

Applied Multivariate Analysis

Prof. Amit Mitra

Prof. Sharmishtha Mitra

Department of Mathematics and Statistics

Indian Institute of Technology, Kanpur

Prologue Lecture

Applied Multivariate Analysis

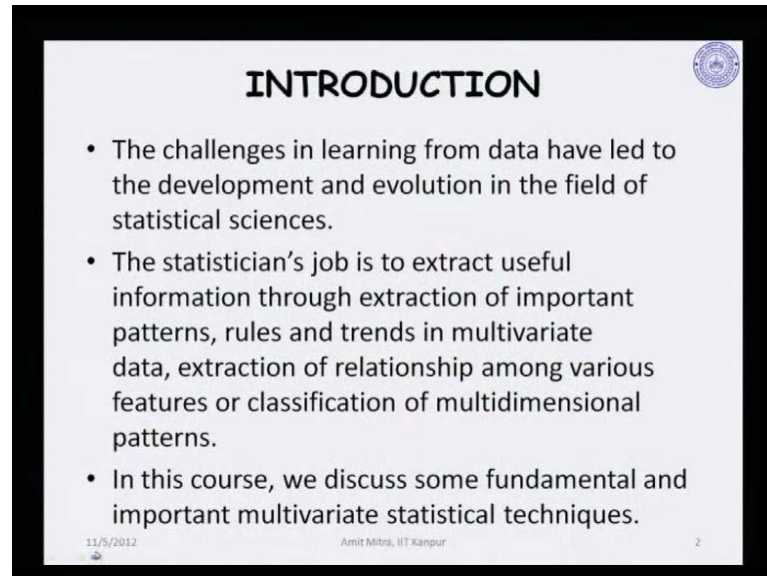
Hello and welcome to this course on applied multivariate analysis. I am Amit Mitra from the department of mathematics and statistics IIT Kanpur. I along with my colleague doctor Sharmishtha Mitra will take you through this particular course on applied multivariate analysis. Now this particular course is divided into two parts. Mainly, in the first part of the course, we will look at basic multivariate distribution theory and look at generalization of various univariate distributions. Say for example, univariate normal distribution, chi square distribution and t distribution through the multivariate counterparts.

Specifically we will look at multivariate normal distribution, its characterization its definition through the Cramer weld's theorem and all. And also look at various important properties of multivariate normal distribution. We will also look at concepts of random sampling from multivariate normal distribution. And, look at various influential issues associated with that. For example, what would be the corresponding sufficient statistic of the unknown set of parameters concerning a multivariate normal distribution? We will talk about estimation of mean vector and covariance matrix of a multivariate normal distribution.

Then, we will look at various derived distributions from multivariate normal distributions. Say for example, we will look at wish art distribution which is multivariate extension of the chi square distribution. We will also look at Hotelling's T square, we will look at all its theoretical justifications various important properties and so on. So, this will comprise roughly the first part of the course wherein the theoretical multivariate distributions, their properties and all will be studied. And in the second part of the course we will look at various important multivariate, applied multivariate statistical techniques.

Their mathematical formulation concepts associated with that and also look at various types of data analysis concerning that.

(Refer Slide Time: 02:27)



INTRODUCTION

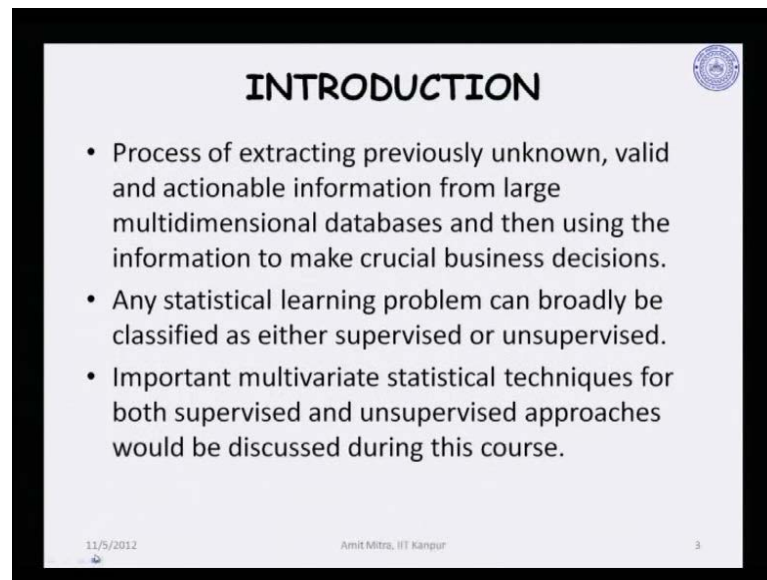
- The challenges in learning from data have led to the development and evolution in the field of statistical sciences.
- The statistician's job is to extract useful information through extraction of important patterns, rules and trends in multivariate data, extraction of relationship among various features or classification of multidimensional patterns.
- In this course, we discuss some fundamental and important multivariate statistical techniques.

11/5/2012 Amit Mitra, IIT Kanpur 2

So, let us start this particular small presentation here, which will actually take us through to what we are going to cover in this particular course. So, this lecture would essentially act as a prologue to this particular course on applied multivariate analysis. Now, to introduce what type of thing we are up to, it may be noted that the challenges in learning from data, have led to large volume of data essentially have led to development and evolution in the field of statistical science. In a nutshell when one tries to see what type of a job is statistician is trying to do?

A statistician's job is to extract useful information, useful and meaningful information through extraction of important patterns rules. And, trends in the multidimensional data, Extraction of relationship among various features or classification of multidimensional patterns. Now, the type of data that is usually encountered is multidimensional in nature. And, thus a proper theoretical foundation of multivariate statistical techniques is not going to recomplete without looking at the multivariate distribution theory as such. And in this course, we will discuss some fundamental and important multivariate statistical techniques.

(Refer Slide Time: 03:36)



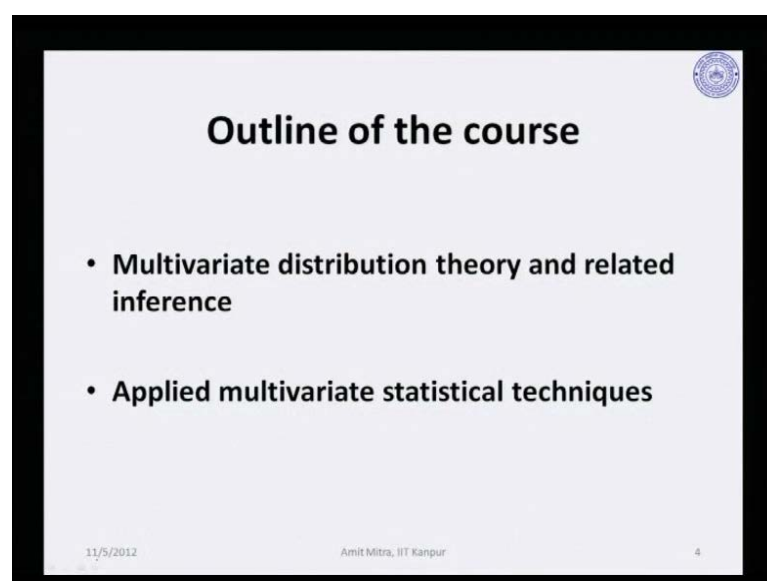
INTRODUCTION

- Process of extracting previously unknown, valid and actionable information from large multidimensional databases and then using the information to make crucial business decisions.
- Any statistical learning problem can broadly be classified as either supervised or unsupervised.
- Important multivariate statistical techniques for both supervised and unsupervised approaches would be discussed during this course.

11/5/2012 Amit Mitra, IIT Kanpur 3

Now, the process of extracting previously unknown valid and actionable information from large multidimensional databases and then, using the information to make crucial business decisions that can actually be performed on two different platforms. One may be on a supervised mode of learning, supervised mode of statistical learning or it can be in an on an unsupervised mode of analysis. In this course specifically, when we talk about applied statistical techniques we will talk both about these supervised techniques, multivariate supervised techniques. And, also a number of unsupervised multivariate data analysis techniques.

(Refer Slide Time: 03:36)



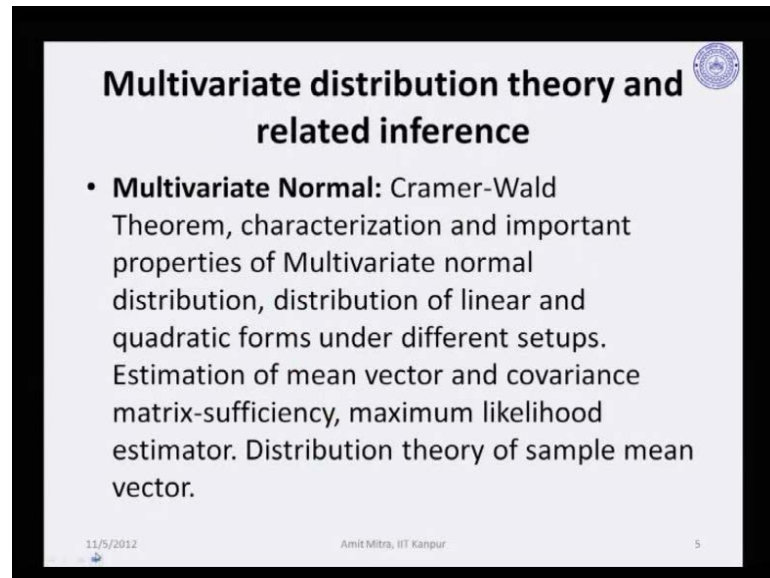
Outline of the course

- **Multivariate distribution theory and related inference**
- **Applied multivariate statistical techniques**

11/5/2012 Amit Mitra, IIT Kanpur 4

Now this is, as I told you in the very beginning that this is a brief outline of the course. In the first part of the course we will look at this multivariate distribution theory and related inference and associated with such multivariate distributions. And in the second part of the course well, it will comprise of applied multivariate statistical technique and also look at various types of data actual data analysis.

(Refer Slide Time: 04:44)



Multivariate distribution theory and related inference

- **Multivariate Normal:** Cramer-Wald Theorem, characterization and important properties of Multivariate normal distribution, distribution of linear and quadratic forms under different setups. Estimation of mean vector and covariance matrix-sufficiency, maximum likelihood estimator. Distribution theory of sample mean vector.

11/5/2012 Amit Mitra, IIT Kanpur 5

Now, what type of multivariate distribution theory are we going to look at? We will first consider a very general framework of looking at multivariate random vector quantities which actually are going to characterize such multivariate vectors. And then introduce multivariate normal distribution. Now while looking at multivariate normal distribution we will first start with a Cramer weld's theorem, which actually tries to look at a multivariate distribution. And, it says that a multivariate distribution is going to be completely known if and only if the distributions of all its linear combinations are known to us.

And that is the basis with which one actually defines a multivariate normal distribution. One defines a multivariate normal distribution through the Cramer weld's theorem as x follows a multivariate normal distribution if and only if every linear combination of x . That is $\alpha'x$, where α belonging to the dimension space associated with that of the multivariate random vector, if we have each and every linear combination following a univariate normal distribution. So, that gives us a characterization of a

multivariate normal distribution the definition we see basically of a multivariate normal distribution. So, we will look at such Cramer Wald's theorem, its statement, its proof will look at characterization of multivariate normal distribution accordingly.

And, then we will look at various important properties of multivariate normal distribution starting from say, deriving the density function, the joint probability density function of a multivariate normal distribution through. Of course, the density of univariate normal distributions is what we are going to look at. We will mainly talk about nonsingular multivariate normal distribution. So, we will look at the covariance matrix to be positive definite. And, hence not look at much although in some of some results that we are going to discuss, will be having certain situations. Wherein sigma may be a singular matrix and in such a situation we might be having a singular multivariate normal distribution.

So, we will derive for a nonsingular sigma covariance matrix, the joint probability density function of multivariate normal distribution. And, keeping in mind that such a density function does not exist, if we have got a singularity in the sigma matrix. We will look at important properties starting from basic properties of multivariate normal distribution, like that of transformation associated with multivariate normal distributions. And also, we will look at a particular type of a section which is important say for example; we will look at quadratic forms derived from multivariate normal distribution. And then, talk about in detail the distributions that one would be getting, when we look at quadratic forms derived from a multivariate normal distribution.

Now, when we talk about quadratic forms derived from multivariate normal distribution it has got huge importance in various fields of applied statistics. For example, if one looks at a linear regression, multiple linear regression model. And then, talks about estimation and hence inference from it say, fundamental theorem of least squares and all. When one tries to actually look at the distribution of residual sum of square, when one is trying to look at distribution of restricted sum of squares, unrestricted sum of squares deriving f statistic. The basis for all such test procedures are based on, distribution of quadratic forms arriving from multivariate normal, normality assumption on the multivariate noise random vector, associated with the multiple linear regression models.

So, it is of fair amount importance actually to look at such distribution of quadratic forms, which will take up in various forms. Now distribution of quadratic form of course, is which comes prior actually to that of quadratic form. So, one looks at a random vector x following a multivariate normal distribution with a mean vector say μ and a covariance matrix σ which is assumed to be positive definite. And then, talks about distribution of $\alpha'x$ which of course, is one such linear combination which has already given characterization to a multivariate normal distribution. And, what would be the distribution of such quadratic forms? And what would be the distribution of such linear forms and also various types of quadratic forms?

Now next we will look at, so, it is up to this particular point that I talked about. And then we will look at the concept of random sampling from a multivariate normal distribution. Now, when we talk about univariate normal distribution, say many population, say are characterized by assuming that the underline population is univariate normal. And then, we know that in the mean and the variance of a normal distribution characterizes normal distribution completely, in case of a univariate normal distribution. When we talk about a multivariate normal distribution, the quantities which completely characterize a multivariate normal distribution are its mean vector.

The p dimensional mean vector, corresponding to a p variate multivariate normal random vector and it is p by p covariance matrix, which is assumed to be positive definite. So, since these are the two quantities of primary importance, if one assumes a multivariate normality on a certain population. And then, tries to find out the parameters or the parameter vector and the parameter matrix and such associated with such a multivariate normal distribution, one talks about random sampling form such a multivariate normal population. So, typically if a multivariate normal distribution, a multivariate population is characterized by a multivariate normal distribution.

Then, we will draw random samples random vectors in this particular situation as x_1 vector, x_2 vector, x_n vector drawn from such a multivariate normal population. And then, based on these x_1 vector, x_2 vector, x_n vector we will address the problem of an inference of the mean vector μ and the covariance matrix σ . So, we will start from say sufficient statistics and we will say that, what is the set of sufficient statistics jointly sufficient in case of μ and σ both are known? And sufficient statistic if one of

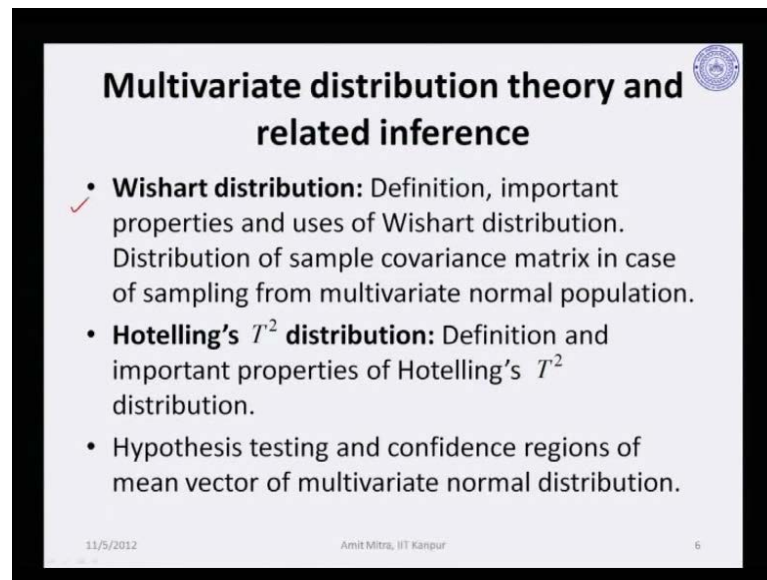
them μ or σ is known to us and talk about the derivation using Neyman Fisher factorization criterion of sufficient statistic.

We will then talk about of course, maximum likelihood method of estimation, associated with a multivariate normal distribution. And also, the distribution theory of the sample mean vector after this particular point. Because the distribution of the sample mean vector, what is going to be the maximum likelihood estimator of the population. Mean vector can be derived using the distribution of linear forms, derived from a multivariate normal distribution. Now, once we are through with that particular multivariate normal distribution theory, random sampling associated with multivariate normal distribution.

We will look at Wishart distribution, which is an important multivariate distribution which is a derived distribution, derived from the multivariate distribution, normal distribution as such. Now it is an extension of the univariate chi square distribution in some sense. Because when one looks at random sampling from an univariate normal distribution. And if the quantity s^2 is one, what one actually tries to look at as an estimator for the unknown σ^2 , associated with that univariate normal population. we know that, $(n-1)s^2/\sigma^2$ follows a chi square distribution or $(n-1)$ degrees of freedom with μ unknown in the univariate normal population.

Now, when we have a multivariate distribution characterized by a multivariate normal distribution with an unknown covariance matrix Σ , we talk about estimation of Σ . And, there we look at as an estimator the sample variance covariance matrix. And the distribution of sample variance covariance with certain constant multipliers, we will derive that such a distribution is that of a Wishart distribution.

(Refer Slide Time: 13:17)



Multivariate distribution theory and related inference

- **Wishart distribution:** Definition, important properties and uses of Wishart distribution. Distribution of sample covariance matrix in case of sampling from multivariate normal population.
- **Hotelling's T^2 distribution:** Definition and important properties of Hotelling's T^2 distribution.
- Hypothesis testing and confidence regions of mean vector of multivariate normal distribution.

11/5/2012 Amit Mitra, IIT Kanpur 6

So, we look at basic definition, how to give the definition of a wish art distribution? In what sense actually the wish art distribution is a distribution, which actually is a generalization of a chi square distribution? So, for p equal to 1 so, to say in such multivariate normal setup we will see that a wish art distribution is what is going to correspond to such a chi square distribution, a central chi square distribution in case of univariate distribution theory. We will talk about important properties and uses of wish art distribution keeping in mind that this wish art distribution is going to play a central role. When we talk about multivariate distribution theory associated with random sampling from a multivariate normal distribution.

So, we will look at distribution of sample covariance matrix in case of sampling from a multivariate normal population. And have the corresponding distribution of the sample variance covariance matrix from such wish art distribution concept. Next in the theory part of this lecture, we will look at what is the Hotelling's T square distribution? We look at the definition and important properties of Hotelling's T square distribution its relationship with univariate t distribution.

So, this is one distribution which is going to be based on multivariate normal and wish art distribution. So, a particular type of combination or a function of multivariate normal and a wish art distribution both of them are recurrent to be independent in that particular setup of framing a Hotelling's T square distribution will be considered. And well one if

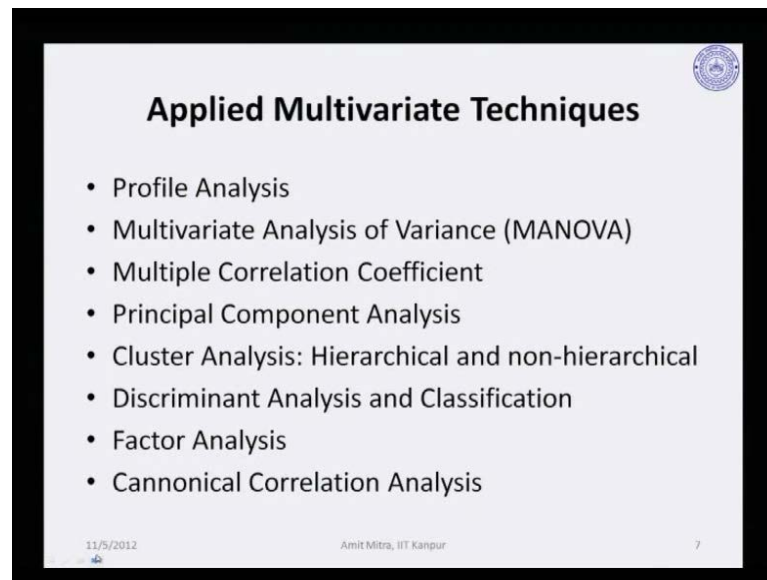
one is looking at a Hotelling's T square distribution the obvious uses of wish art, a Hotelling's T square distribution will be discussed.

We will take up the problem of hypothesis testing, concerning mean vector of a multivariate normal distribution. Remember when we talked about, when we have univariate normal distribution as such and talk about hypothesis testing concerning the mean unknown quantity of that particular of a univariate normal population. We bring in the concept of students t distribution the testing in case μ is unknown σ^2 is unknown is going to be based on that of a univariate students t distribution. Confidence intervals and associated stuff in case of univariate distributions once again are based on students t distribution.

So, when we talk about a multivariate population which is assumed to be having a form of a multivariate normal distribution with an unknown mean vector as μ , we would frequently be interested to actually look at hypothesis testing concerning the mean vector μ . So, we will have say for example, null hypothesis of the form that $\mu = \mu_0$ against $\mu \neq \mu_0$, say as a particular case we can take μ_0 to be equal to a null vector. And, then talk about the framework under which such a hypothesis testing is going to be framed. And, such framing of a testing problem in case of multivariate normal mean vector testing, would be based on a Hotelling's T square statistic.

And, we will see that in detail when we talk about Hotelling's T square distributions. We will also talk about using such concept of Hotelling's T square when we are trying to construct confidence regions. Because we have got mean vector which is p dimensional associated with the dimensionality of the multivariate normal random vector which is assumed to be p dimension. Then, we talk about confidence region concerning that μ vector and setup confidence regions with a confidence coefficient of one minus α . And that would once again make a reuse of Hotelling's T square distribution.

(Refer Slide Time: 17:15)



Now this class would comprise mainly of the theoretical aspects that are going to be covered in this particular post multivariate normal distribution in detail, wish art distribution Hotelling's T square hypothesis testing and other things. Now, the second part of this particular course is going to be having an applied level. So, it would actually try to look at applied multivariate techniques standard applied multivariate statistical techniques, which are of importance. Which would actually give us tools actually a mean order to look at various types of multivariate analysis that may be of interest.

So, in this general we will look at the following applied multivariate techniques. We will look at the techniques of profile analysis; will look at multivariate analysis of variance or the Manova technique. Will briefly look at multiple correlation coefficient as an extension to the bivariate correlation concept, will look at in detail principal component analysis, will look at cluster analysis both hierarchical and non hierarchical modes of cluster analysis. Will look at discriminate analysis and classification, will look at factor analysis and conclude the lecture series with that of canonical correlation analysis.

Now, at the beginning of this lecture I say that we are going to talk about various modes of statistic multivariate statistical analysis. So, various modes would then correspond to different type of approaches say either on a supervised mode or on an unsupervised mode of learning from the data. So, the techniques that I have listed here or we have, we

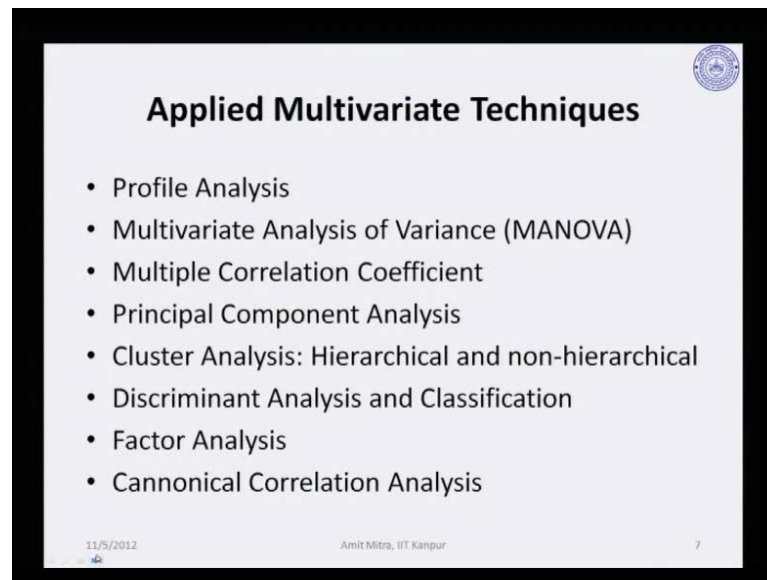
are going to take up in this particular lectures series are going to cover various aspects of such supervised and unsupervised mode of modes of learning.

Say for example, if we look at techniques like principal component analysis a cluster analysis here. They typically would be, comprising of the type of analysis that one usually encounters when one talks about unsupervised mode of learning or trying to extract meaningful information or meaningful pattern from the data. So, it is going to be exploratory in nature, when we talk about such analysis like principal component and that of cluster analysis. On the other hand, if one is trying to look at analysis of the form of discriminate and classification or a canonical correlation analysis or profile analysis or manova technique, then we will have a something in mind.

Then we would actually be looking at building certain type of model or building a framework for hypothesis testing type of problem. And hence we would essentially look at in supervised mode of analysis of multivariate data. Now, specifically when we would consider profile analysis we will look at two or more groups of multivariate populations. And then, look at those two multivariate or two or more multivariate populations being characterized by corresponding unknown mean vectors. And the profile plots that are going to re obtain from such mean vectors. We will look at the sample counterpart of such profile plots and then answer various questions associated with profile analysis.

For example, we would answer questions in sequential order in profile analysis as to the first one whether we have got parallel profiles of the various groups that we are considering. Now if we are having groups, group profiles to be parallel, then we will actually try to look further ahead. And try to see that whether the two parallel profiles if we are able to accept such null hypothesis, against the non null I am sorry with respect to the alternate hypothesis of that of non parallelity of such profile vectors in the population.

(Refer Slide Time: 17:15)



Applied Multivariate Techniques

- Profile Analysis
- Multivariate Analysis of Variance (MANOVA)
- Multiple Correlation Coefficient
- Principal Component Analysis
- Cluster Analysis: Hierarchical and non-hierarchical
- Discriminant Analysis and Classification
- Factor Analysis
- Canonical Correlation Analysis

11/5/2012
Amit Mitra, IIT Kanpur
7

And then, try to look at once null hypotheses of parallelity of profiles are, is accepted. Then we will look further and then try to answer the question whether such parallel profiles are coincident or not. So that we can say that, well in practice in from the data we might see that the profiles are different; however, if the difference in profiles are statistically significant or not. If we are to accept that particular null hypothesis that we have got the two profiles or two or more profiles, when we have more groups we will have more profiles actually.

If we have such peak profiles if we actually go for the null hypothesis that they are coincident profiles. Then the next type of analysis that we would be trying to answer is putting up a null hypothesis of the form, that we will say that the common profile is a level profile. So, there are sequences of such hypothesis testing scenarios that one will be, that will be building up and then looking at with an assumption of multivariate normality in the underlying populations. We will build the framework based on Hotelling's T square statistic once again, in order to do this particular type of analysis which is termed as profile analysis.

Next we will look at this multivariate analysis of variance of the Manova technique. It is a classical technique and it is just an extension of univariate analysis of variance technique while in univariate analysis of variance technique. We talked about, one way analysis of variance two way analysis of variance with various types of assumptions. We

look at corresponding counterparts in the multivariate distribution. And, we look at one way or two way Manova techniques and answer various types of hypothesis testing scenarios associated with such a Manova model.

We will talk about multiple correlation coefficients as I say it is just an extension of bivariate correlation coefficient. So, it is more applicable when we have a group of variables. And then we are trying to actually involve all the variables into one correlation coefficient that is going to be represented as a multiple correlation coefficient. We will talk about principal component analysis which is one important technique when we talk about unsupervised mode of multivariate data analysis. So, this principal component analysis is going to serve doing purpose of projection and visualization of multivariate data.

So, we will look at orthogonal transformation of the original data, what it tries to do is to look at covariance matrix. And then summarize the information that is present in the covariance matrix in terms of a single quantity, which we usually call as a total variation in the triangle vector x . And we will see that we will try to find out from the original set of random vector x will look at a transform set of random variables represented in a vector which would actually be an orthogonal transformation of original set of vector. And, which would be of a special structure which would actually be uncorrelated preserving the total variation that was present in the x vector. And we will also have in such principal component the variances of linear combinations in order of magnitude.

So, that we will be able to say that the first principal component has got the interpretation that it is actually the linear combination of the elements of x . Such that it captures the maximum possible variability that can at all be captured by such linear combinations. Next, we will construct the second principal component which is going to be one, that is going to be uncorrelated with the first principal component. And more so, we will have the second principal component to explain to be one that linear combination which would explain the second maximum variability that can be explained by such linear combinations L prime x .

x being the original random vector such that it is uncorrelated with the first. And like that, a set of principal components p principal components would be constructed starting from the original random vector x . Now, when we have got the principal components,

after transformation from the original set of random vector x through an orthogonal transformation, I said we will have a set of principal components that would be useful in various purposes. We will be able to since we have been able to preserve the total variation in the original set of random variables x_1, x_2, \dots, x_p and we transform the principal components y_1, y_2, \dots, y_p .

We will be able to look at say projection based on the principal components and that projection itself can actually serve the purpose of visualization of multivariate data on a low dimensional visualizable plain say up to the dimension of the of the first three principal components. Now, once such a projection and visualization of multivariate data is obtained through principal components, one can answer various types of questions like detection of outliers from in the multivariate data cloud. One can talk about formation of rough clusters in the data. So, all these type of questions that are going to be answered using principal components are exploratory in nature.

Because there is no supervision as to what is required out of such analysis. It is the data that is going to tell us about what type of say, what type of answers are going to be coming up from such an analysis it is going to be on an unsupervised mode. Now, next we will also consider as I said clustering analysis, classical statistical cluster analysis techniques. In such a genre of course, when one talks about cluster analysis one is trying to look at a group of heterogeneous possibly data. And then from that heterogeneous data one is trying to look at formation of homogenous groups in the object. Now such a cluster analysis can, in its own right can serve as a type of analysis that has to be done prime to any other supervised mode of statistical analysis.

Because if one says is interested about modeling, say modeling of a particular response variable based on a set of independent variables. Then if the group is heterogeneous one would first like to actually divide the heterogeneous group in to that of homogenous sub clusters in the data. And then form a predictive models of responses based on feature vectors on various homogenous groups. So, this may serve as a first step for further analysis or as such can be used in order to find out various important clusters in the data. Now, such statistical cluster analysis mainly of two types.

One is that of hierarchical clustering and the other one is non hierarchical clustering. As the name suggests it is when we talk about hierarchical statistical cluster analysis, we

will have certain hierarchy in which the clusters are going to be formed and are going to be interpreted. When we talk about non hierarchical mode of clustering, in such a situation we will not be having any hierarchy in the formation of the clusters. And, hence the name, non hierarchical clustering. So, we will look at methods for constructing and visualizing such hierarchical and non hierarchical statistical cluster analysis technique.

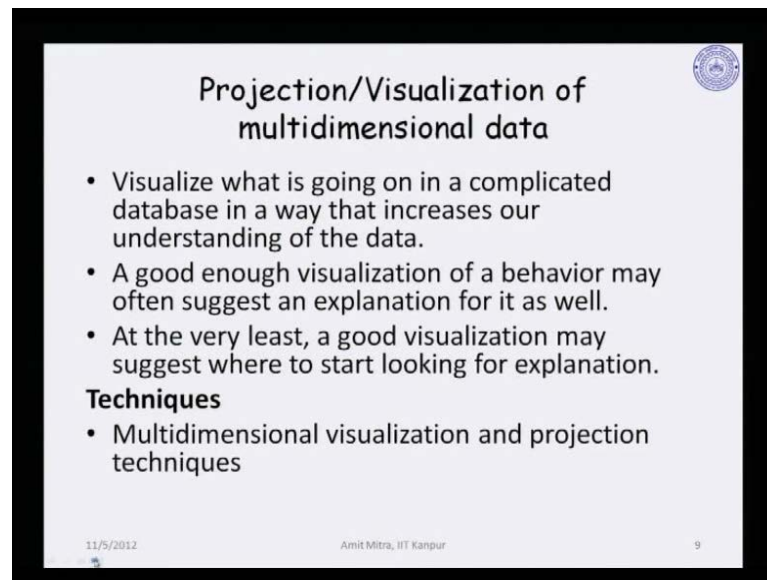
This is also going to be in terms of a non supervised mode of multivariate statistical analysis. We will talk in detail about this discriminate analysis and classification. So, this is one of the classical multivariate statistical analysis types of technique, it talks about a classification of a particular categorical response. And discriminate analysis is going to be based on framing a discriminate function that is going to discriminate between various multivariate populations, in an optimal way that would be defined according to the problem later on.

So, the first task in such an analysis would be to first frame a discriminate analysis. And then once the discriminate analysis which distinguishes distinct populations in an optimal way, will try to use that discriminate function. In order to build up classification models, classification models for looking at various populations. So, suppose we are having c populations as such in the type say c populations, we will try to look at how to use that discriminate function in order to say that his particular feature vector is belonging to a particular class member class.

And hence, prediction of class membership would be of primary importance in such discriminate analysis and classification. We will also look at this factor analysis and canonical correlation analysis which are quite important multivariate statistical technique. So, in the next part of this lecture I will talk about some applied some applications as such of multivariate statistical techniques. As I said is that, this last part of this lecture is going to be comprising of application of some applied multivariate statistical techniques that I have discussed.

And some of these are going to be covered in the lecture series that we are going to start. And some comparisons to some of the methods which are not going to be covered which is beyond the course, beyond the scope of this particular course, will also be talked about in this particular section.

(Refer Slide Time: 30:50)



The slide is titled "Projection/Visualization of multidimensional data" and features a small circular logo in the top right corner. It contains a bulleted list of three points, a section header "Techniques", and a single bullet point under that section. At the bottom, there is a date "11/5/2012", the name "Amit Mitra, IIT Kanpur", and the number "9".

Projection/Visualization of multidimensional data

- Visualize what is going on in a complicated database in a way that increases our understanding of the data.
- A good enough visualization of a behavior may often suggest an explanation for it as well.
- At the very least, a good visualization may suggest where to start looking for explanation.

Techniques

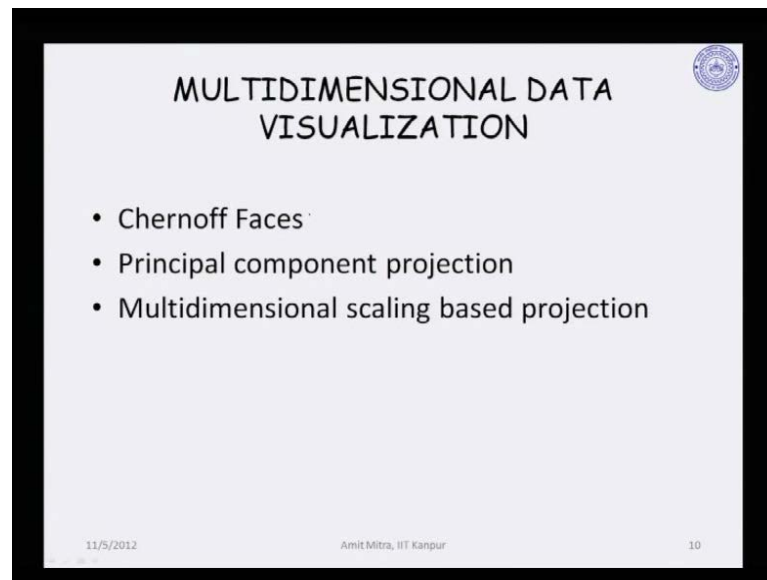
- Multidimensional visualization and projection techniques

11/5/2012 Amit Mitra, IIT Kanpur 9

Now as we said that, when we have got a multivariate data in place the basic thing about basic objective n 1 is looking at such a multivariate data is that we cannot actually visualize such a multivariate data. Unless we do some processing or some analysis of that particular data in order to reduce it to the dimensions up to which we can actually visualize the data. So, this visualization here of multivariate data is an important thing and it is going to be looking at. Suppose we are having a complicated data base and we just want to have a simple understanding of, what is going on in that particular database?

Will look at a simple visualization in order to suggest, what may actually be explanation for further analysis? And at the very least we would say that a good visualization may suggest where to start looking for explanation of such phenomena. Now, the technique that I will just highlight is that multivariate visualization and projection based techniques.

(Refer Slide Time: 31:57)



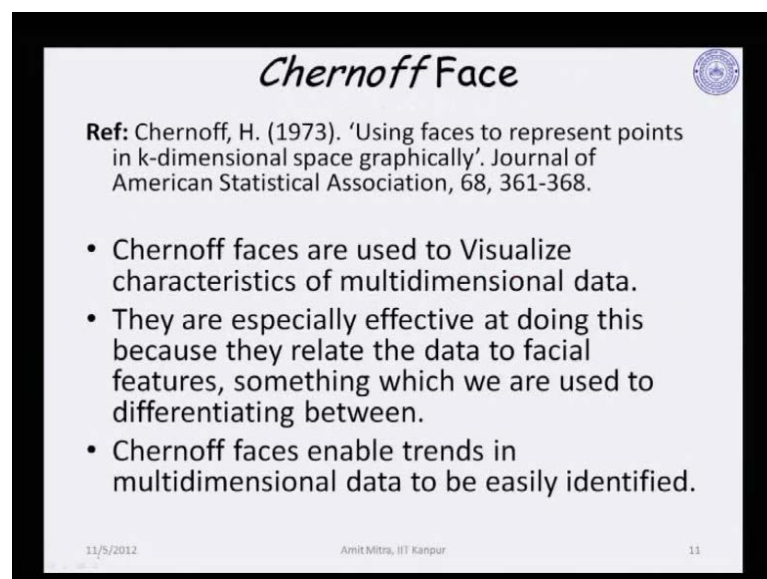
MULTIDIMENSIONAL DATA VISUALIZATION

- Chernoff Faces
- Principal component projection
- Multidimensional scaling based projection

11/5/2012 Amit Mitra, IIT Kanpur 10

So, what sort of visualization as an illustration of multivariate data visualization technique? Look at a standard technique of Chernoff face representation of multivariate data. I will also highlight that of principal component projection which is basically a projection based data and then visualization of the data. Multidimensional scaling based projection also can be a method to look at visualization and projection and hence visualization of the data multi multidimensional data.

(Refer Slide Time: 32:24)



Chernoff Face

Ref: Chernoff, H. (1973). 'Using faces to represent points in k-dimensional space graphically'. Journal of American Statistical Association, 68, 361-368.

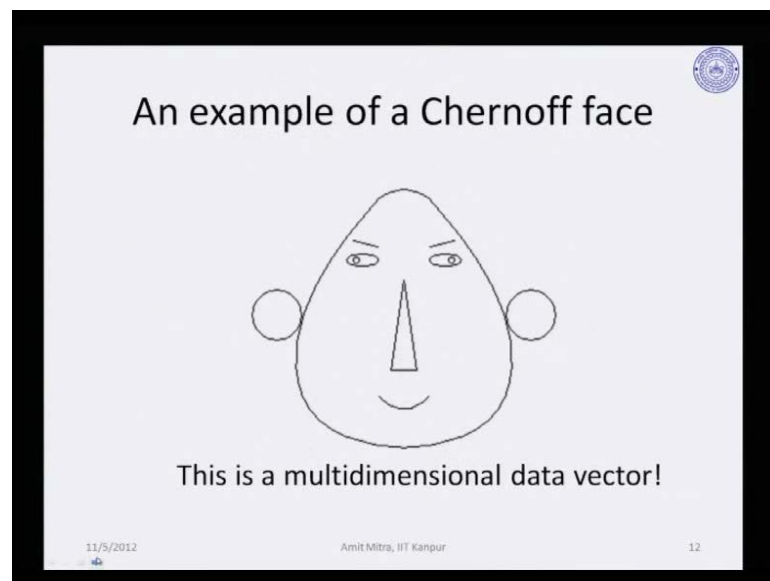
- Chernoff faces are used to Visualize characteristics of multidimensional data.
- They are especially effective at doing this because they relate the data to facial features, something which we are used to differentiating between.
- Chernoff faces enable trends in multidimensional data to be easily identified.

11/5/2012 Amit Mitra, IIT Kanpur 11

Now, let us first look at just briefly, what are Chernoff faces actually? So, the root of this Chernoff face is a method for multidimensional data visualization of course. The root of such Chernoff face multivariate visualization is in a paper in the Journal of the American Statistical Association in way back in 1973 by Chernoff. And what it tries to do is, that it gives us visual representation of multidimensional data. Now they are especially effective at between this multivariate data visualization. Because they relate the data to facial features, something which we are actually used to differentiating between.

So, if we can obtain facial representation of two multivariate data vectors, then we can actually look at what type of whether we can distinguish the two multivariate vectors, through the facial representation of the data. Look at some illustrations shortly Chernoff face enables trend in the multidimensional data also to be easily identified. Because they look at facial representations this is how a Chernoff face actually looks like.

(Refer Slide Time: 33:34)



So, if one looks at such a Chernoff face representation as in this particular case. There are various characteristics as such in a facial representation say curvature of the mouth, separation of the eyes. Then we will have this ear positions, then eccentricity of the lower face, eccentricity of the higher upper face. And, so many other features are associated with a particular facial representation of the data.

(Refer Slide Time: 34:07)

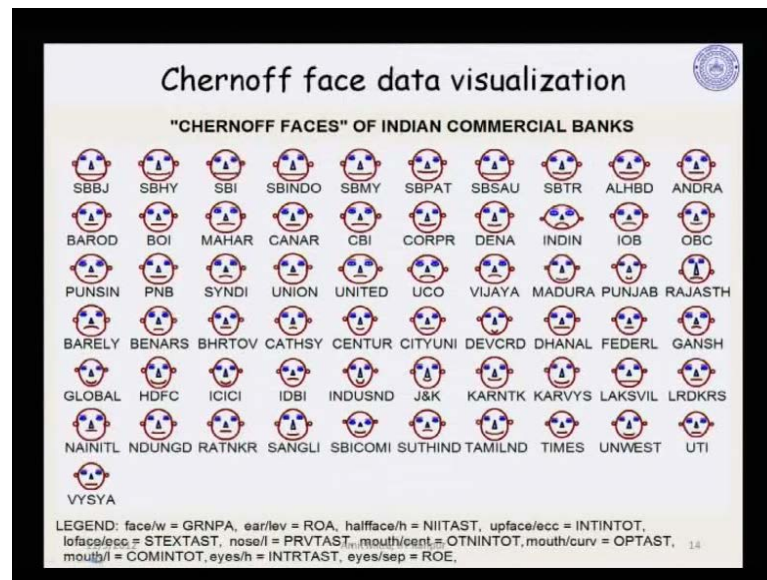
Table 1: Description of facial features of Chernoff face

Dimension	Facial Feature
1	Face width
2	Ear level
3	Half face height
4	Eccentricity of upper ellipse of face
5	Eccentricity of lower ellipse of face
6	Length of nose
7	Position of centre of mouth
8	Curvature of mouth
9	Length of mouth
10	Height of centre of eyes
11	Separation of eyes
12	Slant of eyes
13	Eccentricity of eyes
14	Half length of eye
15	Position of pupil
16	Height of eyebrow
17	Angle of brow
18	Length of brow
19	Radius of ear
20	Nose width

11/5/2012 13

Now, what it is done in the chernoff faces that the dimensions of the data are associated to facial features. Like face width ear level, face half-length, eccentricity of the upper ellipse of the face, eccentricity of the lower ellipse of the face, length of the nose and so on curvature of the mouth. So, in the original formulation of the chernoff face and as extends now, we will have 20 facial features that can be captured in such a facial representation. So, what is done? When we are looking at multivariate data visualization is that from a random vector, x what we have is say a p dimensional random vector $x \in \mathbb{R}^p$. And, then associate each of these variables to important features in the chernoff face representation as in this particular table here.

(Refer Slide Time: 35:11)



And then a multivariate data can be easily visualized through such facial features. Let us look at an actual real life data and see how it actually looks like. This is chernoff face faces of Indian commercial banks for a particular financial year. Now the variables in the data which comprise of the feature vector are these. That it is cross non performing asset, return on assets, net interest income to total assets and so on. So, these are all the features which are characterizing that multidimensional vector and we are trying to have facial representation or a visual representation of such multivariate data.

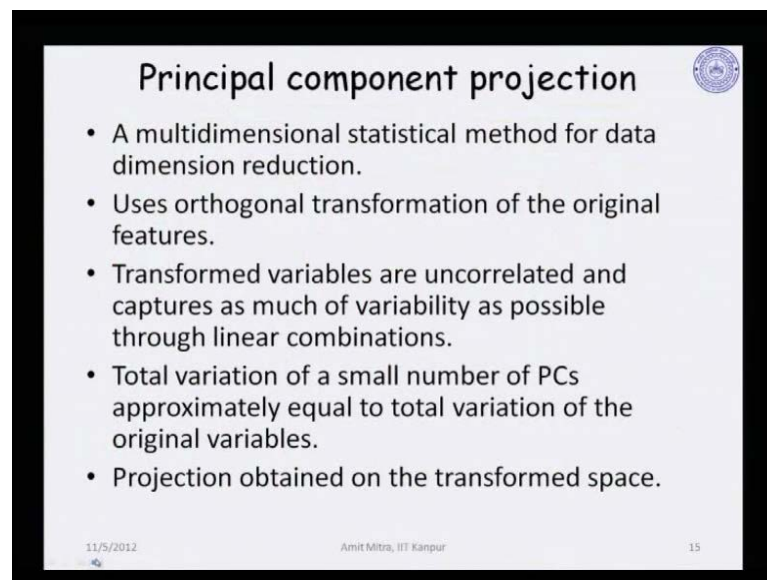
So, in the present case there are 4 5 6 7 8 9 10 and 11 dimensional feature vector. So, for that eleven dimensional feature vector we associate eleven important features from this particular table to that of this multivariate data. And, the features which are associated are highlighted here; say face width here is associated with the gross non performing asset and so on. And, so all these are mentioned here say for example, this mouth curvature is associated with a variable which is actually going to tell us about the health of that particular commercial bank, which is the operating profit to total assets ratio.

And this is what we get as the representation or the visualization of the multivariate data. Say for example, if you look at understanding this multivariate visualized, multivariate data visualization technique. You will be able to say that just by looking at, say facial representation of this data here. If one looks at comparing this multivariate data which is represented by this particular face and this. The differences between the two are going to

be quite obvious, whether we are looking at the multivariate data or its facial representation or the visualization of that multivariate data.

One can easily say that, this does not look that happy a face. And, hence the financial status of this commercial bank at that particular financial year was not in a very happy state as the facial representation says here. On the other hand if one is looking at this particular facial representation of another commercial bank, one would say that this looks pretty happy as such as the financial status of these institutions, like the ones here are quite healthy. Actually they are, they have a smiling face as to that of other type of distorted faces, like that of this institution or this institution or some other institutions.

(Refer Slide Time: 37:48)



Principal component projection

- A multidimensional statistical method for data dimension reduction.
- Uses orthogonal transformation of the original features.
- Transformed variables are uncorrelated and captures as much of variability as possible through linear combinations.
- Total variation of a small number of PCs approximately equal to total variation of the original variables.
- Projection obtained on the transformed space.

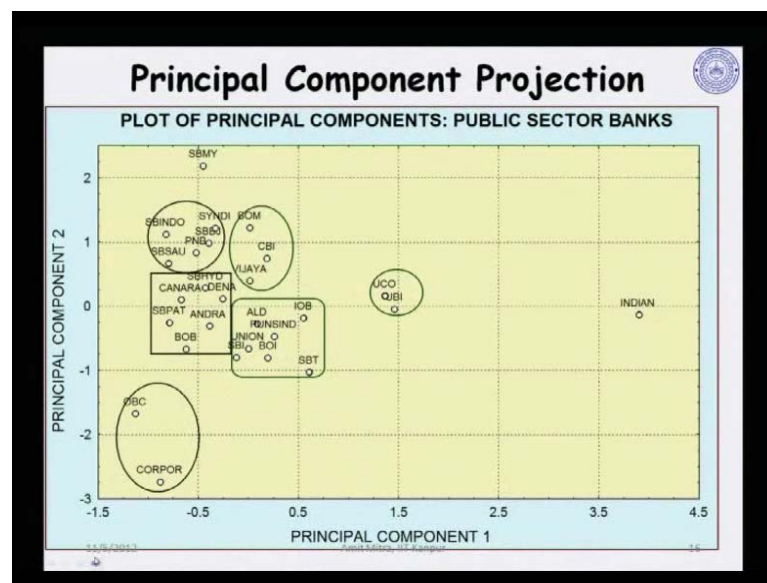
11/5/2012 Amit Mitra, IIT Kanpur 15

So, it is quite appealing actually to look at such exotic multidimensional data visualization techniques. Next we look at a classical approach of looking at principal component projection based method. Now I have discussed already about this principal component method in actual. We are going to look at the theory part of it and its various important properties in this in the course of this lecture. But what it is? What I have said is uses on orthogonal transformation of the original features. And, then we will be looking at transformed variables which are actually uncorrelated and captures as much of variability as possible through linear combinations.

And we are going to try to look at the total variation of a smaller number of principal components. Smaller than that of the original dimension of the random variables,

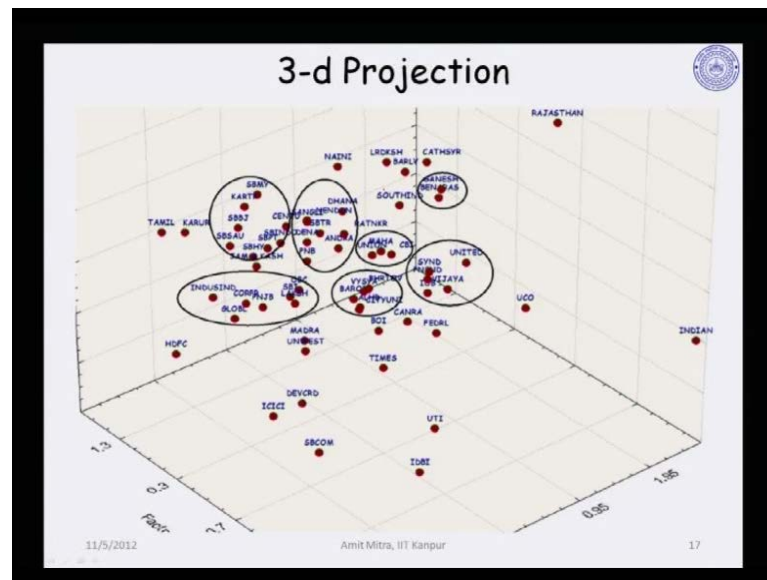
original dimension of the random vector. It is approximately if that is approximately equal to the total variation of the original variables then use of such principal components is going to be most effective. So, when we look at such a scenario, that the total variation of a smaller number of principal components approximately equals to that of the total variation in the original variable. Then the data dimension reduction would be possible through principal components.

(Refer Slide Time: 39:17)



And if that is the case, then we will be able to have effective projection of the multidimensional data to that of the projection, which is in terms of the principal components plain. Let us look at the same type of data that we are used for the chernoff face representation. This is what we get if we look at principal component projection of the financial data that we were looking at its corresponding to the feature vector, which is giving us the financial health of various public sector banks in a particular financial year. And, this is what it looks like on a projected principal component plain. Now, this is a two dimensional projection of such a data of that multidimensional data.

(Refer Slide Time: 39:46)



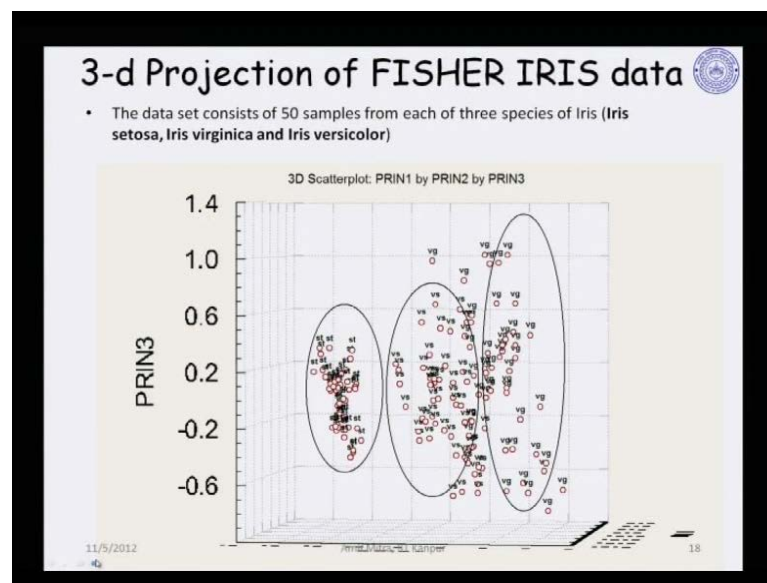
Now, corresponding to the same data we can extend it to the three dimensions. And, if on a two dimensional projection, we were looking at the first two principal components. Because we know that the principal components capture variation in the order in which they actually are constructed. So, the first principal component is going to capture the maximum variability that can be captured by a single linear combination. The first two principal components taken together is going to capture the maximum possible variability. That can be captured by two principal components, two such linear combinations one would say.

And, hence the two dimension projection would naturally be based on the first two dimensions of the principal component plain. And, one can look at the three dimensional projection which is going to be the projection on the first three principal component plain. So, the axes here are the first second and the third principal components, corresponding to each of these multidimensional data. And, each of these multidimensional points are projected in terms of a point, in this three dimensional principal component plain. Now, once we have a, such a projection obtained, I said that it is a way of looking at an exploratory type of data analysis or an unsupervised mode of learning from the data.

Now once we have such a projection many things would come out. Say for example, one would be able to detect multidimensional outliers from such a projection like the

ones that have been, that have come out actually from this particular projection analysis. One would be able to detect rough clusters in the data. Say for example, one has a rough cluster. Now these are subjective clusters these are not clusters that are actually formed. These are subjective clusters, which one actually tries to see when one is looking at such the data in a projective plain. So, one can talk about such cluster formations and outliers been detected from such a projected data.

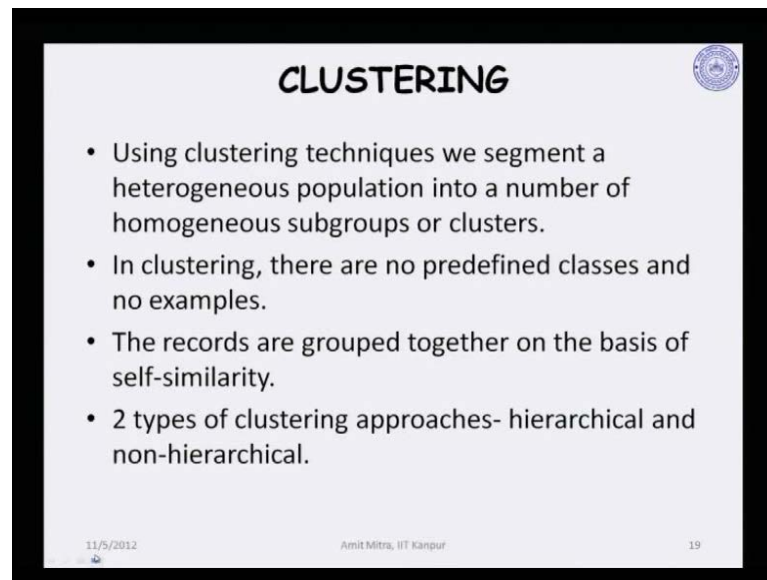
(Refer Slide Time: 41:47)



Now this is another projection that I have included here. This is for a standard data set which is fisher iris data, which is a classical example of or other classical data set. As far as classification models are concerned and also that for an exploratory type of model. So, this it consists of 50 samples, a far from each of the three species of iris flower and the three species are these three here. We use a principal component analysis based projection to project the data along with their class levels, this v g s corresponding to this virginica.

v s is corresponding to this versicolor and this s t, is corresponding to this say setosa species of iris flower. So, the multidimensional data along with their identifications are projected on this three dimensional principal component plain. And, as one can see that after the projection is obtained, these are the set of one species that is clearly getting separated from that particular projection that is obtained on this principal component plain.

(Refer Slide Time: 42:58)



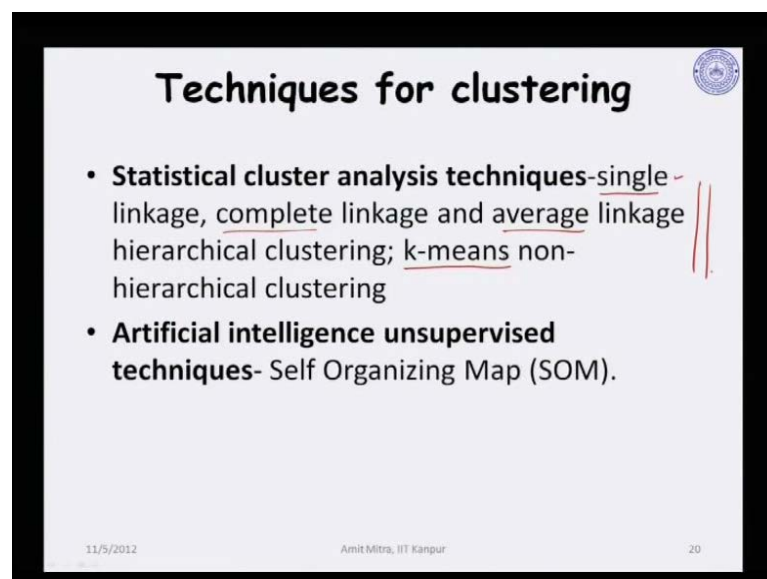
CLUSTERING

- Using clustering techniques we segment a heterogeneous population into a number of homogeneous subgroups or clusters.
- In clustering, there are no predefined classes and no examples.
- The records are grouped together on the basis of self-similarity.
- 2 types of clustering approaches- hierarchical and non-hierarchical.

11/5/2012 Amit Mitra, IIT Kanpur 19

So, it gives a meaningful projection of such a data. Now you talk about clustering. So, I have already given some introduction about the type of clustering techniques that we are going to discuss in this particular lecture. Now in clustering as we, as I said that we would segment a heterogeneous population into that of homogeneous subgroups or clusters. Now in clustering, since it is a type of unsupervised mode of learning, there are no predefined classes or and no predefined, pre classified examples as such. And, one tries to look at the objects and then try to group them, on the basis of self similarity.

(Refer Slide Time: 43:41)



Techniques for clustering

- **Statistical cluster analysis techniques**-single linkage, complete linkage and average linkage hierarchical clustering; k-means non-hierarchical clustering
- **Artificial intelligence unsupervised techniques**- Self Organizing Map (SOM).

11/5/2012 Amit Mitra, IIT Kanpur 20

And, the two type of clustering methods one hierarchical and the non hierarchical methods are what is going to be discussed. I will list here that statistical cluster analysis techniques are the ones that are, that we are going to discuss. Later on these are the type of methods which are single linkage, complete linkage, average linkage, hierarchical clustering all these are comprising of hierarchical clustering based methods. Now the way that these hierarchical clustering methods are going to define, differ rather the single, complete, average are the ways in which, one is actually going to define the distance between groups of objects.

So, when we talk about such cluster formations in the data we will have to introduce a concept of distance between two groups of objects. And then, when we look at various paradigms say look at a single linkage. It talks about quantifying the distance between two groups of objects in a way that, it actually looks at the minimum distance between two groups of objects. So, we look at all possible distances between objects taken from two different groups. And then, try to find out what is the minimum or the shortest distance between objects taken from two different groups that would lead us to single linkage clustering.

On the other hand, if one looks at complete linkage hierarchical clustering one is once again going to look at quantifying the distance between two groups of objects. But in such a situation, one is going to define the distance between two groups of objects as that which is going to be based on the maximum distance. That is possible when one takes one object from one group and another object from the other group. Average linkage is going to look at the average distance of all possible distances between these two groups of objects.

We will also talk about as in, as a method of non hierarchical clustering. The classical k means clustering method and illustrates these cluster formations using real life data. Now it may be a, its worth mentioning at this point that although we are going to talk mainly about statistical cluster analysis techniques in this course. There are other types of methods that are equally or at times better equipped to handle various, other various types of data. Which is, one such method is that of artificial intelligence unsupervised techniques like that of a self organizing map technique.

(Refer Slide Time: 46:03)

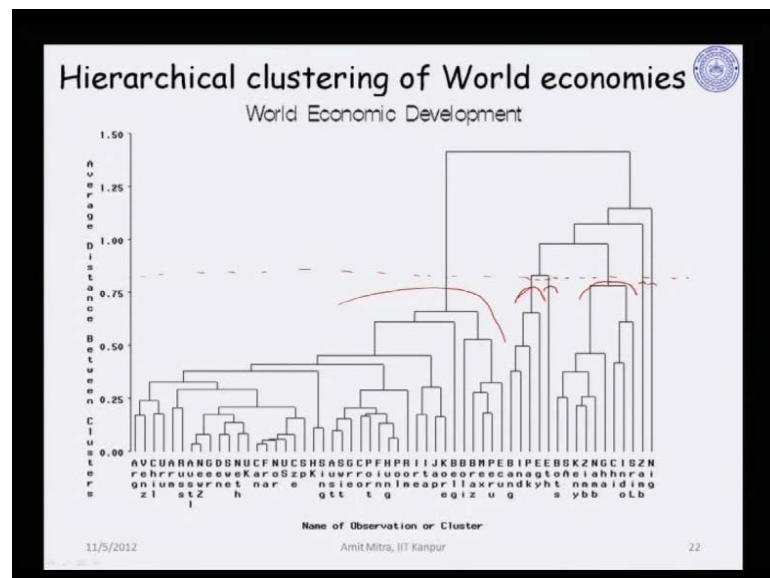
Some Real Life Problems

- Segmentation of customer base: what kind of promotion do customers respond to best; segment first the customer base with similar buying habits
- Identification & influence of financial variables which are causes of low profitability/high non-performing assets in various different segments
- Identification of factors influencing economic growth in different type of economies.

11/5/2012 Amit Mitra, IIT Kanpur 21

Let us look at in this lecture, the type of clustering that we are going to get, these are just to motivate. These are some real life clustering problems say; segmentation of customer base customer segmentation is one of the very basic and important cluster analysis illustrating examples. We can look at such all other examples identification of factors, influencing economic growth of different type of economies. Then one tries to look at other type of problems practical problems.

(Refer Slide Time: 46:33)



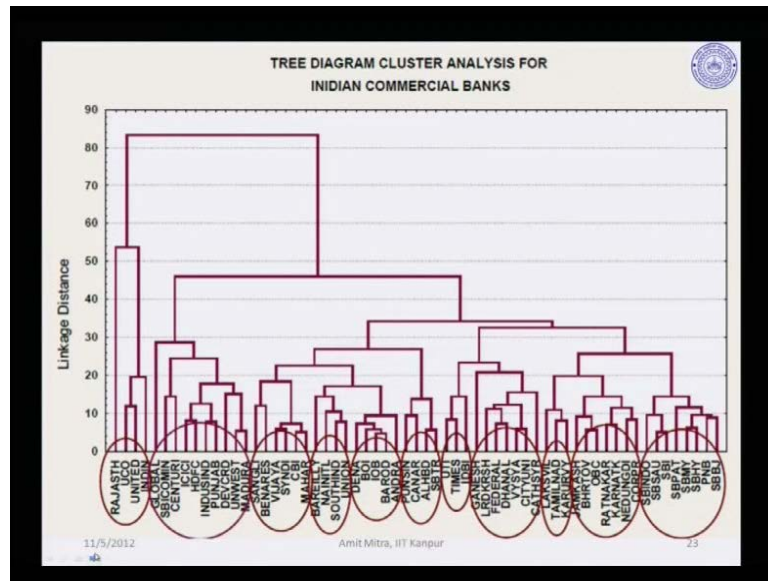
This is what an illustration of a hierarchical clustering technique is. The data is what is corresponding to world economies and each of the world economies that is the countries actually are characterized by multidimensional feature vector. Now those multidimensional feature vectors are actually going to capture, various aspects of economic growth and stability of a particular economy. And, from such a multidimensional feature vector, we obtain this hierarchical clustering of the world economies. And, this is how this icicle plot of hierarchical clustering is going to look like.

So, the interpretation would be that if it depends on the level of resolution at which one is actually trying to look at the cluster formations. So, if one looks at say for example, this particular level of resolution. One will say that all the clusters that are formed below this particular line or the icicles that fall below this particular line are going to be part of different clusters. Like for example, there is only one line which is actually, this line here rather has got all these members below this particular line.

So, they will actually form a cluster of their own. Then we will have another cluster comprising of all the entries which are going to fall within this particular basket here. All the ones that is a singleton actually is going to fall in one single cluster. And then, if the level of resolution cuts at this particular point, we will have all the cases below this particular line falling in one cluster. And, we will have two singleton clusters coming out from this. So, it is going to throw up 1 2 3 4 5 6 clusters in this particular data.

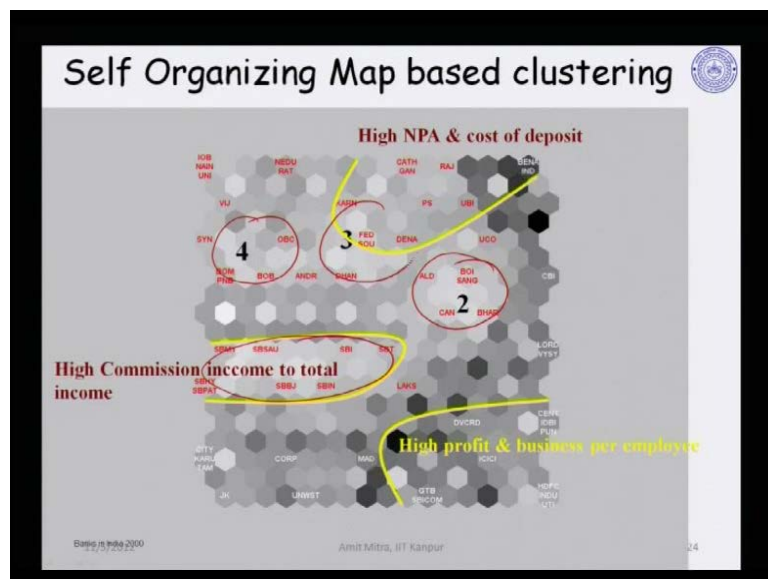
Now, there is a hierarchy in which the clusters are formed actually. Because if we look at this particular level of resolution, we are going to say that this cluster that was there, at this level of resolution is going to be split further into two sub clusters, which are going to be once that are going to hold these elements here. And, there is a hierarchy in which these clusters are going to be joined and finally, going to be placed in one single cluster.

(Refer Slide Time: 48:47)



So, that is how hierarchical clustering is going to be done. This is another illustration of such hierarchical clustering. And, this is for once again for that financial data set that we talked about in chernoff face representation and the principal component visualization. These are once again for those multidimensional feature vectors; these are this is rather the hierarchical cluster formation tree diagram, that we get for such a data.

(Refer Slide Time: 49:14)



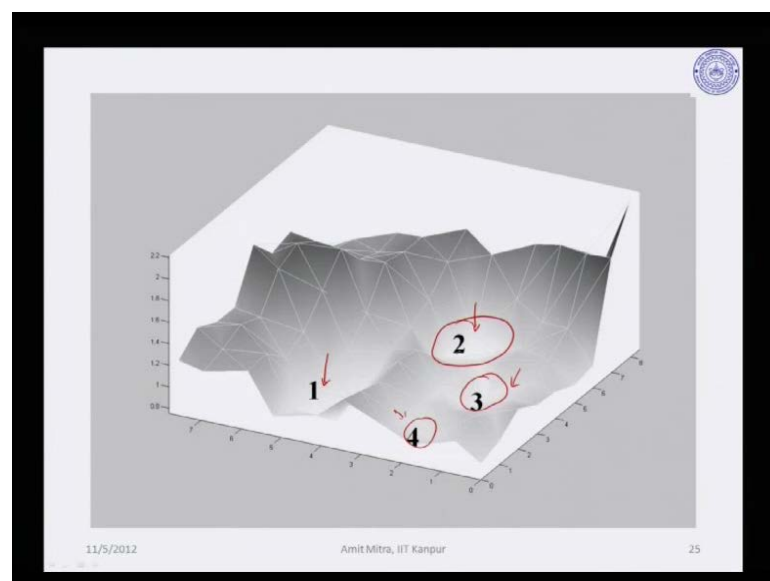
Now, in comparison to that of statistical cluster analysis based techniques, I said that at times artificial intelligence methods give us better visual representation of the data. Now

this is just to illustrate that particular fact that, this is a self organizing map based clustering of the same data as what we considered just now that financial data. This is what is hierarchical clustering of that data and this is what is the self organizing map representation of this data. And from here clusters in the data cloud, multidimensional data cloud can also be detected.

It is to be interpreted in the way that, when we have a such a two dimensional self organizing map representation of the data, we will have the clusters formed in light shaded patches in this particular two dimensional hexagonal grid map here. And hence, the clusters that are rough clusters actually, that are formed are the cases which fall in this particular category. There is a light shade patch present here, there is a light shade patch which is present here, a light shade patch which is present here. These would be the positions of the cluster formations. And, the inter cluster distances between such formed clusters will also be apparent.

If one looks at the shades of separating or rather the shades of hexagons which actually separate these two clusters, it is interesting to look at a three dimensional representation of such a self organizing map. And this is what, is the hill valley surface representation, of this self organizing map representation of this multidimensional data.

(Refer Slide Time: 50:44)

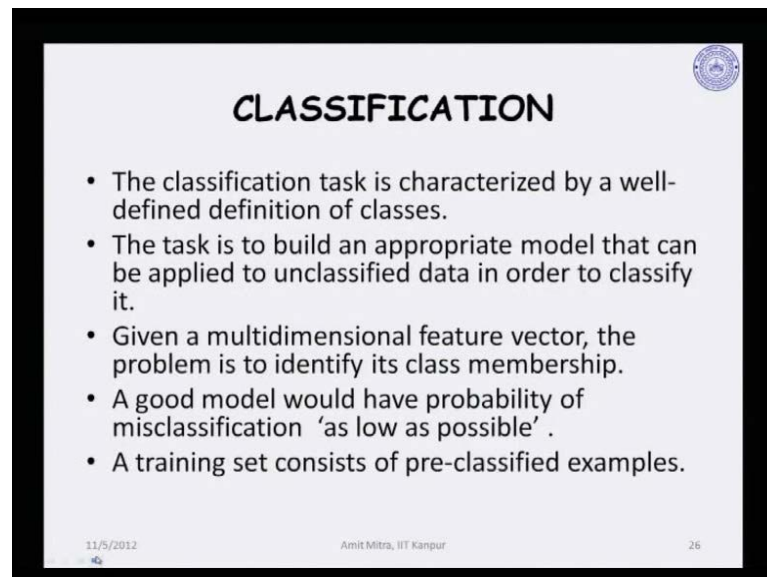


As one actually sees that this is, one location here where there is a formation of a valley, and which has got hills all around. Then that is where actually the clusters are identified

to have formed, there is a second cluster in this depth here of the valley that is formed separated by hills on four sides almost. And, this is two mild the clusters to similar clusters actually, 3 and 4 that are formed on these two locations on this surface here.

Now, these 1 2 3 4 numbers clusters are one, that are corresponding to what the numbers that we have chosen here. There is no number here; this is actually the one the first cluster here which is corresponding to this region in this hill valley plot. If we look at this as the second one here, the corresponding location in this map here is this particular point. So, these are the cluster regions the valleys in this hill valley plot here.

(Refer Slide Time: 51:56)



CLASSIFICATION

- The classification task is characterized by a well-defined definition of classes.
- The task is to build an appropriate model that can be applied to unclassified data in order to classify it.
- Given a multidimensional feature vector, the problem is to identify its class membership.
- A good model would have probability of misclassification 'as low as possible'.
- A training set consists of pre-classified examples.

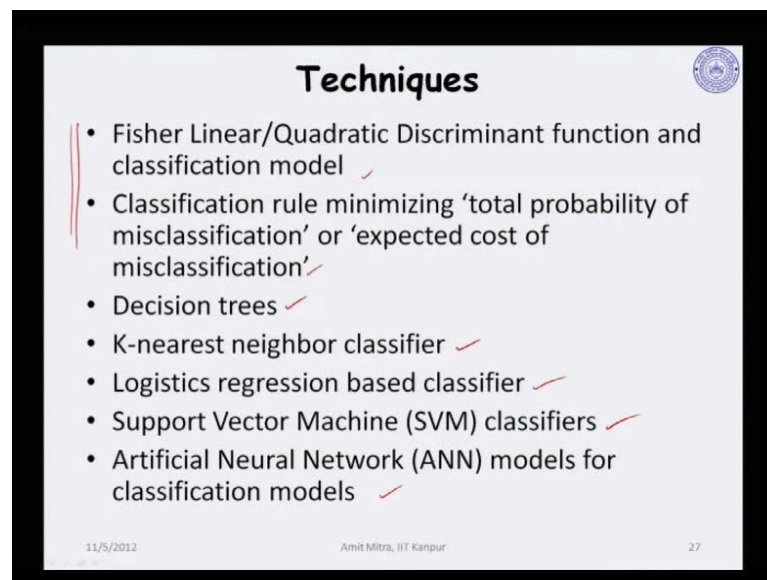
11/5/2012 Amit Mitra, IIT Kanpur 26

Now, classification is the other type of task that we are going to consider in this particular course. It is going to be associated with discriminate analysis and classification task is characterized by a well defined definition of classes. It is going to be different to that of what we have just now discussed, for principal components and that of clustering based methods. Because those two are the type of methods which are non supervised mode of learning exploratory mode of learning. So, this is one method that is on a supervised mode of learning. So, this has got well defined definition of classes and class memberships.

And, we will have a learning sample with pre classified examples. The task would be to build an appropriate model that can be applied to unclassified data in order to classify that particular feature. Now thus, given a multidimensional feature vector the problem is

to identify its class membership, through appropriate statistical technique. Now, since it is a supervised mode of learning we will have to introduce a concept of a good model. A good model would be one that would have probability of misclassification of the patterns as low as possible.

(Refer Slide Time: 53:11)



Techniques

- Fisher Linear/Quadratic Discriminant function and classification model ✓
- Classification rule minimizing 'total probability of misclassification' or 'expected cost of misclassification' ✓
- Decision trees ✓
- K-nearest neighbor classifier ✓
- Logistics regression based classifier ✓
- Support Vector Machine (SVM) classifiers ✓
- Artificial Neural Network (ANN) models for classification models ✓

11/5/2012 Amit Mitra, IIT Kanpur 27

Now, let us look at some of the techniques. Fisher linear discriminate function, quadratic discriminate function based classification models, classification rules which are going to look at a general approach of looking at something which I am going to be define as total probability of misclassification, expected cost of misclassification. Try to look at what type of statistical models would be there in place, when we are trying to look at such paradigms. Now, these are classical methods of discriminate analysis and classification.

Other than that also there are various other types of classification models, like that of decision trees or the classification and decision trees cart, the k nearest neighbor classifier, logistic regression based classifier. Then, we will have support vector machine based classifiers, artificial neural network based methods for classification. Now, in this course we are not going to look at because these are typically covered in courses like that of data mining. So, in this course we are going to look at the classical methods of this classification analysis that of fisher linear discriminate functions. And also that of, this total probability misclassification expected cost of misclassification minimizing classification rules.

(Refer Slide Time: 54:26)

Some Classification Problems

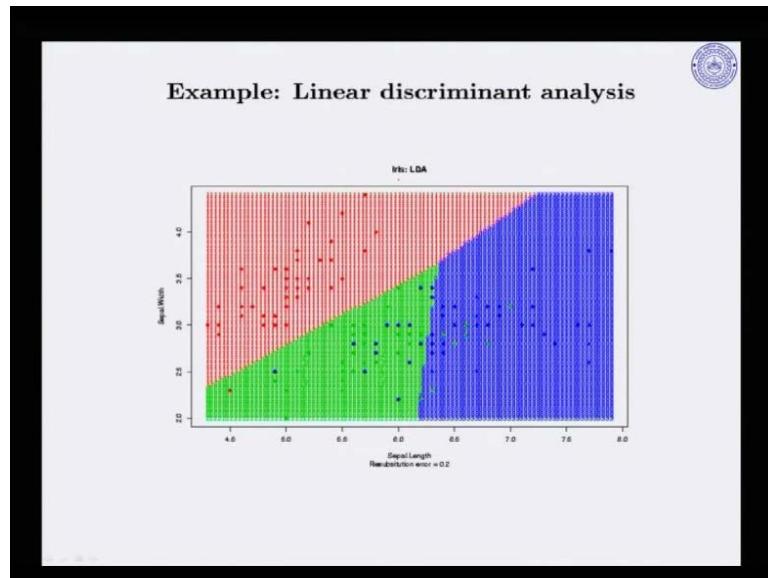
- ✓ Classifying loan applications as low, medium or high risk
- ✓ Classifying candidates applying for a job in a company as future good prospects or otherwise
- ✓ Spotting fraudulent insurance claims $\begin{matrix} \pi_0 \\ \pi_1 \end{matrix}$
- ✓ Identification of correct disease
- ✓ Detection of credit card fraud $\begin{matrix} \pi_0 \\ \pi_1 \end{matrix}$

11/5/2012 Amit Mitra, IIT Kanpur 28

Some practical problems classifying loan applications to a particular financial institution as that of three possible classes, so, this is a three class problem. Then, classifying candidates applying for a job in a company as good prospects or otherwise. So, this is going to be a two class problem. So, based on appropriate features one is going to build a classification model of each of these classification problems, this is another very important and very standard application spotting fraudulent insurance claims.

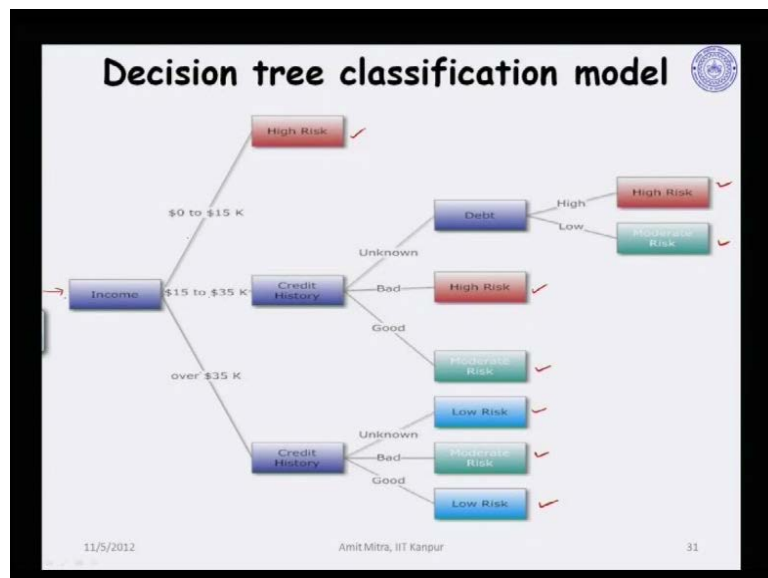
So, it is going to be a two class problem π_0 and π_1 say corresponding to a genuine claim, π_1 to a fraudulent claim. Identification of correct disease, once again it is going to be classification problem. Detection of credit card fraud, once again this is going to be a two class problem, either a genuine transaction or a fraudulent transaction.

(Refer Slide Time: 55:20)



Now, this is how actually this linear discriminate analysis based regions are going to look like. This is for the iris fisher iris data, that we talked about shortly. When we were talking about projection of multidimensional data, this is illustration of linear discriminate analysis and decision boundaries that come, when we have such a three class problem. This is another; this is quadratic discriminate analysis for the same three class problem.

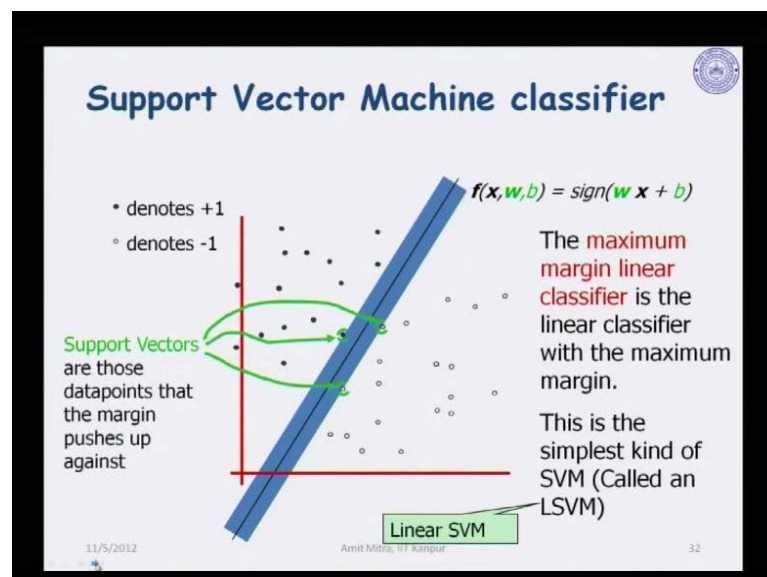
(Refer Slide Time: 55:50)



This is how a decision tree based classification model is going to look like. Now although we are not going to cover this cart models, which I said is typically one that is covered in a course on data mining type of course actually. So, but it is no harm in looking at what is the type of output that one typically gets from decision tree based approach. So, it works in this particular way that, once the input is given at this particular root node of the tree based on certain conditions. And says a feature vector satisfying one of the conditions it takes a particular path in this particular tree.

Now, the nodes of the tree can be branched into two parts, subsequent parts here. Now this part can further be branched into these branches as in this particular case. And then get down to the final root nodes. So, these are going to be what are called as root nodes. And what one looks at is, partition of the feature vector space based on such root nodes. So, these root nodes are going to induce a partition on the feature vector space and then the way that a particular feature vector is going to flow it might land up in a particular roots node. And then, classification of that particular feature vector would be corresponding to the identification level that is associated with such a root node.

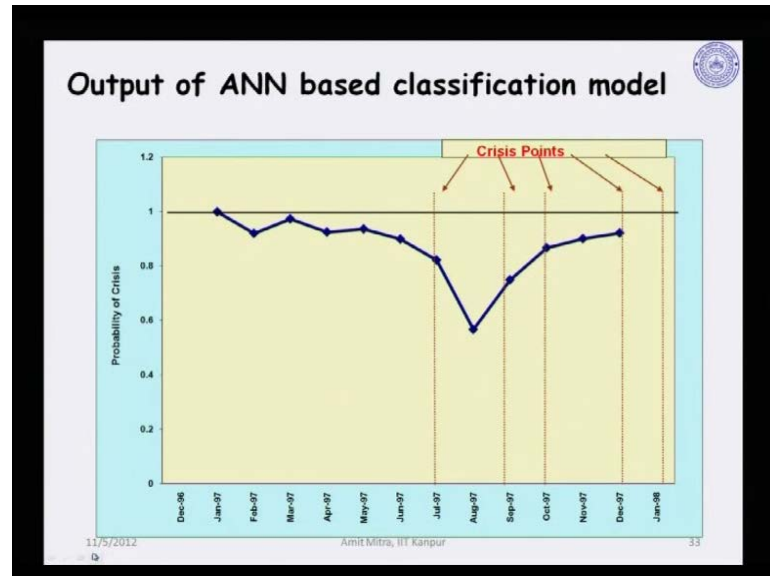
(Refer Slide Time: 57:20)



It is quite an appealing method a graphical way of looking at this particular type of analysis. Now support vector machines based on perception learning approach is, one also quite popular method of building classifiers. So, what it tries to look at is, to look at classes which are linearly separable or otherwise one looks at finding classifier which is

going to be actually maximum margin linear classifier. And that is what is going to lead us to support vector machine classifiers. This is how it looks like.

(Refer Slide Time: 57:48)



This is another example of a classification model. This is an output that would be generated when we look at an ANN based classification model. It actually predicts probabilities of a particular class, here it is a class here of Infosys crisis class. So, it is two class problem crisis or no crisis.

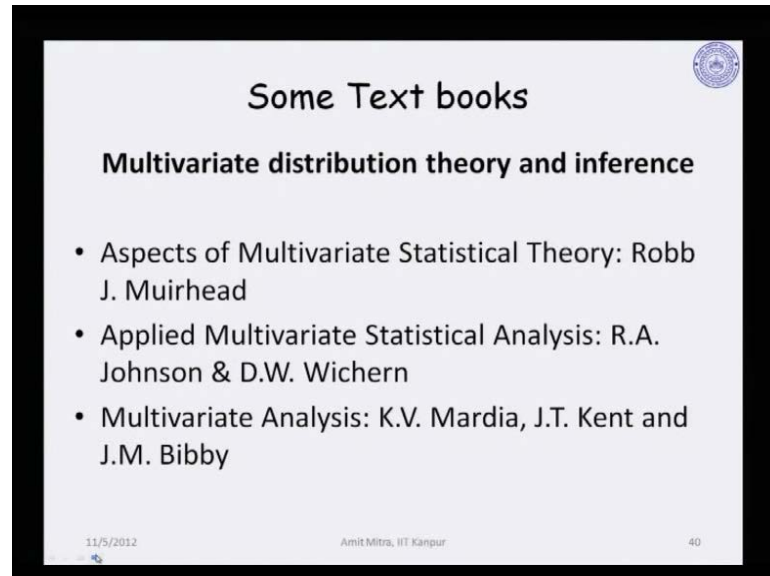
(Refer Slide Time: 58:15)

Multiple Regression Analysis

- A statistical tool for evaluating the relationship of a set of independent variables to a single continuous dependent (response) variable
- Dependent variable: Response variable
- Independent variables: Regressor/explanatory variables-typically multivariate

So, one looks at that as two predicted probabilities coming from an ANN model. Another important application is multiple linear regression models.

(Refer Slide Time: 58:27)



I am not going to talk much about that. I will end this particular lecture with giving you some, say references of some important text books in this particular subject of applied multivariate analysis.

So, as I said that the first section is going to be basically looking at theoretical stuff and multivariate distribution theory. And associated inference three important and very good books as which serve as text books in most of the good universities and institutes across the world, are that of this aspects of multivariate statistical theory, which is by muirhead. That is a very nice book; one can look at a book. This book by Johnson and wincher, applied multivariate statistical analysis and also the book by mardia and (()) the multivariate analysis.

(Refer Slide Time: 59:16)

Some Text books

Applied Multivariate Techniques

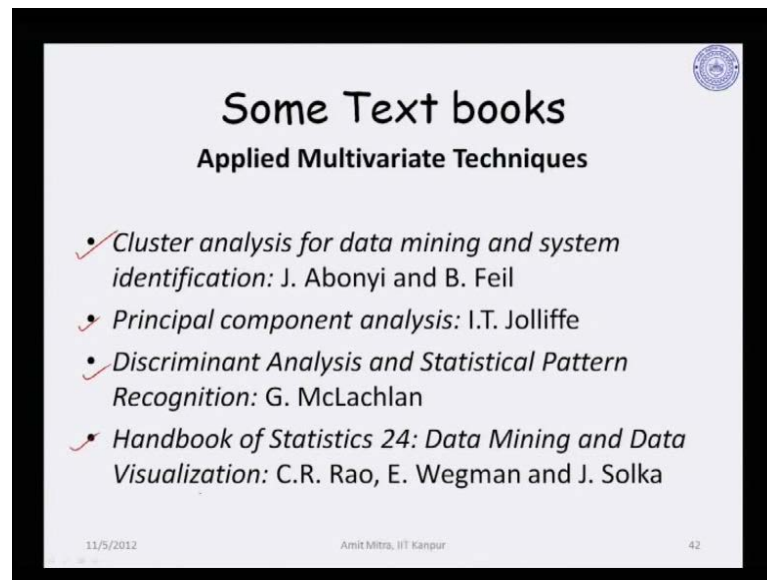
- Applied Multivariate Statistical Analysis: R.A. Johnson & D.W. Wichern
- Applied Multivariate Statistical Analysis: W.K. Hardle and L. Simar
- *Elements of statistical learning: Data mining, inference and prediction*: T. Hastie, R. Tibshirani and J. Friedman

11/5/2012 Amit Mitra, IIT Kanpur 41

These are mainly theoretical books here. When one talks about applied multivariate techniques there is once again whole lot of such books available in the literature. Some of the books which I consider to be good books as text books are listed here. Johnson and wincher's book is also good from an applied multivariate technique point of view. Then we have a nice book by hardle and simar, which is on applied multivariate statistical analysis.

Then, this book by Tibshirani (()) Hastie Tibshirani and Friedman, this elements of statistical learning data mining inference and prediction, although it talks about statistical learning. But it does actually talk about, many such applied multivariate statistical techniques only.

(Refer Slide Time: 01:00)



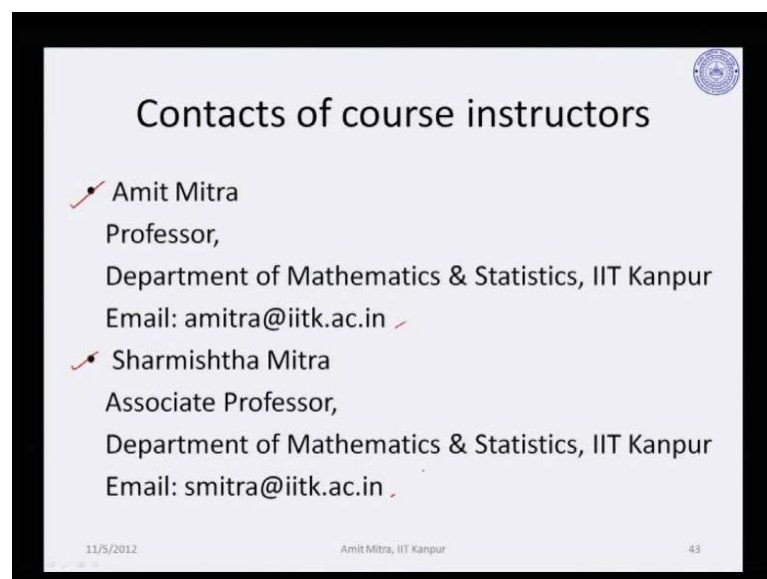
Some Text books
Applied Multivariate Techniques

- ✓ *Cluster analysis for data mining and system identification*: J. Abonyi and B. Feil
- ✓ *Principal component analysis*: I.T. Jolliffe
- ✓ *Discriminant Analysis and Statistical Pattern Recognition*: G. McLachlan
- ✓ *Handbook of Statistics 24: Data Mining and Data Visualization*: C.R. Rao, E. Wegman and J. Solka

11/5/2012 Amit Mitra, IIT Kanpur 42

It is a nice book actually for applied multivariate techniques. And there are other books this book is solely devoted to clustering, cluster analysis for data mining. And system identification, this book is solely devoted on principal component analysis.

(Refer Slide Time: 01:26)



Contacts of course instructors

- ✓ Amit Mitra
Professor,
Department of Mathematics & Statistics, IIT Kanpur
Email: amitra@iitk.ac.in
- ✓ Sharmishtha Mitra
Associate Professor,
Department of Mathematics & Statistics, IIT Kanpur
Email: smitra@iitk.ac.in

11/5/2012 Amit Mitra, IIT Kanpur 43

This is on discriminate analysis and statistical pattern recognition and this is a nice handbook actually, handbook of statistics volume 24 data mining and data visualization by professor c r rao and (()). Now, as I said that I along with my colleague doctor Sharmishtha Mitra is going to take you through the rest of this particular course. Here are

the contacts of the two course instructors. This it is me here and this is my co instructor in this particular course, our email i d s are given here. In case of any queries regarding this particular course, one is free to approach us or with any sorts of problems. Thank you.