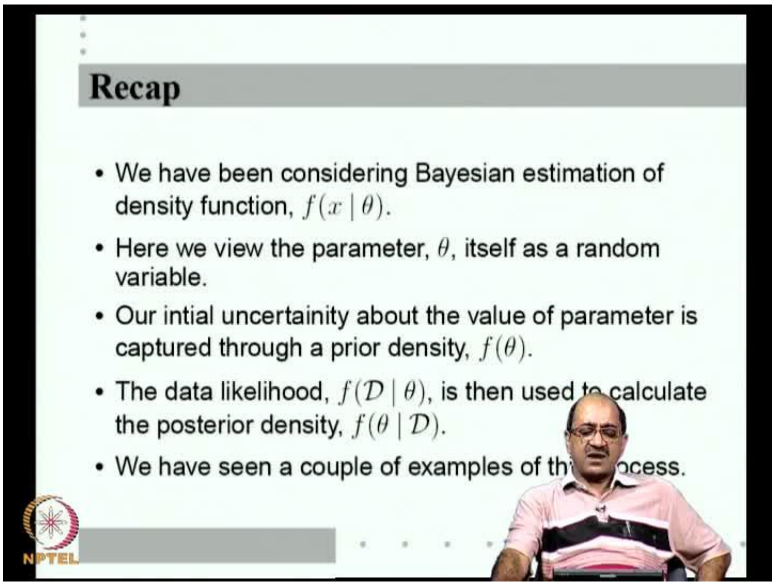
Pattern Recognition
Prof. P. S. Sastry
Department of Electronics and Communication Engineering
Indian Institute of Science, Bangalore

Lecture - 8
Bayesian Estimation examples; the exponential family of densities and ML estimates

Welcome to this next class in Pattern Recognition, we have been looking at density estimation. So, let us briefly recall, what we have been doing in the last couple of classes.

(Refer Slide Time: 00:27)



Recap

- We have been considering Bayesian estimation of density function, $f(x \mid \theta)$.
- Here we view the parameter, θ, itself as a random variable.
- Our intial uncertainity about the value of parameter is captured through a prior density, $f(\theta)$.
- The data likelihood, $f(\mathcal{D} \mid \theta)$, is then used to calculate the posterior density, $f(\theta \mid \mathcal{D})$.
- We have seen a couple of examples of th ocess.

We have been looking at, how to estimate densities for given IID samples, for a particular density. We first looked at the maximum likelihood estimation method. For the last couple of classes, we have been considering the Bayesian estimation of density function. So, given a density function f x given theta where, theta is the parameter, we are considering Bayesian estimate for the parameter theta.

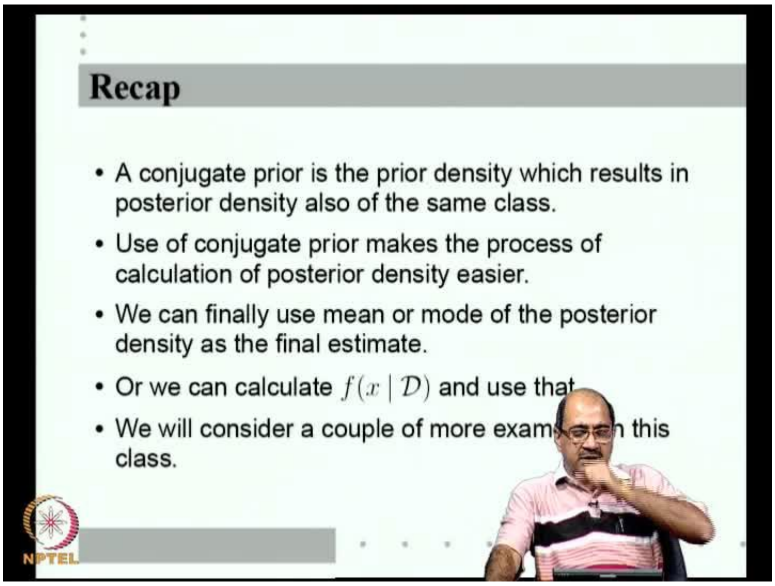
As I have already told you, the main difference between the maximum likelihood estimation and the Bayesian estimation is that, in the Bayesian estimation, we look at the parameter theta, which may be a vector or a scalar, the parameter theta itself is viewed as a random variable. And because, we view it as a random variable, it has a prior density, which captures our initial uncertainty.

Our knowledge or lack of knowledge about the specific value of the parameter is captured through a prior density f theta. So, f theta gives us some idea about, what we think or the possible values for the parameter. Given the given the prior density, we are going to use the data likelihood that is, f D given theta to calculate the posterior density f theta given D.

Once again, I would like to drive your attention to a caution on the notation, for simplicity, the densities of all kind of random variables we are using the same symbol f. So, f of theta, f of D given theta, f of theta given D, all these are densities, purely as mathematical notation looks same because, f is the same function. But, we are we are using f as a notation to denote density and density of, which random variable it is, is clear from context.

Thus, f x given theta is the density of x conditioned on theta, which is the parameter f of theta, is the density of the parameter theta and so on, f theta given D is the conditional density of theta, conditioned on data D. So, even though we are using the same symbol f, for all densities so, I hope you understand that, the f used in different times is a different function. It refers to densities of different random variables and just to keep the notation uncluttered, we are calling it as the f. So, once again, essentially we start with a prior density f of theta, for the parameter theta then, use the data likelihood of D given theta to calculate the posterior f theta given D, we have seen a couple of examples of this process earlier.

(Refer Slide Time: 03:08)



Recap

- A conjugate prior is the prior density which results in posterior density also of the same class.
- Use of conjugate prior makes the process of calculation of posterior density easier.
- We can finally use mean or mode of the posterior density as the final estimate.
- Or we can calculate $f(x \mid \mathcal{D})$ and use that
- We will consider a couple of more exam[illegible]n this class.

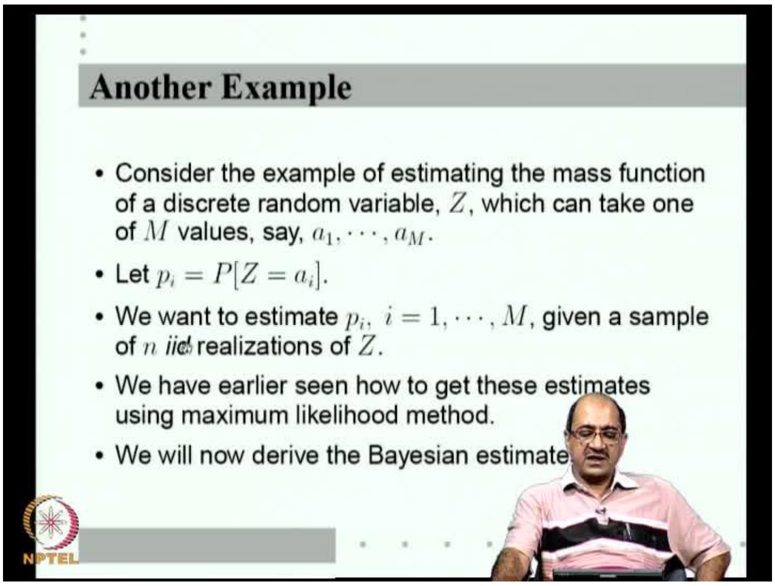
And the idea of Bayesian estimation by now, you would have seen is, to choose a right kind of prior, the right kind of prior for us is, what is called the conjugate prior. The conjugate prior is that prior density, which results in the posterior density belonging to the same class of densities. For example, as we saw when, we are estimating the mean of a Gaussian random variable or mean of a Gaussian density where, the variance is assumed known, we choose Gaussian density for the prior then, the posterior is also Gaussian density.

So, for that particular estimation problem, the prior density happens to be Gaussian similarly, for a Bernoulli problem where, we have to estimate the parameter p namely, the probability of the random variability taken value 1. For parameter p, the appropriate prior density turns out to be beta appropriate in the sense that, if I take prior density to be beta then, the posterior also becomes beta, so such a prior is called a conjugate prior right.

The conjugate prior is that prior density, which results in the posterior density also, to be of the same class of densities, the use is of conjugate prior makes the process of calculation of posterior density easier. As we have seen let us say, in the case of the Bernoulli parameter, that we have seen earlier, if we start with some beta a 0, for the beta a 0 b 0 for the prior then, the posterior is also beta density with possibly some other parameters a and b n.

So, calculation of the posterior density is simply a matter of parameter updation, given the parameters of the prior now, we update them into parameters of the posterior. Having obtained the posterior density, we can finally use either the mean or the mode of the posterior density at the final estimate. We have seen examples of both or we can also calculate f of x given D, that is the actual class conditional density, conditioned on data by integrating the posterior density and we have seen example of that also. So, this class, we will look at a couple of more examples of Bayesian estimation and then, closed Bayesian estimation. As you would have by now seen, Bayesian estimation is a little more complicated mainly because, you have to choose the right kind of prior and different kind of estimation problems make different priors of the conjugate.

(Refer Slide Time: 05:43)



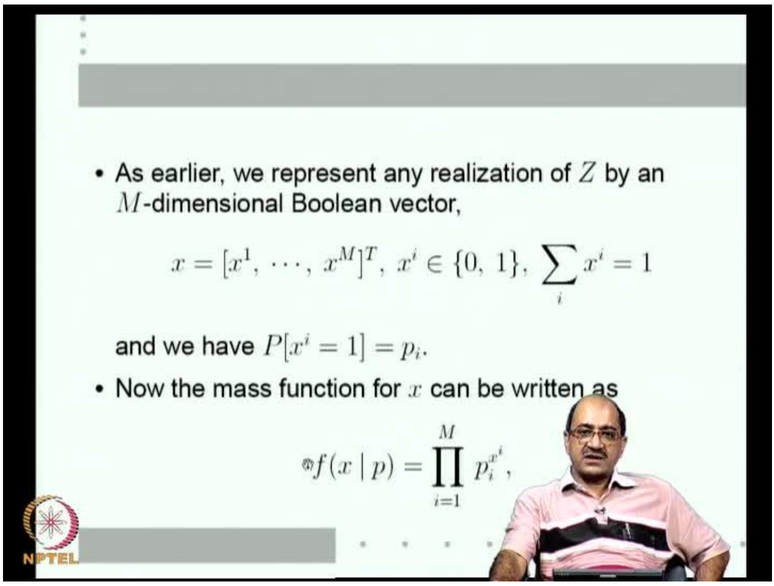
Another Example

- Consider the example of estimating the mass function of a discrete random variable, Z, which can take one of M values, say, $a_1, \cdots, a_M$.
- Let $p_i = P[Z = a_i]$.
- We want to estimate $p_i,\ i = 1, \cdots, M$, given a sample of n *iid* realizations of Z.
- We have earlier seen how to get these estimates using maximum likelihood method.
- We will now derive the Bayesian estimate.

So, we will start with another example, this is the multinomial example that is, we consider estimating the mass function of a discrete random variable, which takes one of M possible values say, a 1 to a M where, p i is probability Z takes the value a i. So, essentially Z takes value a 1 with probability p 1, a 2 with probability p 2, a M with probability p M and we want to estimate this p 1, p 2, p M, given a sample of n iid realizations of Z. We already considered this problem in in the maximum likelihood case and there we told you, this is particularly important in certain class of pattern recognition problem. Especially those to do with say, the text classification and so on where, the discrete random variables for feature that is, features that take only finitely many values are important.

We have seen, how to do the maximum likelihood estimation for obtaining this p 1, p 2's that, p 1, p 2, p M that characterize the mass function of this discrete random variable Z. Now, we will look at, how to do the same thing using Bayesian estimation, I hope the problem is clear, this already is done earlier so, we will we will quickly review the notation that we used earlier.

(Refer Slide Time: 07:06)



- As earlier, we represent any realization of Z by an M-dimensional Boolean vector,

$$x = [x^1, \cdots, x^M]^T, \; x^i \in \{0, 1\}, \; \sum_i x^i = 1$$

and we have $P[x^i = 1] = p_i$.

- Now the mass function for x can be written as

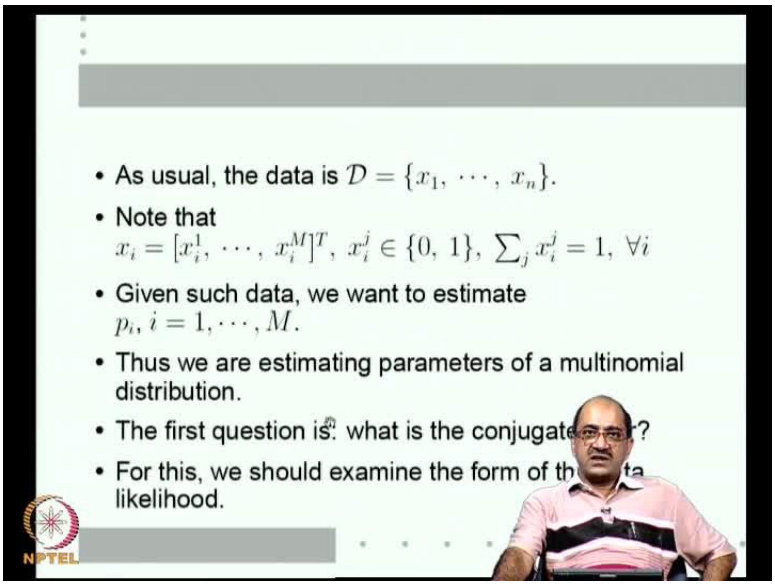
$$f(x \mid p) = \prod_{i=1}^{M} p_i^{x^i},$$

So, as earlier, we represent any realization of Z by M-dimensional Boolean vector x, x has M components x superscript 1, x superscript M. The transpose there is because, as I said, all vectors for us are column vectors, each of these components of this vector x, x superscript i is either 0 or 1, and summation of all of them is 1. That means, essentially x takes only the unit vectors 1 0 0 0, 0 1 0 0's and so on, the idea is that, if z takes the i th value a i then, we represent it by a vector where, i th component is 1 and all others are 0.

We have already seen that, this is a interesting and useful representation for maximum likelihood with this same, we use the same representation here. And now, p i turn out to be, the probability that the i th component of this vector is 1 that is what, I wrote there p x subscript i is equal to 1. Because, p i is the probability with where, Z takes the i th value a i and when Z takes the i th value a i, the i th component of x i becomes 1 where, i th component of x becomes 1.

Also, as I told you last time, the reason, why we use the superscripts to denote the components of x is because, subscripts of x are used as to denote different data or data is x 1 to x n that is why, we are using superscripts to denote the components of particular data. So, we also seen last time that, for this x, the mass function with the single vector parameter p, is product i is equal to 1 to M p i x i because, in any given x, only one component of x is 1 and that is the one that, survives this product. So, if x is 1 0 0 0 then, for that x, f of x given p will become p 1 to the power 1 and p to the power 0 and so on that is, p 1. So thus, this correctly represents the mass function, that we are interested in and p is the parameter of the mass , that we need to estimate.

(Refer Slide Time: 09:06)



- As usual, the data is $\mathcal{D} = \{x_1, \cdots, x_n\}$.
- Note that $x_i = [x_i^1, \cdots, x_i^M]^T$, $x_i^j \in \{0, 1\}$, $\sum_j x_i^j = 1$, $\forall i$
- Given such data, we want to estimate $p_i, i = 1, \cdots, M$.
- Thus we are estimating parameters of a multinomial distribution.
- The first question is: what is the conjugat r?
- For this, we should examine the form of th ta likelihood.

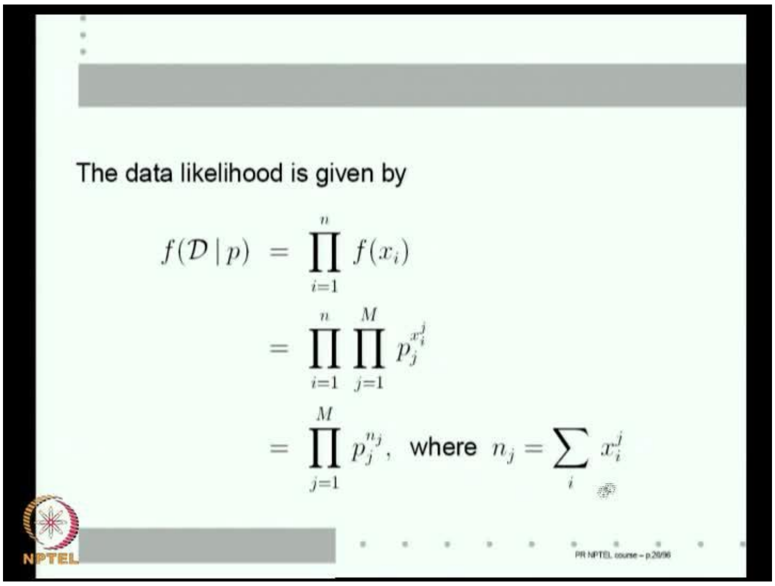
As usual, our data has n samples x 1 to x n, and each sample x i is a vector of M components, the components are shown by superscripts. And each component is either 0 or 1, and in each data it is M x i that is, each M vector x i, if I sum all the components, it becomes 1. That simply means, because, each component is on 0 1 and sum is 1 means, exactly one component is 1 and all others are 0 so, this is the nature of our representation and we have n such data items.

So, given such data, we want to estimate p 1, p 2, p M thus essentially, what we are doing is, we are estimating the parameters of a multinomial distribution. As you know, a multinomial binomial takes only binomial is important, when there is a random experiment that is repeated, which takes only two values success or failure. In the multinomial case, it is the same thing independent realizations of a random experiment that takes more than two values say, M values.

If a random variable takes M different values, I can think of it as a random experiment, which can result in one of M possible outcomes right. So, many samples from that random variable are like, I have a multinomial distribution that is, I repeat a random experiment, that can take one of M possible outcomes, n number of times. So, some of which will be well result in first outcome and so on, some of which will results in second outcome and so on.

So, I know for each repetition, what outcome has come and given those things, we want to estimate the multinomial parameters p 1, p 2, p M. Now because, we are in the Bayesian context, the first question we have to answer is, what is the conjugate prior in this case right. Now, as we have already seen from our earlier examples, for this we should examine the form of the data likelihood. So, we already have our model, we for this x, we have the mass function, that we that we have seen earlier (Refer Slide Time: 11:04) that is the mass function so, given this mass function, what is the data likelihood, that is easy to calculate.

(Refer Slide Time: 11:17)



The data likelihood is given by

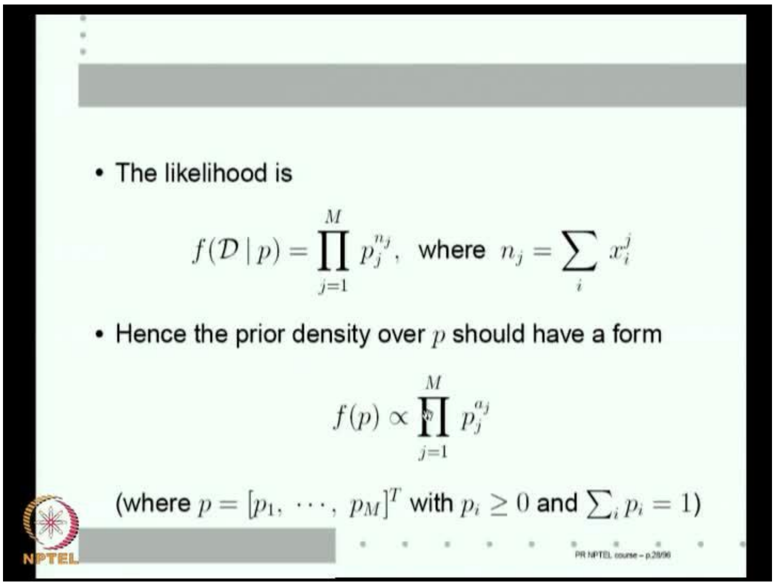
$$
\begin{aligned}
f(\mathcal{D}\,|\,p) &= \prod_{i=1}^{n} f(x_i) \\
&= \prod_{i=1}^{n}\prod_{j=1}^{M} p_j^{x_i^j} \\
&= \prod_{j=1}^{M} p_j^{n_j}, \quad \text{where } n_j = \sum_i x_i^j
\end{aligned}
$$

PR NPTEL course – p.20/86

So, f D given p, p is product of this over i is equal to 1 to n now, we can substitute for f of x i, if I substitute for f of x i, f of x i is once again product p j x i j. Now, if I interchange i and j summation then, p j to the power x i j product over i can be written as product over j p j to the power n j where, n j is summation over i x i j. What does that mean, in any given x i, the j th component is 1, if that particular outcome of Z represents the j th value of Z.

So, this n j tells you, out of the n, how many times Z taken the j th value so, for example, n 1 plus n 2 plus n M will be equal to n, the total number of samples. So, out of n samples, n 1 times the first value of Z has come, n 2 times the second value of Z has come or looking at as a multinomial distribution, n 1 times the first outcome has occurred, n 2 times the second outcome has occurred and so on. So, now, in terms of this n's, the data likelihood is given by product over j, p j to the power n j. Now, if this is the data likelihood, we multiply this with a prior and we should get another expression of the same form, as the prior so, what should be our prior.

(Refer Slide Time: 12:48)



- The likelihood is

$$f(\mathcal{D}\,|\,p) = \prod_{j=1}^{M} p_j^{n_j}, \quad \text{where } n_j = \sum_i x_i^j$$

- Hence the prior density over p should have a form

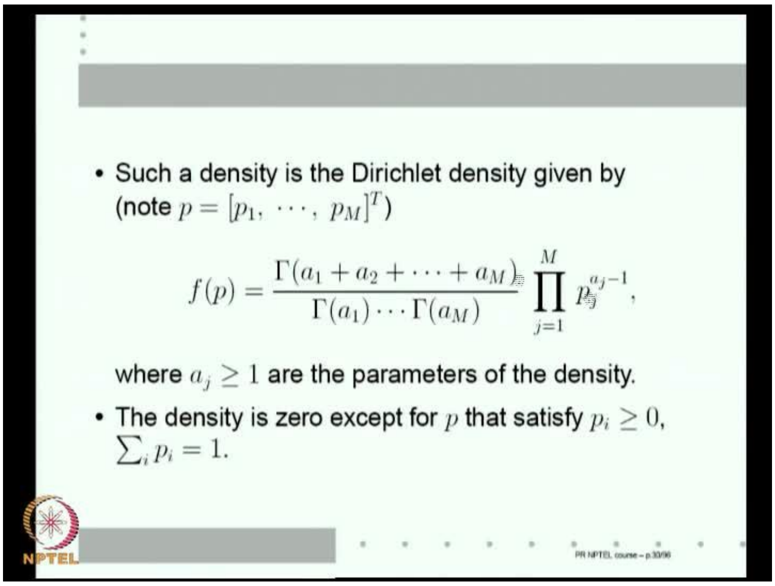
$$f(p) \propto \prod_{j=1}^{M} p_j^{a_j}$$

(where $p = [p_1, \cdots, p_M]^T$ with $p_i \geq 0$ and $\sum_i p_i = 1$)

PR NPTEL course – p.28/96

So, this is the data likelihood so, we expect the prior to have a some density, this proportional to a product of p j to the power a j. If the prior is proportional to product of p j to the power a j and then, I multiply with data likelihood, I get another product of p j to the power some a j prime so, once again that posterior will belong to the same density of the prior right. Let us still remember that, p is a vector parameter, p has M components with all of them are probabilities so, they are greater than or equal to 0. And sum of p a is equal to 1 because, p 1 is the probability of that Z takes first value and so on so, this is needed for the mass function of Z. So, we need a density, which which is a density defined over all p, that satisfy this and that should have a form, which is product p j to the power a j.

(Refer Slide Time: 13:43)



- Such a density is the Dirichlet density given by (note $p = [p_1, \cdots, p_M]^T$)

$$f(p) = \frac{\Gamma(a_1 + a_2 + \cdots + a_M)}{\Gamma(a_1) \cdots \Gamma(a_M)} \prod_{j=1}^{M} p_j^{a_j - 1},$$

where $a_j \geq 1$ are the parameters of the density.

- The density is zero except for p that satisfy $p_i \geq 0$, $\sum_i p_i = 1$.

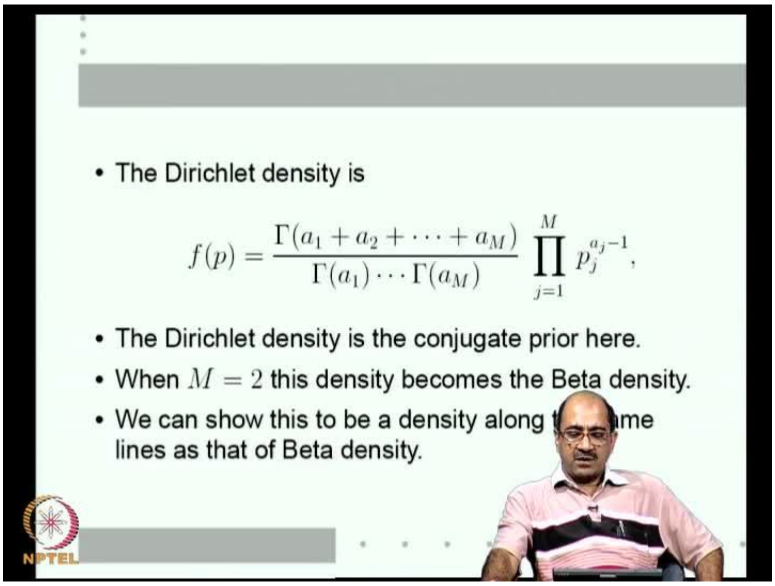
PR NPTEL course – p.30/96

Now, such a density is, what is known as a Dirichlet density, if you remember when there are only two outcomes when you looking at Bernoulli, the the prior happen to be what is called the beta density. So, we will see that, the Dirichlet density is a kind of generalization of the beta density so, the Dirichlet density is given by f of p, is gamma of a 1 plus a 2 plus a M by gamma of a 1 into gamma of a 2 into gamma of a M, product j is equal to 1 to M p j to the power a j minus 1.

Where, this gamma is the gamma function, that we already seen when we discuss the beta function the beta density, gamma function gamma of a is integral 0 to infinity x to the power a minus 1, e power minus x d x. In this, the the parameters of this Dirichlet density or this a j's that is that is, a 1 a 2 a M, all of them are assumed to be greater than equal to 1.

Also, this density has this value, only for those p that satisfy all components greater than or equal to 0 or some of the components is 1, outside of those p's the density is 0. That means, the density is concentrated on that subset of power M, which satisfies p i greater than or equal to 0 and summation p equal to 1, which is essentially called as simplex. Those you know what is simplex is, they are simplex but, anyway even, if you do not know what is simplex is, this density is non-zero only for those p’s that satisfy p i greater than or equal to 0, summation p i is equal to 1 otherwise, the density value is 0.

(Refer Slide Time: 15:21)



- The Dirichlet density is

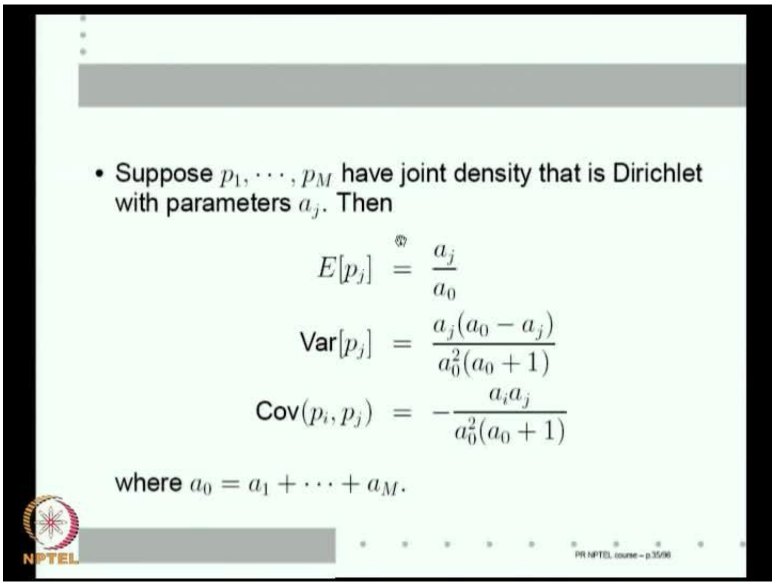
$$f(p) = \frac{\Gamma(a_1 + a_2 + \cdots + a_M)}{\Gamma(a_1) \cdots \Gamma(a_M)} \prod_{j=1}^{M} p_j^{a_j - 1},$$

- The Dirichlet density is the conjugate prior here.
- When $M = 2$ this density becomes the Beta density.
- We can show this to be a density along t[illegible]me lines as that of Beta density.

So, that is the Dirichlet density so, if M is equal to 2, this becomes gamma a 1 plus a 2 by gamma a 1 into gamma a 2, and p 1 to the power a 1 minus 1 and p 2 to the power a 2 minus 1 and that is the beta density. So, when M is equal to 2, this density becomes the beta density and this Dirichlet density happens to be the conjugate prior here. So, as you can see, if I am estimating the parameter of a Bernoulli density then, my conjugate prior happens to be beta.

Whereas, if Iam estimating parameters for a multinomial distribution rather than binomial one then, the prior happens to be Dirichlet, which is a kind of a neat generalization of the beta density to, a to more than two case. Of course, this is a strange expression and we have first show that, this is a density on that particular set of p, that we we mentioned. Using the using similar methods, as in the case of beta density we can show that, this is a density, the other thing that we want is, just like in the beta density case, ultimately because, my posterior will be a Dirichlet density. We need to know the movements of the Dirichlet density so that, I I can correctly use my posterior to get my estimates.

(Refer Slide Time: 16:49)



- Suppose $p_1, \cdots, p_M$ have joint density that is Dirichlet with parameters a_j. Then

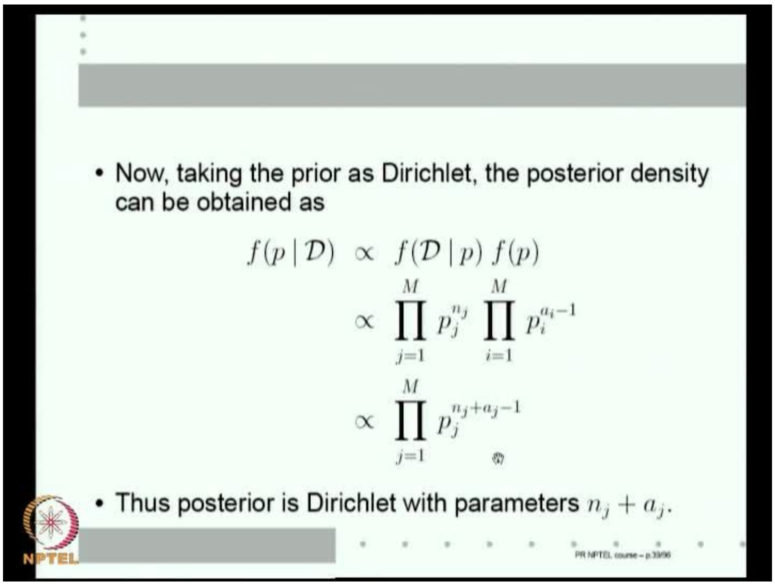
$$\begin{aligned} E[p_j] &= \frac{a_j}{a_0} \\ \text{Var}[p_j] &= \frac{a_j(a_0 - a_j)}{a_0^2(a_0+1)} \\ \text{Cov}(p_i, p_j) &= -\frac{a_i a_j}{a_0^2(a_0+1)} \end{aligned}$$

where $a_0 = a_1 + \cdots + a_M$.

PR NPTEL course – p.35/96

Once again without proofs, I I just put down these movements so, if p 1, p 2, p M have joint densities as Dirichlet, with parameters a 1, a 2, a M then, expected value of any component say, p j is a j by a 0 where, a 0 is a 1 plus a 2 plus a M. Variance of p j happens to be a 0 into a j into a 0 minus a j, by a 0 square into a 0 plus 1 and similarly, this is the co variance. We do not know the co variance but, this kind of gives us all the movements upto the up to order 2. So, for example, if you are going to use the use the mean of the posterior rather or final estimate, we would need this formula.

(Refer Slide Time: 17:38)



- Now, taking the prior as Dirichlet, the posterior density can be obtained as

$$\begin{aligned} f(p\,|\,\mathcal{D}) &\propto f(\mathcal{D}\,|\,p)\, f(p) \\ &\propto \prod_{j=1}^{M} p_j^{n_j} \prod_{i=1}^{M} p_i^{a_i - 1} \\ &\propto \prod_{j=1}^{M} p_j^{n_j + a_j - 1} \end{aligned}$$

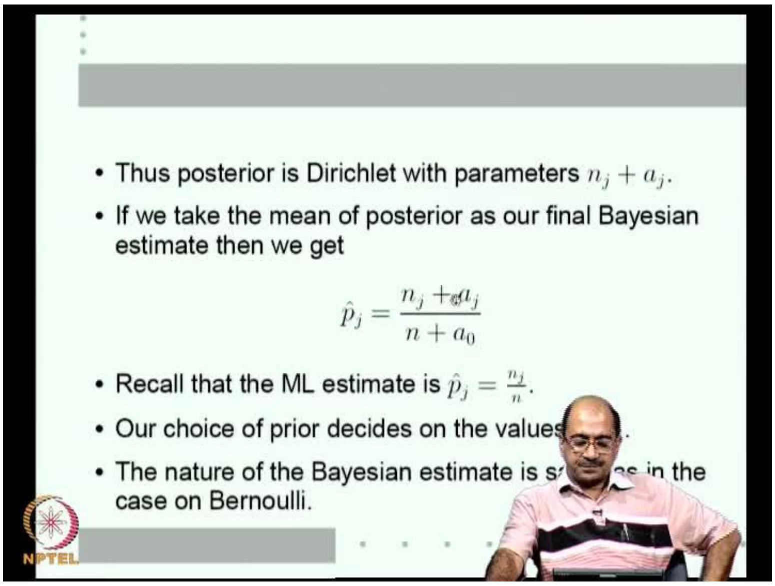
- Thus posterior is Dirichlet with parameters $n_j + a_j$.

PR NPTEL course – p.33/96

So, with this let us now go on in this case, compute the posterior density, if by taking prior as Dirichlet, the posterior density we have, to for the posterior density, as we know is f p given D is proportional to the product of f D given p, into f p f D given p is the data likelihood, f p is the prior we have taken prior to be Dirichlet. We already have an expression for the likelihood so, if you substitute those two so, (Refer Slide Time: 18:06) this is the expression for the likelihood, product p j to the power n j.

And let us say, we have taken the we have taken the prior to be Dirichlet with parameters a 1, a 2, a M then, this becomes the prior so, this product can now be written as, product over p j of n j plus a j minus 1. So, obviously the reason, why we chosen this particular prior is that, the posterior will belong to same class. So, indeed posterior belongs to the same class right, this is product is proportional to product of p j to the power something. So, if my prior is Dirichlet with parameters a 1, a 2, a M then, the posterior is Dirichlet with parameters n j plus a j where, the n j's come from the data, This is once again very similar to, what happened in the Bernoulli case.

(Refer Slide Time: 18:58)



- Thus posterior is Dirichlet with parameters $n_j + a_j$.
- If we take the mean of posterior as our final Bayesian estimate then we get

$$\hat{p}_j = \frac{n_j + a_j}{n + a_0}$$

- Recall that the ML estimate is $\hat{p}_j = \frac{n_j}{n}$.
- Our choice of prior decides on the values .
- The nature of the Bayesian estimate is s in the case on Bernoulli.

Thus, the posterior is also Dirich let with parameters n j plus a j so, if we take for example, the mean of the posterior as our final Bayesian estimate, we already seen, what the mean is a j by sum so, we know summation n j. So, it will be n j plus a j by summation over j n j plus a j, summation over j n j is n, that we have already seen and a 0 is the notation we have given for summation a j's.

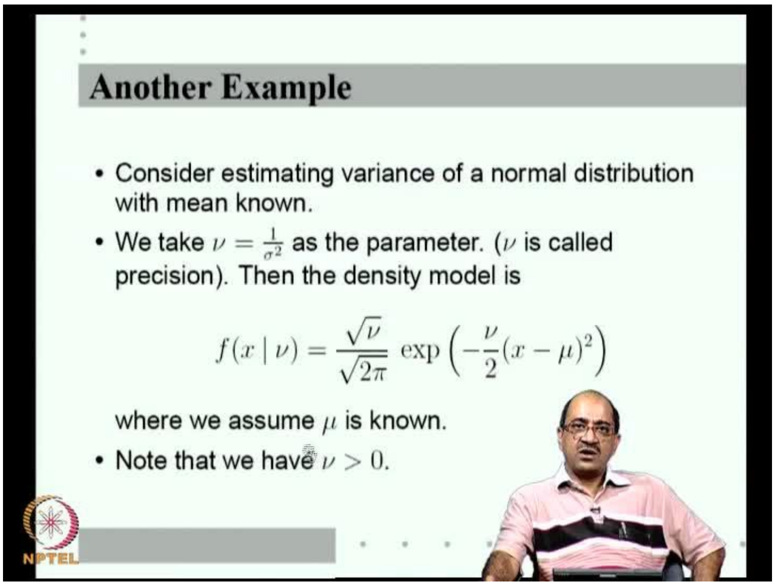
So, the the bayesian density, which is taken as the mean of the posterior turns out to be n j plus a j by n plus a 0. Let us recall, that the ML estimate for this was n j by n right. And the ML estimate is very easy to see, we are asking, what is the probability that Z takes the j th value or what is the probability that, the j th outcome occurs that is, equal to the number of times j th outcome occurred by the total number of samples right, n j means summation over i, x i j is the number of times the j th value has occurred.

So, n j by n was, as we have already derived was the ML estimate here, instead of it being n j by n, it becomes n j plus a j by n plus a 0 where, a j and a 0, which is a 1 plus a 2 plus a M are determined by our choice of the prior right, our choice of prior determines, what the value a j are. So, just like in the case of the Bernoulli parameters, the nature of the Bayesian estimate is same so, you can think of the prior as saying, that before I collect data in my mind because, I have some idea of, what the values of p j's are.

I have coded them to say that, if I have done a zero repetitions of Z, they are fictitious repetitions a 1 of them would give me first value, a 2 of them will give me second value, a M of them will give me third value. So, I can choose the a 0 as well as a 1 to a M based on my idea of, what these numbers p 1 to p M are. So then, the final estimate is the actual in the data, how many times j has occurred plus how many time j has occurred in the fictitious trials, divided by the total number of actual trials and the fictitious trials.

So, when data when the like in the Bernoulli case, if the data is small then, we do not go very wrong because, our paired beliefs we will ensure that, p j's do not go into unnatural values. For example, p j breaking 1 or 0, when data is very small but, as n increases for any fixed a j and a 0, as n increases ultimately, this becomes n j by n. So, asymptotically the Bayesian estimate will be same as the maximum likelihood estimate and hence, it will be consistent. But, once again like in the Bernoulli case, the the prior allows me, to allow my initial beliefs to properly moderate data especially, when data is not very large.

(Refer Slide Time: 22:25)



Another Example

- Consider estimating variance of a normal distribution with mean known.
- We take $\nu = \frac{1}{\sigma^2}$ as the parameter. (ν is called precision). Then the density model is

$$f(x \mid \nu) = \frac{\sqrt{\nu}}{\sqrt{2\pi}} \exp\left(-\frac{\nu}{2}(x-\mu)^2\right)$$

where we assume μ is known.

- Note that we have $\nu > 0$.

Now, let us look at another example, last class we considered the example of estimating mean of a normal distribution where, we assumed the variance to be known right. Now, let us do it the other way round, we want to estimate the variance of a normal distribution and we assume mean to be known. It might look a little strange to you, when we did the ML estimate, we did not have to do so much trouble, we directly did only one example to estimate both mean and variance of a one dimensional Gaussian distribution.

Because, in the ML case, it is a very straight forward thing here, for each kind of parameter, the corresponding prior would be different for example, when we wanted to estimate the mean, the conjugate prior was Gaussian right. Some of you may be thinking that, because we were estimating Gaussian density, the conjugate what happened to be Gaussian, that is not true.

If you want to estimate the variance of a Gaussian where, I assume mean known, the conjugate prior cannot be Gaussian right because, variance as a parameter can take only non-zero values. So, it is density cannot be Gaussian then, we may jump to the conclusion say, may be it is a exponential, exponential is a density that is 0 only only when the random variables takes positive values, it is the density is non-zero, only when the random variable takes positive values.

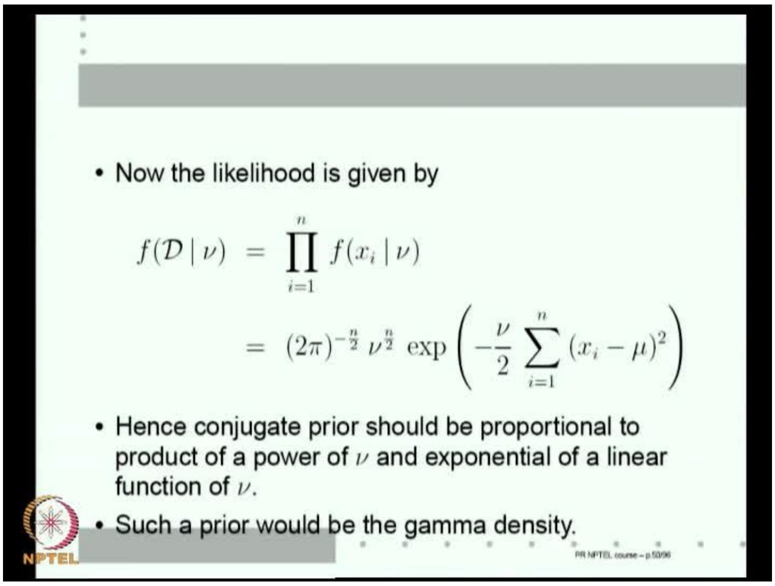
But, exponential is only a special case right, we will see that in general, the the prior is not exponential, exponential is only very special case of the prior. Also, the prior will not be unvariance, as it turns out for this case, is better to take 1 by variance at the parameter, it is often denoted by nu and is often called the precision. While I have not done the vector case, in the vector case, the inverse of the sigma matrix is called the lambda matrix and that is called the precision matrix.

In the in the scalar case of course, we simply take the 1 by variance at the precision, which is often denoted by nu so, in terms of the parameter nu, the normal density model is given by 1 root nu. Because, is normally 1 by sigma root 2 pi, 1 by sigma is root nu so, this root nu by root 2 pi exponential minus half normally, x minus mu whole square by sigma square and 1 by sigma square is nu.

So, it is written as exponential minus nu by 2 into x minus mu whole square. Note that, we are assuming mu is known and that is why, only mu is shown as the conditioning parameter. And we need to find nu said that, nu is always positive so, for example, when we choose a prior density, we choose density that is, the that is 0 on the negative nu and you have to find, what is the prior density, (Refer Slide Time: 25:33) you have to look at the data likelihood.

So, let us look at the data likelihood, the data likelihood is given by in terms of nu, as equal to 1 to n, f of x i given nu, f of x given nu is this. So, if I take a product, this will give me nu to the power root nu to the power n that is, nu to the power n by 2. This 1 by root 2 pi to the power n that is, 2 pi to the power minus n by 2 so, I have a two pi to the power minus n by 2 term, I have a nu to the power n by 2 term. And then, when I take a product of this over x i, it will become exponential minus nu by 2 into sum of this.

(Refer Slide Time: 26:15)



- Now the likelihood is given by

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

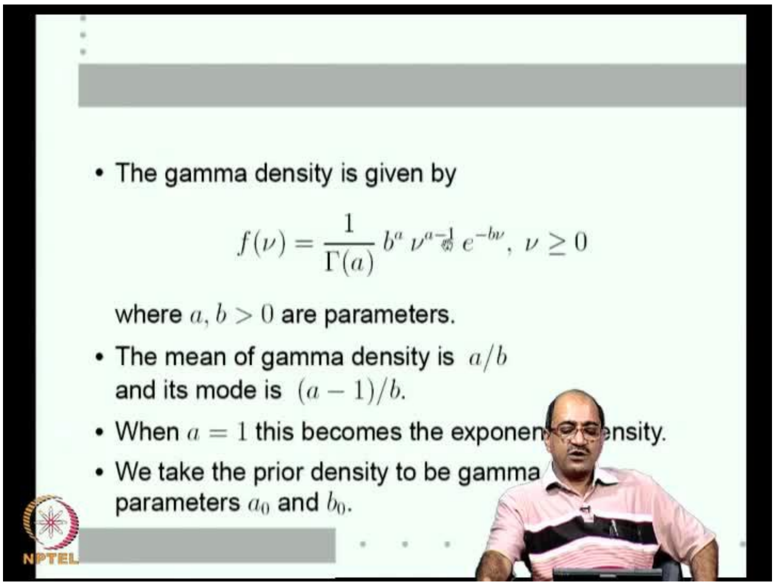
- Hence conjugate prior should be proportional to product of a power of ν and exponential of a linear function of ν.
- Such a prior would be the gamma density.

PR NPTEL course – p.50/96

So, exponential minus nu by 2 into sum over i x i minus mu whole square so now ,to ask what should be the right prior, we should ask what kind of function is this of nu, viewed as a function of nu, what kind of function is this. So, we essentially have an exponential nu into something term and we have a nu to the power something term right. So, the conjugate prior should be something, this proportional to nu power something and that, should be proportional to product of a power of nu and an exponential of a linear function of nu.

Because, the data density is some constant into nu to the power something and exponential minus some some k times nu. So, if the prior also has nu to the power something and exponential some constant into nu, I mean then, the product will once again be nu to the power something into exponentials of constant nu. So, the prior should be proportional to a product of a power of nu and an exponential of a linear function in nu. And such a density transfer to be, what is known as a gamma density, such a prior would be what is called the gamma density.

(Refer Slide Time: 27:36)



- The gamma density is given by

$$f(\nu) = \frac{1}{\Gamma(a)} b^a \nu^{a-1} e^{-b\nu}, \ \nu \geq 0$$

where $a, b > 0$ are parameters.

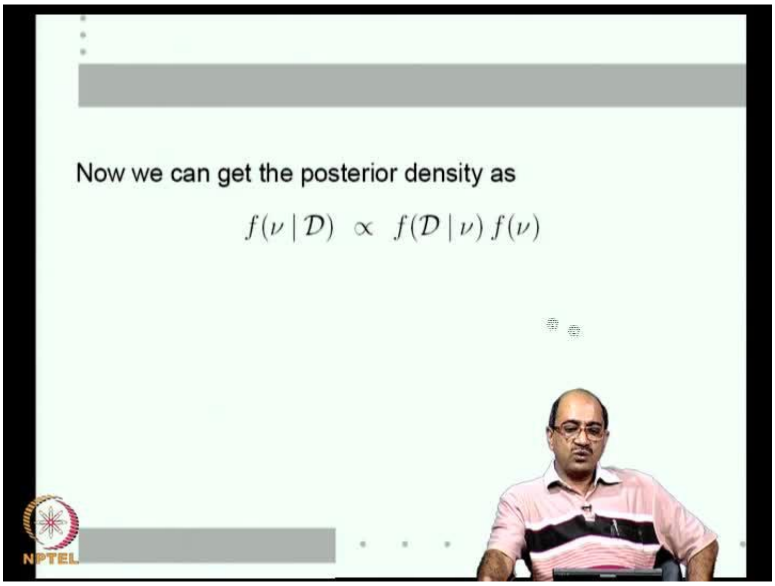
- The mean of gamma density is a/b and its mode is $(a-1)/b$.
- When $a = 1$ this becomes the exponen ensity.
- We take the prior density to be gamma parameters a_0 and b_0.

So, let us look at the gamma density, the gamma density is given by the density function f nu is 1 by gamma a, b to the power of a nu to the power a minus one e to the power of b nu where, a and b are parameters. So, the gamma is once again the gamma function, as a matter of fact, the actual gamma function comes from making this to be a density. By a simple integration, we can show that this to be a density because, this integral will turn out to be the gamma function.

The the gamma density has two parameters a and b, the a comes in nu to the power of a minus one and b comes in e power minus b nu, this b power a is needed so that, the density integrates to 1. So, the nu to the power is controlled by a and exponential of the linear function in nu is controlled by b, these two are the parameters and the mean of gamma density is a by b and the mode is a minus 1 by b.

If I actually choose a to be 1 then, the density turns out to be b e power minus b nu right now, when a is 1, as you know gamma of a is a minus gamma, if one is one turns out to be 1 by straight forward integration. So, when a is equal to 1 is simply b e power minus b nu that is nothing but, the exponential density so, exponential density is a special case of gamma density with a is equal to 1. So, let us take the prior to be gamma with parameters a 0 and b 0 that means, it becomes nu the power of a 0 minus 1 e to the power of minus b 0 nu, those are the two important terms, the rest is constant.

(Refer Slide Time: 29:36)

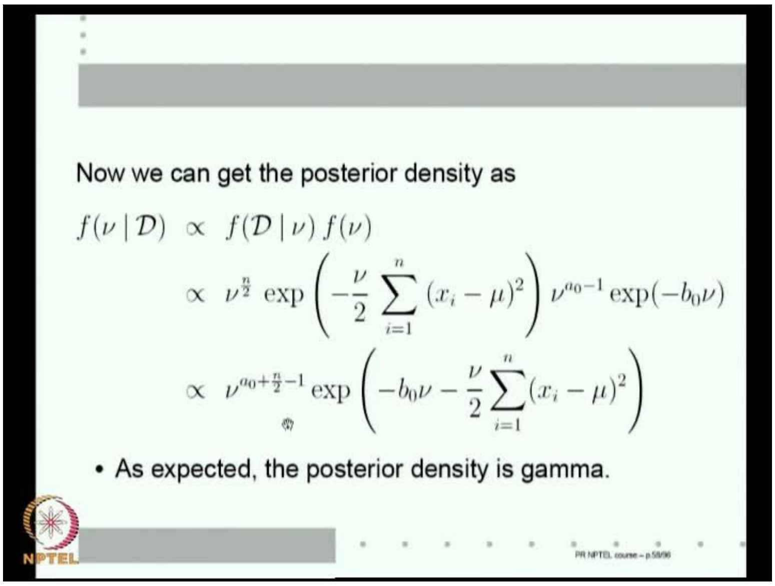


Now we can get the posterior density as

$$f(\nu \mid \mathcal{D}) \;\propto\; f(\mathcal{D} \mid \nu)\, f(\nu)$$

So, the posterior density becomes f of nu given D is proportional f of D given nu into f nu f of D given nu is this and f of nu is this with a as a 0 and b as b 0.

(Refer Slide Time: 29:59)



Now we can get the posterior density as

$$\begin{aligned} f(\nu \mid \mathcal{D}) &\propto f(\mathcal{D} \mid \nu)\, f(\nu) \\ &\propto \nu^{\frac{n}{2}} \exp\left(-\frac{\nu}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \nu^{a_0-1}\exp(-b_0\nu) \\ &\propto \nu^{a_0+\frac{n}{2}-1}\exp\left(-b_0\nu-\frac{\nu}{2}\sum_{i=1}^{n}(x_i-\mu)^2\right) \end{aligned}$$

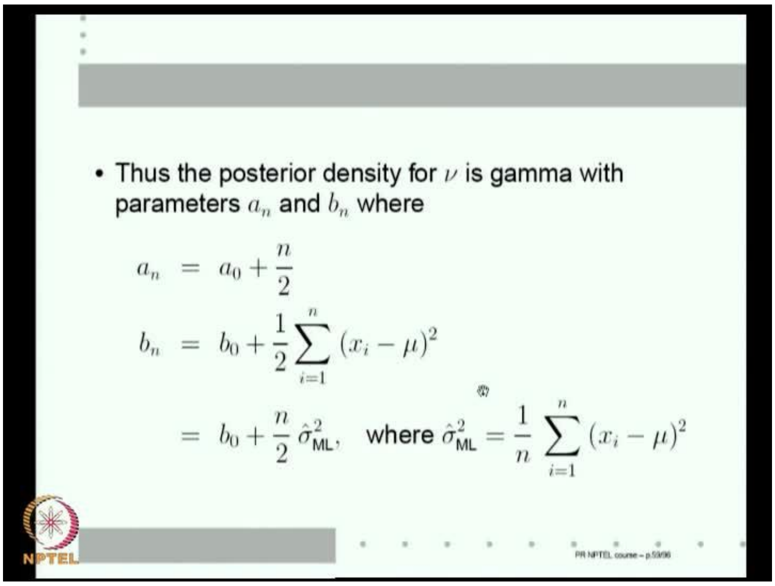
- As expected, the posterior density is gamma.

PR NPTEL course – p.58/96

So, we get this, f of D given nu is nu to the power forgetting the constants keeping only the nu terms; it become nu to the power n by 2 exponential minus nu by 2 into summation x i minus mu whole square; and from f nu, I get nu to the power of a 0 minus 1 exponential minus b 0 nu. For the reason we chose this as the prior is now, these two new terms will become nu to the power something and these exponential terms will become exponential something into nu.

So, if you do that, it becomes nu to the power of a 0 plus n by 2 minus 1 into exponential minus b 0 nu minus nu by 2 into this. So, we once again have nu to the power something exponential minus a linear function of nu so, the posterior as expected, is once again a gamma density. Now, what kind of gamma density is it. The the gamma density is two parameters a and b essentially, forgetting the constants is proportional to nu to the power of a minus 1 e to the power minus b nu right. So for the posterior density, the a parameter is a 0 plus n by 2 and the b parameter is b 0 plus half into this sum right.

(Refer Slide Time: 31:27)



- Thus the posterior density for ν is gamma with parameters a_n and b_n where

$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= b_0 + \frac{n}{2}\hat{\sigma}^2_{\text{ML}}, \quad \text{where } \hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

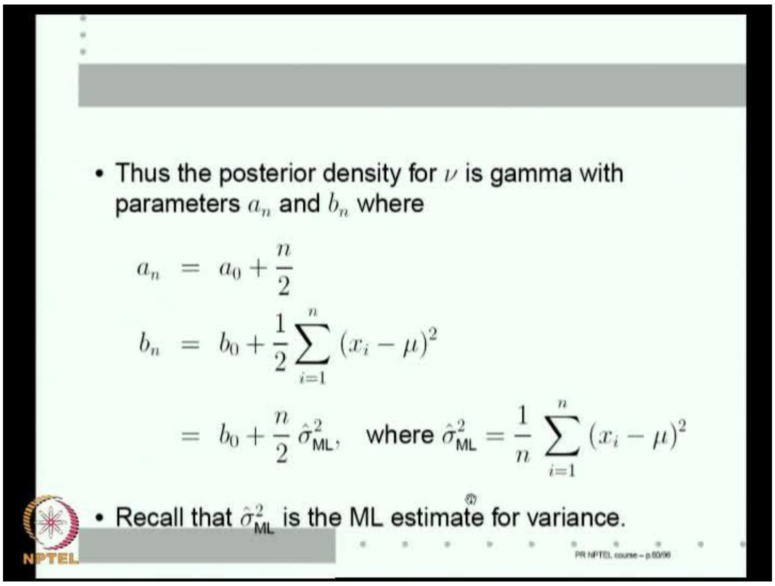
PR NPTEL course – p.53/96

So, if we think that the posterior is a gamma density with parameters a a n and b n right then, this is what we will get. If the posterior is a is a gamma density with parameters a n and b n. Then a n will be a 0 plus n by 2, 0 plus n by 2 and what will be b n, b n will be b 0 plus half into this summation b 0 plus half into the summation. I can write this summation as, I know 1 by n summation is equal to 1 to n, x n minus mu whole square will be the maximum likelihood estimate for variance plus call it sigma square hat ML.

Then, this summation is n times sigma square hat ML so, I can write b n as, b 0 plus n by 2 sigma square hat ML. So, if I chosen the prior to be gamma with parameters a 0 and b 0 then, the posterior will become a gamma with parameters a n and b n. Where a n turns out to be a 0 plus n by 2 and b n turns out to be b 0 plus n by 2 times, sigma square hat

ML where, sigma square hat ML is the maximum likelihood estimator for variance in this case.

(Refer Slide Time: 32:45)



- Thus the posterior density for ν is gamma with parameters a_n and b_n where

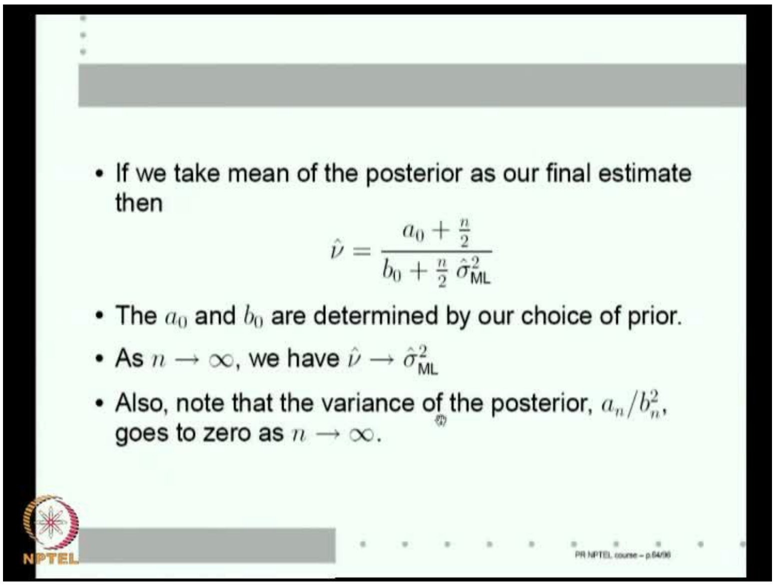
$$\begin{aligned} a_n &= a_0 + \frac{n}{2} \\ b_n &= b_0 + \frac{1}{2}\sum_{i=1}^{n}(x_i - \mu)^2 \\ &= b_0 + \frac{n}{2}\,\hat{\sigma}^2_{\text{ML}}, \quad \text{where } \hat{\sigma}^2_{\text{ML}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2 \end{aligned}$$

- Recall that $\hat{\sigma}^2_{\text{ML}}$ is the ML estimate for variance.

PR NPTEL course – p.63/96

So, recall that sigma square hat ML is the estimate for variance, is the maximum likelihood estimate for variance.

(Refer Slide Time: 32:52)



- If we take mean of the posterior as our final estimate then

$$\hat{\nu} = \frac{a_0 + \frac{n}{2}}{b_0 + \frac{n}{2}\,\hat{\sigma}^2_{\text{ML}}}$$

- The a_0 and b_0 are determined by our choice of prior.
- As $n \to \infty$, we have $\hat{\nu} \to \hat{\sigma}^2_{\text{ML}}$
- Also, note that the variance of the posterior, a_n/b_n^2, goes to zero as $n \to \infty$.

PR NPTEL course – p.64/96

So now, if I want to take the mean of the posterior as the final estimate, as we know for the gamma density with parameters a and b, the mean is a by b. So, here the posterior density is gamma with parameters a n and b n so, the mean will be a n by b n. So, our

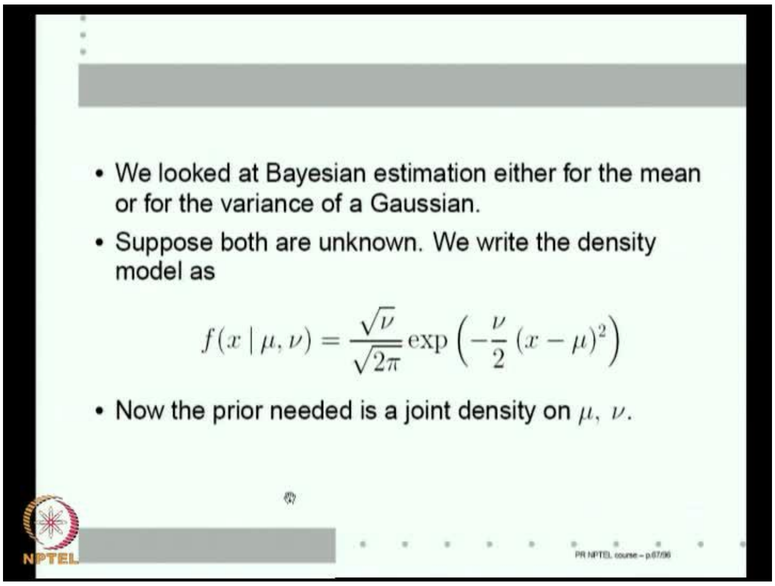
Bayesian estimate for nu nu hat would be a n by b n that is, a 0 plus n by 2 or b 0 plus n by 2 sigma square hat ML right.

Remember that, nu is 1 by sigma square right so, if I do not did not have the a 0 and b 0, nu hat is 1 by sigma square Ml so, it is same as the maximum likelihood estimate. Because, nu is actually 1 by sigma square so, the estimate for 1 by sigma square will be 1 by sigma square hat ML. Now, the a 0 and b 0 are determined by our choice of prior right, we are choosing a gamma density at the prior so, the kind of gamma density we want, is what determines the values a 0 and b 0.

Now, what can we say about this density, once I have this estimate, once again as n tends to infinity, nu hat converges to sigma square ML right because, as n tends to infinity, n by 2 will be greater than both a 0 and b 0. So, this fraction essentially becomes n by 2 by n by two sigma square hat ML so, I am I am sorry about to the typo nu hat converges to 1 by sigma square hat ML. I am sorry, it is not nu hat converges to sigma square ML but, nu hat converges to 1 by sigma square hat ML.

Also note, that the variance of the posterior right for a gamma density with parameters a n and b n, the posterior the variance is a n by b n square. So, this is a n, this is b n so, if you take the square, the numerator goes as n whereas, denominator increases as n square so, that the variance goes to 0, as n tends to infinity. So, as n tends to infinity, the posterior essentially becomes same as the mean and the mean is 1 by sigma square hat ML so, once again just as we expect, the the Bayesian estimate is consistent. But, at any small sample size, it is not only determined by the data thus, 1 by sigma square hat ML but, is also determined by the initial a 0 and b 0, we choose for the prior prior density, which is gaussian with parameters a 0 and b 0.

(Refer Slide Time: 35:29)



- We looked at Bayesian estimation either for the mean or for the variance of a Gaussian.
- Suppose both are unknown. We write the density model as

$$f(x \mid \mu, \nu) = \frac{\sqrt{\nu}}{\sqrt{2\pi}} \exp\left(-\frac{\nu}{2}(x-\mu)^2\right)$$

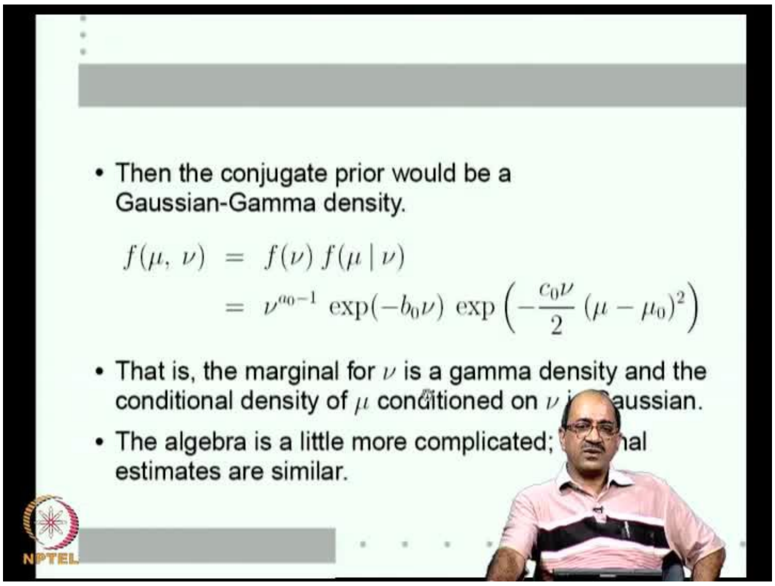
- Now the prior needed is a joint density on μ, ν.

PR NPTEL course – p.67/96

So, we have seen both Bayesian estimation for either the mean or the variance of the Gaussian, mean we seen last time, for variance we are seeing just now. So, when I want to estimate only the mean, assuming that the variance is known then, the prior turns out to be a a Gaussian. When I want to estimate only the variance, assuming that the mean is known, the prior turns out to be gamma. So, if I want to estimate both mean and variance now, I have two parameters once again, from my experience in estimating variance, we will choose nu as the parameter, for parameters in the density model.

So, my density model now is f of x given mu nu is this remember, that nu is 1 by sigma square so, if both mu and nu are unknown, we need a prior, there is a joint density on mu and nu. We already know that, if nu is known, only mu is unknown then, the prior density is Gaussian, if mu is known and nu is unknown then, I know the prior density is gamma. So, the joint density should be some combination of Gaussian and gamma, the the algebra turns out be a little cumbersome so, I do not give you all the algebra, I will just give you the final expression.

(Refer Slide Time: 36:47)



- Then the conjugate prior would be a Gaussian-Gamma density.

$$\begin{aligned} f(\mu, \nu) &= f(\nu)\, f(\mu \mid \nu) \\ &= \nu^{a_0 - 1} \exp(-b_0 \nu) \exp\left(-\frac{c_0 \nu}{2} (\mu - \mu_0)^2\right) \end{aligned}$$

- That is, the marginal for ν is a gamma density and the conditional density of μ conditioned on ν is Gaussian.
- The algebra is a little more complicated; final estimates are similar.

So, then the conjugate prior would be, what is called as Gaussian gamma density, the Gaussian gamma density this, any joint density of any two random variables mu and nu here, can be written as a product of the marginal of nu multiplied by the conditional of mu given nu that is, true of anything. So, this is how, we will model this so, the the Gaussian gamma density model is given here, as you can see what we are saying is, f nu is the first term here, upto here first meaning, the first two terms so, this is the density what we already seen, this is a gamma right.

So, f nu is gamma with parameters a 0 and b 0 and the density mu given nu is essentially a Gaussian density with nu as it is precision or 1 by nu as it is variance. So, the conditional of mu given nu is Gaussian, with this is a Gaussian in the in the variable mu with it is own mean mu 0 and the variance being a function of the conditioning random variable. That is, this is not just directly 1 by nu but it is 1 by c 0 nu so that is, the marginal for nu is a gamma density and the conditional density of mu, conditioned on nu is a Gaussian. Actually, by looking at the at the data likelihood, we can find that, this is the kind of dependence we need that, nu can always be expressed in terms of nu to the power something and exponential linear in nu.

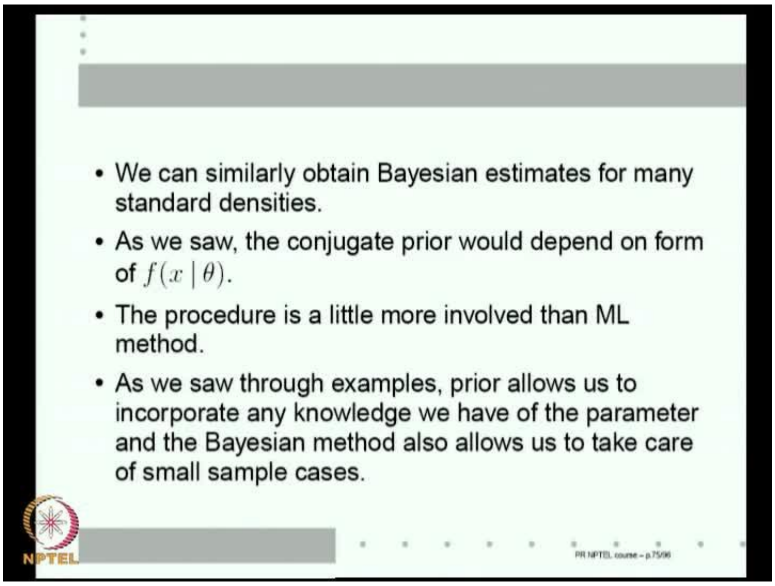
Whereas, the mu dependence can only be expressed by something that is the function of both nu and mu that is why, we have to model this in this kind of a factorization. When I put equality hereby now, we know, that we do not need the actual constants, we are only looking at the form of this densities. So, by now, we have seen enough examples so, I started misusing the abusing the notation, I just put equality even though, it is not really equal.

Because, this thing is not really a density, there will always be a, in the second thing there will be some normalizing constant. There will be one constant to make this gamma density a density, another constant to make this into a proper normal density so, there will be some constant. But by now, we know this constants do not matter so, I just, we are abusing notation by not putting that constant. And also, actually we would have some relation between c 0 and b 0 and a 0 but, it really does not matter, we can choose a slightly bigger class of densities at the conjugate prior.

Of course, this prior density is quite involved and doing the Bayesian estimation with this prior density is not easy so, I will skip the details, you people can sit and do the algebra, the algebra is more complicated. But ultimately, you get similar looking final (( )) essentially, what we get is a convex combination of the sample mean plus something that depends on the on the prior parameters for the for the estimate of nu.

And similarly, for the estimate of nu right, it will be some factor involving 1 by sigma square hat ML and something that, depends on your a 0 b 0 in such a way that, as n tends to infinity, once again the ML estimates and the Bayesian estimates will be same. So, we will we will we, I have just given you the prior for this but, we will not actually derive the final estimates. I have not done any of multidimensional examples, they are not conceptually any more difficult than the one dimensional normally we did. But obviously, as you can see compared to maximum likelihood, obtaining Bayesian estimates has lot more algebra so, just the algebraic notation will be more cumbersome so, we will we will skip that.

(Refer Slide Time: 41:21)



- We can similarly obtain Bayesian estimates for many standard densities.
- As we saw, the conjugate prior would depend on form of $f(x \mid \theta)$.
- The procedure is a little more involved than ML method.
- As we saw through examples, prior allows us to incorporate any knowledge we have of the parameter and the Bayesian method also allows us to take care of small sample cases.

PR NPTEL course – p.75/96

So, we will simply say that, we can obtain Bayesian estimates like this for many standard densities but, there is one part that is, by now evident, obtaining maximum likelihood estimates and Bayesian estimates is not the same. For maximum likelihood estimates, for almost mechanically, I can calculate the likelihood function, differentiate, equate to 0, find the maximum and I get the estimates. For Bayesian estimate, I have to properly choose the right kind of prior, which is the conjugate prior for that particular problem and only then, right the the expressions are amenable to simplification.

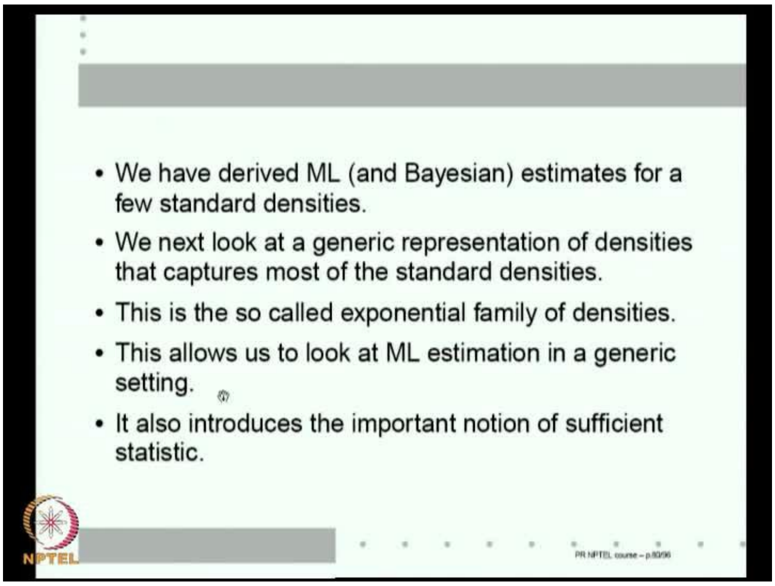
And then, I have to look at the look at the parameters of the posterior density and based on that, I I have to obtain my Bayesian estimates. As we have seen the conjugate prior would depend on the form of x given theta as a matter of fact, for the same density in our mind say, Gaussian depending on what is the parameterization that we think, what are the unknown parameters that we think, the prior changes.

For example, if we think only the mean of the Gaussian is unknown then, the conjugate prior is Gaussian, if we think only the variance is unknown and we choose the variance in terms of the precision parameter then, the density happens to be gamma. If we think both mean and the precision are unknown then, the prior density, the conjugate prior density turns out to be that gaussian gamma, I told you. So, the conjugate prior would depend very much on the form of f of x given theta, the procedure little more involves certainly than the maximum likelihood estimate.

So, what is that we gain with it, why is not maximum likelihood estimate sufficient, that we have already seen, when we when we started the Bayesian estimate, that the reason why we came to Bayesian estimate is that, maximum likelihood estimate blindly believes the data. So, if we have some prior information about the kind of values the parameter can take or because, our first few data are bad and we have very little data.

There is no way, we can make any incomplete information we have about the parameter to to bear on the final estimate we get. Whereas, the prior density allows us this the this flexibility so, essentially the prior density allows us to incorporate knowledge, that we may have about the parameter that is, in the form of a conjugate prior. And as we seen in the final expressions, it always comes up with an expression whereby, this small sample problems are automatically handled by trading the part that I get only from data, not trading by combining the part, that I get only from the data, with the part I get from the prior. So, at small sample, our beliefs, kind of moderators in not jumping to too drastic at conclusions based on data, that is the essence of the Bayesian estimation.

(Refer Slide Time: 44:39)



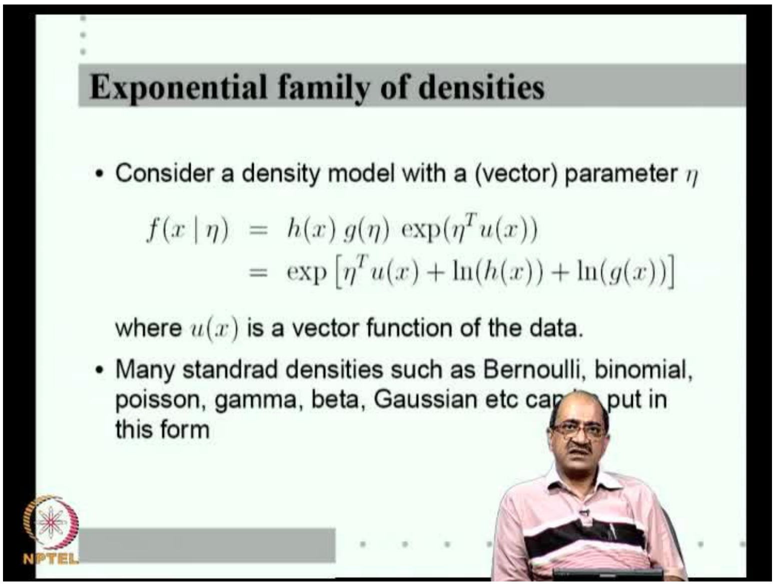
- We have derived ML (and Bayesian) estimates for a few standard densities.
- We next look at a generic representation of densities that captures most of the standard densities.
- This is the so called exponential family of densities.
- This allows us to look at ML estimation in a generic setting.
- It also introduces the important notion of sufficient statistic.

Now, we will slightly move to a few more related issues in estimation, we seen two specific methods of estimation, we have seen how to do the estimation for different densities. For example, we have derived ML and Bayesian estimates for a few standard densities now, let us look at a few more generic issues about estimation. The first thing that we will do is, to look at what is what is a generic representation for many densities so, there is one form for a density.

By by now, I suppose you become familiar to this, that as i said right in the beginning, when we started on our estimation, we use this the word density to mean either density or mass function. Depending on the random variable is discrete or continuous so, we use density in a generic sense so, we are saying, we will look first at at the representation of a density function in terms of some parameters, that captures most of the standard densities, such a form is called the exponential family of densities.

It is a it is a very important thing because, as we shall see later on, for exponential family of densities ML estimates become very straight forward. So, we get generic ML estimates for all densities within the exponential family, given the exponential family, for all of them, we can write one kind of generic set of equations to solve, to get the ML estimates, we do not have to do individually. And equally importantly, looking at this, kind of generic notion of a density function allows us to introduce an important notion estimation, which is called the sufficient statistic.

(Refer Slide Time: 46:28)



Exponential family of densities

- Consider a density model with a (vector) parameter η

$$\begin{aligned} f(x \mid \eta) &= h(x)\, g(\eta)\, \exp(\eta^T u(x)) \\ &= \exp\left[\eta^T u(x) + \ln(h(x)) + \ln(g(x))\right] \end{aligned}$$

where $u(x)$ is a vector function of the data.

- Many standrad densities such as Bernoulli, binomial, poisson, gamma, beta, Gaussian etc can be put in this form

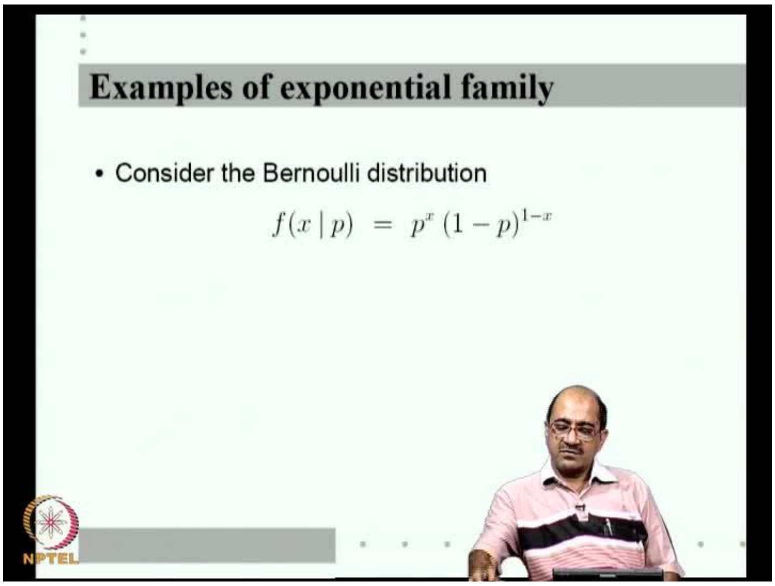
So, first let us look at, what we call exponential family of densities suppose, you have a density model for a random variable x with parameters eta. Eta could be a single parameter or many parameters, with many parameter we will think of it as a parameter vector. We will write the the density model as f x given eta as h of x that is, some function with only of x multiplied by some function g of eta, is some function only of the parameter eta multiplied by the exponential of eta transpose u x where u x is a vector of functions of the data.

So, given data I can make some new functions of data u 1 x, u 2 x for example, data is x 1, x 2, x n, u 1 x could be summation x i, u 2 x could be summation x i square and so on, you have sorry u 1 x could be x, u 2 x could be x square and so on. So, u x are some vector functions of x; so if the density can be written as a product of a time involving only x and a time involving only eta and a time that involves both eta and u and x in a very special way, exponential eta transpose u x where u is a vector of predefined, a vector of given functions of x.

The reason, why it is called the exponential family is that, I can always write it as exponential of something all right. I can write the exponential of eta transpose u x plus this h x factor can be brought inside the exponential by writing it as l n h x similarly, this as l n g x. Because, exponential of l n h x will be h x, exponential of l n g x will be g x because, I can write it like this, it is called a exponential family.

The important thing is that, many standard densities Bernoulli, binomial, Poisson, gamma, beta, Gaussian, exponential everything can be put in this form. Of course, among these standard densities, the notable one that cannot be written in this form is, uniform density except from uniform density, almost all the standard densities can be put in this form.

(Refer Slide Time: 48:50)



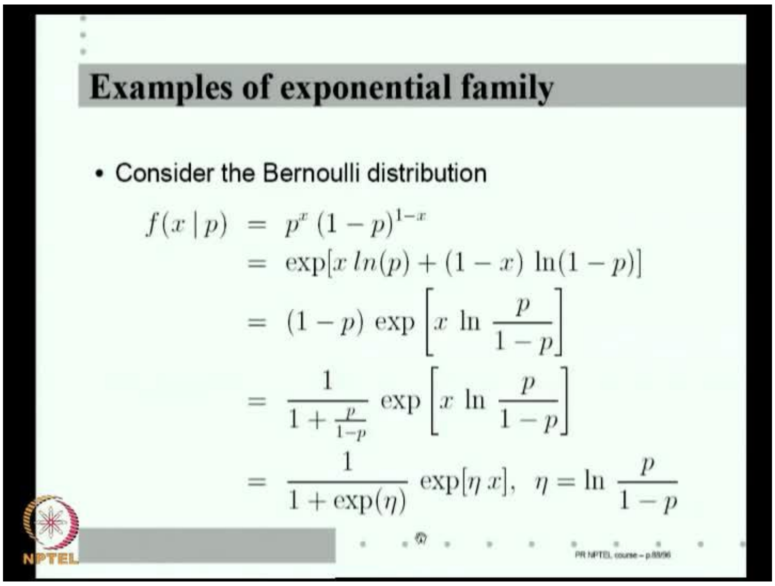
Examples of exponential family

- Consider the Bernoulli distribution

$$f(x \mid p) = p^x (1-p)^{1-x}$$

So, let us look at a simple example of, how to put a density in in the standard exponential let us consider, the Bernoulli distribution so, the mass function with parameter p is given in terms of p power x into 1 minus p to the power 1 minus x. This is obviously, not in the form of a factor involving only a x, multiplied factor involving only the parameter multiplied by a factor like this right. But the point is, by not thinking of p as the parameter but, something else as the parameter, we would be able to put it in this form.

(Refer Slide Time: 49:29)



Examples of exponential family

- Consider the Bernoulli distribution

$$\begin{aligned} f(x \mid p) &= p^x (1-p)^{1-x} \\ &= \exp[x\, ln(p) + (1-x)\, \ln(1-p)] \\ &= (1-p)\, \exp\left[x\, \ln \frac{p}{1-p}\right] \\ &= \frac{1}{1+\frac{p}{1-p}}\, \exp\left[x\, \ln \frac{p}{1-p}\right] \\ &= \frac{1}{1+\exp(\eta)}\, \exp[\eta\, x], \quad \eta = \ln \frac{p}{1-p} \end{aligned}$$

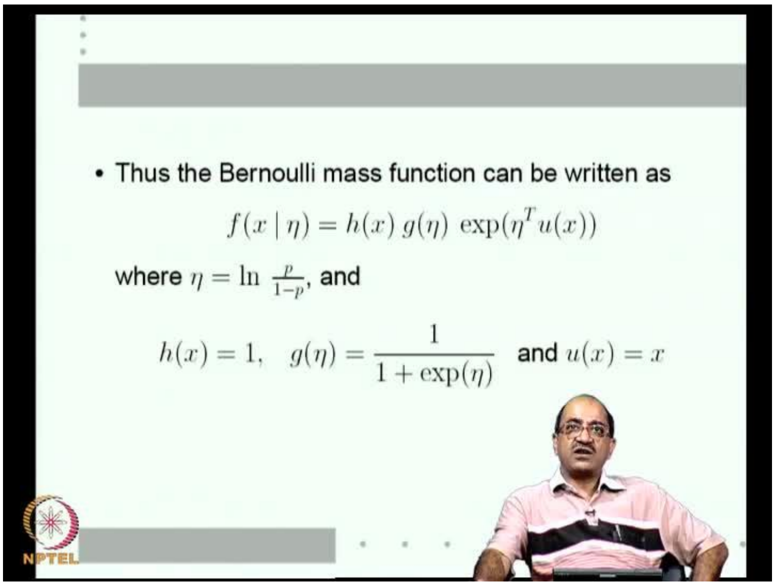
PR NPTEL course – p.89/96

So, let us look at that so, starting with this, we can write this as, I can always write anything so, I want you, I can write this p x into1e minus p to the power 1 minus x as exponential l n of that. So, if I if I take l n and put inside exponential, the l n of this will become x l n p plus 1 minus x l n 1 minus p so, that is what I did, exponential x l n p plus 1 minus x l n 1 minus p. Now, there is, this 1 into l n 1 minus p, exponential of l n 1 minus p will be 1 minus p so, let us let me take that factor out that is, 1 minus p.

Now, I have got x l n p and minus x l n one minus p, I can write it as, x into l n p by 1 minus p so, I can write this as 1 minus p exponential of x l n p by 1 minus p. Now, this I can further write as, 1 by 1 plus p by 1 minus p right. So this now, one can immediately see, if I think of p by 1 minus p as a parameter then, I can write this as let us say, that is what I want to call eta then, this is some factor that is dependent only on eta.

And this is a factor that depend exponential of let us say, l n p into 1 minus p is my eta then, eta times a function of x namely, x. So, I have to somehow, write this as also l n p by 1 minus p, this is very easy I can always write p by 1 minus p as exponential l n p by 1 minus p. So, I can write this as, 1 by 1 plus exponential eta into exponential eta x where, eta is l n p by 1 minus p right. So, I can write my Bernoulli mass function in as, 1 by 1 plus exponential eta into exponential eta x where, eta is l n p by 1 minus p.

(Refer Slide Time: 51:30)



- Thus the Bernoulli mass function can be written as

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

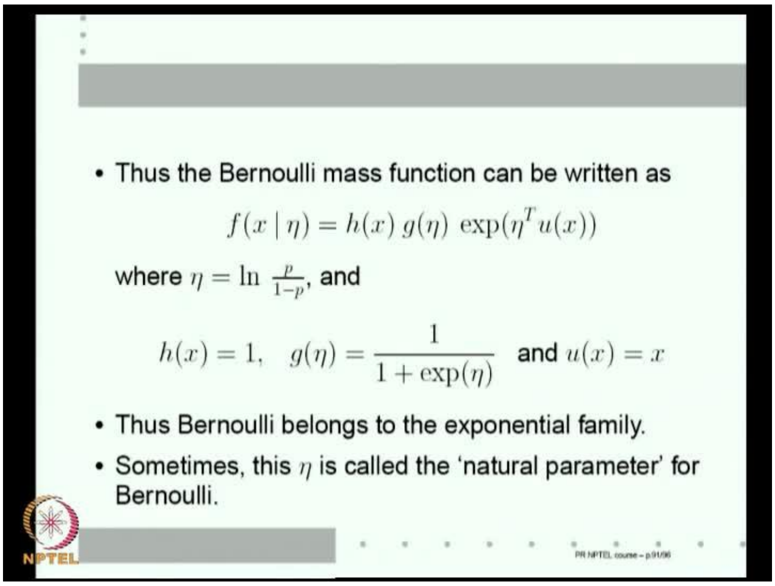
where $\eta = \ln \frac{p}{1-p}$, and

$$h(x) = 1, \quad g(\eta) = \frac{1}{1+\exp(\eta)} \quad \text{and } u(x) = x$$

So, what does this mean, this is exactly in the form h x into g eta into exponential eta transpose u x right h x is 1, there’ i no factor, there is only dependent on x right so, h x is

1. What is g eta, g eta is this factor 1 by 1 plus exponential eta right that is, g eta and I want exponential eta transpose u x, I have got eta times x. So, eta is a scalar here so, I can simply take u x to be x right. So, if I take eta as l n p by 1 minus p, h x as 1, g eta as 1 by 1 plus exponential eta and u x is equal to x then, it is in this form where, this transpose is of course, is redundant here because, eta happens to be one dimensional.

(Refer Slide Time: 52:26)



- Thus the Bernoulli mass function can be written as

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

where $\eta = \ln \frac{p}{1-p}$, and

$$h(x) = 1, \quad g(\eta) = \frac{1}{1 + \exp(\eta)} \quad \text{and } u(x) = x$$

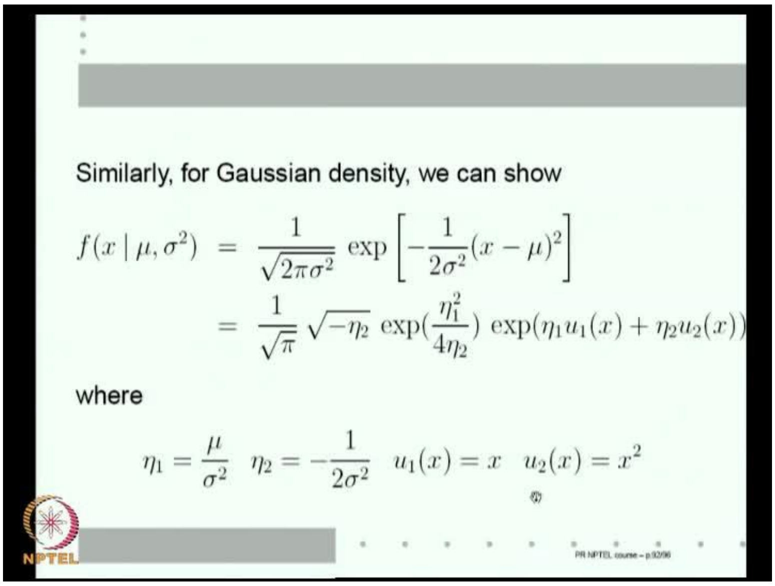
- Thus Bernoulli belongs to the exponential family.
- Sometimes, this η is called the 'natural parameter' for Bernoulli.

PR NPTEL course – p.91/96

Whereas this means, that the Bernoulli density belongs to the exponential family in the same way, for all the standard densities, we can put them in this general frame work. Sometimes see for example, if I want to represent the Bernoulli mass function with p as the parameter then, it is not in this generic form, h x into g eta into this. But, if I u instead of using p as the parameter, I use l n p by 1 minus p as the parameter then, I can put in this.

After all, if you give me eta that is, l n p by 1 minus p, I can calculate p or if you give me p, I can calculate l n p by 1 minus p so, the eta to p transformation is one to one and invertible. So, by that I think of eta as the parameter, as p as the parameter it does not matter but, if I think of eta as the parameter, the mass function comes to a very standard form so, sometimes eta is called the natural parameter for Bernoulli, this particular eta.

(Refer Slide Time: 53:26)



Similarly, for Gaussian density, we can show

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu)^2\right]$$
$$= \frac{1}{\sqrt{\pi}} \sqrt{-\eta_2} \exp(\frac{\eta_1^2}{4\eta_2}) \exp(\eta_1 u_1(x) + \eta_2 u_2(x))$$

where

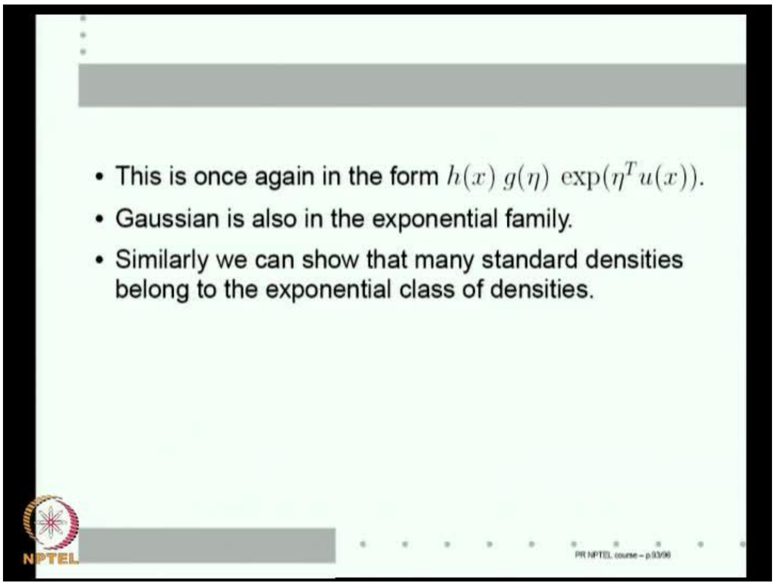
$$\eta_1 = \frac{\mu}{\sigma^2} \quad \eta_2 = -\frac{1}{2\sigma^2} \quad u_1(x) = x \quad u_2(x) = x^2$$

PR NPTEL course – p.32/96

In the same, many other densities can be put in the exponential form so, for example, for the Gaussian, the normally I will write it with, mu and sigma square as the two parameters like this. But we can also write it like this, some function represents only acts essentially a constant function, some factor that depends only on some parameters, which I call eta 1 and eta 2.

And then, exponential eta 1 time some function of x plus eta 2 times some function of x where, eta 1 happens to be mu by sigma square, eta 2 happens to be minus 1 by 2 sigma square, u 1 x happens to be x and u 2 x happens to be x square. The algebra involved is a little more than the algebra involved in showing this by the Bernoulli density but, once again it is just algebra. So, starting from this expression, one can show that, this is same as this expression where, I make this following changes eta 1 is mu by sigma square, eta 2 is minus 1 by sigma square, u 1 x is equal to x, u 2 x is equal to x square.

(Refer Slide Time: 54:31)

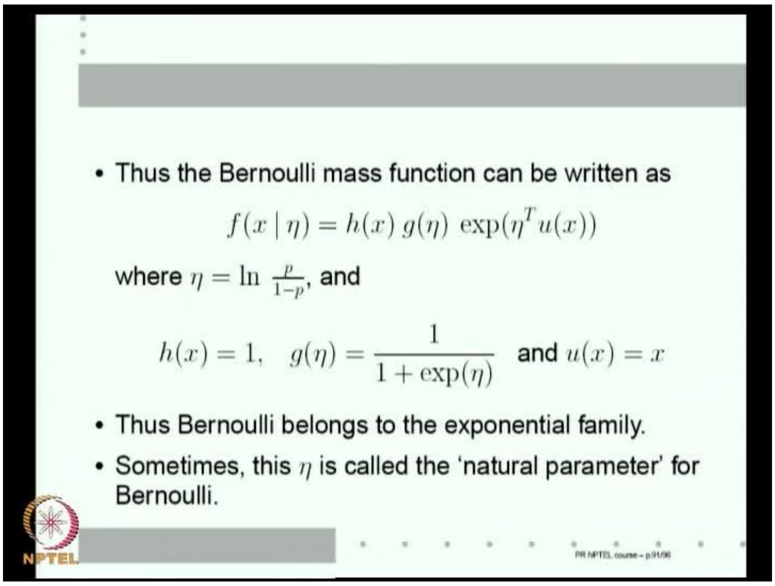


- This is once again in the form $h(x)\ g(\eta)\ \exp(\eta^T u(x))$.
- Gaussian is also in the exponential family.
- Similarly we can show that many standard densities belong to the exponential class of densities.

PR NPTEL course – p.33/96

So, under these things, this density once again recommends the form x h g eta exponential eta transpose u x right. As I said h x can be thought of as it is constant function, this is the g eta function and this is exponential eta 1 u 1 x plus eta 2 u 2 x where, eta 1, eta 2, u 1, u 2 are given right. So, once again these are the form h x g eta exponential eta transpose u x so, Gaussian is also in the exponential family and like this, we can show that almost all standard densities belong to exponential class of densities. Now, what is the use of showing many of these densities belong to the exponential family of densities, the the main utility is that, as I said, we get a very standard generic form for the maximum likelihood estimate.

(Refer Slide Time: 55:32)



- Thus the Bernoulli mass function can be written as

$$f(x \mid \eta) = h(x)\, g(\eta)\, \exp(\eta^T u(x))$$

where $\eta = \ln \frac{p}{1-p}$, and

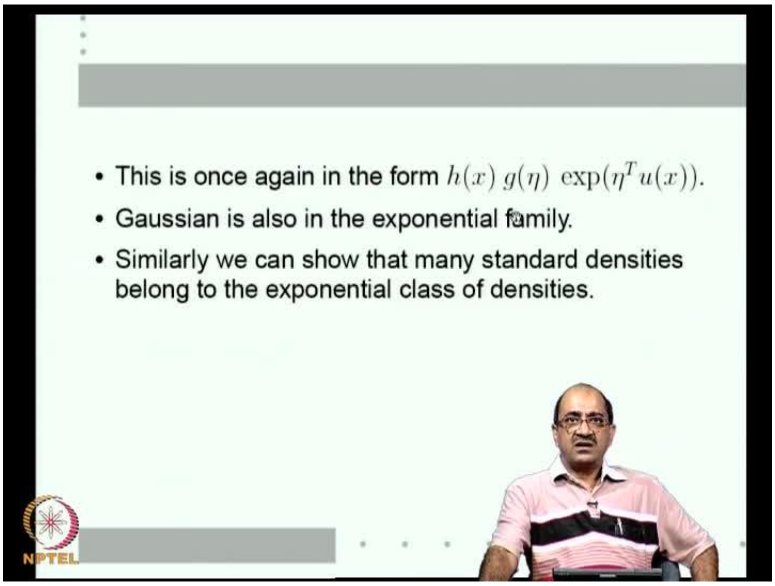
$$h(x) = 1, \quad g(\eta) = \frac{1}{1 + \exp(\eta)} \quad \text{and } u(x) = x$$

- Thus Bernoulli belongs to the exponential family.
- Sometimes, this η is called the 'natural parameter' for Bernoulli.

PR NPTEL course – p.91/96

And also, what it would mean is the following now. When a density is in this form, if i take the data likelihood right, the data likelihood will depend on n fold product of this that is, product of x h i g eta to the power n. And when I multiply exponential eta transpose u x 1 into exponential transpose u x 2 and so on ultimately, I get exponential eta transpose summation u x i. So, these functions u x or the quantity summation u x i are the only way, the data affects the data likelihood. So, this form gives us a very interesting generic way, in which data affects the data likelihood and hence, gives us a standard method for calculating maximum likelihood estimates for all such densities.

(Refer Slide Time: 56:29)



- This is once again in the form $h(x)\, g(\eta)\, \exp(\eta^T u(x))$.
- Gaussian is also in the exponential family.
- Similarly we can show that many standard densities belong to the exponential class of densities.

So, in the next class, we will we will look at a few of the examples of the exponential family of densities. And how, looking at all of them as exponential family of densities, allows us to obtain maximum likelihood estimates in a in a generic fashion and then, we will introduce the notion of, what is called a sufficient statistic.

Thank you.