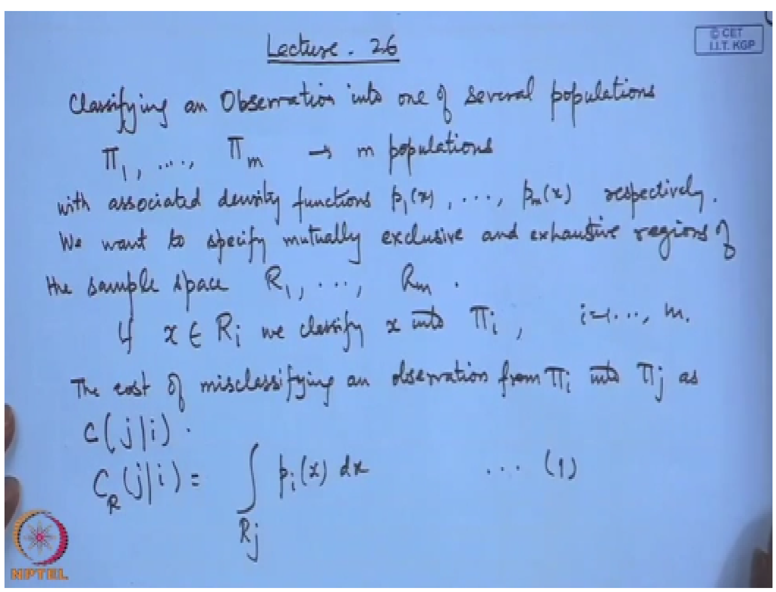
Statistical Methods for Scientists and Engineers
Prof. Somesh Kumar
Department of Mathematics
Indian Institute of Technology – Kharagpur

Lecture – 26
Multivariate Analysis – XI

In the last few lectures, I have introduced the problem of classifying an observation into one of the 2 populations. I discussed various procedures. In particular, I showed that one can define Bayesian classification rules or what you can say as good classification rules, so we call them the admissible rules or minimal complete class, that means the rules beyond which you need not discuss. In particular, we consider applications to the classification for an observation into 2 multivariate normal populations.

The first case was when all the parameters are known and then second case we discussed when the parameters are unknown. Now, I will generalize this concept to the problem of classification of one observation into several populations. So, let me introduce the concept of optimal rules here and how to derive these rules.

(Refer Slide Time: 01:20)

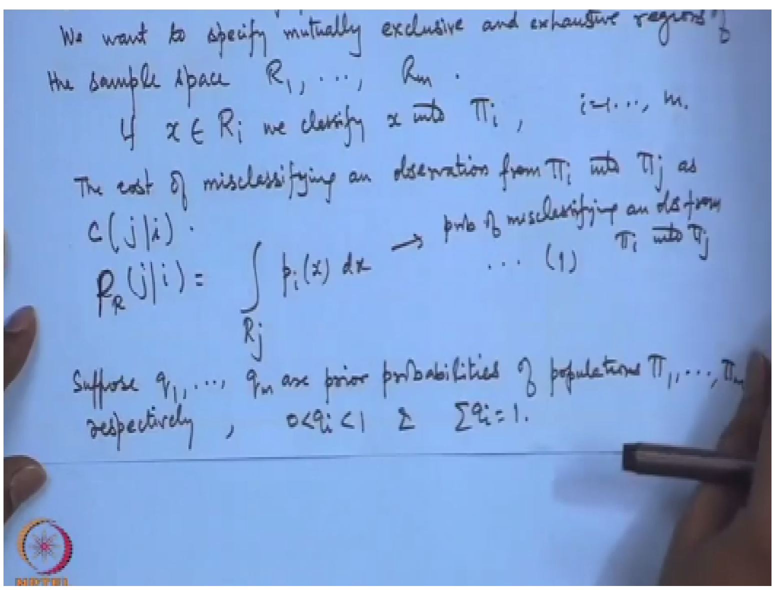


So, classifying an observation into one of several populations. So, supposed we are having populations pi1, pi2, … pi m, these are m populations and we are considering the associated density functions say P1X, P2X, … PmX, etc., okay. So, we wish to classify or we can find out m mutually exclusive regions. We want to specify say mutually exclusive and exhaustive regions of the sample space, say R1, R2, … Rm.

So, if the observation x belongs to say Ri, we classify x into the i population for i = 1….m. We can also consider the cast of misclassification as we have done earlier. The cost function we can introduce the cost of misclassifying and observation which is actually from say pi i but we classify into pi j. Then, we call this function as Cji. Now, we can define it like this Crj given i=see the observation is initially from the i-th 1 but we have classified it as into Jth 1.

So, the probability of misclassification or the cost of misclassification can be considered like this. Now, if you remember the case of classification into 2 populations, I had considered one particular case when we fix the initial proportions of the population as q1 and q2 where q1+q2 = 1. In a similar way, if I have m populations I may consider the case when the initial proportions of these populations are known. We call them prior probabilities say q1, q2, ….qm.

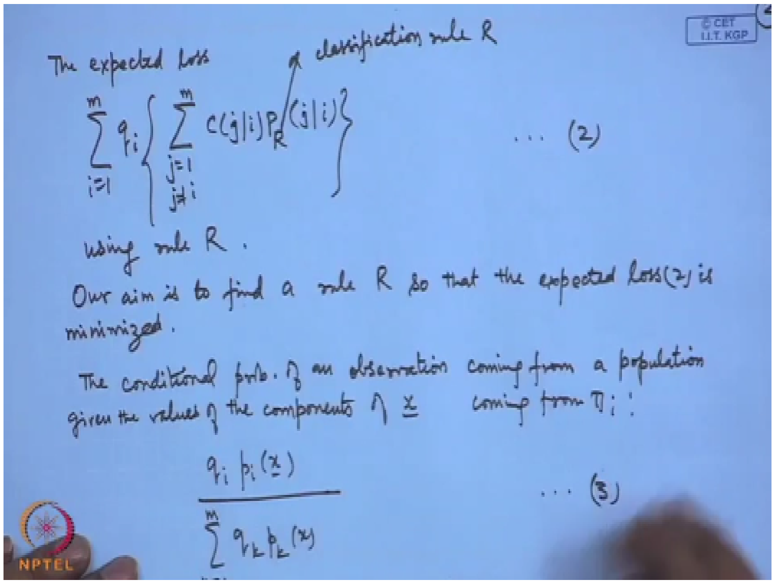
(Refer Slide Time: 04:35)



Suppose q1, q2, …qm are prior probabilities of populations by pi1, pi2, …pi m respectively, that means 0 < qi < 1 and sigma of qi = 1. If I am considering the expected loss of classifying i*j then I can consider based on the prior probabilities the total expected loss. So, the total expected loss can be defined, so let us consider see we are having. Okay, I just made a small mistake here. This should be Pr, that is the probability of misclassification. This is the cost.

So, this is actually the probability of misclassifying an observation from pi i*pi j. So, Prj given i and the cost of misclassifying is Cji. So, if I consider Cji*Prji, then this will become the expected cost.

(Refer Slide Time: 06:10)



Now, let us consider the total expected loss. So, let us consider this, Cj given i, Prj given i. This r denotes the classification rule. This is the classification rule. Now, what we have done here is that observation is from i-th population and we are classifying it as into the Jth population. So, this is the expected cost or expected loss you can consider. Now, you may put it into any of the remaining populations. So, we vary j from 1 to m where j is != i. Now, this term which is written here.

So, now the observation is from the i-th 1 and we can put into any other population, other than the i-th population. So, this the total expected loss that is coming there. Now, the probability or the proportion of the i-th population that is qi. So, I multiplied by that and then I sum over all i from 1 to m, then this is the expected loss using classification rule R, so this is the expected loss that we can consider.

You can remember the value which I wrote for the case of 2 populations. In the case of 2 populations, let me show you the expression here which we discussed earlier. I will show you the thing and then you can compare, so it will be clear that how this has been obtained here. If you remember the case of 2 populations, we had only 2 value C21 and C12. In this case, we have several values Cji where i am j both can vary from 1 to m where i is != j.

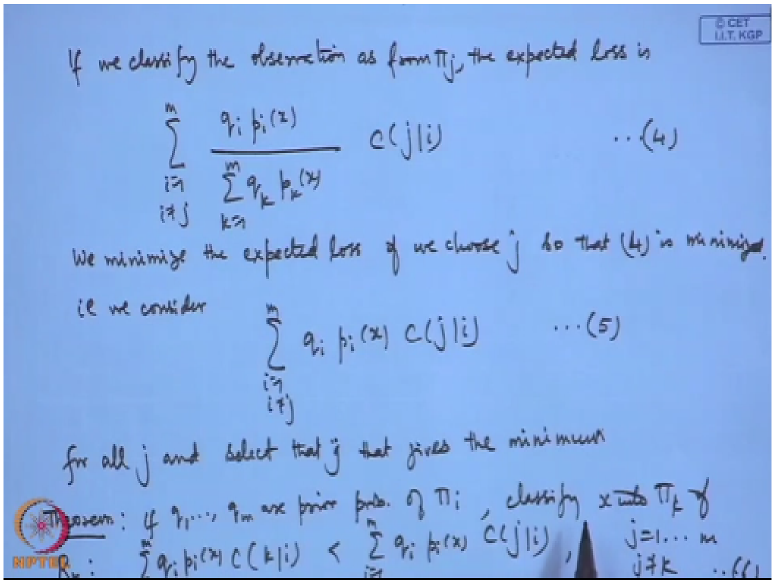
So, this is the difference that is coming here. In place of 2 values C21 and C12, now I have Cji for all j != i and for all i. So, total number of values will be m*m-1 that you will get here. Let me also show you the expected loss that we had considered here. So, the term that I wrote here Prj given i. In the case of 2 populations, we had only 2 values Pr2 given 1 and Pr1 given 2.

Now, I have m*m-1 values once again that will be there and the expected loss of misclassification was only C21 Pr21 q1+C12 Pr1 given 2*q2. Now, you compare this with the value that I wrote just now because any observation from the i-th 1 can get into any of the other than the i-th 1 one, then you consider all such cost then you add them. Then you look at the i-th 1 and multiply by the prior probability of that and then you sum over.

So, this is the expression that you will be getting. So, this is the full explanation of the expected cost as compared to the case of 2 populations. So, as you can see here the expression becomes much more complex here. However, our aim or the motive remains the same, that is to minimize the expected cost of classification, expected loss by misclassification. So, our aim is to find a rule R so that the expected loss 2 is minimized.

As again in the case of 1, we had considered q1p1xyq1p1x+q2p2x, in a similar way, we can consider the conditional probability of an observation coming from a population given the values of the components of x. So, given that it is coming from pi i. So, that is defined as an qipix/sigma qkpkx. Earlier it was q1p1x/q1p1x+q2p2x or q2p2x divided by the same term, but here now I will have all the m terms in the denominator.

(Refer Slide Time: 11:52)



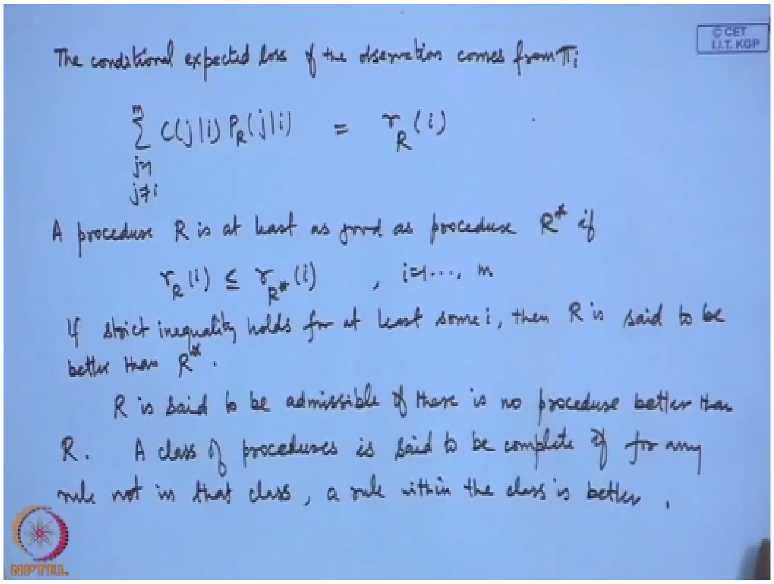
If we classify the observation as from pi j, then the expected loss is qipix/sigma qkpkxk=1 to m multiplied by Cji, I = i to m, i != j. We minimize the expected loss if we choose j, so that 4 is minimized, that is we consider the term qipix Cji given i because the denominator is common here for all j and select that j that gives the minimum. So, in principle if you look at this is a direct extension of the case of 2 populations.

If of course there may be a case when 2 different j give you the same value, in that case you can choose the one, well it does not matter because then whichever you choose it will give the same minimizing constant. So, now I consider this procedure assigns the value. So, we are assigning towards a j. So, that is region Rj here.

So, we consider then the following result then that if qi, ….qm are prior probabilities of pi i and the cost function is given here, then the region of classifying Rk is given, so Rk region is sigma qipix Ck given i < sigma qipix Cj given i, here i =1 to m, i != j and here it is i = 1 to m, i != k, that means for the kth 1 if this is the minimum, then you are getting the rule, that is you should classify X into pi k if this is happening.

I will not get into the proof of this. In fact, the proof is almost the generalization of the proof for the 2 population, that means if I consider any other rule which is minimizing, then I can consider the expected loss from the 2 given rules and write down the difference and as in the case of 2 populations, you can consider the conditions for greater than or equal to thing. So, it will come immediately. So, I am skipping the proof here.

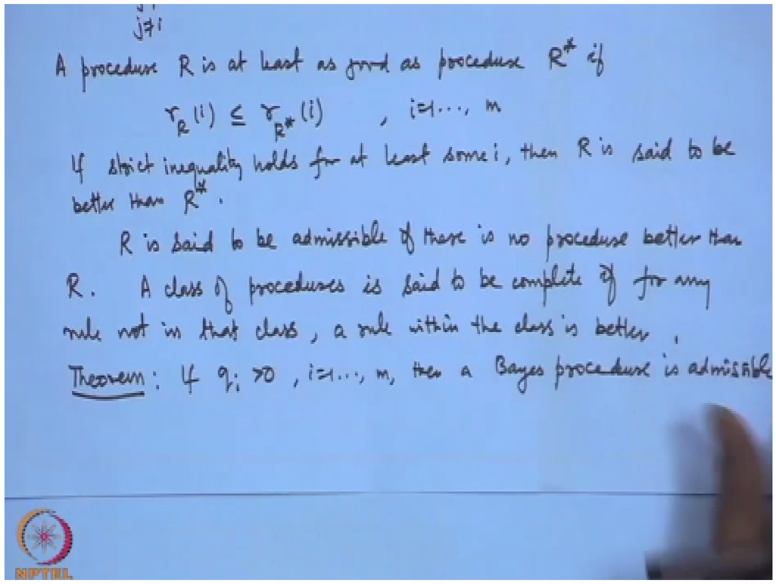
(Refer Slide Time: 15:47)



Like the case of 2 populations, we can consider the optimality criteria like admissibility, Bayes rules etc. Let me just formally define that here. We can consider the conditional expected loss if the observation comes from pi i, that is sigma Cj given i, Prj given i, this term I wrote earlier, this one basically. So, I am writing this one. So, this is for j = 1 to m, j != i. I will use the notation rR i here.

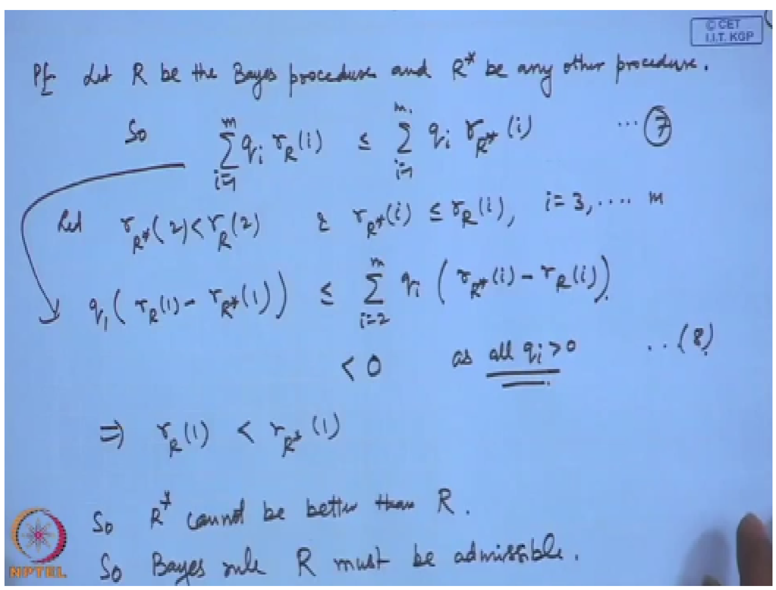
So, a procedure R is at least as good as procedure R star if and only if we are having rR i <= rR star i for i = 1 to m. If strict inequality holds for at least some i, then R is said to be better than R star. R is said to be admissible if there is no procedure better than R. A class of procedures is said to be complete if for any rule not in that class, a rule within the class is better. So, these definitions are similar to the one which I gave for the case of 2 populations.

(Refer Slide Time: 18:49)



We can consider that this result is also similar as we had earlier that if qi is positive, then a Bayes procedure is admissible. Once again, the proof is almost the same. Let me exhibit at least this proof here.

(Refer Slide Time: 19:29)

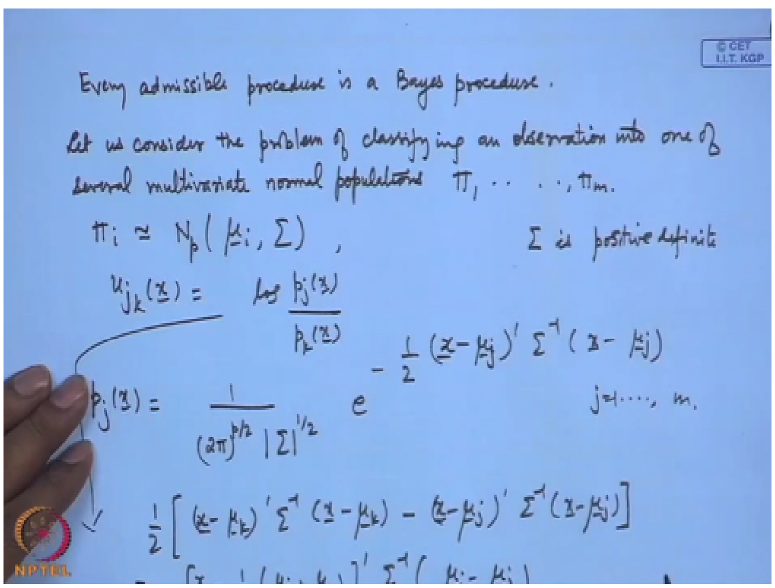


Let R be the Bayes procedure and R star be any other procedure. Now, Bayes procedure will give you the expected loss <= the expected loss corresponding to the rule R star, okay. Now, let us take say one of the components to be strictly smaller than the corresponding component for the rule R. So, what I am saying is that actually if I want to show that R star is better than R, then we should have R star i <= rR i for all i and strict inequality for at least one value.

So, for the time being I am assuming less than or equal to for 2…m and strict inequality at least for the 2, then let us see what happens. So, if we substitute it here, so what we get q1rR1-rR star 1, this will be <= sigma i = 2 to m, so from here actually I am writing her, qirR star i-rRi. Now, let us look at these 2, I am writing for 2 to m, so for 3 to m, it is less than or equal to and at least there is one strict inequality and all qi are positive.

Therefore, this will be strictly < 0 as all qi are positive. If this is happening, this is implying that rR1 is strictly < rR star 1, so R star cannot be better than R. So, that means the Bayes rule R must be admissible. Actually, you can see that the proof is similar to the case for the case of 2 populations, that is m = 2. There I had taken only strict inequality for 2 and then for 1 I got the reverse one and similarly for the other case. Now, if the cost functions are given then also the Bayes procedure are admissible.

(Refer Slide Time: 22:55)

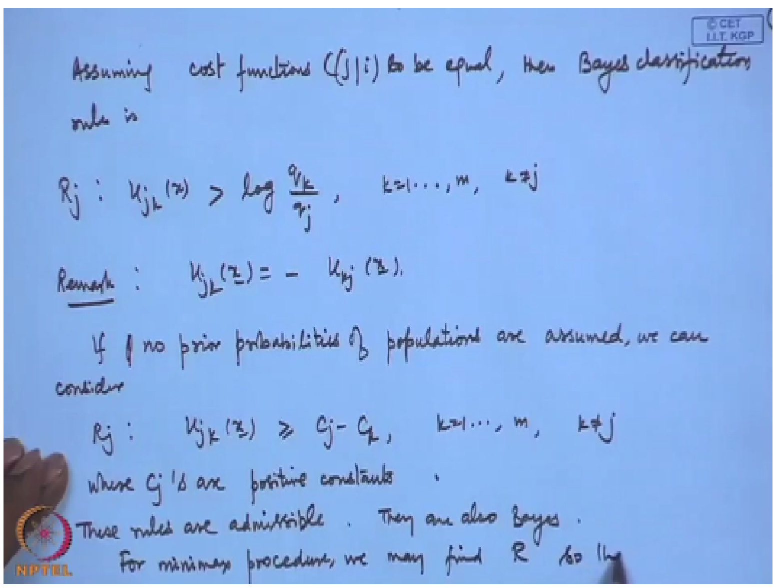


The converse of this result is also true, that is every admissible procedure is also a Bayes procedure. I will not give the proof of this also as the proof is similar to the case of 2 populations. Now, what we want to do is that let us consider the classification of multivariate normal population. So, let us look at this problem which had considered for the case of m = 2. So, let us consider the problem of classifying an observation into one of several multivariate normal populations.

The populations are pi1, pi2, …. pi n where pi i is the population normal mu i sigma. So, let us define say Ujk function that is equal to log of Pjx/Pkx. Actually hear, Pjx will be 1/2pi to the power P/2 determinant of sigma to the power 1/2 e to the power -1/2x-mu j, j I am writing so it should be j here, prime sigma inverse x-mu j. So I am assuming sigma is positive definite because I am writing down the existence of the density function for j = 1 to m. So, this quantity that is the ratio here log of Pjx/Pkx.

You can consider here, see this term will get cancelled out. So, you will get e to the power 1/2 and the term will be corresponding to k in the numerator and j in the denominator and then you take log. So, e will go way, so you will get basically 1/2 of x-mu k prime sigma inverse x-mu k-x-mu j prime sigma inverse x-mu j, that is equal to after simplification x-1/2 mu j+mu k prime sigma inverse mu i-mu j.

(Refer Slide Time: 26:49)



Then, the regions of classification, if we apply this method that is we are considering sigma qirRi <= this. Basically, we have mentioned here that the one which is minimizing this. So, the procedure will become, if we assume the cost functions that is Cji to be say equal, then the Bayes classification rule is Rj that is classify the observation into jth population if Ujkx is > a log of qk/qj for k equal to 1 to m, k is != j.

We can notice here that this function Ujk they satisfy the symmetry property. So, actually what kind of regions are these. If you see it carefully, they are nothing but the RI are actually bounded by the hyperplanes type of region because x-1/2 mu j+mu k prime sigma inverse mu y-mu j. So, what kind of regions you will be getting. You will be getting the regions of the type of the hyperplane.

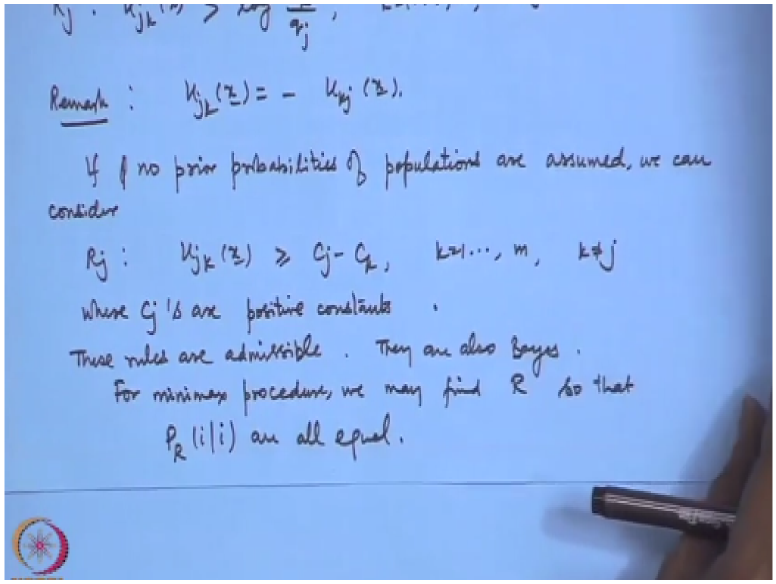
So, if the mean (()) (28:55) m-1 dimensional hyperplane, then Ri is bounded by m-1 hyperplanes that you will be getting, because the X value that will give you a hyperplane, X greater than something or less than something and if the prior probabilities are not given, then in place of log of qk/qj, you can put some value and in order to maintain some sort of symmetry of representation, actually what is the value of log qk-log of qj because these are probabilities.

So, basically you are getting them to be negative because they are lying between 0 and 1. So, we can write log of qk-log of qj. So, with minus signs were are getting, so we can put a log of qj before because there is a minus sign, so we can put it in terms of the non-negative values. If no prior probabilities of populations are assumed, we can consider Rj as Ujkx >= Cj-Ck for k=1 to m, k to 1 to m, k != j, where Cj's are positive constants.

Actually, any rule of this type is a Bayes rule. So, in case of prior probabilities, if we are putting some other numbers we can actually define respective probabilities in such a way that they will be equal to something. So, all such procedures, they will be giving you the Bayes admissible rules. So, basically, this is you can say minimal complete class of the classification procedures for classifying into one of several populations. These rules are admissible rules.

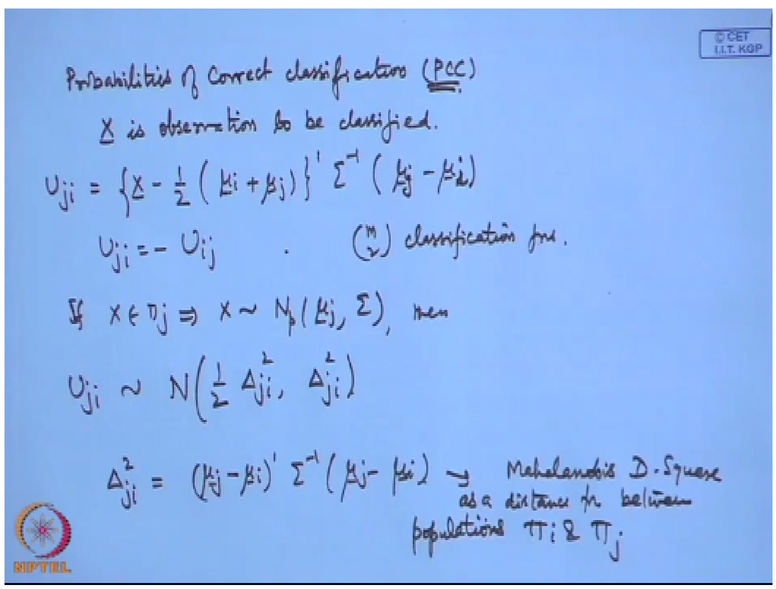
They are also Bayes because the class of Bayes rules and admissible rules is the same here. Now, if you want to find out a minimax procedure, then we can consider say probability of the correct classification and we can make them to be equal.

(Refer Slide Time: 32:00)



For minimax procedure, we may find R so that Pri given i are all equal.

(Refer Slide Time: 32:13)

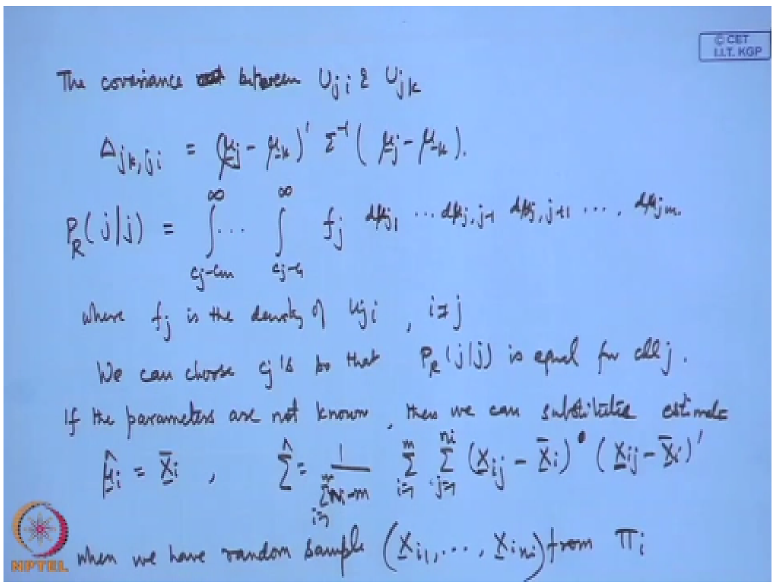


Let us look at what are the probabilities of correct classification which we can also call PCC. So, x is the observation to be classified, then we can consider say Uji=x-1/2 mu i+mu j transpose sigma inverse mu i-mu j. This is the classification function that we got and of course Uji-Uij they are related here, so that means basically we can consider MC2 classification functions, because we do not have to consider both here, MC2 classification functions are there.

Of course, this is because if they span in m-1 dimensional hyperplane. So, now if x belongs to pij, that means x is having Np mu j sigma distribution, then what is the distribution of Uji. See this is normal and we can apply the linearity properly. So, this will become actually mu j-mu y and here also you are having this thing here. So, this is mu j-mu i prime sigma inverse mu j-mu i 1/2 can be taken outside.

If I define the term say delta ji square which is a generalization of the Mahalanobis D square function which I wrote in the case of 2 populations, then this is equal to mu j-mu i prime sigma inverse mu j-mu y. So, in terms of this, this is actually Mahalanobis D square as a distance function between populations pii and pij. So, then this is equal to normal 1/2 delta ji square, delta ji square. Also, we can look at the covariance of Uji and Ujk.

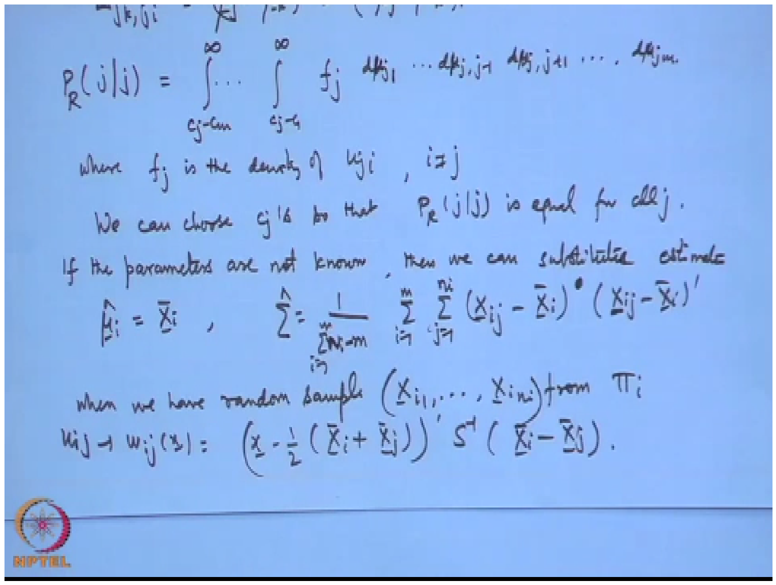
(Refer Slide Time: 35:36)



So, this is a scalar because Uji is a scalar function. The covariance matrix between Uji and Ujk, so that is equal to I use a notation delta jkji=mu j-mu k prime sigma inverse mu j-mu k. In the classification rule when the prior probabilities are not fixed in advance, then we have to determine constant Cj, Ck, etc which I mention that we can choose them to be nonnegative. So, we can consider the probability of classifying in 2j when it is from j=Fj that is the observation is from the jth 1, duj1 and so on…d mu jj-1, d mu jj+1 and so on d mu jn and these are Cj-C1 and so on…Cj-Cn the because upper side is infinity here.

Where Fj is the density of Uji for i != j. So, we can choose Cj so that Prj given j is equal for all j. Now, the another situation arises if the parameters are not equal, they are not known then we can substitute estimates. For example, mu i head can be xi bar and similarly you can have sigma head = 1/sigma ni-m, i = 1 to m and Xij-Xi prime, j is equal 1 to say ni, i = 1 to m. When we have random samples xi1, xi2, …xini from pii then we can consider these estimates.

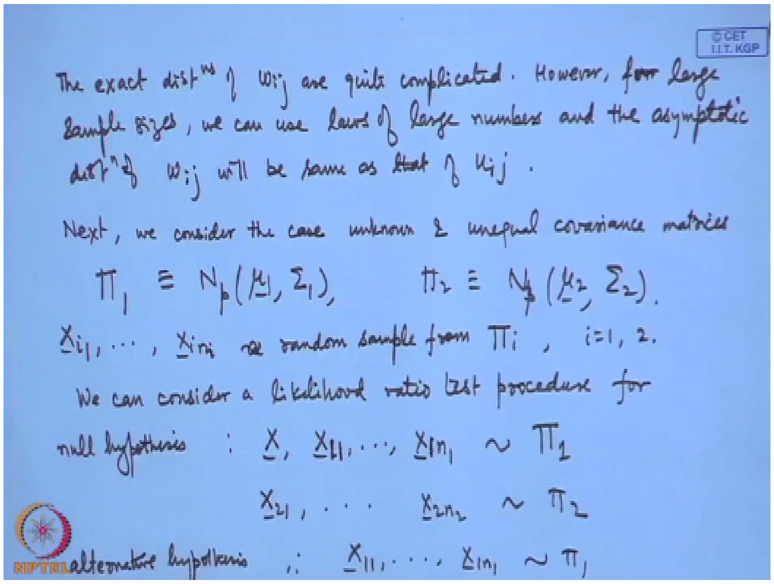
(Refer Slide Time: 39:36)



The analog of Uij, this will become Uij that will be x-1/2 xi bar+xj bar prime S inverse xi bar-xj bar. Now, as we discussed the case of 2 normal populations, the distribution theory for this part is somewhat more complicated; however, the exact distribution theory is not very difficult because strong of large numbers will hold and if I take here a large sample sizes and then you can consider here that xi bars will converge to corresponding mu i, xj will go to mu j, S inverse will go to sigma inverse in probability and so on.

Therefore, the asymptotic distribution of Uji, Wij will be almost the same as the Uji. So, the problem can be handled.

(Refer Slide Time: 40:54)



So, the exact distributions of Wij are quite complicated; however, for large sample sizes we can use laws of large numbers and the asymptotic distribution of Wij will be same as that of Uij. So, this problem can be solved. Now let us also go back to one of the problems that I discussed earlier that is classifying into 2 multivariate populations when the variance covariance matrices were unequal.

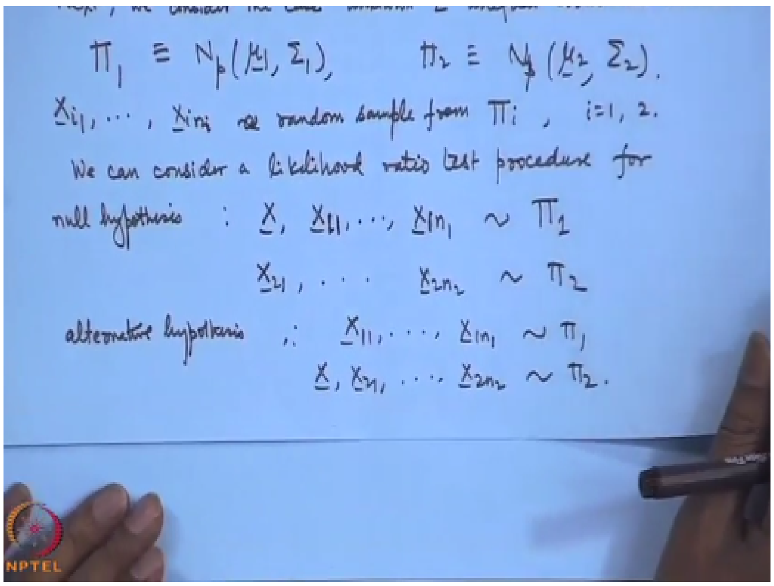
I discussed the rule when the 2 populations had known parameters, so if you remember the rule that I had mentioned here. It was given by this that is when sigma1 and sigma2 are different, I mentioned that in place of hyperplanic regions, you are actually getting much more complex regions because I mention that one of them becomes central chi-square distribution. So, this region becomes much more complicated.

Now, we also consider this case for the unknown sigma1 and sigma2 case. In that case, we have to substitute the estimators. So, let me briefly discuss this case also. Next, we consider the case of unknown and unequal covariance matrices. So, in particular let us consider say pi1 as the population Np mu1 sigma1 and pi2 is the population Np mu2 sigma2. So, let me go back to the expression that I derived earlier.

The expression that we obtained was in fact there was a power here which I would have missed at that time. It should be power half here and power half here also and it is e to the 1/2 x-mu2 prime sigma2 inverse x-mu2-x-mu1 prime sigma inverse x-mu1. So, we can consider a likelihood ratio procedure. We are having the samples here say xi1 and so on…xini. This is from a random sample from pii, i is equal 1 to 2.

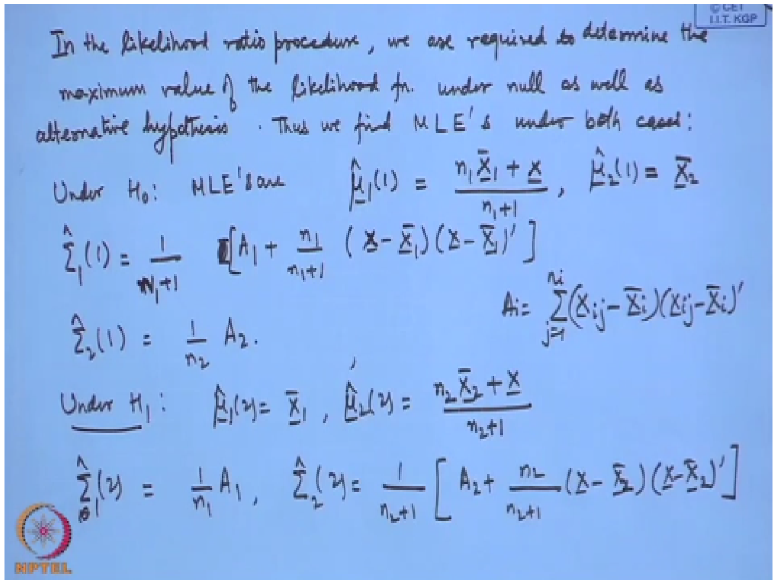
We can consider a likelihood ratio test procedure for null hypothesis, that is the observation x, x11, x12, …x1n1, they are from pi1 and x21, x22,…x2n2 this is from pi2 and the alternative hypothesis will be that is x11,… x1n1 this is from pi1.

(Refer Slide Time: 46:02)



x21 and so on…x2n2 this is from pi2. Now, in the likelihood ratio procedure, I have to consider the maximization of the likelihood function under both null and alternative hypothesis. So, in the null hypothesis, I will have n1+1 observations from pi1 and n2 observations from pi2. In the alternative hypothesis, I will have n1 observations from pi1 and n2+1 observations from pi2. Since all the parameters are unknown and unequal, this is simply reducing to the problem of finding out maximum likelihood estimators for these cases.

(Refer Slide Time: 46:57)



So, we can easily write down the maximum likelihood estimators. In the likelihood ratio procedure, we are required to determine the maximum value of the likelihood function under null as well as alternative hypothesis. Thus, we find MLE’s under both cases. So, this we can write easily because the procedure for finding out the MLE is known in the case of multivariate normal distribution.

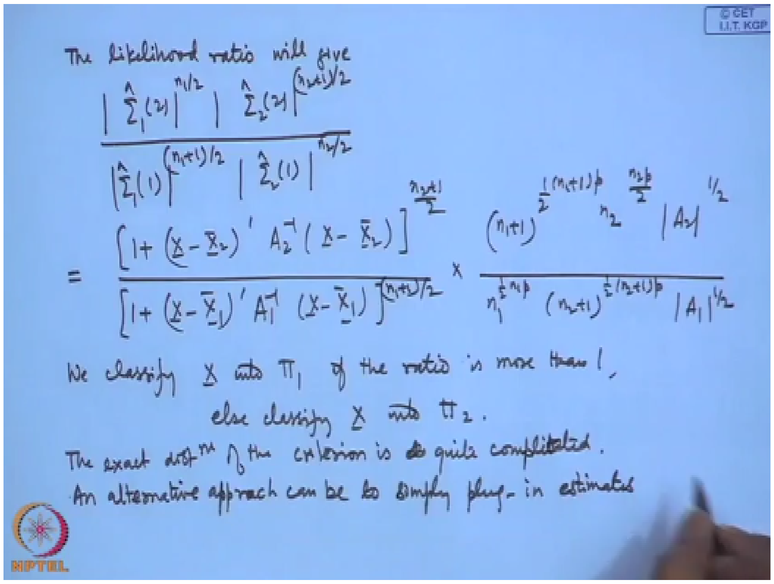
We know actually that the sample mean and the sample covariance matrix, they are the maximum likelihood estimators. So, under the null hypothesis MLE's are, so we write it as mu1 head 1=n1x1 bar+x/n1+1 because this is a sum of all the observations from the first sample plus x because we are saying that it is coming from the first population, mu2 head 1, so 1 means basically under the null hypothesis.

This is given by x2 bar and sigma1 head 1 that will be 1/n1+1 sigma, so we call it say A1+n1/n1+1 x-x1 bar x-x1 bar transpose and sigma2 head 1=1/n2 A2. Here Ai are Xij-Xi bar Xij-Xi bar transpose, j = 1 to ni. Under H1 that is the alternative hypothesis, actually this null hypothesis I am calling H knot and this alternative hypothesis I am calling H1. So, under H1 this will turn out to be mu1 head 2 that will become X1 bar and mu2 head 2. Now here I will have n2+1 observations from the second one, so it will be n2x2 bar+X/n2+1.

For sigma1 head, this will become equal to 1/n1 A1 and for sigma2 head, this will become

1/n2+1 A2+n2/n2+1 x-x2 bar x-x3 bar transpose.

(Refer Slide Time: 51:25)

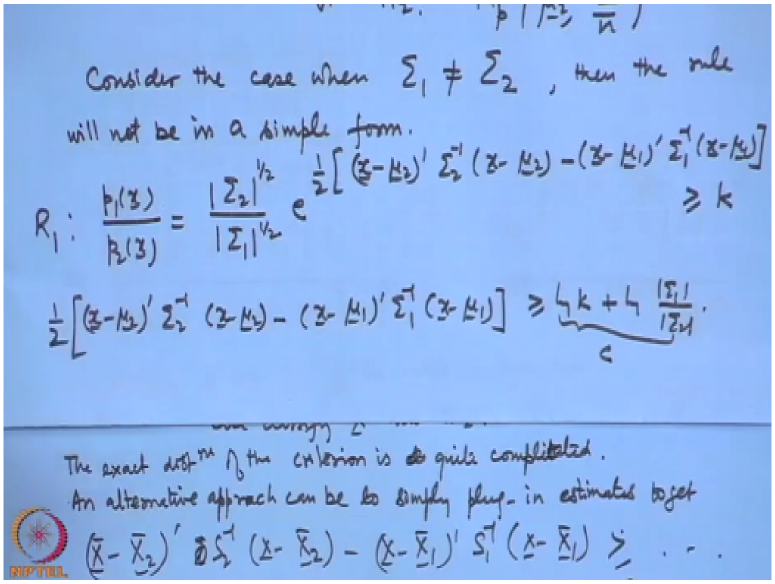


So, if we consider the likelihood ratio criteria that will give us the likelihood ratio. So, the exponent term will get cancelled out and we will get sigma1 head to the power n1/2 sigma2 head 2 to the power n2+1/2 divided by sigma1 head 1 to the power n1+1/2 and sigma2 head 1 to the power n2/2 which we can also write after simplification as 1+x-x2 prime A2 inverse x-x2 bar whole to the power n2+1/2/1+x-x1 bar prime A1 inverse x-x1 bar to the power n1+1/2*n1+1 to the power 1/2 n1+1p n2 to the power n2p/2, determinant of A2 to the power 1/2/n1 to the power 1/2 n1p n2+1 to the power 1/2 , n2+1p A1 to the power 1/2.

If we consider the costs of misclassification to be the same and the prior probabilities to be equal, we can consider this ratio to be > 1, then you classify into pi1. So, we classify X*pi1 if the ratio is more than 1. Else classify X*pi2. Now, this is one criteria that is the likelihood ratio criteria. Let me also again come back to this original observation that I got here. Another thing could be that I substitute direct estimates, that means here I put x2 bar, here I put S2 inverse, here I put S1 inverse, here I put x1 bar.

So, in both the cases, the exact distributions of the criteria are not easy. The exact distribution of the criteria is quite complicated.

(Refer Slide Time: 55:04)



An alternative approach can be to simply plug-in estimates into this function, that is x-x2 bar prime sigma, that is S2 inverse x-x2 bar-x-x1 bar prime S inverse x-x1 bar. So, you put this as some greater than or equal to something and less than something. So, this could be another one. Once again, the distribution of the criteria is quite complicated.

Even if I look at the asymptotic distribution for the large sample sizes by applying the laws of large numbers, I will get this one. I have already discussed that if x belongs to pi1, then this is non-central chi-square. This will be something like a central chi-square. So, the difference of 2 and it is going to be quite complicated. In case it is from pi2, then this one is central and this one is non-central. Once again the exact distribution of these things are difficult to obtain.

So, in particular we are saying is that when the variance covariance matrices are unequal, the classification rules no doubt can be easily found out but in order to obtain desirable rules such as a minimax procedure among them is a difficult task. Because the probabilities of correct classification are the probabilities of misclassification will be quite complicated. Friends, so we have actually discussed so many classification rules.

In fact, I framed a general decision theoretic approach to the classification problem by considering the Bayes decision rule and the criteria of admissible rules, the minimax classification procedure, the minimal complete class, etc and in particular, I showed applications to the classification of multivariate normal populations. We have considered 2 normal populations and multivariate normal population.

So, I actually wind up the discussion on the problem of classification now. One can consider some other classification procedures which are available nowadays but that can be a subject of full-fledged discussion. I will move over to another topic that is the problem of principal components. So, in the next lecture I will briefly introduce the problem how to determine the principal components and also maybe I will touch up on the canonical correlations. So, that will be the topic of next lecture.