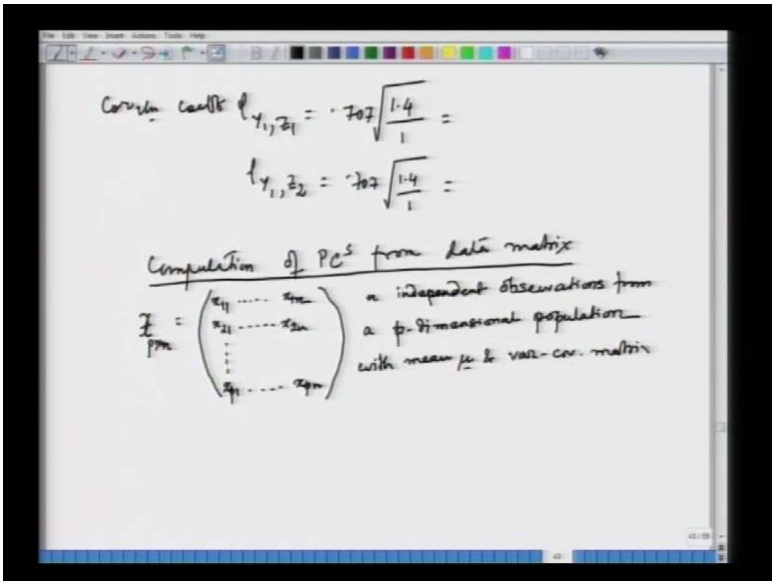
Applied Multivariate Analysis

Prof. Amit Mitra
Prof. Sharmishtha Mitra
Department of Mathematics and Statistics
Indian Institute of Technology, Kanpur

Lecture No. # 25
Principal Component Analysis

The examples that we have considered till now, we saw that the data at our disposal, what the population variance covariance matrix or for that matter population correlation coefficient matrix, that is the variance covariance matrix of the standardized variables. But in most of the situations in reality, what we will have at our disposal is basically the data matrix. We may have P variables with as x 1, x 2 to x p, and we shall have n independent observations on each of these P variables, and from that data from the scratch, we have to calculate the principal components - the sample principal components to do a complete analysis. So, now we take up the computation of principal components from such given data matrix, that is from the given data on a multi-dimensional random vector.

(Refer Slide Time: 01:10)



Corrln coeffs $\rho_{Y_1, Z_1} = \cdot 707 \sqrt{\frac{1\cdot 4}{1}} =$

$\rho_{Y_1, Z_2} = \cdot 707 \sqrt{\frac{1\cdot 4}{1}} =$

Computation of PC^s from data matrix

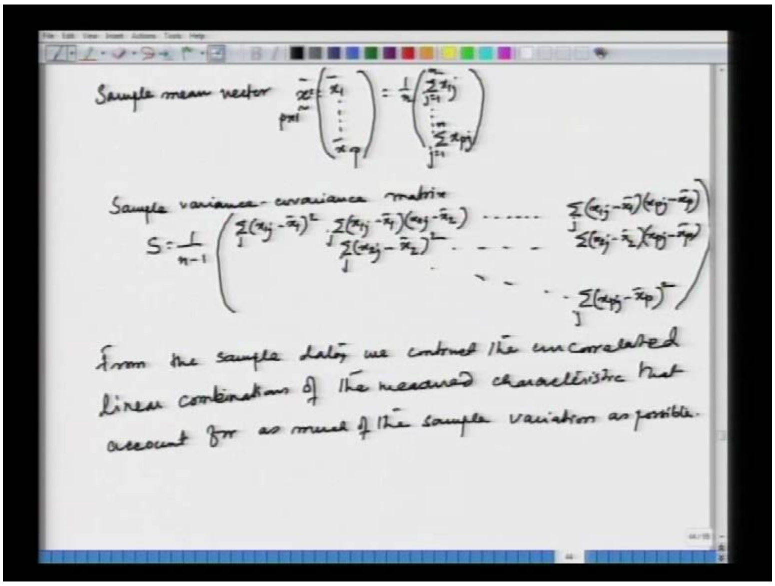
$$\underset{p \times n}{X} = \begin{pmatrix} x_{11} & \cdots & x_{1n} \\ x_{21} & \cdots & x_{2n} \\ \vdots & & \\ x_{p1} & \cdots & x_{pn} \end{pmatrix}$$

n independent observations from a p-dimensional population with mean $\underset{\sim}{\mu}$ & var-cov. matrix

So, that is what we have with us now, that is computation of the principal components from a data matrix. We have at our disposals as we said that we have P variables on which we have on each of which we have n independent observations. So, we prepare the data matrix say script X which is of the dimension P by n which means that we have the P random variables and n observations on each of these. And hence, the elements are arranged in this way x 11 up to x 1n, x 21 to x 2n, and the last row being x P1 to x Pn. So, what we have is the first row is giving the first random vector, the sample of that random vector of the first random variable, and n observations of the first random variable x 1; these are the observations coming from that variable x 1. So, these are x 11 to x 1n , the first subscript 1 is talking me, telling me about the random variable that we have at this stage; that is x 1, and the second subscript 1 to n is tell me about the number of observations that I have on this variable.

Similarly, in the second row, I have the second data and n observations on that variable and P th row is giving me the information on the P th variable. If you consider the first column, it is you have all the values of all the variables x 1, x 2, x P at the first instance at the first case that is the second subscript here is 1. So, we have, what we have is n independent observations is n independent observation that we have arranged in this way from a P-dimensional that is how we have the first dimension of the data matrix, P-dimensional population with mean vector mu and variance matrix - variance-covariance matrix sigma.

(Refer Slide Time: 03:40)



Sample mean vector $\underset{p\times 1}{\bar{\underset{\sim}{x}}} = \begin{pmatrix} \bar{x}_1 \\ \vdots \\ \bar{x}_p \end{pmatrix} = \frac{1}{n}\begin{pmatrix} \sum_{j=1}^{n} x_{1j} \\ \vdots \\ \sum_{j=1}^{n} x_{pj} \end{pmatrix}$

Sample variance-covariance matrix

$$S = \frac{1}{n-1}\begin{pmatrix} \sum_j (x_{1j}-\bar{x}_1)^2 & \sum_j (x_{1j}-\bar{x}_1)(x_{2j}-\bar{x}_2) & \cdots & \sum_j (x_{1j}-\bar{x}_1)(x_{pj}-\bar{x}_p) \\ & \sum_j (x_{2j}-\bar{x}_2)^2 & \cdots & \sum (x_{2j}-\bar{x}_2)(x_{pj}-\bar{x}_p) \\ & & \ddots & \\ & & \cdots & \sum_j (x_{pj}-\bar{x}_p)^2 \end{pmatrix}$$

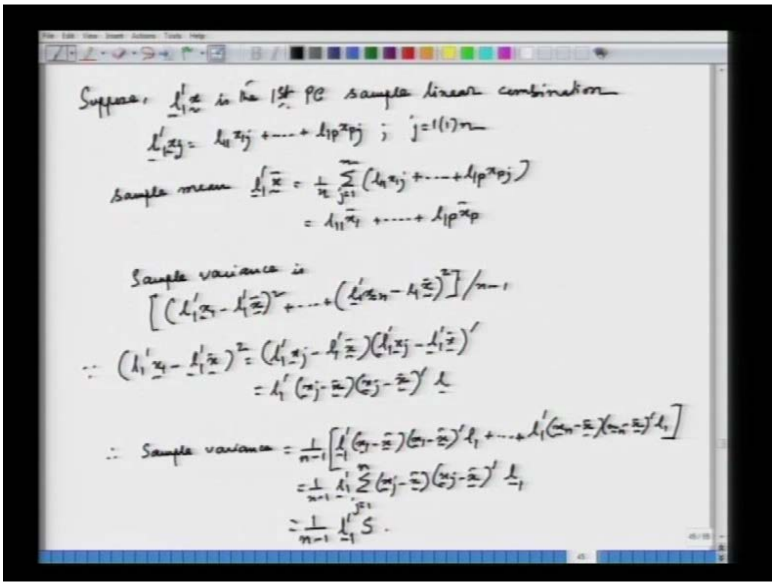
From the sample data we construct the uncorrelated linear combinations of the measured characteristic that account for as much of the sample variation as possible.

From here, what we can have is the sample mean vector. So, let us say the sample mean vector is denoted by x bar and it is nothing but the elements, since x is P-dimensional, the data vector. So, we have this also as P dimensional and we have x 1 to x P; obviously, the first x1 is the sample mean of the first random variable that is, x 1 j from 1 to n that is, the first row elements of the first row from the matrix and similarly, go down to the P (( )) search random observations, j of them j from 1 to n takes some of these and divide by n to get x P j. So, these are not vectors, we have to write them. In fact, we have the mark that is x 1 to x P and these are the sample means from each of the cases.

So, we have an upper bar actually, these are all elements making up the P-dimensional sample mean vector. Similarly, the sample variance covariance matrix have S and this is 1 by n minus 1 and then the first row is nothing but x 1 j minus x 1 bar square and then the covariance between the first and the second; So, consider x 1 j minus x 1 bar x 2 j minus x 2 bar and similarly, the last one is x 1 j minus x 1 bar with x P j minus x P bar. The second the (( )) element is the sample variance for the case of the second variable; that is x 2 j minus x 2 bar square, all these summations are over j from 1 to n and the last one in this row is x 2 j minus x 2 bar with x P j minus x P bar and the P (( )) are the last element which is the variance corresponding to P, the variable; that is x P j minus x P bar whole square.

So, this is the sample variance covariance matrix, I have with divisor n minus 1. From the sample data, our objective we may either have the complete data or we may have concise data like the mean vector and the sample variance covariance matrix whatever; now from the sample data, our objective is to construct the principal components, the uncorrelated linear combinations. Basically these are principal components of the measured characteristic, that account for that is the purpose, that account for as much of the sample variation as possible. So, this is the crux of the calculation of the sample principal components as possible.

(Refer Slide Time: 07:49)



Suppose, $\underset{\sim}{l}_1' \underset{\sim}{x}$ is the 1st PC sample linear combination

$$\underset{\sim}{l}_1' \underset{\sim}{x}_j = l_{11} x_{1j} + \cdots + l_{1p} x_{pj} \; ; \quad j = 1(1)n$$

Sample mean $\underset{\sim}{l}_1' \bar{\underset{\sim}{x}} = \frac{1}{n} \sum_{j=1}^{n} (l_{11} x_{1j} + \cdots + l_{1p} x_{pj})$

$$= l_{11} \bar{x}_1 + \cdots + l_{1p} \bar{x}_p$$

Sample variance is

$$\left[(\underset{\sim}{l}_1' \underset{\sim}{x}_1 - \underset{\sim}{l}_1' \bar{\underset{\sim}{x}})^2 + \cdots + (\underset{\sim}{l}_1' \underset{\sim}{x}_n - \underset{\sim}{l}_1' \bar{\underset{\sim}{x}})^2 \right] / n-1$$

$$\because (\underset{\sim}{l}_1' \underset{\sim}{x}_j - \underset{\sim}{l}_1' \bar{\underset{\sim}{x}})^2 = (\underset{\sim}{l}_1' \underset{\sim}{x}_j - \underset{\sim}{l}_1' \bar{\underset{\sim}{x}})(\underset{\sim}{l}_1' \underset{\sim}{x}_j - \underset{\sim}{l}_1' \bar{\underset{\sim}{x}})'$$

$$= \underset{\sim}{l}_1' (\underset{\sim}{x}_j - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}})' \underset{\sim}{l}$$

$$\therefore \text{Sample variance} = \frac{1}{n-1} \left[\underset{\sim}{l}_1' (\underset{\sim}{x}_1 - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_1 - \bar{\underset{\sim}{x}})' l_1 + \cdots + l_1' (\underset{\sim}{x}_n - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_n - \bar{\underset{\sim}{x}})' l_1 \right]$$

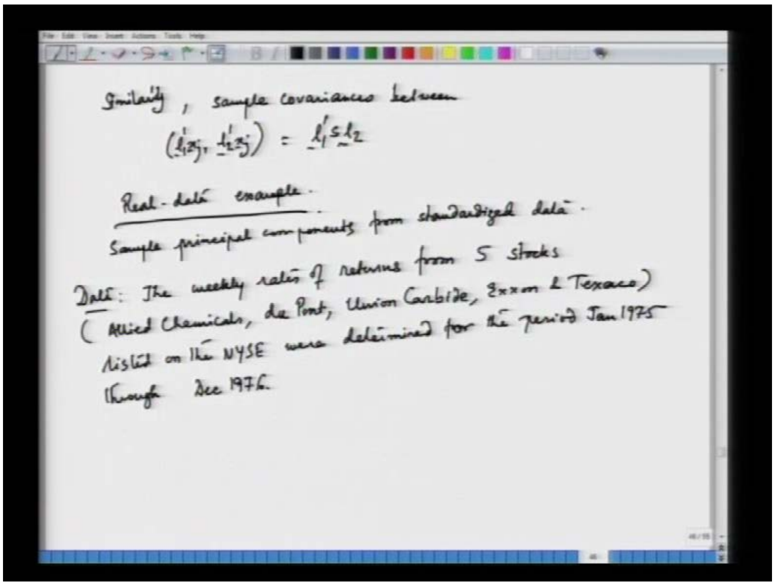
$$= \frac{1}{n-1} \underset{\sim}{l}_1' \sum_{j=1}^{n} (\underset{\sim}{x}_j - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_j - \bar{\underset{\sim}{x}})' \underset{\sim}{l}_1$$

$$= \frac{1}{n-1} \underset{\sim}{l}_1' S .$$

Now, suppose first principal component, l 1 transpose x is the first sample PC or first principal component sample linear combination same thing ; let us say that l 1 transpose x j is nothing but l 11 with x 1 j up to l 1 transpose x j in this manner ; put this is l 1 transpose x j and this is l 1 sum up to l 1 P x P j and this is defined for all j from 1 to n and in that case, the sample mean of this component as l x l 1 transpose x bar, and this is nothing but 1 by n sum of l 1 x 1 j up to l 1P x P j sum over j from 1 to n and this basically the way we have defined. So, these are nothing but l 11 x1 bar up to l 1P x P bar and this l 1 transpose x bar is the notation that we have used and the sample variance is l 1 prime x 1 minus, you have to consider the mean out of it.

So that is, l 1 prime x 1 bar whole square up to l 1 x n minus l1x bar whole square and this is divided by the divisor n minus 1. Also equivalently this l 1 x 1 bar minus l 1 transpose x bar square as l 1 transpose x j minus l 1 transpose x bar along with l 1 transpose x j minus l 1 transpose x bar transpose, and this is nothing but l 1 transpose x j minus x bar, x j minus x bar transpose with l. So, sample variance is also equal to 1 by n minus 1 with l 1 transpose x 1 minus x bar x 1 minus x bar transpose l 1 up to l 1 transpose x n minus x bar x n minus x bar transpose with l 1. This gives 1 by n minus 1 l 1 transpose sum of x j minus x bar x j minus x bar transpose, j from 1 to n and I have one l 1 here. So, this is actually giving the definition of the sample variance covariance matrix, this is nothing but 1 by n minus 1 l 1 transpose S l 1.

(Refer Slide Time: 12:26)



Similarly, sample covariances between

$$(\underset{\sim}{l}_1'\underset{\sim}{x}_j, \underset{\sim}{l}_2'\underset{\sim}{x}_j) = \underset{\sim}{l}_1' S \underset{\sim}{l}_2$$

Real-data example.

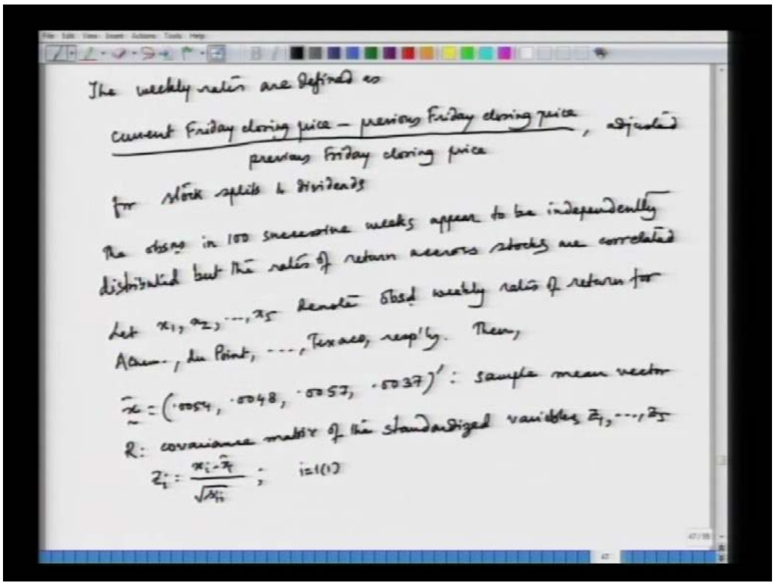
Sample principal components from standardized data.

Data: The weekly rates of returns from 5 stocks
(Allied Chemicals, du Pont, Union Carbide, Exxon & Texaco)
listed on the NYSE were determined for the period Jan 1975
through Dec 1976.

Similarly the sample covariance between l 1 prime x j and l 2 prime x j can be expressed as in the same manner l 1 transpose S l 2. So, in this way calculate the Eigen values from the sample covariance matrix or the sample correlation matrix, and then calculate the associated orthonormal eigenvectors to obtain the principal components. So, the major task is to calculate the sample variance covariance matrix and obtain the Eigen values. Now let us take an example from real life data. So, real data example has the calculation of sample principal components from standardized data. Because that is what you will encounter mostly in practical situation, your variables will be from whole gamete where they have different sorts of description, interpretation, different units.

So, the best thing here will be to consider standardized variables. So that, we calculate Eigen values not from the sample variance covariance matrix; but from the sample correlation matrix. So, that is what we are going to do here; we are calculating sample principal components from standardized data and the data in this example are the weekly rates of returns from 5 stocks and these stocks are allied chemicals, du punt, union carbide, Exxon and Texaco. So, these are listed on the NYSE, the New York stock exchange and the weekly rates of all these 5 stocks were determined for the period Jan 1975 through; it old data does not matter; through December 1976.

(Refer Slide Time: 15:55)



The weekly rates are defined as

$$\frac{\text{current Friday closing price} - \text{previous Friday closing price}}{\text{previous Friday closing price}}, \text{ adjusted}$$

for stock splits & dividends

The obsns in 100 successive weeks appear to be independently distributed but the rates of return across stocks are correlated

Let $x_1, x_2, \ldots, x_5$ denote obsd weekly rates of return for Allied, Du Point, ..., Texaco, resp'ly. Then,

$\underset{\sim}{\bar{x}} = (.0054, .0048, .0057, .0037)'$: sample mean vector

R: covariance matrix of the standardized variables $z_1, \ldots, z_5$

$z_i = \frac{x_i - \bar{x}_i}{\sqrt{s_{ii}}}$; $i = 1(1)$

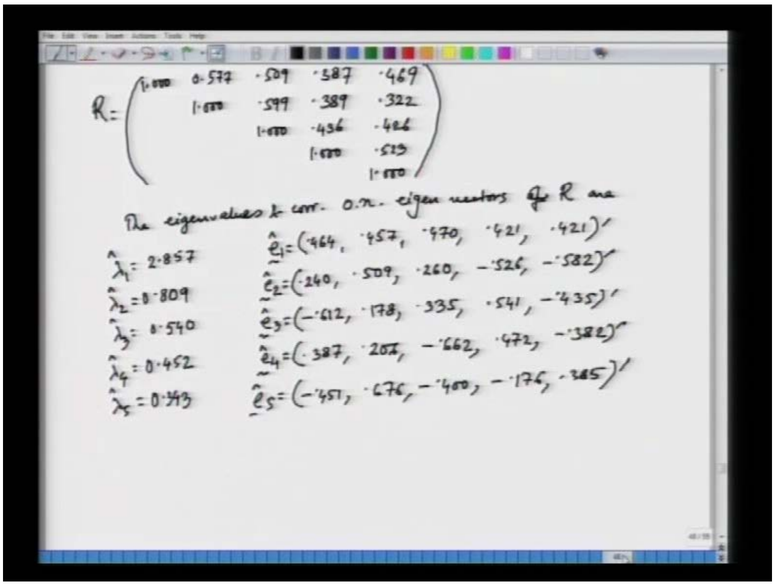
So, we have daily data and from here we have calculated the weekly rates. The weekly rates are defined as current Friday closing price minus previous Friday closing price divided by previous Friday closing price; this is one example just to tell you how the weekly rates are calculated from the daily data. There are some final issues like these are adjusted for stock splits and dividends. We may not go to the detail of this just assume that we have this data at our disposal to calculate the sample principal components. So, this is how we have calculated the weekly rate from the daily returns of the 5 stocks and then we have to calculate that we have a multivariate data. We basically have a five-dimensional data and number of observations ofcourse the number of weeks that we have in this period.

So, it is a multidimensional data and we are going to look in to the sample calculation of principal sample components and give some interpretation to those principal components. Now you see the observations. We have the observations in hundred successive weeks. So, that the period that we said it has hundred weeks. So, hundred successive weeks appear to be, now what we are doing here. We are sort of trying to justify our initial setup that we say that we have l independent observations. Now, meaning to this setup what we will say that for each day the stock returns are independent; these are independent observations. But across the stocks they are not independent.

By which we mean that since these are playing in the market, the returns over each of these stocks; they are sort of interrelated with each other. The price of one will influence the price of the other and so on. But over different, over the period these are all independent observations. So, the observations and hundred successive weeks appear to be independently distributed; but the rates of return across stocks are correlated for a fixed time point. So, then we have let x 1, x 2 to x 5 denotes observed weekly rates of return for the 5 stocks that is allied chemicals, du punt, etcetera up to Texaco respectively. Then suppose we have this information at our disposal, the five-dimensional sample vector is given to us, if not the full data.

If the full data is given to us, we can easily calculate the sample mean for each of these stocks and then give you the sample mean vector. So, whatever be the situation if the data is given to us, we will have to calculate it. Otherwise, the sample mean vector as a whole may be given to us and this is given by it is a P dimensional vector. Its values or the elements are 0.0054, 0.0048, 0.0057 and 0.0037; this is our sample mean vector. We have correlation coefficient matrix of these stocks. So, covariance matrix of the standardized variables are denotes by Z. So, this is Z 1 to Z 5 where Z i is nothing but x i minus x i bar by root x i, this is true for i from 1 to 5 and this matrix is given by 5 by 5 symmetric matrix.

(Refer Slide Time: 21:12)



$$R = \begin{pmatrix} 1.000 & 0.577 & .509 & .387 & .469 \\ & 1.000 & .599 & .389 & .322 \\ & & 1.000 & .436 & .426 \\ & & & 1.000 & .523 \\ & & & & 1.000 \end{pmatrix}$$

The eigenvalues & corr. o.n. eigen vectors of R are

$\hat{\lambda}_1 = 2.857$ $\quad \hat{\underset{\sim}{e}}_1 = (.464, .457, .470, .421, .421)'$

$\hat{\lambda}_2 = 0.809$ $\quad \hat{\underset{\sim}{e}}_2 = (.240, .509, .260, -.526, -.582)'$

$\hat{\lambda}_3 = 0.540$ $\quad \hat{\underset{\sim}{e}}_3 = (-.612, .178, .335, .541, -.435)'$

$\hat{\lambda}_4 = 0.452$ $\quad \hat{\underset{\sim}{e}}_4 = (.387, .205, -.662, .472, -.382)'$

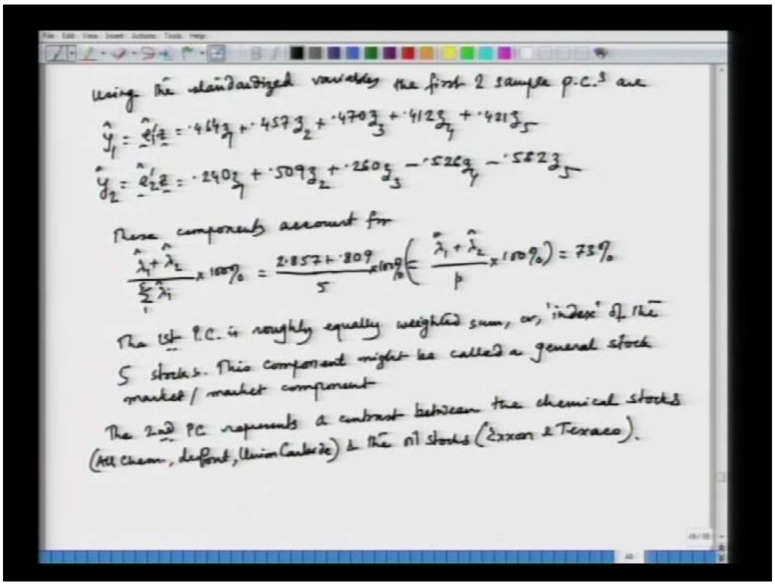
$\hat{\lambda}_5 = 0.343$ $\quad \hat{\underset{\sim}{e}}_5 = (-.451, .676, -.400, -.176, .385)'$

So, this is given by the variances are the measure of correlation coefficient. So, diagonals are all 1 and then we have the off diagonals, the correlation coefficients. So, these are

0.577, 0.509, 0.387, 0.469; again we have 1 and then 0.599, 0.389, quite close values 0.322, 1 and then 0.436, 0.426, 1 again in this position; 523 and the last element is ofcourse, 1.So, this is our sample variance covariance of the standardized variables or the correlation matrix of the original variables and we calculate the Eigen values from this matrix. So, the Eigen values and corresponding orthonormal eigenvectors, that is all we need to get the sample principal components.

Corresponding Ortho normalized eigenvectors of this matrix are lambda 1 hat is 2.857, lambda 2 hat is ofcourse what we said in the very beginning; they are in decreasing order. So, that we are listing the values in this manner; 0.809, lambda 3 hat is 0.540, lambda 4 hat is 0.452 and lambda 5 hats is 0.343. I am getting the corresponding Ortho normalized eigenvector are e 1 hat is 0.464 0.457 0.470 0.421 0.421 this is my first eigenvector. The second eigenvector is given by 0.240, 0.509, 0.260, minus 0.526, minus 0.582; e 3 hat is minus 0.612, 0.178, 0.335, 0.541, minus 0.435; e4 hat is 0.387, 0.206, minus 0.662, 0.472, minus 0.382 and the last one e5 hat is minus 0.451, 0.676, minus 0.4, minus 0.176 and 0.385.

(Refer Slide Time: 24:50)



using the standardized variables the first 2 sample p.c.s are

$$\hat{y}_1 = \hat{\underline{e}}_1'\underline{z} = \cdot 464 z_1 + \cdot 457 z_2 + \cdot 470 z_3 + \cdot 412 z_4 + \cdot 421 z_5$$

$$\hat{y}_2 = \hat{\underline{e}}_2'\underline{z} = \cdot 240 z_1 + \cdot 509 z_2 + \cdot 260 z_3 - \cdot 526 z_4 - \cdot 582 z_5$$

These components account for

$$\frac{\hat{\lambda}_1 + \hat{\lambda}_2}{\sum_1^5 \hat{\lambda}_i} \times 100\% = \frac{2\cdot857 + \cdot809}{5} \times 100\% \left(= \frac{\hat{\lambda}_1 + \hat{\lambda}_2}{p} \times 100\% \right) = 73\%$$

The 1st P.C. is roughly equally weighted sum, or, 'index' of the 5 stocks. This component might be called a general stock market / market component

The 2nd PC represents a contrast between the chemical stocks (All Chem, duPont, Union Carbide) & the oil stocks (Exxon & Texaco).

So, I have the first principal component using the standardized variables; because we have used the correlation coefficient matrix. So, we will mention it that using the standardized variables, the first two sample principal components are; the first one is y1 hat and that is e1 hat transpose Z and that is 0.464 Z1 directly from the first Eigen vector, 0.457 Z2, 0.470 Z3, 0.412 Z4 plus 0.421 Z 5. And the second one y 2 hat is from using the second eigenvector that is e 2 hat transpose Z gives 0.240 Z 1 plus 0.509 Z 2 plus 0.260 Z 3 minus 0.526 Z 4 minus 0.582 Z 5. We are interested in calculating the proportion of the variability explained by the sample principal components.

We have stopped that two principal components; because for us, it is the easiest to manage two principal components. We can easily project the data. There are tw0-dimensional plane and try and sort of interpret the data. But we must see before that how much of the total proportion of variability, these two principal components are explaining. So, this is given by the first two sample Eigen values. So, these two components account for lambda 1 hat plus lambda 2 hat by sum of lambda i hat that is 1 to 5; this into hundred percent is 2.857 plus 0.809. You can see here that this is the trace of the matrix, sum of the Eigen values of the r matrix. Now, trace of the matrix is nothing but sum of 1 up to P times.

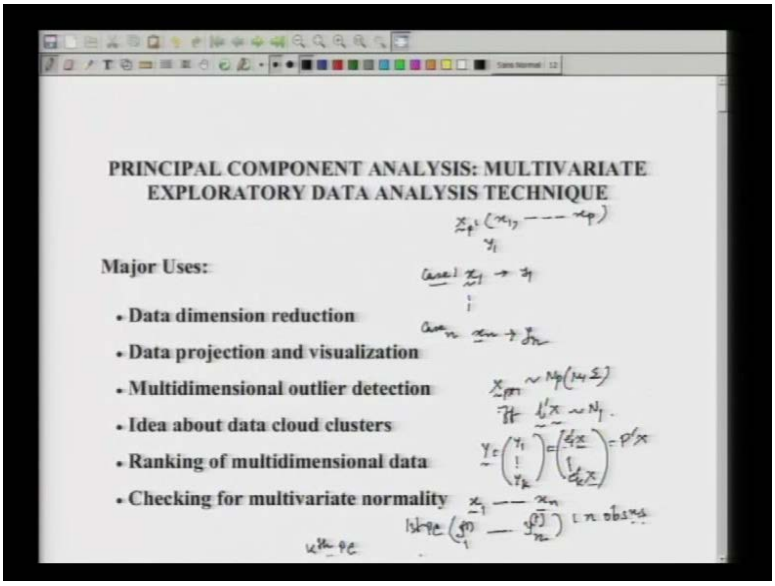
So, in case we are using the standardized variables, the summation lambda I hat as simply by lambda 1 hat plus lambda 2 hat by P which is basically the dimensionality of the data. So, this is 2.857 plus 0.809 by 5 into 100 percent. So, this let us keep in bracket, this is equal to 73 percent. So, from real life data, it is about 73 percent of the total variability that, the first two sample principal components are explaining. Now, if you look at the interpretation of these principal components, it is quite interesting to see that the first principal component is roughly giving a weighted index to all the stocks .It is a 0.464 with the first 0.457 with Z 2 in this way. So, this y1 hat, this first principal component can be thought of a sort of a market indicator, a market stock.

It may tell you that it is basically giving you the weighted average of all the stocks prices. The second one is grouping the first 3 and then it is grouping the last two. So, basically the first three stocks which are coming from the chemical companies. It is weighing these in one fashion and then it is sort of segregating out from the other two which are sort of together industry stocks of Exxon and Texaco. So, let us give this interpretation briefly. I will write this interpretation the first (( )) sorry the first component, the first PC is roughly equally weighted sum. If you look at the weights, the Eigen values are almost equal, roughly equally weighted sum or sort of a market index or index of the 5 stocks.

This principal component that is, y 1 hat can be thought of a might be called a general stock market or just a market component and the second principal component represents a contrast between the chemical stocks; that is the first three allied chemicals, du punt and the third one is union carbide and the oil stocks, the last two that is Exxon and Texaco. You may also go on calculating the other principal components, the third one will also come and the fourth one will come. Now, these may have positive negative signs with any of these variables and these may not be so easy to interpret. But ofcourse, if you are not satisfied with 73 percent of total variability been explained, you can go to the third principal component, and together you can see that how much of the total variability the first three are explaining.

Now strictly speaking, when we are handling real life data, we should stop at three-dimensional; otherwise the whole purpose is lost. We are trying to visualize the picture, the multidimensional data if it goes beyond three-dimensional; it is really of no use to us. So, and if we can give such nice interpretations or sort of it gives us some satisfaction, that there is some nice interpretation to the principal components that we obtain. So, we end this example here. Next we take up the situations where we start from the scratch that is from the data and I shall show you how you can use the sass software to calculate the principal components with given data. If you recall that, initially we had talked about the various uses of principal component analysis; that is, after the data dimension reduction what all we can achieve through that and I am going to explain you in detail on those aspects of the uses of principal component analysis.

(Refer Slide Time: 32:58)



PRINCIPAL COMPONENT ANALYSIS: MULTIVARIATE EXPLORATORY DATA ANALYSIS TECHNIQUE

$\underset{\sim}{x}_p : (x_1, ---- x_p)$

y_1

Major Uses:

- Data dimension reduction
- Data projection and visualization
- Multidimensional outlier detection
- Idea about data cloud clusters
- Ranking of multidimensional data
- Checking for multivariate normality

Case 1 $\underset{\sim}{x}_1 \to y_1$

⋮

Case n $\underset{\sim}{x}_n \to y_n$

$\underset{\sim}{X}_{p\times 1} \sim N_p(\mu, \Sigma)$

iff $\underset{\sim}{l}'\underset{\sim}{x} \sim N_1$.

$\underset{\sim}{Y} = \begin{pmatrix} y_1 \\ \vdots \\ y_k \end{pmatrix} = \begin{pmatrix} l_1'\underset{\sim}{x} \\ \vdots \\ l_k'\underset{\sim}{x} \end{pmatrix} = P'x$

$\underset{\sim}{x}_1 —— x_n$

1st PC $(y_1^{(1)} — y_n^{(1)})$ ← n obsns

kth PC

So, this is the principal component analysis. It is a part of multivariate exploratory data analysis technique. If you recall, we have told initially that the major uses of principal component analysis are data dimension reduction. We start with a P dimensional random vector. But ultimately we come to a k dimensional principal components; k ideally two or at the most three and what does this reduction in data dimension do. It helps us to project the multivariate data and visualize the various characteristics of the data. Now, once we project the multivariate data in the two-dimensional or three-dimensional planes, it serves some other purposes we can see whether there is any outlier present in the data. Hence we have multidimensional outlier detection.

We have an idea about the data cloud clusters that is the grouping of the various data; some of them will form one group; some other will form another group in this way. We can also do ranking of multidimensional data. How will we have the first principal component; on the basis of the first principal component, if you have now the ranking of multidimensional data. So, this is our example where we start from the scratch and use the sass software to calculate the principal components. It is a branch of multivariate exploratory data analysis and if you recall, the major uses of principal components what we have mentioned in the very first session on this p c a those were that data dimension reduction where this is the crux of the backbone of the whole thing.

We start with a P dimensional data vector. But we come down to a k dimensional principal components vector; k ideally is 2 or at the most 3. So, we can visualize the multidimensional data in a two or a three-dimensional plane. Now, once we can project the data, so the reduction of data helps us in the projecting the data and visualize it. Once we are able to project the data, we can view certain other characteristics of the data; major of these being that we can see whether if there is any outlier present in the multidimensional data or if the observations or variables are forming any groups. So, that is idea about data cloud clusters. We can also rank the multidimensional data, if we have a p dimensional data, you have x p.

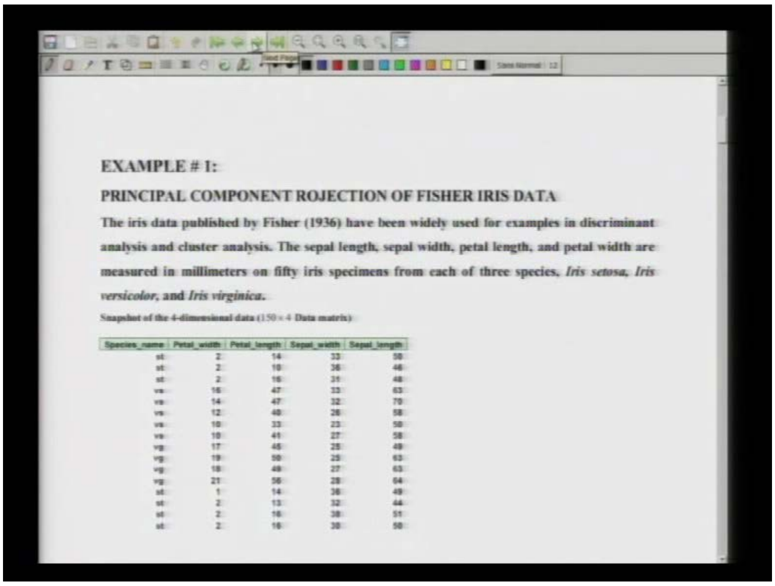
So, how you rank this you have x 1 to x p. So, that is p such variables and you have n observations on each of these variables, how you rank such data. Now, what we do is we calculate the first principal component y 1, which is the first principal component. For the first case, that is the first variable we have the first group of observations that is, if we call it x 1 that is first cases you have x 1; that is the first observation on all the variables x 1, x 2, x p. From here we get y 1 and then we go up to such n cases; that is we have total of n cases, case n and I combine that in the nth vector from where we have y n. So, basically these are all y 1 to y n are the values of the first principal component and then I can rank among these variables; because these are now reduced to one-dimensional data.

I can easily rank these one-dimensional data and obtain a ranking of the multidimensional data. Similarly checking for multivariate normality, by definition we have that x is going to follow a multivariate normal distribution say normal P mu sigma, if and only if every linear combination of x that is in prime; x follows univariate normal. Now, we see that the principal component again y, we have k such principal components y 1, y 2 to k and these are nothing but e 1 transpose x up to e k transpose x. These are the principal components; these are all linear combinations of x. So, this is our p transpose x matrix.

So, we check for the univariate normality of each of this y 1, y 2 to y k and if anyone of these are not univariate normal; we have reasons to believe that this x data vector is not coming from a multivariate normal population. You also may have observational sample of size n. So, you have x 1, x 2 to x n. From this multivariate normal population, you have first principal component; you check it the first principal component what are the observations you have y 1. So, say for y 1 the first one you have these n observations y 1 giving n. So, basically you have a sample of size n from univariate normal distribution; you have to check it that whether it is actually coming from univariate normal distribution which we all know.

There is certain ways blocks and certain other things from which you can check it and then if it is satisfied then you say that it is coming the first principal component; that is a 1 linear combination is coming from the univariate normal distribution. Similarly, you go and checking up to the k, the principal component that is n observations, again the sample of size n from a univariate normal distribution. Checking that and if all these are satisfied, we can believe that x is coming from these observations are coming from a multivariate normal population. If one of them is not satisfying the univariate normality, we may say that these may not come; this data x 1, x 2, x n may not come from the multivariate normal population.

(Refer Slide Time: 39:42)



EXAMPLE # 1:

PRINCIPAL COMPONENT ROJECTION OF FISHER IRIS DATA

The iris data published by Fisher (1936) have been widely used for examples in discriminant analysis and cluster analysis. The sepal length, sepal width, petal length, and petal width are measured in millimeters on fifty iris specimens from each of three species, *Iris setosa, Iris versicolor,* and *Iris virginica.*

Snapshot of the 4-dimensional data (150 × 4 Data matrix)

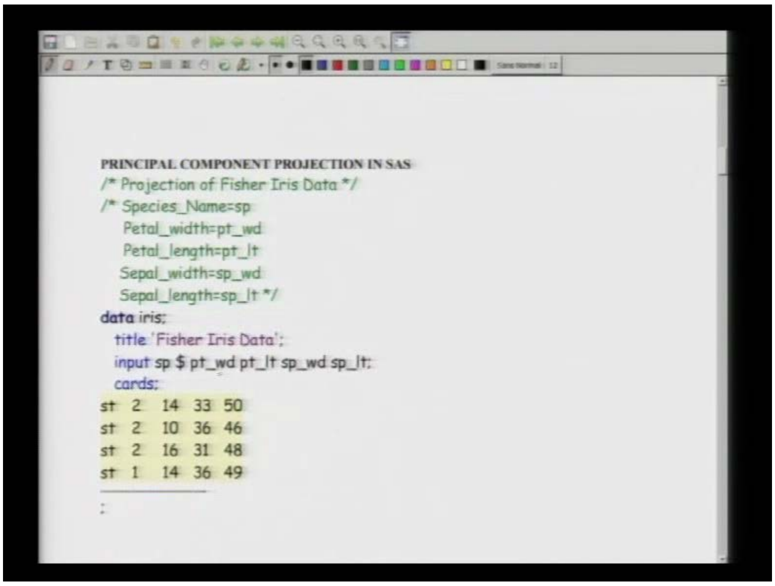
Species_name	Petal_width	Petal_length	Sepal_width	Sepal_length
st	2	14	33	50
st	2	10	36	46
st	2	16	31	48
vs	16	47	33	63
vs	14	47	32	70
vs	12	40	26	58
vs	10	33	23	50
vs	10	41	27	58
vg	17	45	25	49
vg	19	50	25	63
vg	18	49	27	63
vg	21	56	28	64
st	1	14	36	49
st	2	13	32	44
st	2	16	38	51
st	2	16	30	50

Now, the example that we do the first one is we have this is the famous iris data published by fisher; very old data 1936, it has been widely used for examples in other branches of multivariate analysis also like discriminant analysis and cluster analysis. Now, what has been collected here is the data on the iris specimen. This is a type of a flower and there are three spaces of this iris flower are the iris setosa, iris versicolor and iris virginica and four characteristics on each of these three types, data on the four characteristics are collected. So, this is a just a snapshot of this four-dimensional data and you can say this is a 150 by 4 data matrix. So, in all we have 150 observation vectors.

Out of which how many are for this first species, how many for the second that is we can see if it is fine there; otherwise we have to find it out from this data. Here we have only a part of the data. So, the first characteristic is the petal width, the second is petal length, the third is sepal width and the fourth is sepal length; st stands for iris setosa. So, for this we are using st, for iris vesicular we are using vs and for iris Virginia we are using vg and these are the data, these are the length in millimeters of this 50. It gives that there are 50 specimens from each of the three spaces. So, that makes it 1; 50 it is given here; this adds the part of the data, let us see what we can do with this data.

Now what are we going to do easily you can guess it that if you can project this four-dimensional data on a two-dimensional plane and see that if there is any link of these with the measures of the characteristics and the type of the species. From this snapshot, you can see only with the first characteristic that is petal width. Whenever it is iris setosa, you can see that it is a small value; it is about two millimeters. So, petal width is really small with respect to the other two spaces; but what happens to the other three. When you consider the other three characteristics, very readily you cannot say anything and it is just a four-dimensional data in front of you. So, to get something from this data, we try to project it in a two-dimensional or atmost a three-dimensional plane.

(Refer Slide Time: 42:46)

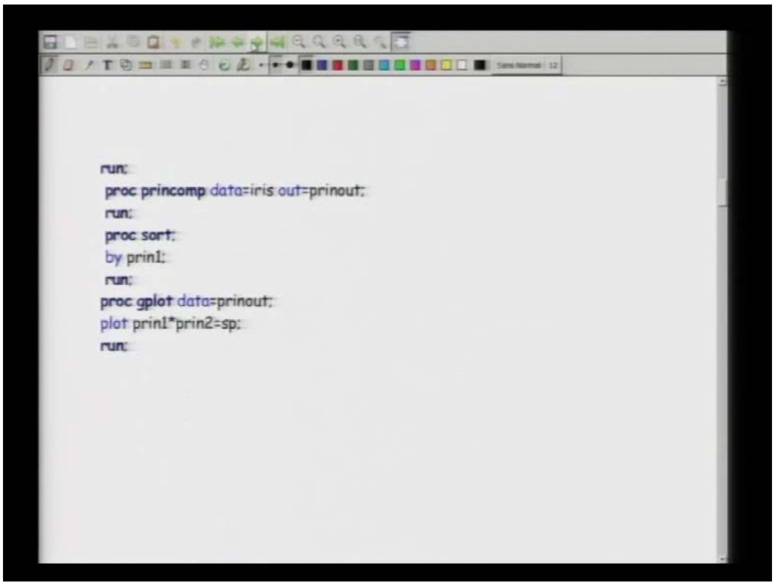


PRINCIPAL COMPONENT PROJECTION IN SAS

```
/* Projection of Fisher Iris Data */
/* Species_Name=sp
    Petal_width=pt_wd
    Petal_length=pt_lt
   Sepal_width=sp_wd
   Sepal_length=sp_lt */
data iris;
  title 'Fisher Iris Data';
  input sp $ pt_wd pt_lt sp_wd sp_lt;
  cards;
st  2   14  33  50
st  2   10  36  46
st  2   16  31  48
st  1   14  36  49
______________
;
```

So, the next thing is the sass code for calculating the principal component. So, this is how we have named these are comments. So, these are put within the slash star and we have stored this data. I named it as iris; it is given in this editor itself. So, it is with card system that we have put the data and these are the four characteristics the petal width, petal length, sepal width and sepal length.

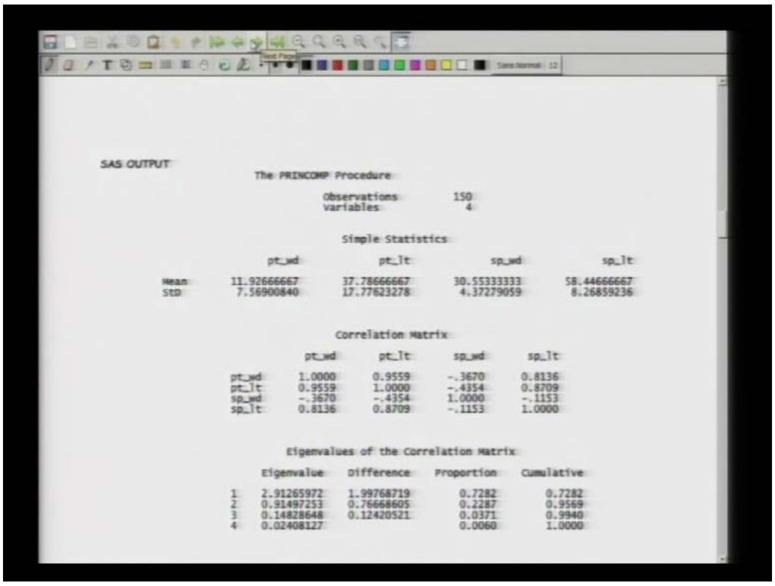
(Refer Slide Time: 43:16)



```
run;
 proc princomp data=iris out=prinout;
 run;
 proc sort;
 by prin1;
 run;
proc gplot data=prinout;
plot prin1*prin2=sp;
run;
```

So, we have entered the data run it and then we call proc princomp that is the sass subroutine for it; that is procedural princomp. We have data is iris; that is how we have named our data and out (( )) is printout. We sort by the first principal component; that is, we rank the data by the first principal component. We also plot the data and we obtain a two-dimensional plot of principal component one by principal component two.

(Refer Slide Time: 43:51)



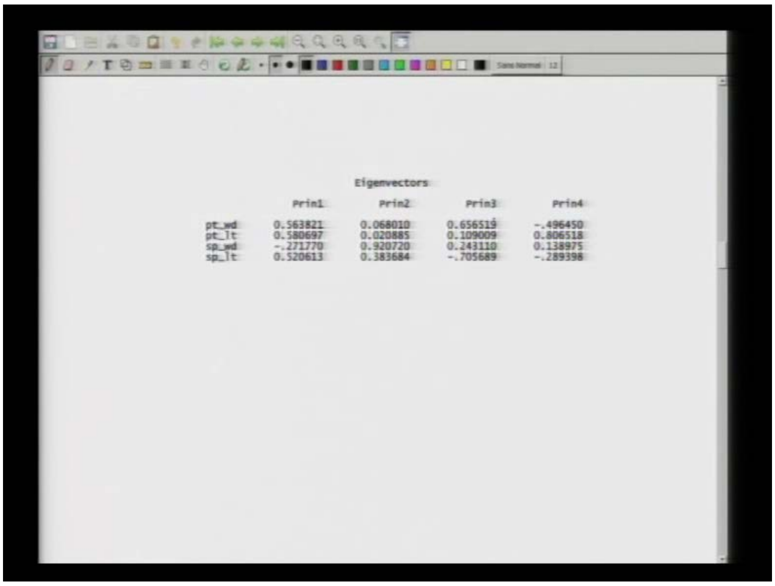
```
SAS OUTPUT
                    The PRINCOMP Procedure

                         Observations        150
                         Variables             4

                              Simple Statistics

                   pt_wd            pt_lt            sp_wd            sp_lt

Mean         11.92666667      37.78666667      30.55333333      58.44666667
StD           7.56900840      17.77623278       4.37279059       8.26859236


                              Correlation Matrix

                        pt_wd        pt_lt        sp_wd        sp_lt

          pt_wd        1.0000       0.9559       -.3670       0.8136
          pt_lt        0.9559       1.0000       -.4354       0.8709
          sp_wd        -.3670       -.4354       1.0000       -.1153
          sp_lt        0.8136       0.8709       -.1153       1.0000


                     Eigenvalues of the Correlation Matrix

              Eigenvalue     Difference     Proportion     Cumulative

          1   2.91265972     1.99768719         0.7282         0.7282
          2   0.91497253     0.76668605         0.2287         0.9569
          3   0.14828648     0.12420521         0.0371         0.9940
          4   0.02408127                        0.0060         1.0000
```

This is the number of observations, number of variables that there are four characteristics. So, that is four and the mean of the different characteristics that is x 1, x

2, x 3 and x 4, the four characteristics; these are the sample mean. So, that is the first one was petal width, this has the sample mean; next is petal length, sepal width and sepal length. Similarly, the sample standard deviation, this is the sample correlation matrix. We have calculated that is for the standardized values and then, Eigen values are calculated from the correlation matrix; these are the 4 Eigen values ordered. So, in the decreasing order, the first one being 2.91 then 0.91, 0.15 and 0.02.

This is the successive difference and this is the proportion of variability. So, this is important to us the proportion of variability that each of these Eigen values is explaining and this is the cumulative part. So, this is the first variability explained by the first principal component. First two principal components explained 96 percent of the variability and when I take the first three, almost the total variability is getting explained; because you see here it is .994. So, I think in this situation, the first two principal components are good enough; because the two together are explaining 96 percent of the total variability.

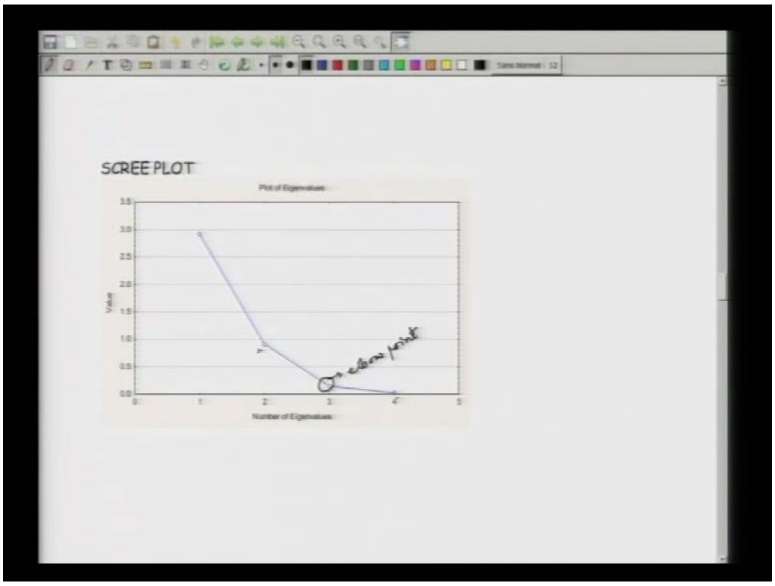
(Refer Slide Time: 45:32)



```
                         Eigenvectors

             Prin1         Prin2         Prin3         Prin4

pt_wd     0.563821      0.068010      0.656519      -.496450
pt_lt     0.580697      0.020885      0.109009      0.806518
sp_wd     -.271770      0.920720      0.243110      0.138975
sp_lt     0.520613      0.383684      -.705689      -.289398
```

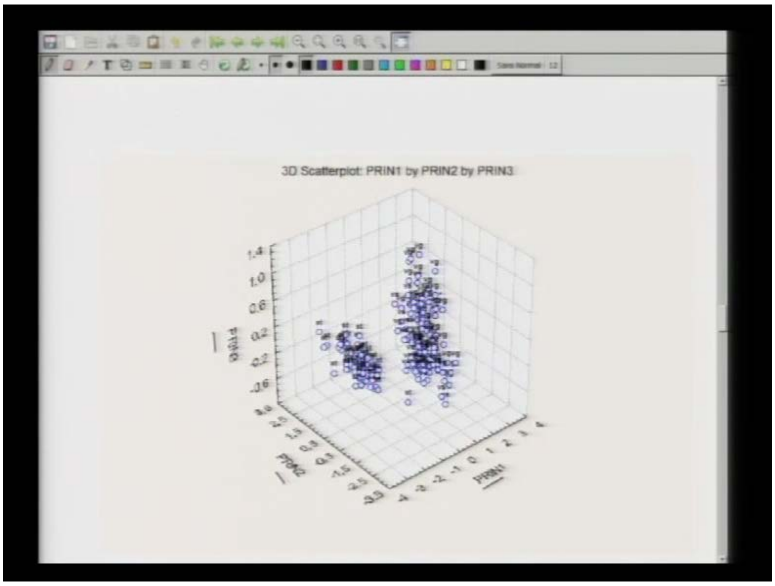
These are the principal components in the sense that we have the constants; that is the coefficients given to you. So, these are the values, the first principal component; these are the coefficient of this variable. We have to calculate the principle the case by case value of the principal components with the given value of these characteristics; these are the coefficients given for each case.

(Refer Slide Time: 46:01)



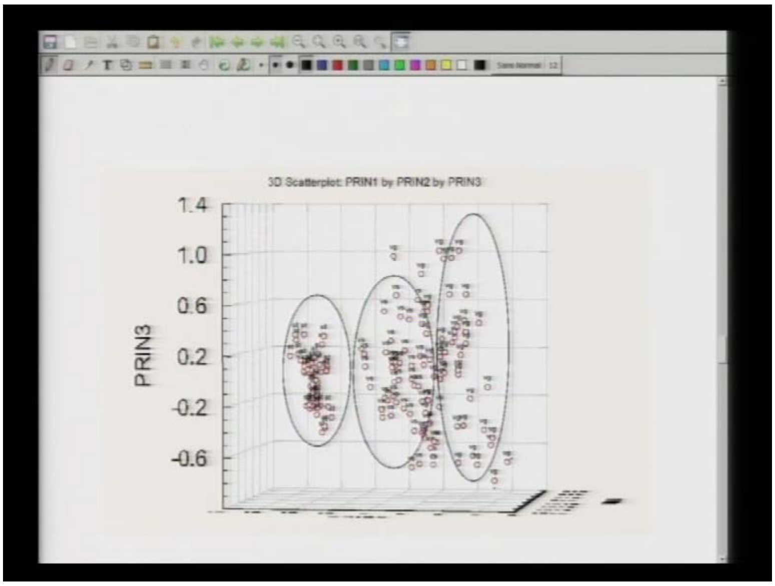
This is the screen plot. It gives you that the number of Eigen values that is the first Eigen value is close to 32. 9, something the second one is 0.91. So, that is this is the value, the third one has this value and the fourth one is this. Now, the screen plot comes down in the shape of an elbow may be called the elbow point of the screen plot. So, this is sort of the elbow point here and then it flattens out. So, this elbow point tells out that we can stop at this point, at this number of Eigen values, at this number of principal components. There is no reason to go beyond; actually in this case, even if we stop here, it is good enough. So, this is the significance of this screen plot. It actually plots the values of the Eigen values and tries to see that at which point you have the elbow point.

(Refer Slide Time: 47:24)



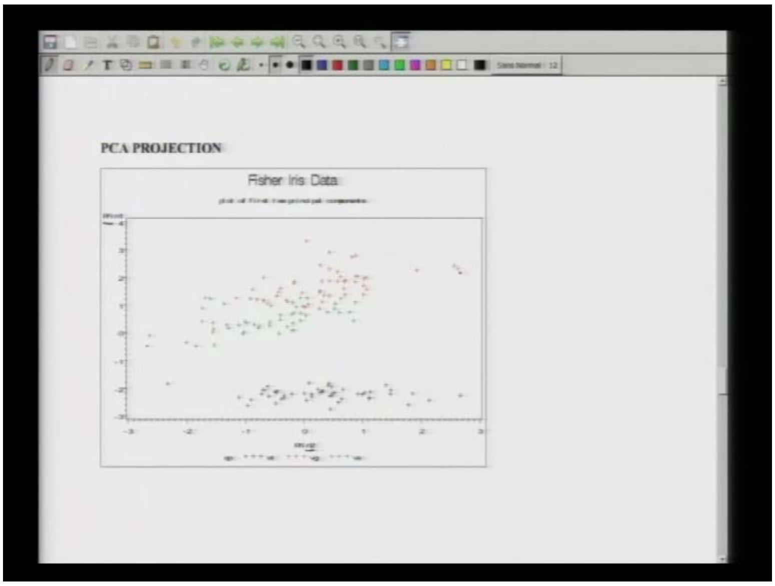
This is just four-dimensional; but you may have much higher dimensional data and in that case, it sort of useful to us and this is the three-dimensional scatter plot. We look at the three-dimensional scatter plot of the data. We have all the three axis; here principal component 1, 2 and 3.We try to see that if we can see any grouping, any outlier; this is a sort of an outlier; this is the first one st and these are vs and vg. So, you see vs and vg are sort of clocked together and st the first group, the first species is forming a cloud of its own. These are totally mixed up; it is very difficult. But still you see that on this side, you have some of the vg s which have separated out here. On the lower part, some of the vs have separated out and the middle portion is this vg s and vs s are totally mixed up.

(Refer Slide Time: 48:22)



So, let us go to the next plot and this is again a 3D scatter plot. But it is given from another orientation; here it is sort of much clearer about. We can see clearly the three groups; the first one that come nicely in this type of 3D plot.

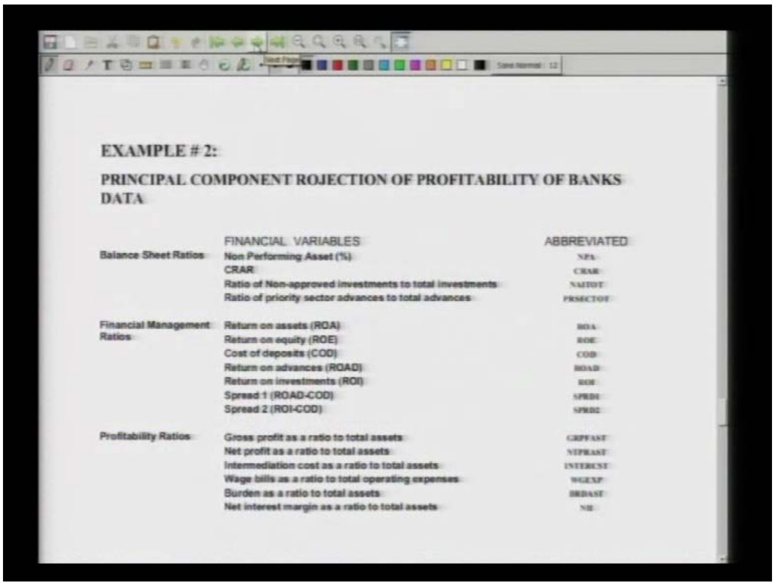
(Refer Slide Time: 48:54)



PCA PROJECTION

You may say that some of these like here one and here mainly the third species, some of them are outliers. If we do it in the two-dimensional plot that is we are only considering a principal component one and principal component two, these are the two access. Here also the grouping comes quite nicely. The lowest one is giving you the first species. The red color one is the second one vg and the third that is the green color is giving third species vs or again we can see some data outliers in these situations. So, these are the principal components for the case by case, you put the value of the characteristics, obtain the data and plot these here to get these plots.

(Refer Slide Time: 49:44)



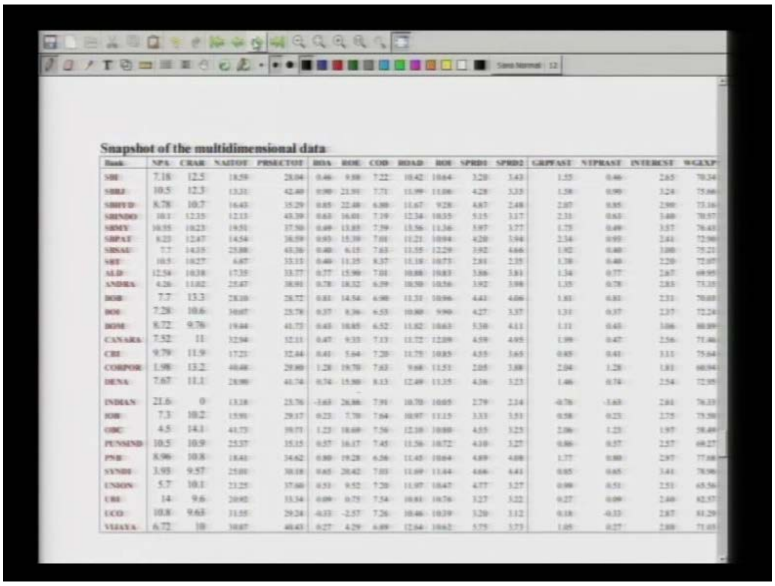
EXAMPLE # 2:

PRINCIPAL COMPONENT ROJECTION OF PROFITABILITY OF BANKS DATA

	FINANCIAL VARIABLES	ABBREVIATED
Balance Sheet Ratios	Non Performing Asset (%)	NPA
	CRAR	CRAR
	Ratio of Non-approved investments to total investments	NAITOT
	Ratio of priority sector advances to total advances	PRSECTOT
Financial Management Ratios	Return on assets (ROA)	ROA
	Return on equity (ROE)	ROE
	Cost of deposits (COD)	COD
	Return on advances (ROAD)	ROAD
	Return on investments (ROI)	ROI
	Spread 1 (ROAD-COD)	SPRD1
	Spread 2 (ROI-COD)	SPRD2
Profitability Ratios	Gross profit as a ratio to total assets	GRPFAST
	Net profit as a ratio to total assets	NTPRAST
	Intermediation cost as a ratio to total assets	INTERCST
	Wage bills as a ratio to total operating expenses	WGEXP
	Burden as a ratio to total assets	BRDAST
	Net interest margin as a ratio to total assets	NII

Our next example is about some financial variables. This is a feasible component projection of profitability of banks data what we have is the sort of a number of financial variables on the banks balance sheet ratios, financial management ratios and profitability ratios. These are some of the very important characteristics of the performance of banks.

(Refer Slide Time: 50:13)



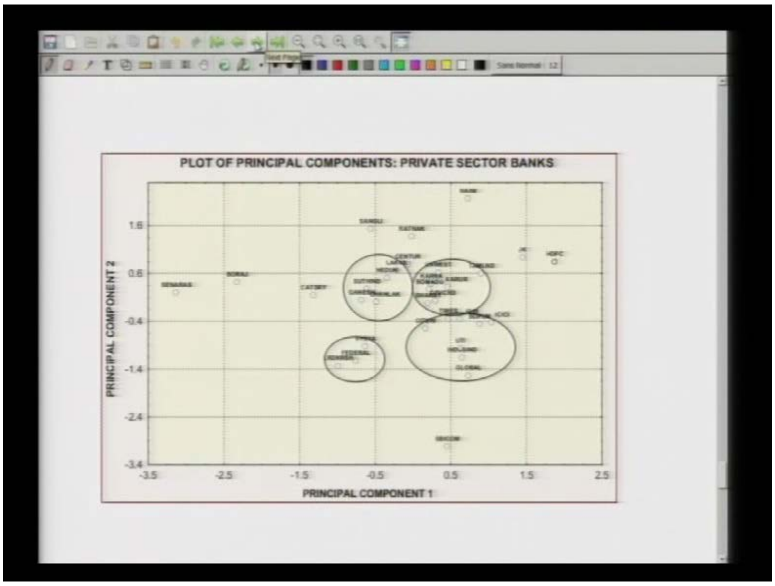
Snapshot of the multidimensional data

Bank	NPA	CRAR	NAITOT	PRSECTOT	ROA	ROE	COD	ROAD	ROI	SPRD1	SPRD2	GRPFAST	NTPRAST	INTERCST	WGEXP
SBI	7.18	12.5	18.59	28.04	0.46	9.88	7.22	10.42	10.64	3.20	3.43	1.55	0.46	2.65	70.34
SBBJ	10.5	12.3	13.31	42.40	0.90	21.91	7.71	11.99	11.06	4.28	3.35	1.58	0.90	3.24	75.66
SBHYD	8.78	10.7	16.43	35.29	0.85	22.48	6.80	11.67	9.28	4.87	2.48	2.07	0.85	2.90	73.16
SBINDO	10.1	12.35	12.13	43.39	0.63	16.01	7.19	12.34	10.35	5.15	3.17	2.31	0.63	3.40	70.57
SBMY	10.55	10.23	19.51	37.50	0.49	13.85	7.59	13.56	11.36	5.97	3.77	1.75	0.49	3.57	76.43
SBPAT	8.23	12.47	14.54	38.59	0.93	15.39	7.00	11.21	10.94	4.20	3.94	2.34	0.93	2.41	72.90
SBSAU	7.7	14.35	25.88	43.36	0.40	6.15	7.63	11.55	12.29	3.92	4.66	1.92	0.40	3.00	75.21
SBT	10.5	10.27	6.67	35.13	0.40	11.35	8.37	11.18	10.73	2.81	2.35	1.38	0.40	2.20	72.07
ALD	12.54	10.38	17.35	33.77	0.77	15.90	7.01	10.88	10.83	3.86	3.81	1.34	0.77	2.67	69.95
ANDHRA	4.26	11.02	25.47	38.91	0.78	18.32	6.59	10.50	10.56	3.92	3.98	1.35	0.78	2.83	73.35
BOB	7.7	13.3	28.10	28.72	0.81	14.54	6.90	11.31	10.96	4.41	4.06	1.81	0.81	2.31	70.03
BOI	7.28	10.6	30.07	23.78	0.37	8.36	6.53	10.80	9.90	4.27	3.37	1.31	0.37	2.37	72.24
BOM	8.72	9.76	19.44	41.73	0.43	10.85	6.52	11.82	10.63	5.30	4.11	1.11	0.43	3.06	80.99
CANARA	7.52	11	32.94	32.11	0.47	9.33	7.13	11.72	12.09	4.59	4.95	1.99	0.47	2.56	71.46
CBI	9.79	11.9	17.23	32.44	0.41	5.64	7.20	11.75	10.85	4.55	3.65	0.85	0.41	3.11	75.64
CORPOR	1.98	13.2	40.48	29.80	1.28	19.70	7.63	9.68	11.51	2.05	3.88	2.04	1.28	1.81	60.94
DENA	7.67	11.1	28.90	41.74	0.74	15.90	8.13	12.49	11.35	4.36	3.23	1.46	0.74	2.54	72.95
INDIAN	21.6	0	13.18	23.76	-3.63	26.86	7.91	10.70	10.05	2.79	2.14	-0.76	-3.63	2.81	76.33
IOB	7.3	10.2	15.91	29.17	0.23	7.70	7.64	10.97	11.15	3.33	3.51	0.58	0.23	2.75	73.50
OBC	4.5	14.1	41.73	39.77	1.23	18.69	7.56	12.10	10.80	4.55	3.25	2.06	1.23	1.97	58.49
PUNSIND	10.5	10.9	25.37	35.15	0.57	16.17	7.45	11.56	10.72	4.10	3.27	0.86	0.57	2.57	69.27
PNB	8.96	10.8	18.41	34.62	0.80	19.28	6.56	11.45	10.64	4.89	4.08	1.77	0.80	2.97	77.68
SYNDI	3.93	9.57	25.01	30.18	0.65	20.42	7.03	11.69	11.44	4.66	4.41	0.85	0.65	3.41	78.96
UNION	5.7	10.1	21.25	37.60	0.51	9.52	7.20	11.97	10.47	4.77	3.27	0.90	0.51	2.51	65.56
UBI	14	9.6	20.92	33.34	0.09	0.75	7.54	10.81	10.76	3.27	3.22	0.27	0.09	2.40	62.57
UCO	10.8	9.63	31.55	29.24	-0.33	-2.57	7.26	10.46	10.39	3.20	3.12	0.18	-0.33	2.87	61.29
VIJAYA	6.72	10	30.87	40.43	0.27	4.29	6.89	12.64	10.62	5.75	3.73	1.05	0.27	2.88	71.03

So, these are the data basically we have and these are the abbreviated forms of the data and what we have is data on all those characteristics for some private sector and public sector banks of the Indian economy. This is again snapshot of the multidimensional data;

this is huge data we have. We have looked in to a number of financial variables for a number of banks. So, both P and n is quite high in this case and this is a snapshot of that data.

(Refer Slide Time: 50:42)

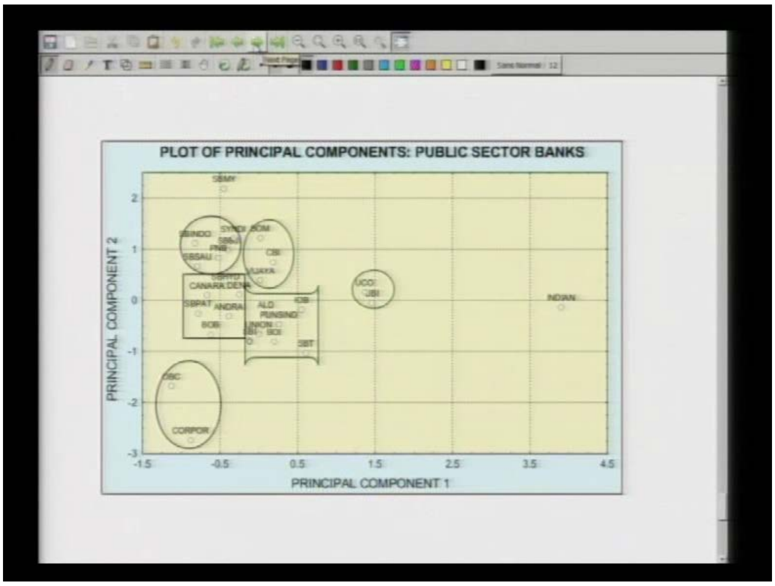


Now, what we do in this case is we separate out the private sector banks and we have calculated the principal components from the sample data that we have. So, we have first segregated the private sector banks and the public sector banks and then calculated the sample covariance matrix or the correlation matrices for each of the two cases. Then we have done the grouping separately ofcourse, you can do it in a same plot also; that is without segregating private sector and public sector banks. Basically, it depends on the purpose on your privative what you were trying to see. So, here we are trying to see, how the variability of the private sector banks among themselves; they are getting explained by the characteristics.

Similarly, how these characteristics are going to explain the variability present in the public sector banks. If it is for the whole banking system of the country, you will do it for the combined data. So, this is the first one for the private sector banks and we have plotted that, the first two principal components were sort of good enough for us .The exact proportion of the variability etcetera are not stated here; the main thing is we are looking at this sort of groups and clusters here. However the first two principal components served our purpose and then we projected the data on a two-dimensional plane and then we sort of trying to see that based on the characteristics whether we could form groups of banks; that is performance of the banks.

Based on the variability of the characteristics that are given, these can be said to be they are more or less same; they fall in the same group. So, in this way we have some of the banks, which are totally falling outside the clusters. So, they can be termed as the outlier and some of them are coming in one group. So, we can say the performance solely on the variability present on those financial ratios that we are considering. So, based on that, these banks come under one group; these private sector banks come under one group and so on.

(Refer Slide Time: 53:06)



The second one is for the public sector banks. You can see that you have similar sort of grouping possibly for the public sector banks. Here also we are projecting it on the two dimensional plane. So, it means that we are satisfied with the variability explained by the first two principal components. We see that here, you can see the State bank of my sore is an outlier here and the Indian bank probably for that data period had a major setback. So, this is one outlier here; otherwise you can see the UBI and UCO bank; they are in one group doing not too well at that time. And here, we have one group of Allahabad bank, Punjab and Sind bank, overseas bank, SBI and Bank of India.

Similarly, some other banks here, we have a group of Canara Dena, Bank of Baroda, Andhra bank and again a small group here for the OBC and the Corporation bank and Vijaya Central bank and one other is forming a group here. These are mainly the SBI group, the SBI Saurashtra and then we have SBI Indore and another one SBI. This is some other group of the state bank. So, this is about the public sector banks of India. The data period is probably not mentioned here. This is again with respect to that data period it is not mentioned; it is also an old data about a may be a data of the 1990s. These are the financial variables that we have here.

So, number of them are the characteristics giving you the dimension of the data and the observations, you have the number of banks. So, this exercise can be repeated for the whole banking industry of the Indian economy, if you combine the private sector and public sector banks, and then project the whole data. But for this type, we were mainly interested in the analysis where we were comparing the different sectors within themselves and hence, we obtained this sort of plots. So, this is how we can make some very nice use practical applications of principal component analysis.

With the help of this technique, we can really do some intelligent analysis of the multidimensional data which otherwise is just a group of haphazard data to us. We cannot do any ranking. We cannot do any projection of that data. We cannot make out any meaningful conclusion or interpretation of such data. It is just like a jungle of data in front of us; out of which you cannot make out anything; but if you can reduce the dimension of the data that is the crux of it. If you can reduce the dimension to k equal to 2 or 3, then with the help of the projection of the data, you can not only see the different groups, you can also see if there is any outlier.

The ranking of the data is also possible; ranking ofcourse makes sense if everywhere we are saying that it is ranking based on the first principal component. So, obviously we are assuming that the first principal component is explaining about atleast above 70 percent of the total variability. Otherwise ranking on the first principal component does not make much sense and ofcourse if that does not make much sense, then going to the second one also is not much helpful to us. So, in this way we can make some nice use of this technique of multivariate analysis.