density over 0 1 and a, b are constants which are assumed to be greater than or equal to 1. Where the gamma is the gamma function, the gamma function is given by gamma of z is 0 to infinity x power z minus 1 e power minus x d x. I hope all of you have have come across gamma functions sometime

There are many, many places where you could have studied gamma function including your probability course but, any way with this as the gamma function this is the density. While I am assuming that all the people using this course, know basic probability which means you know about density, joint densities you know all the standard densities, such as exponential, Gaussian, Bernoulli, binomial, Poisson and so on.

(Refer Slide Time: 35:57)



• The Beta(a, b) density is

$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}, \quad p \in [0, 1], \quad a, b \geq 1$$

• This is an important density over $[0, 1]$.

• When $a = b = 1$ it reduces to the uniform density.

• To show that this is a density we need to show

$$\Gamma(a)\Gamma(b) = \Gamma(a+b) \int_0^1 p^{a-1} (1-p)^{b-1} dp$$

PR NPTEL course - p.05/130

This beta density is not something that one often comes across in first course in probability. So, we will spend a little time on understanding the beta density first. So, beta density beta a comma b is beta density with parameter a and b is a density given by $f(p)$ is gamma of a plus b by gamma a into gamma b into p to the power a minus 1 1 minus p to the power b minus 1, this is the density.

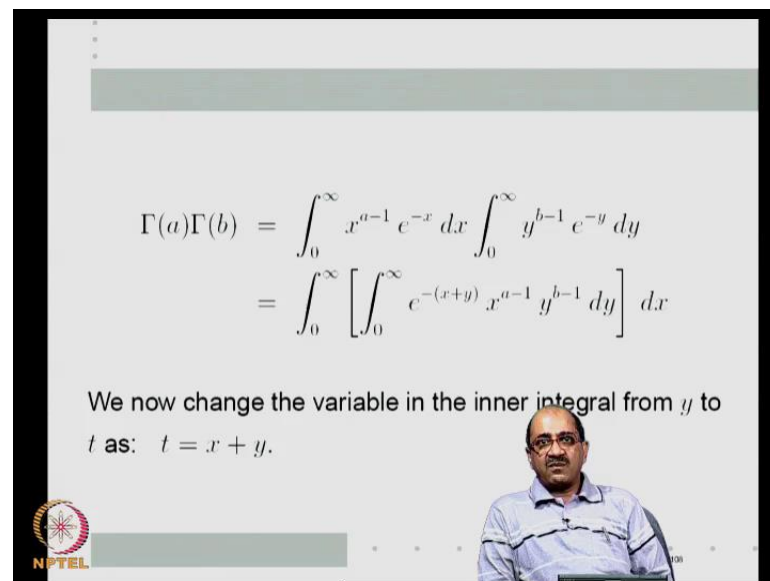
This is a very important density over $0, 1$ many other say $0, 1$ all of you know for example, uniform is 1 density over $0, 1$ that everybody knows. So, after uniform density possibly, this is 1 of the very important continuous densities, which are defined over $0, 1$ where the random variable takes values only over $0, 1$. If a is equal to, b is equal to 1, if I put a is equal to 1, b is equals to 1 $f(p)$ is a constant, for all p belonging to $0, 1$. That is nothing but, the uniform density the constant should turn over to be the right constant, namely 1 but, that that it will will currently show that this is a density.

So, if a is equal to b is equal to 1, it reduces to uniform density but, if a is equal to b is if a and b are different. Then this is a density which is not uniform over. So, that is how we would be able to to give some bias. To the give some some view to our initial views about the values of p . We will come back to that again later but, just right now let us remember that a is equal to b is equal to 1, it reduces to uniform density.

Now, first let us show that this is the density, what do you have to show for for this to be a density? A gamma function is obviously, positive and p is between 0 and 1. A and b

are greater than equal to 1. So, this entire expression is positive. So, the only thing we have to show is integrated over 0 to 1 with respect to p this expression should be 1. So, which is same as gamma a gamma of a plus b integral 0 to 1 with p power a minus 1 1 minus p b minus 1 d p this should be equal to gamma a into gamma b. So, that if I bring gamma a gamma b this side that whole thing will become 1.

(Refer Slide Time: 38:17)



$$\Gamma(a)\Gamma(b) = \int_0^\infty x^{a-1} e^{-x} dx \int_0^\infty y^{b-1} e^{-y} dy$$

$$= \int_0^\infty \left[\int_0^\infty e^{-(x+y)} x^{a-1} y^{b-1} dy \right] dx$$

We now change the variable in the inner integral from y to t as: $t = x + y$.

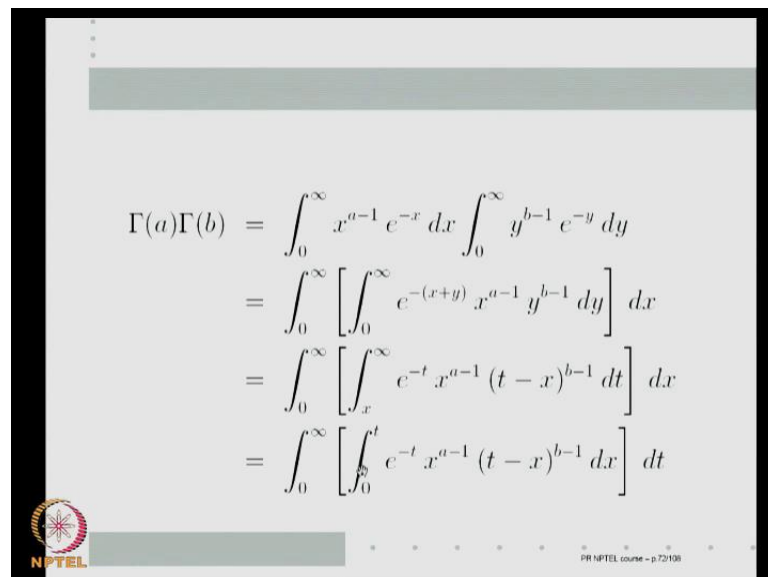
So, this is what I have to show, to show that this is a density, this is a little non trivial thing. So, let us just show it, so first. So, this is what we want to show. So, let us take the LHS gamma a into gamma b and ask what is that. So, this is the gamma function. So, because I want product of two gamma functions, I have taken different dummy variables. So, gamma a is x to the power a minus 1 e power minus x d x over 0 to infinity gamma b is y to the power b minus 1 e power minus y d y over 0 to infinity because, I have taken variables I can write it as a double integral. So, 0 to infinity 0 to infinity e power minus x plus y x to the power a minus 1 y to the power b minus 1 d y d x. I need some order of integration, it really does not matter they are independent details here. So, I just chose d y and then d x.

Once I do that in the inner integral we will make a small change of variable, so the inner integral with respect to y. So, let us change change y to t using the transformation t is equal to x plus y. What will this mean d t is d a d y is d t. So, that does not change when y takes value 0 t takes value x when y takes infinity t takes infinity. So, the limits 0 to

infinity for y will become x to infinity for t and x plus y will be substituted by t and x and y itself will become t minus x.

Now, this integral will now become, x plus y has become t x power a minus 1 as it is y has become t minus x the variable has become d t and limits have changed from x to infinity. So, this is a d t d x integral x is the outer integral that goes from 0 to infinity, t is the inner integral that goes to x to infinity. So, let us say I want to change the order of integration that is I want to first integrate with respect to x and then integrate with respect to t.


(Refer Slide Time: 39:32)



$$\begin{aligned}
 \Gamma(a)\Gamma(b) &= \int_0^\infty x^{a-1} e^{-x} dx \int_0^\infty y^{b-1} e^{-y} dy \\
 &= \int_0^\infty \left[\int_0^\infty e^{-(x+y)} x^{a-1} y^{b-1} dy \right] dx \\
 &= \int_0^\infty \left[\int_x^\infty e^{-t} x^{a-1} (t-x)^{b-1} dt \right] dx \\
 &= \int_0^\infty \left[\int_0^t e^{-t} x^{a-1} (t-x)^{b-1} dx \right] dt
 \end{aligned}$$


So, how do the limits change, here for each x, t goes from x to infinity, which means for each t x goes from 0 to t. So, if I change the order of integration this expression will become the outer outer integral is now d t integral, right the inner integral is d x integral. So, it becomes 0 to infinity, the t goes from 0 to infinity, now and x goes from 0 to t earlier x goes from 0 to infinity t goes from x to infinity which means for each x t goes from x to infinity which means very simple geometry will tell you which means for each t x goes from 0 to t. Now, let us start with this.

(Refer Slide Time: 40:46)




$$\Gamma(a)\Gamma(b) = \int_0^\infty \left[\int_0^t e^{-t} x^{a-1} (t-x)^{b-1} dx \right] dt$$

Now, in the inner integral we change the variable from x to u as: $x = tu$. (When x goes from 0 to t , u goes from 0 to 1. Also, $dx = t du$).



So, we come up till here gamma a gamma b is given by this expression, now once again we will make a change of variable in the inner integral right, we change x to tu where u is the new variable.

(Refer Slide Time: 41:23)



$$\begin{aligned} \Gamma(a)\Gamma(b) &= \int_0^\infty \left[\int_0^t e^{-t} x^{a-1} (t-x)^{b-1} dx \right] dt \\ &= \int_0^\infty \left[\int_0^1 e^{-t} t^{a-1} u^{a-1} t^{b-1} (1-u)^{b-1} t du \right] dt \\ &= \int_0^\infty \left[\int_0^1 e^{-t} t^{a+b-1} u^{a-1} (1-u)^{b-1} du \right] dt \\ &= \int_0^\infty e^{-t} t^{a+b-1} dt \int_0^1 u^{a-1} (1-u)^{b-1} du \end{aligned}$$

PR NPTEL course - p.77/138

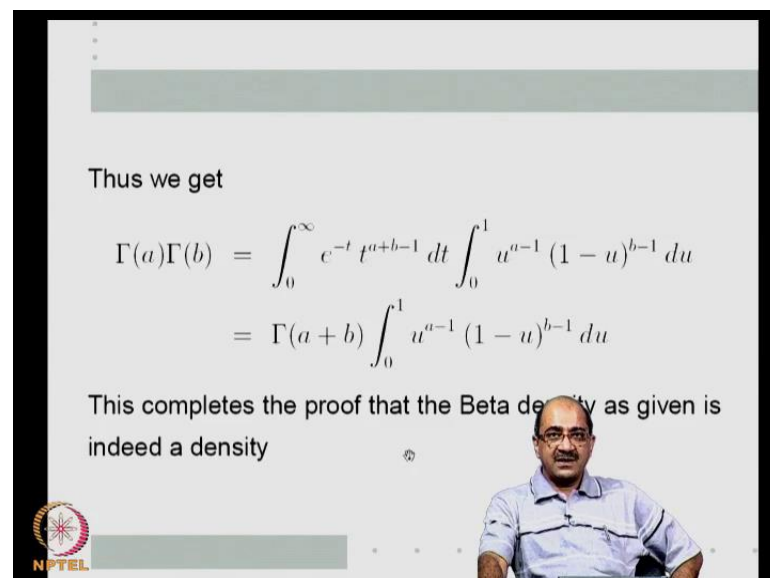
What does this give us, when x goes from 0 to t what happens to u , when x is equal to 0, u is 0. When x is equal to t u is equal to 1. So, as x goes from 0 to t , u goes from 0 to 1 and dx is $t du$. So, using that we change the variable from x to u in the inner integral. So, that will give me a change from x to u this becomes a du integral e power minus t

stays x has become $t u$. So, that is t to the power of $a - 1$ t to the power of $b - 1$ and u to the power of $a - 1$.

Now, once again x has become $t u$, so I can take t common. So, it becomes t to the power $b - 1$ into $1 - u$ to the power $b - 1$ and dx has become $t du$. Now, I can gather to get all the t term this is a t power $a - 1$ here, t power $b - 1$ here and a t here. So, that gives me t power $a + b - 1$ because $1 - 1$ will cancel with this t . So, that will give me this 0 to infinity 0 to t e^{-t} t to the power $a + b - 1$. Then u to the power of $a - 1$ $1 - u$ to the power $b - 1$ du . Now, as you can see in this double integral, right the limits do not depend on each other and the t and u terms are separable. So, I can write this as a product of two integrals right.

So, I bring all the t terms out 0 to infinity e^{-t} t to the power $a + b - 1$ dt . And all the u terms separately 0 to 1 u to the power $a - 1$ $1 - u$ to the power $b - 1$ du . Now, are we done, can we recognize this as 0 to infinity e^{-t} t to the power something dt . This is nothing but, a gamma function, gamma function of $a + b$ right.

(Refer Slide Time: 42:53)



Thus we get

$$\Gamma(a)\Gamma(b) = \int_0^\infty e^{-t} t^{a+b-1} dt \int_0^1 u^{a-1} (1-u)^{b-1} du$$

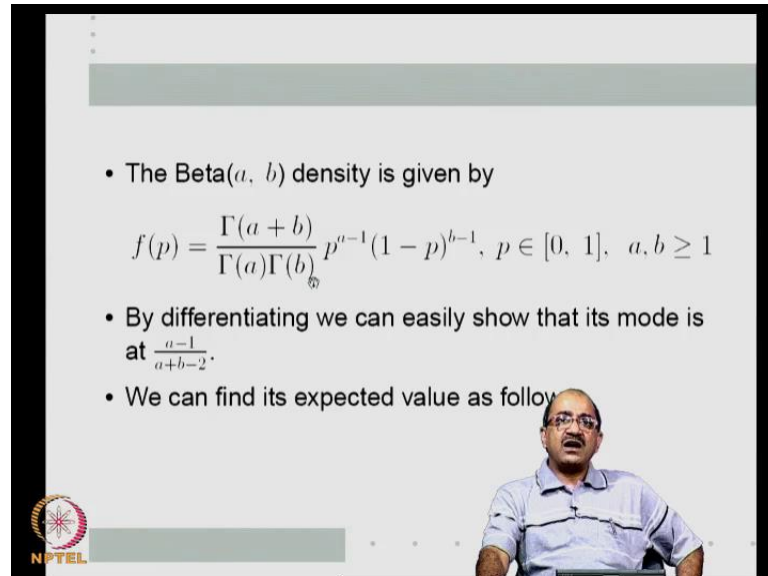
$$= \Gamma(a+b) \int_0^1 u^{a-1} (1-u)^{b-1} du$$

This completes the proof that the Beta density as given is indeed a density

So, what have we got, this is a gamma function $a + b$. So, is gamma of $a + b$ 0 to 1 u to the power of $a - 1$ $1 - u$ to the power $b - 1$ du . So, if I taken u to be p , then this is exactly what we wanted gamma a gamma b is gamma of $a + b$ into 0 to 1 p

to the power of a minus 1 into 1 minus p to the power b minus 1 d p. So, this completes the proof that the beta density is indeed a density.

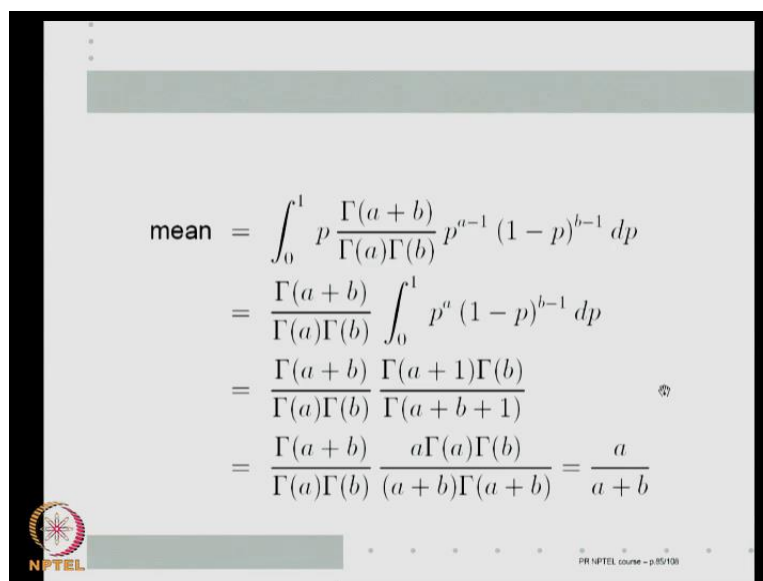
(Refer Slide Time: 43:23)



- The Beta(a, b) density is given by
$$f(p) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1}, p \in [0, 1], a, b \geq 1$$
- By differentiating we can easily show that its mode is at $\frac{a-1}{a+b-2}$.
- We can find its expected value as follows

There are some interesting things about the beta density, this is the beta density. Suppose I want the mode of the beta density that means, I want the value of p at which f (p) is maximum that is very simple you differentiate this with respect to p equate to 0 using that 1 can easily calculate that the mode occurs at a minus 1 by a plus b minus 2. Once again if a is equal to 1 and b is equal to 1, this is an undefined quantity 0 by 0; obviously, the uniform density has no more. Similarly, we can also find the mean value of this.

(Refer Slide Time: 44:01)



$$\begin{aligned}
 \text{mean} &= \int_0^1 p \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} p^{a-1} (1-p)^{b-1} dp \\
 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^a (1-p)^{b-1} dp \\
 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(a+1)\Gamma(b)}{\Gamma(a+b+1)} \\
 &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{a\Gamma(a)\Gamma(b)}{(a+b)\Gamma(a+b)} = \frac{a}{a+b}
 \end{aligned}$$


How do I find the mean value of this? If I want to find mean this is the $f(p)$ density. So, I have to do p into $f(p) dp$ from 0 to 1. This is also easy to do. So, let us do this. So, bring $\Gamma(a+b)$ this constant out. So, I have 0 to 1 this p goes with p of a minus 1 to give me p of $a+1$ minus p power b minus 1 dp . Now, because you already shown beta density to be a density right. We know if I multiply this with see it is just another beta density, where in place of a I have $a+1$ in place of b I have b .

So, I know this integral can be written in terms of numbers. So, that will be $\Gamma(a+1)\Gamma(b)$ by $\Gamma(a+b+1)$ right. Because, if I had given 1 by this at the constant here, this would have been 1 because that is a beta density. Now, I suppose all of you know the standard property of the gamma function which can be obtained by integrating by parts the gamma function.

That is for any x $\Gamma(x)$ is $x\Gamma(x)$, $\Gamma(x+1)$ is $x\Gamma(x)$. So, $\Gamma(a+1)$ is $a\Gamma(a)$ and $\Gamma(a+b+1)$ is $(a+b)\Gamma(a+b)$ right. So, if I write that it becomes $\Gamma(a+b)$ by $\Gamma(a)\Gamma(b)$ into $a\Gamma(a)\Gamma(b)$ by $(a+b)\Gamma(a+b)$. So, cancel all the gamma terms we get a by $a+b$ at the mean. So, for the beta density a by $a+b$ is the mean and $a-1$ by $a+b-2$ is the mode so, much about the beta density let us get back to the problem that we are interested in. The original problem was we wanted to estimate the parameter of a Bernoulli density and we taken the beta density to be the prior. So, what is the

posterior density f of. So, we are doing the Bayesian estimation of the Bernoulli parameter.

(Refer Slide Time: 45:53)



- Now getting back to Bayesian estimation of Bernoulli density, the posterior is given by

$$\begin{aligned}
 f(p | \mathcal{D}) &= K f(\mathcal{D} | p) f(p) \\
 &= K_1 p^{\sum x_i} (1-p)^{n-\sum x_i} p^{a-1} (1-p)^{b-1} \\
 &= K_1 p^{\sum x_i + a - 1} (1-p)^{n+b-\sum x_i - 1}
 \end{aligned}$$

- Hence the posterior is Beta($\sum x_i + a, n + b - \sum x_i$) density

PR NPTEL course - p.89/101

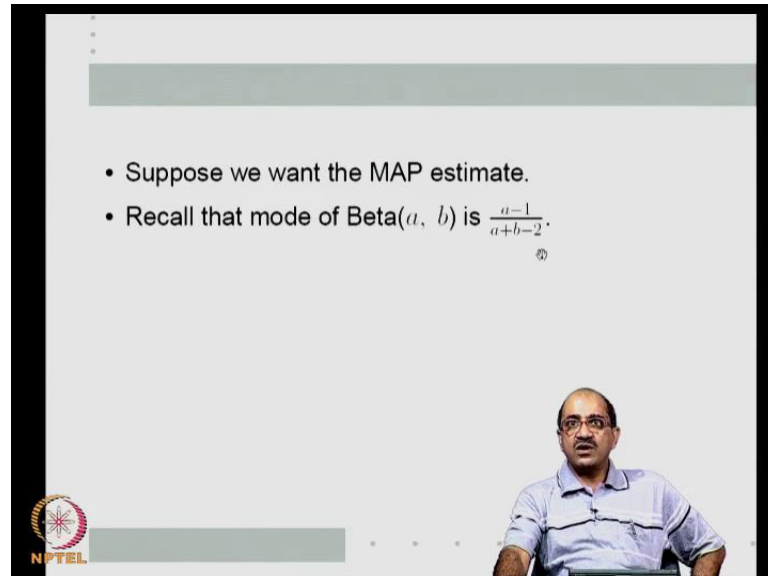
So, we want to calculate the posterior density. So, posterior density is given by f of p given \mathcal{D} is f of \mathcal{D} given p f by some normalizing constant. So, I just put that as K . So, $f(p)$ given \mathcal{D} is some constant K times above \mathcal{D} given p $f(p)$. Now, I can substitute f of \mathcal{D} with p and $f(p)$, $f(p)$ is the beta density. So, it has some constant in terms of gamma functions. This we already calculated earlier.

So, forgetting about the constants I change the constant K to K_1 . f of \mathcal{D} given p where $(())$ to p , p to the power summation x_i $1 - p$ to the power $n - \text{summation } x_i$ and we had taken the prior to be beta. So, p power $a - 1$ $1 - p$ power $b - 1$. So, the constant that comes in beta is subsumed into this that is why that K has been changed to K_1 . So, what is this K_1 times if we gather to get all the p terms I get p to the power summation x_i plus $a - 1$ $1 - p$ to the power of n plus b minus summation x_i minus 1 .

So, this is once again a beta density only thing is the parameters have changed, we started with the prior which is beta with parameters a comma b . Then the posterior turns out to be beta with parameters. Summation x_i plus a and n plus b minus summation x_i . So, this is the posterior density, posterior density happens to be beta. Now, once you have posterior density we have as we seen earlier, we have various choices for the

estimate, we can use the mean of the posterior density, we can use the mode of the posterior density and we can actually calculate $f(x)$ given D .

(Refer Slide Time: 47:39)



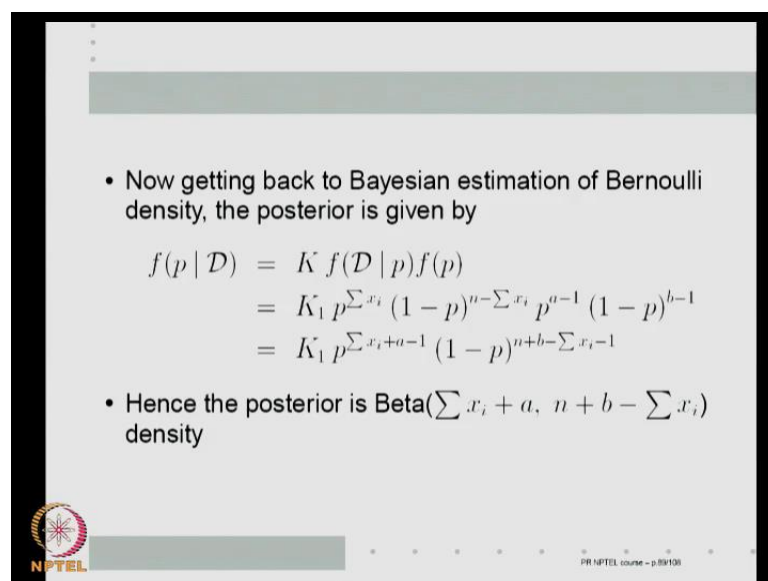
• Suppose we want the MAP estimate.

• Recall that mode of $\text{Beta}(a, b)$ is $\frac{a-1}{a+b-2}$.

NPTEL

Once again let us do all this suppose we want the MAP estimate what is the MAP estimate is the maximum of the posterior density that is we take the mode of the posterior density as our estimate. So, just now we shown that for a posterior density with parameters a and b the mode is $a - 1$ by $a + b - 2$.

(Refer Slide Time: 45:53)



• Now getting back to Bayesian estimation of Bernoulli density, the posterior is given by

$$\begin{aligned} f(p | \mathcal{D}) &= K f(\mathcal{D} | p) f(p) \\ &= K_1 p^{\sum x_i} (1-p)^{n-\sum x_i} p^{a-1} (1-p)^{b-1} \\ &= K_1 p^{\sum x_i + a - 1} (1-p)^{n+b-\sum x_i - 1} \end{aligned}$$

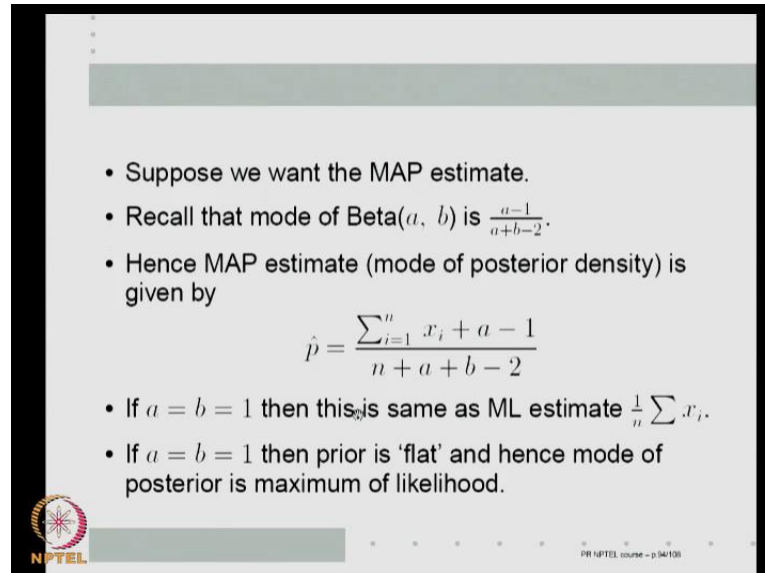
• Hence the posterior is $\text{Beta}(\sum x_i + a, n + b - \sum x_i)$ density

NPTEL

PR NPTEL course - p.89/130

Here I have my parameters summation x_i plus a and n plus b minus summation x_i where, remember n is the number of samples we have, so these are the parameters.

(Refer Slide Time: 48:44)



- Suppose we want the MAP estimate.
- Recall that mode of $\text{Beta}(a, b)$ is $\frac{a-1}{a+b-2}$.
- Hence MAP estimate (mode of posterior density) is given by
$$\hat{p} = \frac{\sum_{i=1}^n x_i + a - 1}{n + a + b - 2}$$
- If $a = b = 1$ then this is same as ML estimate $\frac{1}{n} \sum x_i$.
- If $a = b = 1$ then prior is 'flat' and hence mode of posterior is maximum of likelihood.


So, which means my MAP estimate will be summation x_i plus a minus 1 what was a is summation x_i plus a and what was b is when I do a plus b what happens is now a is summation x_i plus a my b parameters n plus b minus summation x_i . So, a plus b the summation x_i will cancel. So, my map estimate will be i is equal to n summation x_i plus a minus 1 by n plus a plus b minus 2.

The one thing that we can immediately see in this, is the following, if I take a is equals to b is equals to 1 that becomes 1 by n summation x_i which is nothing but, the sample mean for Bernoulli also we know the sample mean is the ML estimate. So, if I take a is equal to b is equals to 1, we get the sample estimate no matter what n is we get ML estimate. Why is that so? When a is equal to b is equal to 1, the prior is flat, the posterior f of D given θ of D given θ is given by the product of of D given θ into f of θ the likelihood and the posterior by some normalizing constant.

So, if I am maximizing the posterior and f of θ is a constant then maximizing of the posterior will be same as the maximizing of the likelihood and that is the ML estimate. So, if I take a is equal to b is equal to 1 for all n the MAP estimate is same as the maximum likelihood estimate as you are expect because for a is equal to b equal to 1 the prior is flat. So, if a and b are different then of course, we get a different this thing.

See this is a Bernoulli random variable. So, x_i 's are 0 or 1. So, we can think of n as the number of trials and each x_i is either 1 or 0 depending on the rate of success or failure. Then ML estimate is nothing but, the fraction of success. Here instead of, just making the fraction of success I have added some a 's and b 's right based on my initial belief. So, that is how my initial belief changed the final estimate, we will come to that I will explain those a 's and b 's in a minute.

(Refer Slide Time: 50:38)



- As earlier, we can compute $f(x | \mathcal{D})$ and use it as the class conditional density.
- Since $x \in \{0, 1\}$, we need only $P(x = 1 | \mathcal{D})$.

$$\begin{aligned}
 P[x = 1 | \mathcal{D}] &= \int_0^1 P[x = 1 | p] f(p | \mathcal{D}) dp \\
 &= \int_0^1 p f(p | \mathcal{D}) dp \\
 &= \frac{\sum_{i=1}^n x_i + a}{n + a + b}
 \end{aligned}$$

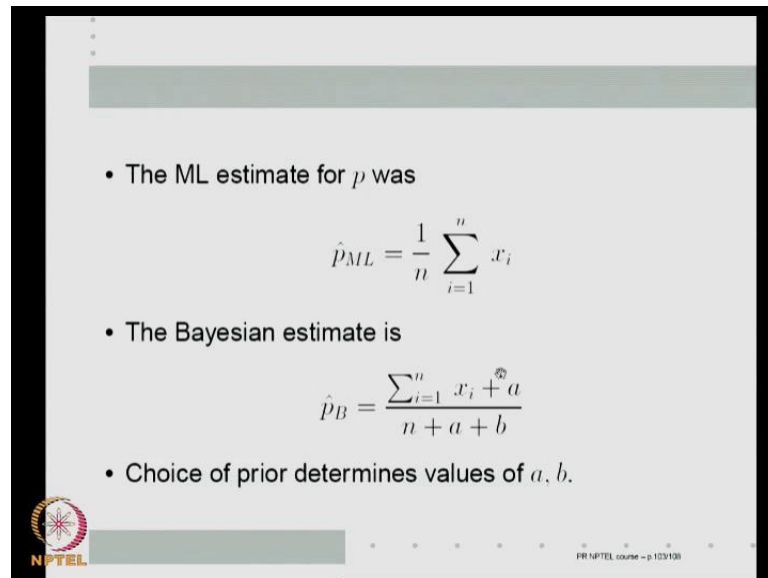
- This turns out to be simply the mean of the posterior.*

But, let us continue this what MAP estimate would have been now let us ask what would be if we calculate $f(x)$ given \mathcal{D} as we did last time. Now, this is a Bernoulli this x is a binary random variable right. So, x takes only values 0 and 1 which means, there are only basically these are mass function and there are only two values for it $f(1)$ given \mathcal{D} $f(0)$ given \mathcal{D} . Let us write that as probability x is equal to 1 given \mathcal{D} and probability x is equal to 0 given \mathcal{D} .

They sum to 1. So, I need only 1 of them right. So, to calculate $f(x)$ given \mathcal{D} , I need to only calculate $P(x \text{ is equal to } 1 \text{ given } \mathcal{D})$. That is once again given by 0 to 1 $P(x \text{ is equal to } 1 \text{ given } p) f(p \text{ given } \mathcal{D})$. This is the posterior integrate over P . Now, given our model p is equal to p probability x equal to 1 given p is p and this is the posterior. So, it becomes 0 to 1 $p f(p \text{ given } \mathcal{D}) dp$ $f(p \text{ given } \mathcal{D})$ is the density over p . So, this is nothing but, the mean of the posterior density.

So, if I actually calculate $f(x)$ given D that is determined by the mean of the posterior density and we already know the mean of the posterior density this is $a + b$. So, that is I is equal to $1 + n \times \hat{p}$ plus $a + b$. So, that is the reason why I earlier slide I have only shown the example of using MAP estimate I did not show the example of using the mean because using mean or $f(x)$ given D is the same.

(Refer Slide Time: 52:21)



- The ML estimate for p was

$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^n x_i$$

- The Bayesian estimate is

$$\hat{p}_B = \frac{\sum_{i=1}^n x_i + a}{n + a + b}$$


- Choice of prior determines values of a, b .

NPTEL

PR NPTEL course - p.103/108

So, if I actually calculate want to calculate $f(x)$ given D in this case that turns out to be the mean of the posterior density. Otherwise, I can use the MAP estimate. So, let us take a look we could have we can discuss either of them. So, let us look at the mean, so this turns out to be simply the mean of the posterior density. So, let us look at the ML estimate the the two estimates, this is the ML estimate with the sample mean with the fraction of success, number of success by total number of trials. This is the bayes estimate I is equal to $1 + n \times \hat{p}$ plus $a + b$. Essentially the the Bayesian estimates value depends on a and b which are the prior parameters of the priors. So, our choice of prior determines the values of a and b what does that mean once again.

(Refer Slide Time: 52:57)


$$\hat{p}_{ML} = \frac{1}{n} \sum_{i=1}^n x_i$$
$$\hat{p}_B = \frac{\sum_{i=1}^n x_i + a}{n + a + b}$$

- We can say we have started with $a + b$ 'fictitious' trials of which a were successes.
- This is how our 'prior beliefs' affect final estimate.

PR NPTEL course - p.106/106

Let us take this the easy two estimates, we already know what maximum likelihood estimate represents, the summation x_i is the number of successes. Because x 's are 1 or 0 and n is the total trials by the total number of success by total number of trials. Now, I can interpret the the the Bayes as follows, I can think of. So, I have added some a to summation x_i . So, it is like I have just added a few more successes, and added a plus b to total number of trials. So, it is like I have added a plus b more trials, of which I decided a or successes right. So, we can say before any data is seen we already have started with a plus b fictitious trials of which a were successes. We choose a and b to give some value for p like this.

Of course I can it is not dependent only on the on the fraction of a by a plus b , because the the the actual shape of the density depends on the individual values of a and b but, essentially it is like saying that I am I am tossing the coin I want to find the probability of heads. If I am looking only at the data, then if I toss if I toss the coin only once then my only possible maximum likelihood estimates are either 1 or 0 which is ridiculous. But, if I think that the coin is fair and I think that I need at least a sample size of five.

Then I can take some say a sample size of six, I take a is equal to b is equal to 3. Then even if I have only one data, then it becomes 3 plus 1 by 3 plus 3 plus 1. So, it does not go too far away. As the number of trials increases right. Then no summation x_i will be much larger than a and n will be much larger than $a + b$. So, essentially, this will be

same as the ML estimate but, for small sample case it is like we started with a plus b fictitious trials, of which a were successes right.

If you remember last class we looked at the philosophical stand point of Bayesian approach. There I gave you the example of three scenarios, all of them the data says that it is perfect somebody wanted to guess, whether coin turns heads or tails five or three out of three times he guessed correct. In ML estimate I have to estimate the parameter as 1 but, in the Bayesian estimate depending on my prior right till this(())being evidence right. If few lucky runs will not throw my estimate away, that is essentially how the prior effects my final estimate. So, this is how prior effects the final estimate.

So, in both these examples, both in the normal example and in the in the Bernoulli example you can see how the prior has played a role in in shaping the final estimate in view of the data right. As a matter of fact, when I look at the mean of the posterior density at the estimate in the Bernoulli case, the the role of the prior density parameters a and b is particularly clear. So, we will stop here today and next class we will look at a couple of more examples of Bayesian estimation and then move on to other topics.

Thank you.