# Applied Multivariate Analysis

**Prof. Amit Mitra**

**Prof. Sharmishtha Mitra**

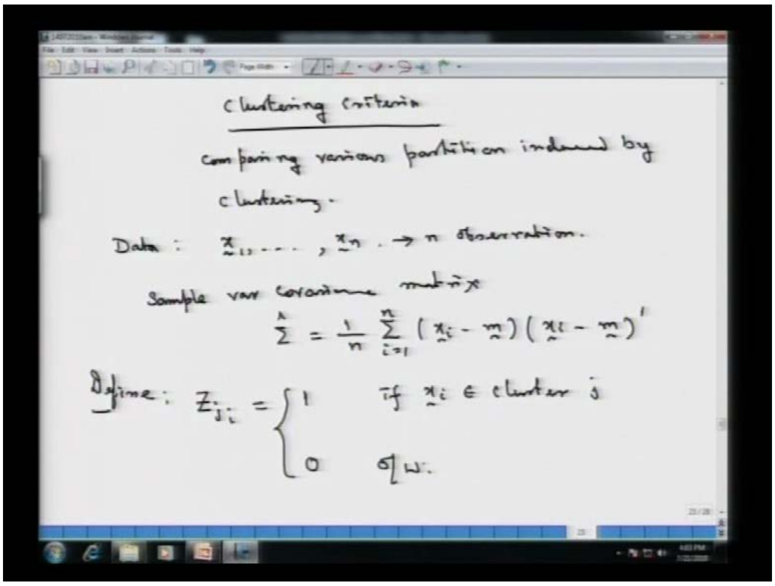
**Department of Mathematics and Statistics**
**Indian Institute of Technology, Kanpur**

## Lecture No. # 29

## Cluster Analysis

In the last lecture, we were discussing about the clustering criterion actually. In that, we are going to compare various partitions that are induced by clustering.

(Refer Slide Time: 02:21)



Clustering criteria

comparing various partition induced by clustering.

Data : $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n \rightarrow$ n observation.

Sample var covariance matrix

$$\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^{n}(\underset{\sim}{x}_i - \underset{\sim}{m})(\underset{\sim}{x}_i - \underset{\sim}{m})'$$

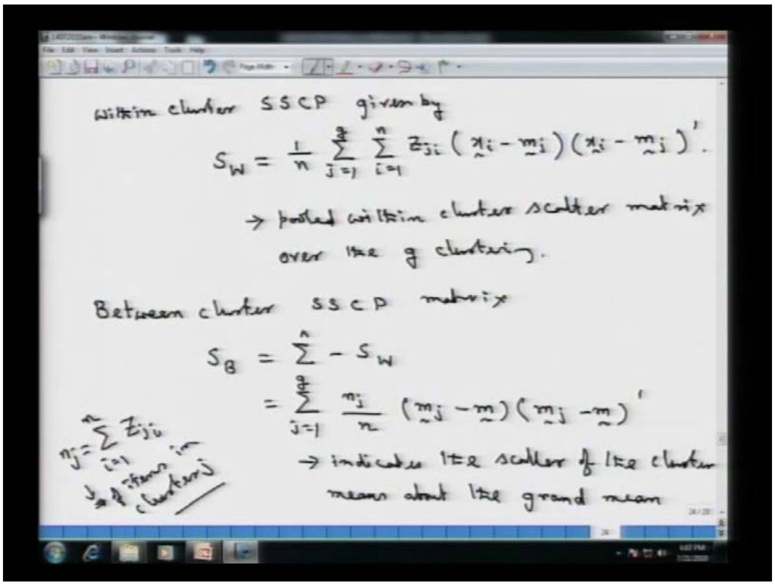
Define : $Z_{j_i} = \begin{cases} 1 & \text{if } \underset{\sim}{x}_i \in \text{cluster } j \\ 0 & \text{o/w.} \end{cases}$

We had come up to this point that let me just recollect clustering criteria is what we were looking at?  This we are trying to do is to compare or comparing various partitions induced by clustering. And in order to derive the criterion, we had said that, suppose we have the data in the following form. That we have some dimensional data say p dimensional data we have x 1, x 2, x n these n observations. These are basically the n observations on the multidimensional data.

And they and we had introduced this sample variance, covariance matrix as sigma hat with a divisor say n that is a maximum likelihood estimator as 1 upon n i equal to 1 to n.

x i minus n, which is the grand mean x i minus m transpose. We had defined m to be the grand mean in the last lecture. With this sigma hat we had also defined a random variable. This I will say that rather than definition define this z j i that was defined to be a binary variable taking values 1 and 0. One if x i then i the observation belongs to cluster j and is equal to 0 if it is otherwise.

(Refer Slide Time: 05:29)



Within cluster SSCP given by

$$S_W = \frac{1}{n} \sum_{j=1}^{g} \sum_{i=1}^{n} z_{ji} (\underset{\sim}{x}_i - \underset{\sim}{m}_j)(\underset{\sim}{x}_i - \underset{\sim}{m}_j)'.$$

→ pooled within cluster scatter matrix over the g clustering.

Between cluster SSCP matrix

$$S_B = \hat{\Sigma} - S_W$$
$$= \sum_{j=1}^{g} \frac{n_j}{n} (\underset{\sim}{m}_j - \underset{\sim}{m})(\underset{\sim}{m}_j - \underset{\sim}{m})'$$

$n_j = \sum_{i=1}^{n} z_{ji}$ # of items in cluster j

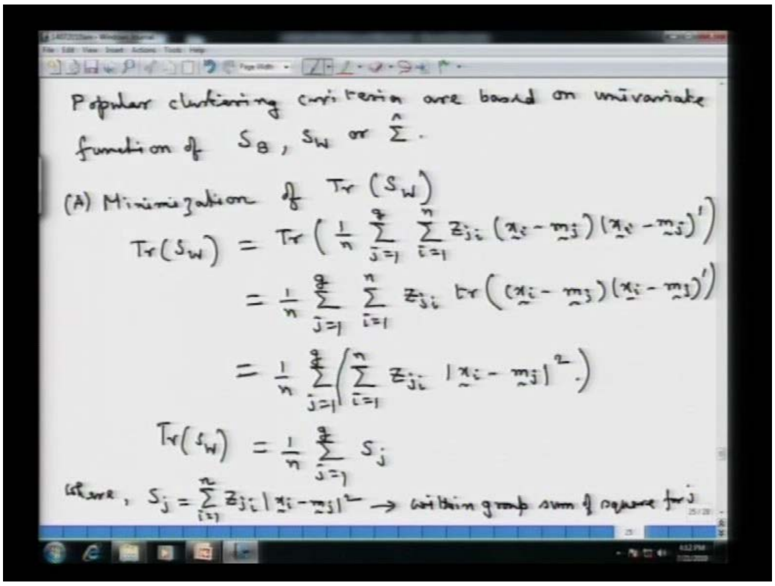
→ indicates the scatter of the cluster means about the grand mean

With this z j i indicator variable defined. We had the within cluster sum of squares and cross product scatter matrix given by the following. We had S W in the last lectures notation. then summation j equal to 1 to up to g, where g is number of clusters and i equal to 1 to up to n. z j i that multiplied by x i minus m j. m j is the cluster mean corresponding to the j cluster x i minus this m j vector it is transpose. So, this we interpret as a pooled within cluster scatter matrix over the g clusters. Suppose we have in a clustering algorithm giving us g clusters then, S W the sum of squares and cross product scatter matrix is going to be defined by this particular quantity.

And we also defined this between cluster sum of squares and cross product scatter matrix which is nothing but, S B matrix which is given by this sigma hat matrix minus what is the within cluster sum of squares and cross product matrix and it is elementary to see that this actually reduces to j equal to 1 to up to q n j divided by n and then we have this as m j minus m the grand mean that multiplied by m j minus m its transpose. This basically indicates the scatter of the cluster means about the grand mean which is m. where in here this n j term is nothing, but summation of z j i terms for i equal to 1 to up to n.

Because z i j is equal to 1. If x i belongs to the j th cluster and hence this particular sum of the indicators for i equal to 1 to n all the data is going to give us the number. This is the number of items in cluster j. So, that is the interpretation what we have. Now, having defined this within clusters sum of squares cross n cross product matrix as S W and the between clusters sum of squares and cross product scatter matrix to be S B. What we have actually the popular criterion for comparing various partitions are based on univariate functions of these quantities.

(Refer Slide Time: 09:35)



Popular clustering criterion are based on univariate function of $S_B$, $S_W$ or $\hat{\Sigma}$.

(A) Minimization of $Tr(S_W)$

$$Tr(S_W) = Tr\left(\frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} z_{ji}(\underset{\sim}{x_i} - \underset{\sim}{m_j})(\underset{\sim}{x_i} - \underset{\sim}{m_j})'\right)$$

$$= \frac{1}{n}\sum_{j=1}^{g}\sum_{i=1}^{n} z_{ji}\, tr\left((\underset{\sim}{x_i} - \underset{\sim}{m_j})(\underset{\sim}{x_i} - \underset{\sim}{m_j})'\right)$$

$$= \frac{1}{n}\sum_{j=1}^{g}\left(\sum_{i=1}^{n} z_{ji}\, |\underset{\sim}{x_i} - \underset{\sim}{m_j}|^2.\right)$$

$$Tr(S_W) = \frac{1}{n}\sum_{j=1}^{g} S_j$$

Where, $S_j = \sum_{i=1}^{n} z_{ji}|\underset{\sim}{x_i} - \underset{\sim}{m_j}|^2$ → within group sum of square for j

The popular clustering criterion are based on univariate functions of this S B matrix or S W matrix or this sigma hat matrix or a combination of these matrices. Now one such measure or one such criterion is what is given by the following quantity. It looks at minimization of the trace of one of these matrices which is S W minimization of trace of s w. what is trace of S W? Recall that what S W is from the previous slide here. S W is this particular quantity. We are looking at trace of this quantity.

That is 1 upon n summation j equal to 1 to up to g. Summation i equal to 1 to up to n then, we have z j i that into x i minus m j into x i minus m j transpose. We are looking at trace of this, now trace can be taken inside because trace of a b equal to trace of a plus trace of b. one can write this in the following way that is 1 upon n j equal to 1 to up to g.

Then we take trace inside the second sum here. So, it is z j i and then trace of this quantity which is x i minus m j into x i minus m j transpose.

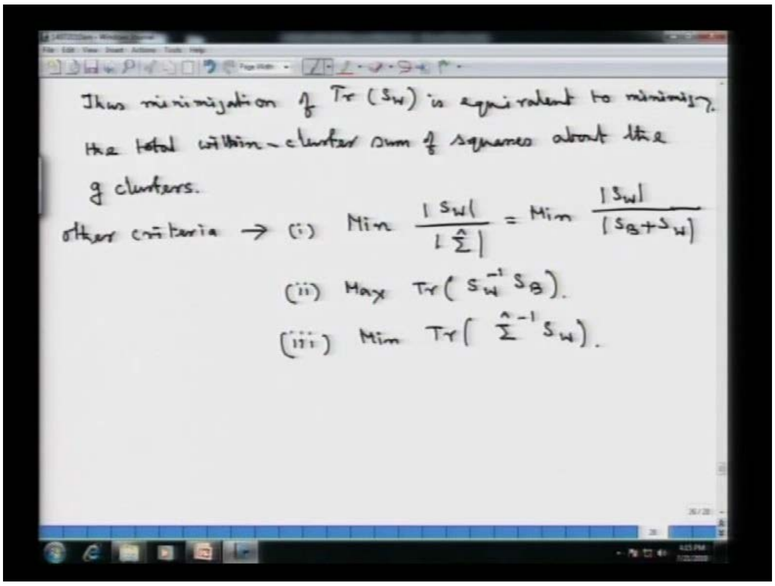
Now trace of a b equal to trace of b a. So, we can write that as summation j equal to 1 to up to g summation over i equal to 1 to up to n. then z j i into this quantity is nothing but x i minus m j square. So, this term is this only. What we have or what we can write is 1 upon n summation g equal to 1 to up to q and then this entire term here, which is a function of  j which is depending on j. for the particular j, we write that as S j where we for completion write that this S j is nothing, but this particular quantity which is z j i x i minus m j square this i equal to 1 to up to n.

Now, what is this quantity? If you look carefully at this particular quantity this is going to be n j number of terms. For which this z j i will be equal to 1 for the rest of the small n minus n j terms. This z j i is equal to 0. So, corresponding to those what we are doing is, looking at the vector x i item and then looking at its deviation from its mean. This basically is giving us, the within group sum of squares for cluster j sum of squares for cluster j. And then we are looking at, when we are looking at trace of S W we are looking at sum of that basically average 1 upon n for each of these observations.

This trace of S W is this quantity, and for different say prospective clusters or competing clusters corresponding to different partitions of the data what one is looking at is to look at which of those partitions, which of those possible partitions is giving us the minimum value of trace of this S W. And that is what is the desirable quantity is. Because minimization of this particular sum here, minimization of this summation j equal to 1 to up to g of s g. where, s g terms are within group sum of squares for that cluster g is basically trying to look at, what is the minimum quantity of that because we are looking at the minimum quantity of these terms in each of these clusters. And what are these? It is basically trying to ensure that cluster number j is as compact as possible.

If that particular j cluster is as compact as possible we will be having a minimum possible value of that S j corresponding to the jth cluster. And we look at over all possible such clusters and that thus becomes a criterion.

(Refer Slide Time: 13:23)



Thus minimization of $Tr(S_W)$ is equivalent to minimizing the total within-cluster sum of squares about the g clusters.

other criteria → (i) $\text{Min} \frac{|S_W|}{|\hat{\Sigma}|} = \text{Min} \frac{|S_W|}{|S_B + S_W|}$

(ii) $\text{Max} \; Tr(S_W^{-1} S_B)$.

(iii) $\text{Min} \; Tr(\hat{\Sigma}^{-1} S_W)$.

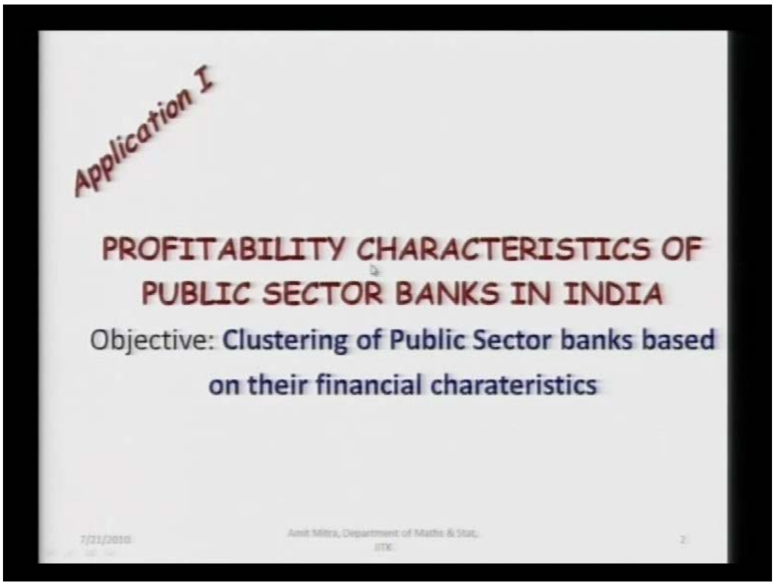
From the discussions that we are looking at this minimization of this trace of S W matrix is equivalent to the following, is equivalent to minimizing the total within cluster sum of squares about the g clusters. So, that is what the interpretation of this trace of s w is. It is a valid that we can take this trace of S W as one criterion for choosing among possible partitions among possible clusters in the data.

Because if we are looking at minimizing this particular quantity, we are looking at minimizing this quantity and hence for every cluster j we are looking at minimizing this quantity, which is the within group sum of squares for the jth cluster. Because it is looking at the deviation from that particular cluster mean. There are other criteria, I will say that other criterion the following or other type of criterion which are say looking at minimum of the following that it is determinant of S W by determinant of sigma hat this is actually minimum over all possible such partitions S b plus S w.

So, it is looking at this particular ratio, which is the ratio of the determinant of the within cluster sum of square and cross product matrix S W. That divided by the total, which is sigma hat. Some other criterion are to look at maximum of trace of S W inverse times S b or to look at minimum of the trace of sigma hat inverse times s w. There are other types of criterion also for looking at. What is the best? That one can actually look at while comparing various possible partitions of the data, various possible clusters that are coming.

Say for example, if one is considering a hierarchical clustering method based on the threshold distance. one can have different types of partitions of the data and then a criteria based, on which one can actually look at the optimum partition of the data is to look at which partition is basically looking at the minimum or the maximum. Whichever be the optimum criterion is corresponding to that particular setup. Now, tet me look at some actual real life examples, in order to look at the clustering analysis, what we have learnt so far. We look at some practical applications, practical data analysis concerning this cluster analysis technique.

(Refer Slide Time: 14:22)



Application I

PROFITABILITY CHARACTERISTICS OF PUBLIC SECTOR BANKS IN INDIA

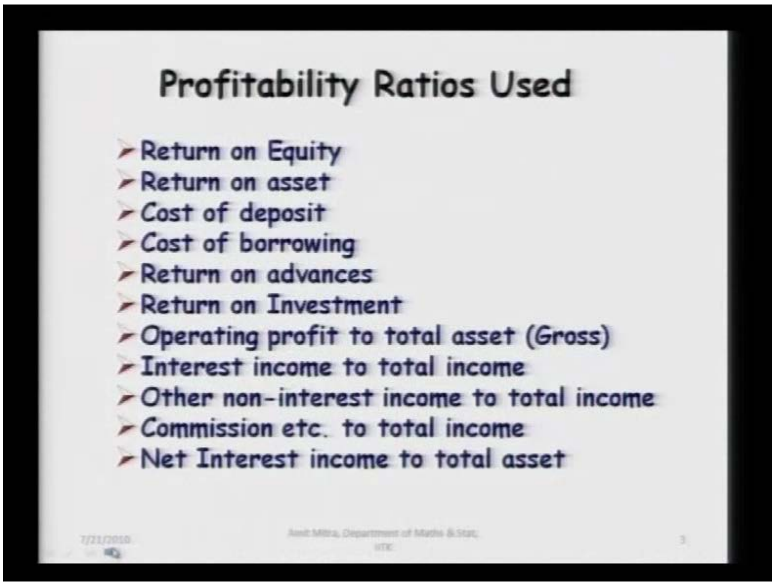
Objective: Clustering of Public Sector banks based on their financial charateristics

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 2

We will look at the standard technique. While discuss about two applications, the first application is looking at the profitability characteristics of public sector banks in India. Now, the objective in doing that is we are looking at clustering of public sector banks. Not only public sector banks, actually we are looking at, we will also look at other private sector banks and all banks put together. What sort of clustering does one can actually have, when one is looking at all public sector banks operating in India based on their financial characteristics.

Now, when we say that, we are looking at financial characteristics of financial companies. It is not that we look at, one just one variable it is usually a very high dimensional multivariate characteristic vector. That is what is usually looked at, and thus we have the following case. That we have say k number of financial institutions which in this case are public sector banks operating in India. And each of them has got a characteristic vector; the characteristic vector is based on, say financial characteristics. Now there are various types of financial characteristics that one can define.

(Refer Slide Time: 15:33)



## Profitability Ratios Used

- Return on Equity
- Return on asset
- Cost of deposit
- Cost of borrowing
- Return on advances
- Return on Investment
- Operating profit to total asset (Gross)
- Interest income to total income
- Other non-interest income to total income
- Commission etc. to total income
- Net Interest income to total asset

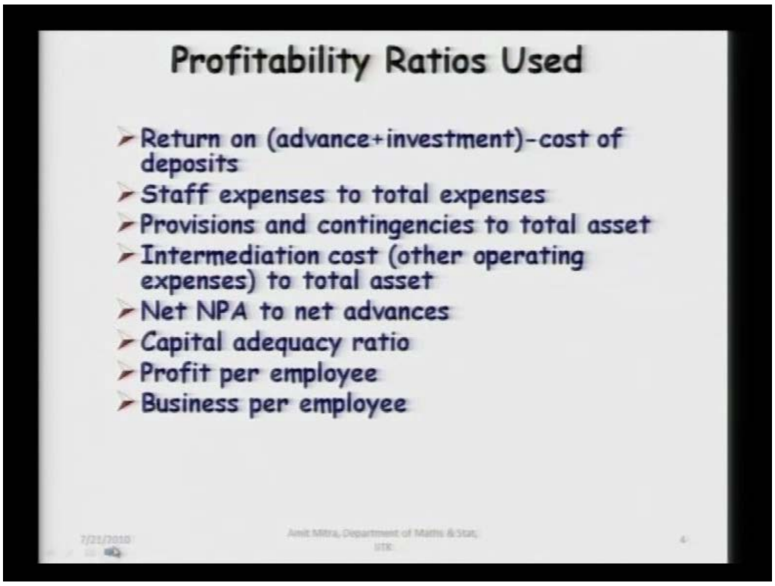
7/21/2010

Amit Mitra, Department of Maths & Stat, IITK

3

In this analysis, we have used following profitability ratios. Some of them are derived from the balance sheet of the respective banks themselves. We have these listed variables, as the profitability ratios like return on equity, return on asset cost of deposit cost of borrowing, return on advances, return on investment operating profit to total assets, interest income to total income, other non-interest income to total income, ratio, commission etcetera. To total income net interest income to total assets ratio and various other important financial indicators.

(Refer Slide Time: 16; 05)



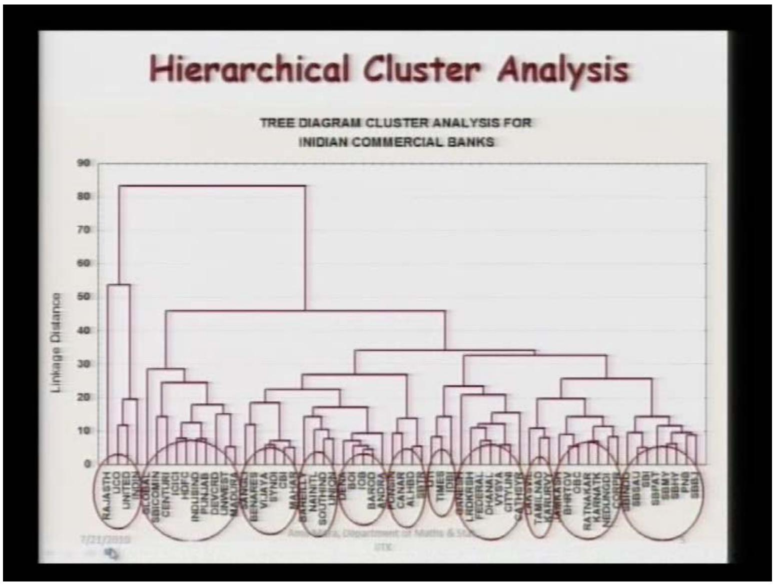
Profitability Ratios Used

- Return on (advance+investment)-cost of deposits
- Staff expenses to total expenses
- Provisions and contingencies to total asset
- Intermediation cost (other operating expenses) to total asset
- Net NPA to net advances
- Capital adequacy ratio
- Profit per employee
- Business per employee

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 4

It is corresponding to each of the financial institutions we will be having such a multidimensional data, where the dimensions of the data are these profitability ratios right. Suppose, we consider 25 such financial ratios, then each of the financial institutions they are those are the items or the cases here they are having a 25 dimensional data.

(Refer Slide Time: 16:38)



Hierarchical Cluster Analysis

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 5

And is what the problem is? And we are looking at that multidimensional data and looking at, what sort of clustering does actually emerge from such a data. We have learnt in the theory part, what a hierarchical cluster analysis is? Remember, there are two approaches of constructing a hierarchical cluster analysis. One was the agglomerative algorithm, agglomerative hierarchical clustering analysis. The other one was the divisive clustering algorithm. We had discussed about different measures, by which one can look at distances between various clusters. Which is an important consideration, when we look at construction of such hierarchical clustering?

Now, that theory is implemented in the following figure. We construct a hierarchical cluster analysis or a dendrogram tree. From these Indian commercial banks, what we have here is a similar type of picture. That is what we were discussing in the theory part. If we look at a on the y axis, we have as before the linkage distance or the merger or the fusion levels. If we look at a very high distance, then all the banks appear to be belonging to one single cluster and if we make the resolution level. So, fine that the distance is very small, then we will have all the cases to be members of singleton member, singleton unit clusters.

Now, what we have here at the end is that we can see that say for example, this is a branch here of the dendrogram tree. Which actually have these following banks which is the bank of Rajasthan, united commercial bank, united bank of India and Indian bank. Based on the data that is, what we had for mid 2000 actually 2005 or around 2006. The data was from that particular period. We see that clearly, there is a cluster formation of these four banks which is indicated from the branch that is, what we have from this if we cut off the tree, the dendrogram tree at this particular level we will see that mainly there are two branches. This is one branch under which there are four banks and all other banks are in one single branch.

If we get our resolution level down to this particular level, we will now find that there are 1, 2, 3 and 4 such branches of the tree. And hence four clusters in the data at that particular level as we go down in the resolution we actually we make our resolution level finer and finer. We will have more and more clusters emerging. If you consider the branch cut off at this particular linkage distance, you will find that all these cases which are listed here and inside this particular figure, they are belonging to this particular branch. Those are sub branches of this particular main branch.

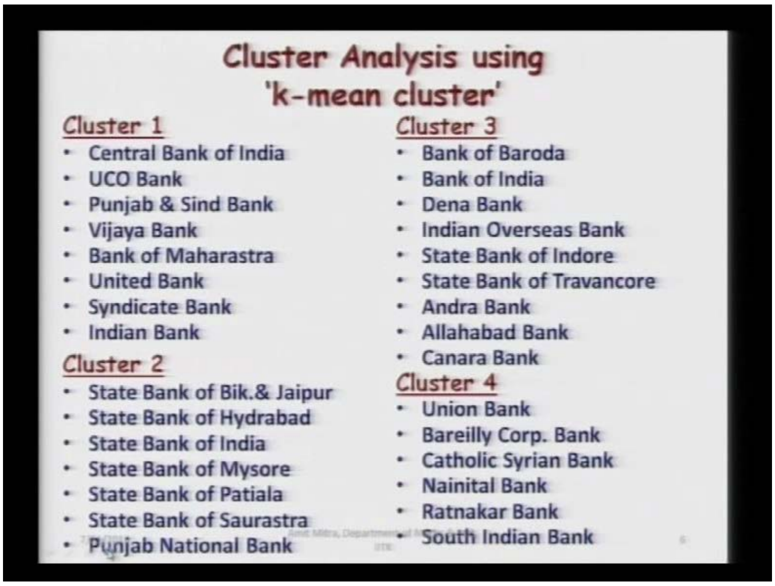
And hence all the cases below this particular branch are belonging to one single cluster. The banks are indicated, which are the basically new private sector banks. As you can see ICICI bank, HDFC bank inducing bank and development credit bank and all those banks, centurion bank are also present in this particular cluster. Now the big cluster, here are the other banks which we have here. Now one can also slice it along this particular level. And then look at various clusters that are formed. I have just as an indicative indication of these clusters that are formed I have put these figures. So, that these banks belong to one single cluster. These banks belong to one cluster which is coming under this particular branch of the tree.

Then these banks come under this branch here, these branches these banks come under this branch here and so on. So, you will find that the clustering that, we have obtained at the end of the day the hierarchical clustering. They naturally have a hierarchy in the formation of the clusters and very interesting clusters do actually emerge. If you look at the last cluster that we have under this particular branch here. So, this branch of the dendrogram tree branches out to include all these banks here. What are these are state bank of India SBI and all its associates.

It is basically the cluster of all those state bank of India and associates group which of course, have a similar structure of their business. And hence they belong to one single cluster out here. This is how one actually gets to a hierarchical cluster analysis. Now a point that I have made, when we were looking at theoretical discussions of this particular subject that such a hierarchical cluster analysis technique is not advisable. If we have a large data set that is if we have more and more cases the x axis will become so crowded. That it would be difficult for us to have such nice clusters as are formed in this particular data.

And hence in such a situation, one usually does not use a hierarchical clustering. If the data size is too large, one looks at a k means clustering which is a non hierarchical clustering technique.

(Refer Slide Time: 21:46)



## Cluster Analysis using 'k-mean cluster'

Cluster 1
- Central Bank of India
- UCO Bank
- Punjab & Sind Bank
- Vijaya Bank
- Bank of Maharastra
- United Bank
- Syndicate Bank
- Indian Bank

Cluster 2
- State Bank of Bik.& Jaipur
- State Bank of Hydrabad
- State Bank of India
- State Bank of Mysore
- State Bank of Patiala
- State Bank of Saurastra
- Punjab National Bank

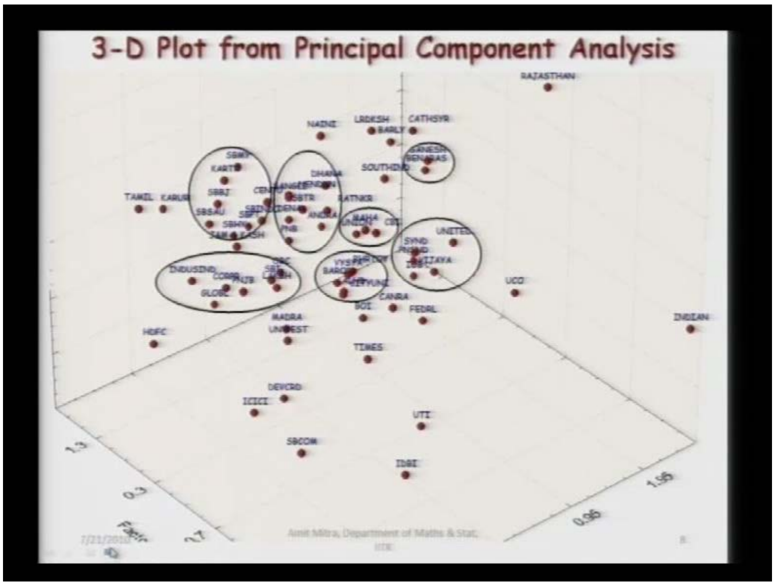
Cluster 3
- Bank of Baroda
- Bank of India
- Dena Bank
- Indian Overseas Bank
- State Bank of Indore
- State Bank of Travancore
- Andra Bank
- Allahabad Bank
- Canara Bank

Cluster 4
- Union Bank
- Bareilly Corp. Bank
- Catholic Syrian Bank
- Nainital Bank
- Ratnakar Bank
- South Indian Bank

Now, if we apply a non hierarchical clustering to the same data. We get such clusters this is the output of a k means clustering algorithm. For the same data we have in cluster one these banks. In cluster two, state bank of India and mostly its associates. You may note that this Punjab national bank later on merged with the state bank of India associates. Cluster number three consisting of all these banks here. So, each of these banks were initially from the data set were characterized by that high dimensional multivariate data and based on the algorithm. they have either been classified using, say the statistical non hierarchical clustering method. Like the k means clustering or the hierarchical clustering method leading us to the previous figure of that dendrogram tree. You see a cluster 4 of these banks here. Then we have a cluster five of these banks, cluster 6 for these banks. These are the clusters emerging not only from public sector banks all commercial banks in India including those. What are those called Private sector, public sector and corporate banks.

Then, cluster number seven are mostly good performing private sector banks ICICI, HDFC time's bank many of these banks later on merged, and then cluster eight is the cluster of these banks.

(Refer Slide Time: 23:09)



3-D Plot from Principal Component Analysis

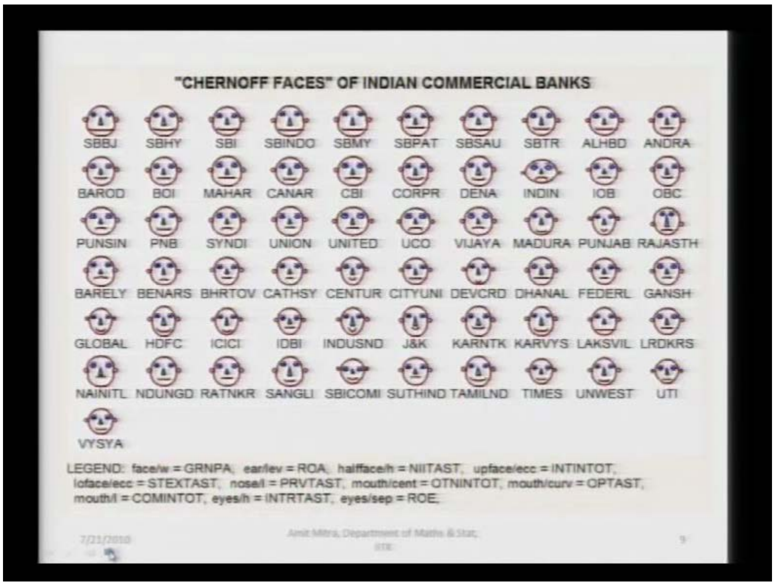
Now, in the theoretical lectures, one have we have also looked at the principal component analysis. Now principal component analysis is a very powerful technique for data dimension reduction, not only data dimension reduction it also looks at multidimensional projection of multidimensional data. And if we look at the data structure that is what we have it is basically a multidimensional characteristic data.

And then from a principal component projection of the same multidimensional data, one can also look at some rough clustering in the data. If we have the principal component projection applied to the same data, we have this three-dimensional figure. The dimensions are the first three principal components. This is the first principal component axis, the second principal component axis and the third principal component axis. Each of these cases, each of these multidimensional data is now represented by the first three-dimensions in the principal component.

And then looking at that three-dimensional figure, one can once again come up with different clusters emerging. Now, these are subjective clusters of course. One can also have idea about outliers in the data as was discussed in the theoretical lectures of principal component analysis. If you see that there is bank, which is in the Indian bank. This is bank of Rajasthan and these banks are clearly an outlier bank which does not actually is included in the main data cluster in this principal component projection of the data.

These are some subjective clusters that are taken out from this principal component projection. Once again we see that there are some meaningful clusters not as well defined clusters. As the type of output that one gets from a cluster analysis technique like the ones that I have discussed just now. So, this is a three-dimensional principal component analysis projection and the corresponding clustering.

(Refer Slide Time: 25:03)



"CHERNOFF FACES" OF INDIAN COMMERCIAL BANKS

LEGEND: face/w = GRNPA, ear/lev = ROA, halfface/h = NIITAST, upface/ecc = INTINTOT, loface/ecc = STEXTAST, nose/l = PRVTAST, mouth/cent = OTNINTOT, mouth/curv = OPTAST, mouth/l = COMINTOT, eyes/h = INTRTAST, eyes/sep = ROE.

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 9

Now, this is an interesting aspect of looking at this multidimensional data something which is called a chain of face. Chain of faces a way to look at visualization of a multidimensional data vector and it is a powerful technique actually. Because, if we look at dimensions two, three. One can visualize how the data is looking like? But if we have dimensions more than three, then the data cannot actually be visualized. However, there are various techniques for visualizing those multidimensional data. And to me, chain of faces. One of those methods which give us a nice pictorial representation of a multidimensional data with the same multidimensional data used for the cluster analysis. In this example, we obtain the chain of faces of all those banks.

Now to a layman if you look at a face here, the chain of face actually represents multidimensional data through faces. Now why faces? Because face representation of face characteristics is one that one can easily actually decipher. And then looking at just the face of that multidimensional data what one can see is that corresponding to the health of that particular multidimensional data if one is at all concerned about that. Now, in our example, we have the financial characteristics or financial institutions mainly the Indian commercial banks.

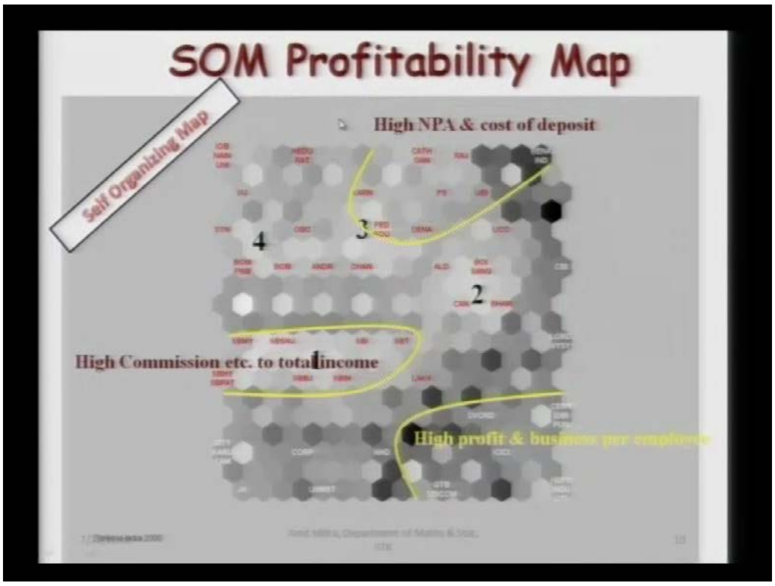
Now, if a bank looks smiling like for example, ICICI bank you will actually understand that particular banks entire multidimensional characteristics have been now represented by this particular figure and in the faces smiling, you will associate the interpretation that the bank is performing well. You see that the ICICI bank, the HDFC bank, the global trust bank, inducing bank J and K bank and some of the other banks have a smiley face. Some of the banks on the other hand does not look at all smiling like for example, the Indian bank.

The financial characteristics of this particular financial institution, the Indian bank is not at all good actually. And thus the facial representation of the multidimensional data gives a sad looking face of that particular financial institution. You can one can also look at multidimensional trends in the data through such chain of faces by looking at multidimensional data over the period of time. And look at how the faces are evolving over time and thus leading us to detect some sort of a multidimensional trend in the data.

This is not exactly cluster analysis, but this is a multidimensional visualization or visualization of multidimensional data. Now there are other ways of looking at clustering what we have discussed in the theoretical lectures where, statistical cluster analysis techniques. Now apart from the standard classical statistical cluster analysis techniques, the clustering can also be obtained under different philosophies. Like for example, if you look at artificial intelligence approach, then self organizing map is another method which actually leads us to looking at clusters in the data.

They have several advantages actually, over classical statistical cluster analysis techniques in various senses that they not only give clusters in the data. They also look at how compact the clusters are.

(Refer Slide Time: 28:34)



SOM Profitability Map

Self Organizing Map
High NPA & cost of deposit
4
3
2
1
High Commission etc. to total income
High profit & business per employee

Just by looking at the figure itself, the self organizing map. What we have tried to produce here this type of map actually not only lead us to understanding clusters in the multidimensional data, but also lets us understand the compactness of each of those clusters detection of multidimensional outliers and how well the clusters are separated from one another.

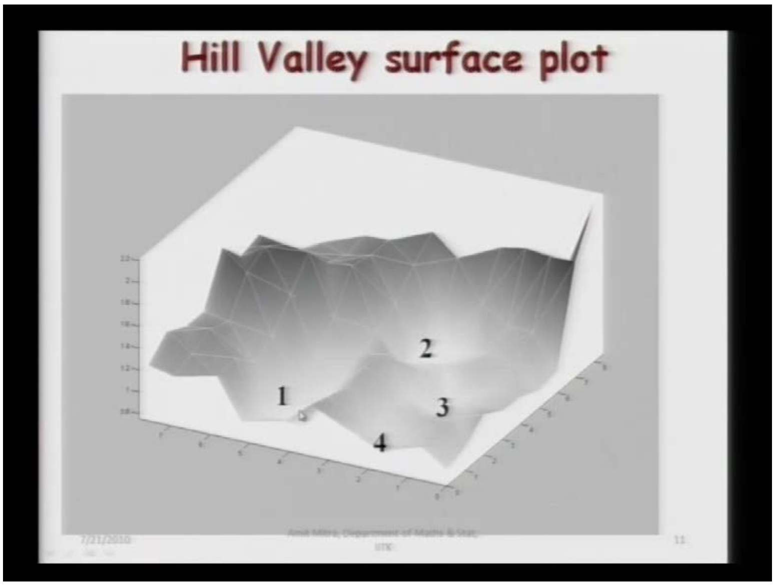
The same multidimensional data is used. In order to get to this self organizing profitability map of the data, now which is represented in terms of this bee hive actually. So, this hexagonal blocks here one after the other which is this three dimensional, two-dimensional representation. Now we have well defined clusters as what we have the banks in this region here. High profit, high business per employee and they are basically the new private sector league of banks.

Now, we have a means cluster. This is emerging out here which are the high commission income to the total income banks and all the banks here in this particular region are state bank of India and its associates. So, they form a very compact cluster among themselves. Now what is common? And what is any way high commission, commission is basically these state bank of India and its associates. They perform all the document businesses and the government says for example, they take in an income tax challans and etcetera.

These are government businesses that these groups of banks actually perform. And then they get commissions from the government of India. And that is basically a characterization of this particular sector. Here, we have a cluster being formed here. these are the banks which are high nonperforming assets and cost of deposit. So, they are basically banks which do not perform that particular well. There are two clusters here cluster number 4 , cluster number 3 and cluster number 2 here, which are formed from the self organizing map here.

Now, the way that clusters are detected from such a self organizing map is that one looks at say patches of shades of white leading us to believe that. There say these if two units here are separated by units which have a lower shade on a gray scale that indicates that those two units are say some sort of closely associated and hence can be put in one single cluster. And thus patches of light shades, in such a self in such a SOM plain leads us to having an interpretation of the data. That such area actually is associated with formation of clusters in the data. On the other hand, if we find cases which are in the dark patches we will understand that those are outliers.

(Refer Slide Time: 31:31)



Hill Valley surface plot

7/21/2010

Amit Mitra, Department of Maths & Stat, IITK

11

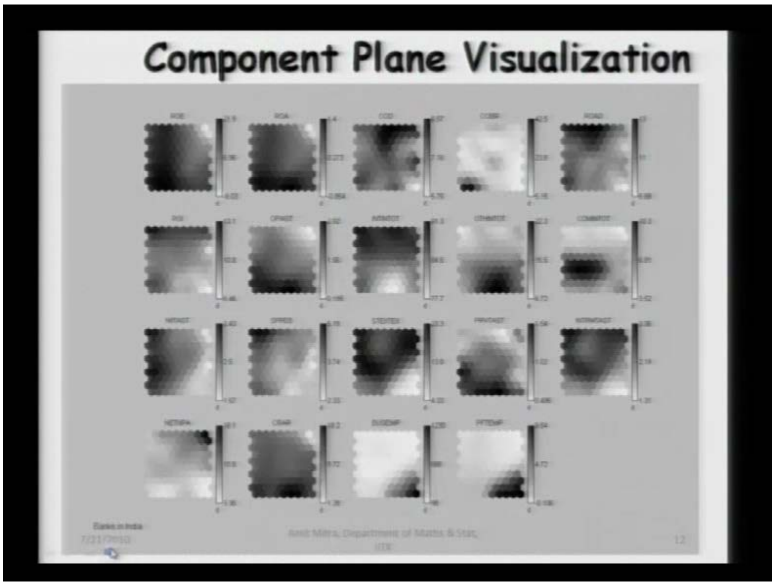
Now this type of figure also is sometimes visualized using a three-dimensional so, called hill valley surface plot representation of this self organizing map, where in we actually look at a three-dimensional visualization of this where the dimension of say color says either light shade or a dark shade of gray is now having a third dimension here. And thus here the valleys or places where the clusters are formed and the peaks here are the points where outlying observations are present in the data.

Corresponding to the previous figure, when we had one here 1, 2, 3 and 4. The same figure is represented in terms of that three-dimensional hill valley plot. So, whatever cases were, there in that one sector which was state bank of India and its associates they are in this particular valley, in this hill valley plot here. Now people in the valley, they we say that they are likeminded people which reside in a particular valley and the like minded cases are basically put in this particular valley and which form a closed cluster clearly separated from the other clusters.

Because, the height of the hill lying in between cluster number one and the remaining clusters is quite well demarcated and hence this is a very close compact cluster. And so is cluster number two however, the demarcation between cluster number two and cluster number three is not  that well defined as that is for one at the rest of the clusters although there is some height between the two clusters two and three. As regards to cluster number three and four, we have the two or the rather the characteristics of the two clusters are almost the similar. However, they are quite different as this particular dark patch here is indicating.

Well, where is this particular location here. This particular location here is on a higher plain which is located in this particular drop down there. If we look at this in this particular other side of the hill valley plot. We will have the other banks which were represented here high profit and high business per employee basically the more efficient private sector banks.

(Refer Slide Time: 33:56)



Component Plane Visualization

Banks in India

7/21/2010

Amit Mitra, Department of Maths & Stat, IITK

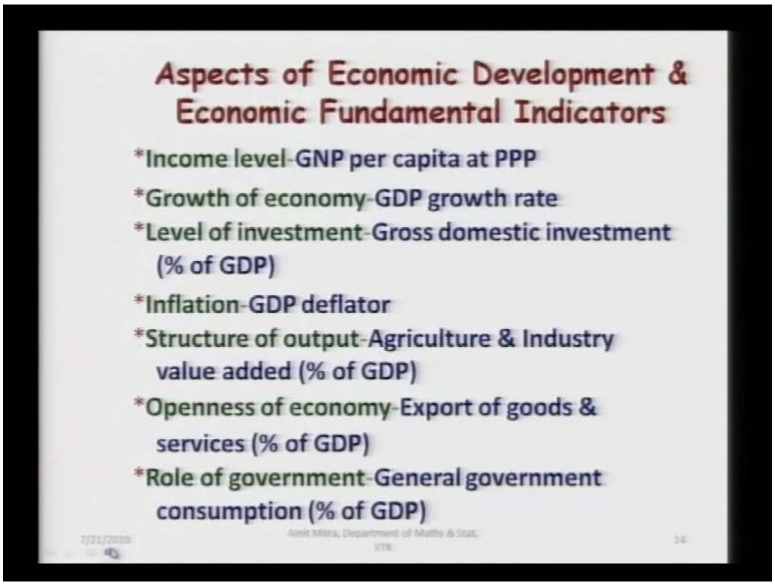
12

Clustering can also be obtained from such an approach of self organizing map which is an approach, of an artificial intelligence technique now, the same data when we had the SOM representation and the hill valley real representation or the visualization of the self organizing map.

This is another way to look at that particular clustering, which is called the component plain visualization. This actually looks at each of the constituent variables. Now these remember the variables, which actually were used in that multidimensional data. So, we have this fifteen plus four nineteen such variables which had led to the all the previous visualizations in the data now looking at this also, one get to nice interpretation of the clusters that have been formed in the data? For example, if you look at this particular region, that was the region corresponding to the high. Let me just go back to that particular figure, this was the region where we had the high profit and high business per employee. And if you look at the component plain visualizations, you have this business per employee profit per employee. So, this on a gray scale, this is on a higher scale and that is a clearly a demarcating factor as far as the cluster that is formed on those banks.

Now, if you look at another variables and this is net non performing asset. So, if these institutions were good from the point of view of business per employee, profit per employee on a high scale. Then also good from the point of view from this non performing asset keeping it down to a very low level right, we will also look at another example, which is a tie we are looking at socio economic development of world economies. Now objective in such an application is to cluster world economies, according to the level of their economic or socio economic development.

(Refer Slide Time: 35:54)



## Aspects of Economic Development & Economic Fundamental Indicators

*Income level-GNP per capita at PPP

*Growth of economy-GDP growth rate

*Level of investment-Gross domestic investment (% of GDP)

*Inflation-GDP deflator

*Structure of output-Agriculture & Industry value added (% of GDP)

*Openness of economy-Export of goods & services (% of GDP)

*Role of government-General government consumption (% of GDP)

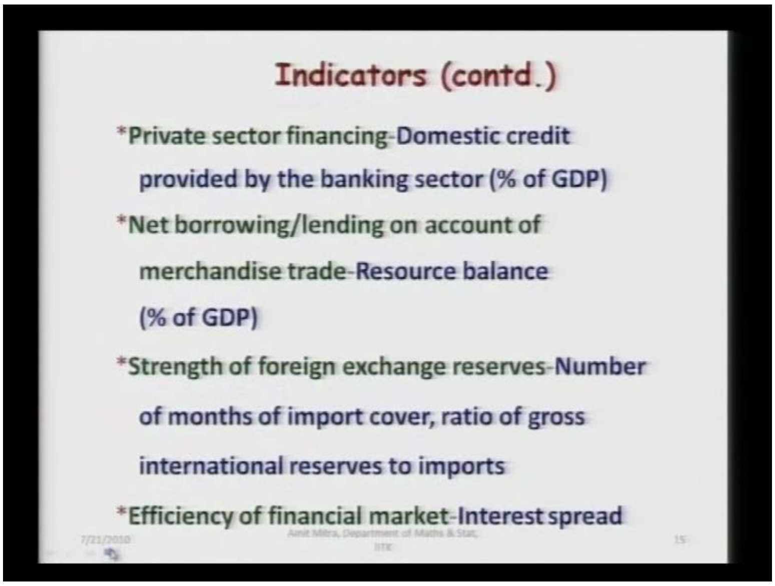
7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 14

In this example, we are trying to look at clusters of economies according to their levels of economic or socio economic development. So, what is the data? The data structure is like that we have a world economy that is countries basically. So, each country now in the previous example, we were looking at financial status of financial institutions. There were a different set of variables on which the financial health of a financial institution is determined. When we are looking at countries. there is a different set of variables naturally, but the data is still multidimensional.

We look at the following aspects of economic development and economic fundamental indicators. For example, for each of the countries in this particular study we look at income level which is given through the gross national product per capital at purchasing power parity rate growth of the economy characterized by the gross domestic product GDP growth rate the level of investment of a particular country measured by gross domestic investment as percentage of the GDP. Then inflation measured through GDP deflator then structure of output agriculture and industry value added percentage of GDP both as percentage of GDP.

Then as a measure of openness of the economy, we look at the export of goods and services which is taken as a percentage of GDP once again. Then, role of government the general government consumption percentage of GDP.

(Refer Slide Time: 37:14)



## Indicators (contd.)

*Private sector financing-Domestic credit provided by the banking sector (% of GDP)

*Net borrowing/lending on account of merchandise trade-Resource balance (% of GDP)

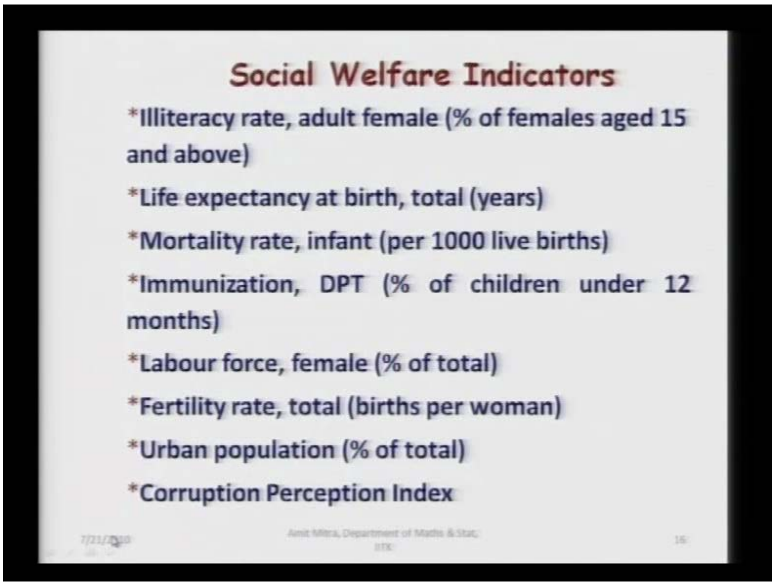
*Strength of foreign exchange reserves-Number of months of import cover, ratio of gross international reserves to imports

*Efficiency of financial market-Interest spread

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 15

Then, the letter indicators are private sector financing domestic credit provided by the banking sector measured as a percentage of GDP net borrowing lending on account of merchandise trade. that is, what is measured through the resource balance. Then strength of the foreign exchange reserve that is the number of months of import cover of that particular country as a and the ratio of gross international reserves to imports.

Efficiency of the financial market measured through interest spread of that country. In addition to these previous indicators, which basically are economic fundamental indicators. (Refer Slide Time: 37:52)



## Social Welfare Indicators

*Illiteracy rate, adult female (% of females aged 15 and above)

*Life expectancy at birth, total (years)

*Mortality rate, infant (per 1000 live births)

*Immunization, DPT (% of children under 12 months)

*Labour force, female (% of total)

*Fertility rate, total (births per woman)

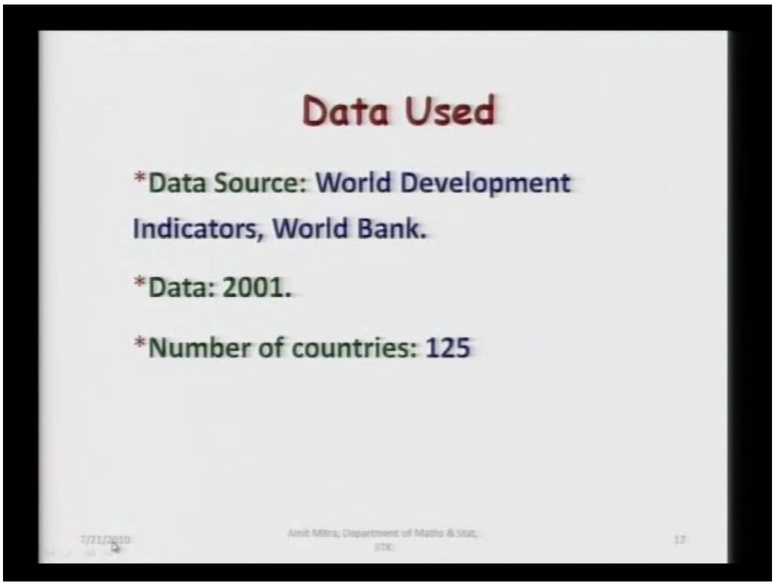
*Urban population (% of total)

*Corruption Perception Index

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 16

We also consider some social welfare indicators, because the health of a particular country is also looked upon when one is also concerned about the social welfare level of that particular country. Now these are standard variables which are used in order to measures social welfares of various countries illiteracy rates, adult female, life expectancy at birth and various other labor force female fertility rate, urban population, corruption perception index and so on.

(Refer Slide Time: 38:22)



## Data Used

*Data Source: World Development Indicators, World Bank.

*Data: 2001.

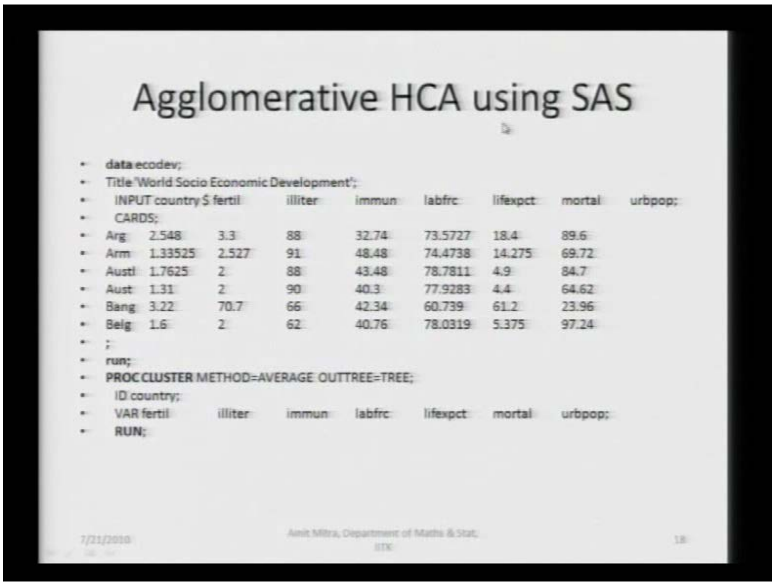
*Number of countries: 125

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 17

The majority of the data has been collected from world development indicators. I forgot to tell you, what is the source of the data for the previous application. The source of the data reserve bank of India the data source for this is the world development indicators of say indicators publication of the World Bank. So, the data corresponding to 2001 have been chosen and the number of countries that is the cases is 125.

What is the dimension of the data? That is, what we have if we have 25 indicators, 25 indicators of economic development and socio economic development and 125 such cases, we have in all 125 by the number of dimensions say 25. That is the dimension of the data for this particular application. Now once again, we are trying to find clusters of economies in the data which countries are similar in nature and sSo on. What we are first looking at an agglomerative hierarchical clustering analysis using SAS?

(Refer Slide Time: 39:24)



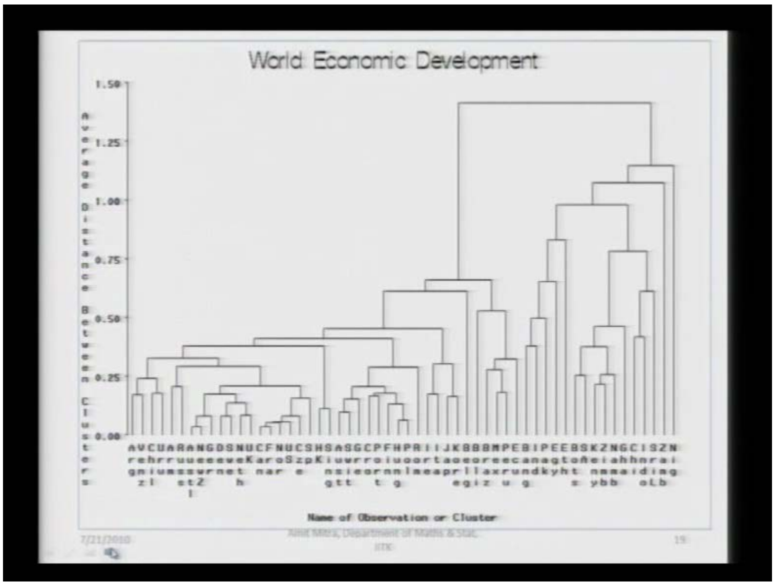
## Agglomerative HCA using SAS

```
•   data ecodev;
•   Title 'World Socio Economic Development';
•     INPUT country $ fertil      illiter     immun      labfrc      lifexpct     mortal      urbpop;
•     CARDS;
•   Arg    2.548      3.3        88          32.74      73.5727     18.4         89.6
•   Arm    1.33525    2.527      91          48.48      74.4738     14.275       69.72
•   Austl  1.7625     2          88          43.48      78.7811     4.9          84.7
•   Aust   1.31       2          90          40.3       77.9283     4.4          64.62
•   Bang   3.22       70.7       66          42.34      60.739      61.2         23.96
•   Belg   1.6        2          62          40.76      78.0319     5.375        97.24
•   ;
•   run;
•   PROC CLUSTER METHOD=AVERAGE OUTTREE=TREE;
•     ID country;
•     VAR fertil      illiter     immun      labfrc      lifexpct     mortal      urbpop;
•     RUN;
```

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 18

In this example, I also give you this particular part of the SAS code which is not given in the previous application. So, this is basically the SAS code this is how the data actually looks like. This is for the indicators, which are actually for the social welfare indicators. These are the countries ARG for example, is Argentina, Australia, Austria, Bangladesh, Belgium and all the data records 125 records are there in alphabetical order.

(Refer Slide Time: 39:53)



7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 19

We will we use a hierarchical clustering method and obtain this hierarchical clustering dendrogram tree for the world economic development.

Now as you can see that, there are various types of clusters that emerge in the data at various levels of resolution. So, broadly speaking there are two types of clusters which comes under these two branches. So, all the countries below this particular main branch are coming in one cluster and all the countries in this particular block here coming under this particular branch comes in one cluster. Now that is a very gross representation, gross clustering of the data. We should actually make the level of resolution finer and then get to various types of groups of countries.

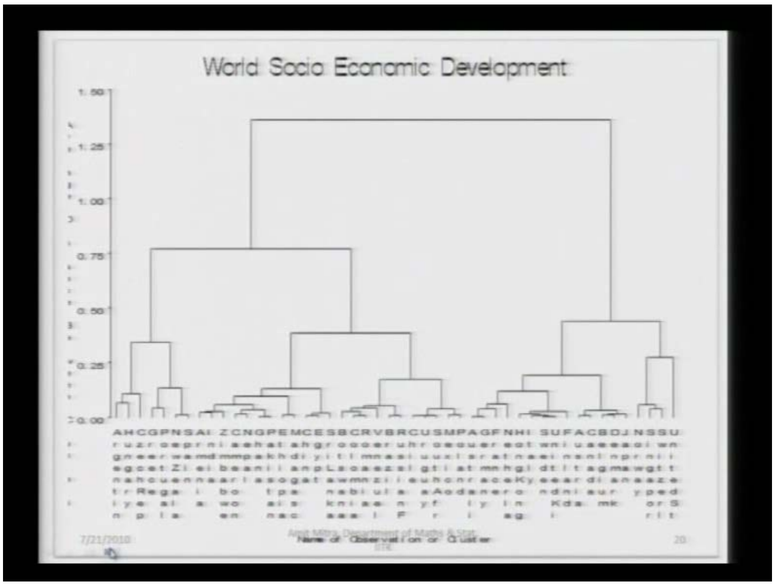
If one looks at carefully at some of these groups here say this particular cluster, here is cluster of mostly developed countries. As you see the countries fere are Australia, New Zealand, Germany, Denmark, all these are abbreviated. GER corresponds to Germany, DEN we have Denmark, this is Sudan, this is Netherland, this is UK, Canada, France, Norway, the United states, the Czech republic, Spain. So, these are basically that cluster of countries which have nicely emerged when we have applied this hierarchical cluster analysis to this particular data.

Now, there are two countries which have a very close nature of association. This is Hong Kong and Singapore now these of course, are two countries which depend very heavily on export and import and hence they have a very close nature of association. Now, there is another cluster of country which is very close which is Austria, Switzerland this is Greece, Persia, Portugal, Finland, Hungary and so on which have a very means good level of social welfare. It is interesting to find out where actually India is sitting? So, as you can see that IND is India and we are sitting right next to Pakistan where PAK Pakistan here.

This is basically the cluster here BANG is the abbreviated form of Bangladesh. So, you see that from this particular branch here. There are two sub branches, one branch containing Bangladesh and India. And the other branch contains Pakistan. At this particular level of resolution, we have the socio economic development level of India Pakistan and Bangladesh almost in the same level. There are other groups, other interesting groups of countries that one gets there is a group of country which is corresponding to say very under developed African nations actually. That also is somewhere here unable to see from this particular figure.

One can find out that quite easily, this is Zimbabwe, Kenya, Namibia, and Ghana and So on. They also Botswana, that is the group of that African clusters. Botswana and all those they belong to this particular close cluster out here.

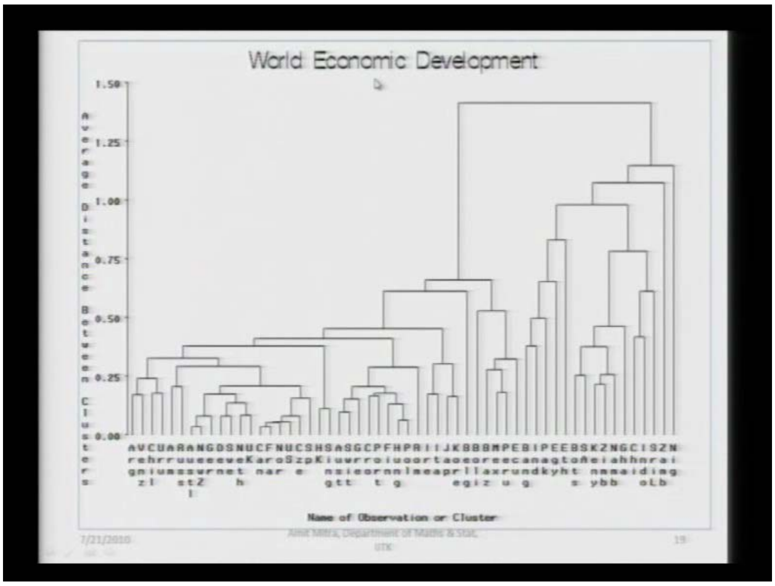
(Refer Slide Time: 43:08)


World Socio Economic Development

7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 20

Now, if we look at the socio economic development data. Now the previous one was just economic development, now this has changed over the years this is the data corresponding to 2001 this picture has changed.

(Refer Slide Time: 43:16)


World Economic Development

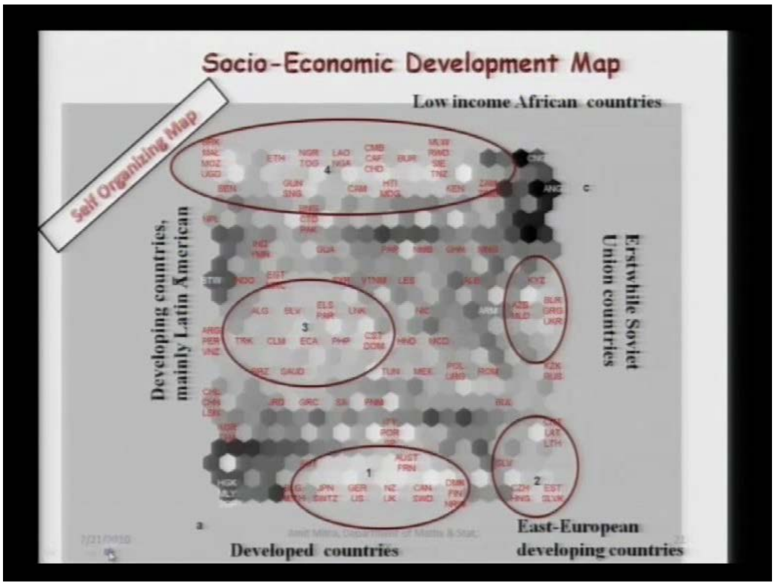
7/21/2010 Amit Mitra, Department of Maths & Stat, IITK 19

This is considering only the economic variables economic fundamental variables in the multidimensional data. And in the next figure, what we are going to see is the overall socio economic development feature of this. Now we see even more clearer clusters emerge. Actually, in the data there are three major clusters this is one cluster, this is the second cluster and this is the third cluster which emerges from this particular data. If you go closer into this level of resolution, you will find that there are mainly two clusters. Now what is this cluster? This is Australia, Germany, France, Netherlands, Hong Kong, Italy, Sweden, United Kingdom, Finland Australia, Austria, Canada and so on.

They are basically the league of developed nations. United States is sitting here and so, will be the United Kingdom and Japan Norway and all those countries are present there. And apart from this, there is also a close association of type of countries. Now where is India is is now belonging here. It is basically in the group of developing countries along with this Armenia, Zimbabwe, and Cameroon, Nepal, Ghana, Pakistan, etcetera. So, this is the group of that socio economic clustering that India belongs to that is coming from the output of this hierarchical clustering method.

There is also a close group of countries, which are which were asked while Soviet Union countries and they broke up after the Soviet Union actually broke into pieces. You will find that too in this particular figure Armenia and those types of countries. They are also located somewhere here, we see a group of countries here, which are basically European countries which are not as developed as this united states, United Kingdom, Spain, New Zealand, Portugal, Greece, and so on. Czech Republic, Hungary, Argentina, also belonging to this particular cluster.

(Refer Slide Time: 45:31)



Socio-Economic Development Map

So, nice type of clusters actually does emerge from real life data. Now as I was talking that statistical cluster analysis is not the only method to look at clustering or rather to look at the type of cluster analysis hierarchical or non hierarchical clustering. That emerges from a data one can also look at self organizing feature map which comes under the philosophy of artificial intelligence. So, for the same multidimensional data the socio economic development data one can obtain a SOM representation of the data and then the clusters in the data can also be found out actually.

So, what we look at in the SOM is that patches of lighter shades gives rise to formation of clusters in those regions. And patches of dark shades actually separate those clusters. And cases which are belonging to those regions are outliers like this are an outlier CNG is Congo, ANG is angular. So, these are outliers in the world economic development socio economic development map and we see a clear formation of a clear cluster out here. Because we have a lighter patch in gray scale and the countries all belonging to that particular patch here form one single cluster. And so, there is a cluster formation out here which is written as cluster number 4.

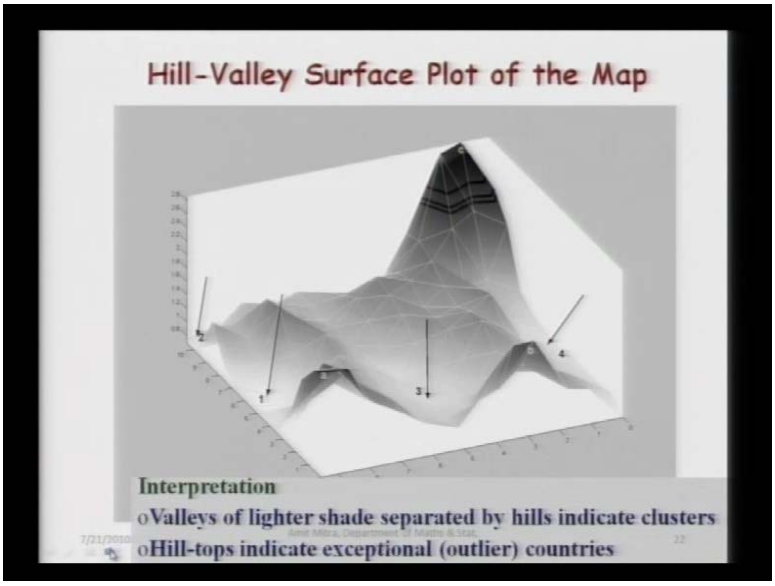
These are all actually low income African countries under developed countries actually. So, all those are belonging to this particular cluster. Now, what is this cluster custer number 3 are developing countries mainly Latin American countries. As you see here El Salvador, Paraguay, Bolivia, and then you will have equator Brazil and so on and Saudi

Arabia etcetera are also present in that particular cluster. Argentina, Peru, Venezuela, they are on the boundary of that particular cluster.

Now, there is a strong cluster formation out here on the right side of this particular map. What are those,? those were erstwhile Soviet Union countries which broke up. Then these are Azerbaijan, Moldova I think MLD Belarus, Georgia, Ukraine, and those type of countries Kazakhstan, Russia are also in the boundary of that. Now we have a very deep cluster here very close compact cluster out here. What are those? Those basically are the countries which are developed countries Austria, France, Denmark, Finland, Norway; these are Scandinavian countries Canada, Sweden, New Zealand, UK, Germany, US, Japan, Switzerland.

Now, on the boundary of that, is also sitting Austria, Belgium, Netherlands, and on the right hand side we have a block of countries which are forming cluster number 2. Which are East European developing countries like Latvia, Curasia, Slovenia, and Estonia, Slovakia, and all those So, this is a self organizing map cluster of the same multidimensional data .as we had seen in the previous example, this self organizing map the previous representation is a is on the two- dimensional SOM plain.

(Refer Slide Time: 48:36)



Hill-Valley Surface Plot of the Map

Interpretation

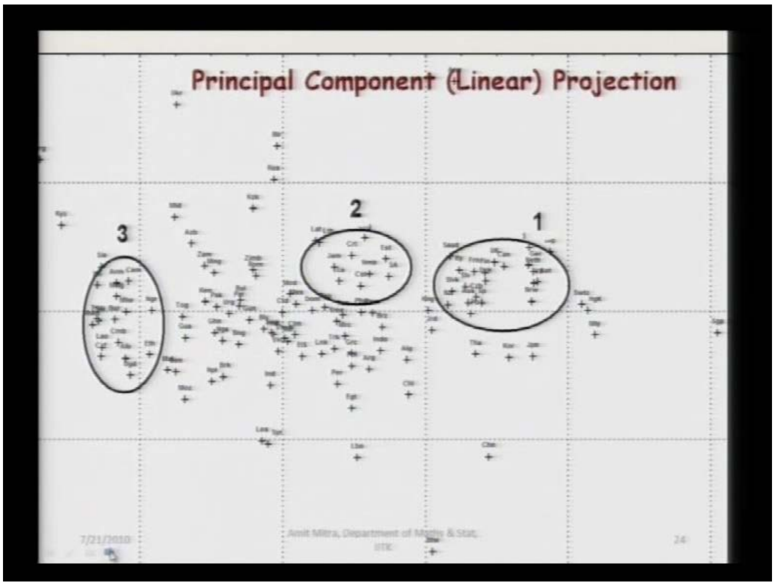
- Valleys of lighter shade separated by hills indicate clusters
- Hill-tops indicate exceptional (outlier) countries

7/21/2010 Amit Mitra, Department of Maths & Stat. 22

One can look at the three dimensional hill valley surface plot of the same map, same clustering tendencies and wherein the clusters are indicated by the valleys. So, this is cluster number 1, which was present in the previous map. In this location this is the cluster of the developed countries. You see that basically is sitting in this particular valley very well separated by a high hill around that particular valley. So, the type of socio economic development pattern is quite different of these countries than the rest of the world economies.

The cluster number 2 which was on the bottom right side here is now represented in this particular corner here. This is a cluster 3; here this is a cluster 4. Here the location of India is somewhere, here India, Yemen, and with Egypt, Morocco, and all those countries sitting here. Now Pakistan, Bangladesh of course, are very nearby to the type of patterns that India is exhibiting. Now, here we once again, we are this interpretation the valleys of lighter shades separated by hills indicated clusters. The hill tops as we had discussed this is a hill top this is another small hill top they are basically cases which are exceptional a typical observations which are associated with outliers in the data.

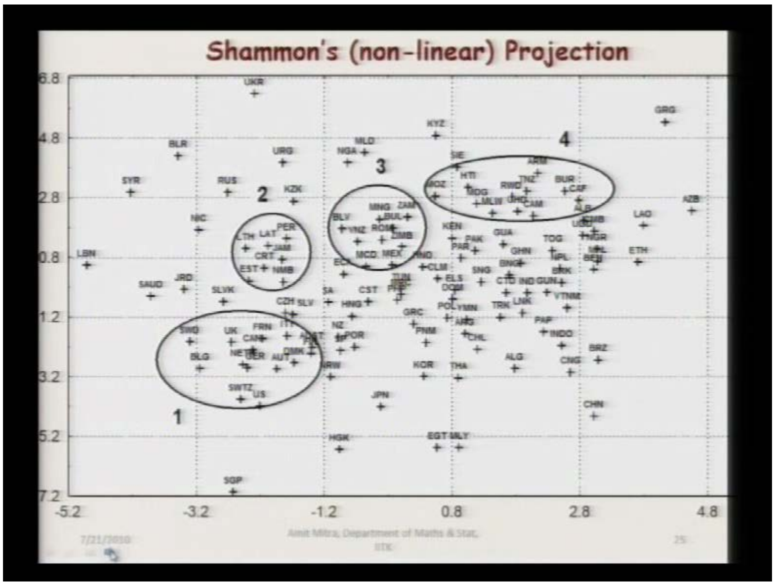
(Refer Slide Time: 50:02)


Principal Component (Linear) Projection

This is the component plain visualization. What type of things that we were discussing in the previous example, also of this clustering tendency. Now, as I said the principal component also is capable of giving us rough clusters in the data this is a two dimensional principal component representation wherein we have to access as the first two principal components in this data. and then if we have the projection of the data in the principal component plain, then we can also make out this rough clusters in the data which are these are of course, subjective clusters that we have here.

Well the clusters that we have formed here they are also quite subjective, but they have quite well demarcated boundaries. So, you have cluster number this and cluster number this being separated by dark patches and similarly cluster this and this separated by these dark patches indicating that members in cluster number one are quite different than the members in cluster number three or two  or this particular point here.

(Refer Slide Time: 51:04)



Shammon's (non-linear) Projection

7/21/2010

Amit Mitra, Department of Maths & Stat, IITK

25

So, we have this as the principal component projection of the same data and we also have a non-linear projection the shaman’s projection of the same data giving us also some rough clusters in the data.

As we have learned that, there are various ways of looking at these clustering patterns in the data. Once we have multidimensional data, we can either look at various options are available. We can apply standard statistical clustered analysis techniques and obtain hierarchical clustering leading us to dendrogram type of trees. And then make out clusters from that dendrogram tree at a desired level of resolution in the data. We can look at a non hierarchical way of looking at that particular data and then forming non hierarchical clusters using a k means algorithm.

And then have clusters in the data, we can adapt a different approach. We can look at principal component analysis, we can look at the type of projection that a principal component analysis is capable of. And then once the multidimensional data is projected on to a lower dimensional visualizable plain say up to three dimensions we can look at that three dimensional or two dimensional projections of the data. And then make out rough clusters in the data or we can actually go completely orthogonal to what the statistical techniques have to offer. We can look at artificial intelligence approach, we can look at a self organizing map.

Now, well I had given in this particular say presentation in of looking at this real life data and we had discussed about the self organizing map feature self organizing map clustering techniques, but we have not looked into the theory part of it. That is beyond this particular multivariate statistical analysis course. These types of techniques are usually covered in courses on data mining or artificial intelligence applications. However, they are capable of giving a nice clustering of multidimensional data. The task basically is same in we are looking at multidimensional data and we are looking at possible clusters that are emerging from the data.

The approach may be different on an artificial intelligence, more we are looking at self organizing map to be the technique and the output of that also leads us to same clustering of the data. Now thus clustered analysis actually is usually which is a form of exploratory data analysis is considered to be a step before we go on to other type of statistical analysis. Say for example, if you are looking at say regression analysis. One would actually try to form a huge data set. One would look at homogeneous patches in the data homogeneous cases in the data and then apply further statistical analysis.

For example, if one is interested in building regression type of models. One will not actually take into that particular one single model various heterogeneous blocks or groups of data. One would like to first have an idea about the number or rather patches in the data heterogeneous blocks in the data. And then for each of those homogenous patches the clusters in the data, one can look at regression models in those different homogeneous patches. So, cluster analysis to me is very important applied multivariate statistical method and that is what we try to learn in theory and also in practice.

So, from next time, next lecture onwards what we will try to look at is another important multivariate applied multivariate technique of discriminate analysis and classification. Thank you.