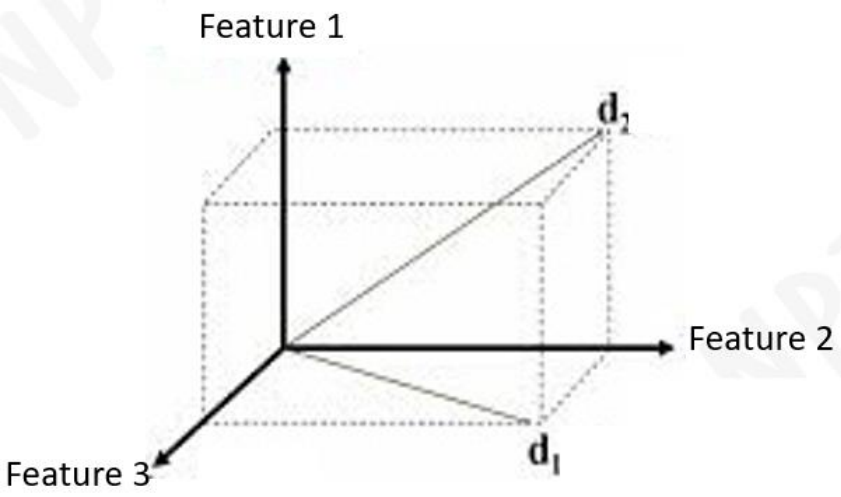
Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 02
Lecture - 07

Lecture 07: Dimensionality Reduction

Hello everyone. Welcome to the recommender system course. Now, we are in module 2 and lecture 7. Today, we will be talking on dimensionality reduction. Moving on, we are continuing our discussion on data pre-processing where we have many activities to perform. One of them is dimensionality reduction. Under this, we are going to talk about principal component analysis and singular value decomposition. So, here as we discussed in the last class, this data reduction has many aspects data cube aggregation, attribute subset selection, numericity reduction, reduction and dimensionality reduction. So, in case of dimensionality reduction, some encoding mechanism is to be decided to reduce the data set size and this has to be done in a automated manner. Let us once again look at the data.

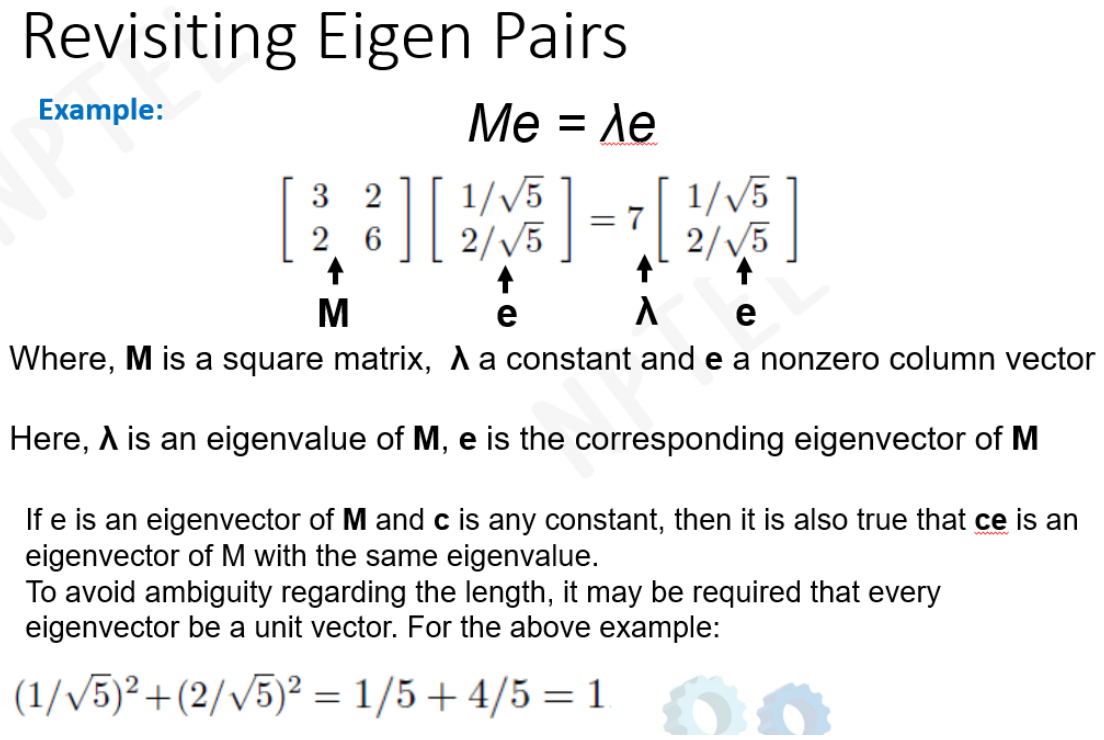
When we talk about the data, specifically multivariable data, it has many dimensions or features. So, this in this particular example, we see the data has 3 features, feature 1, 2 and 3. So, this point D, D 1 or D 2 which is represented in this space are in fact, can be represented in the form of a matrix. For example, here the matrix is for the data point 1, data point 2 will be the row elements and the columns we have feature 1, feature 2 and feature 3. So, this putting these values what we get is a matrix.



So, in this data matrix, now our aim is to represent these features in some manner. So, that instead of this let us say here of course, we have 3 features because we are we wanted to visualize it, but we can go up to n features. So, this n features we would like to reduce and represent in terms of some m number of features or some p number of features where p is less than n. So, how it is to be achieved? Before we talk about how it is to be achieved, let us look at the data that

we just now discussed, how to characterize this data. If you remember while talking about characterizing the data, we understood that this feature thus this features that are n features we can represent in the form of some covariance matrix.

So, if we have n features the covariance matrix is going to be of size n cross n and this covariance matrix is a square matrix as well as a symmetric matrix with the diagonal elements as the variances and individual elements are the covariances. So, now, if we have a covariance matrix or in that sense some kind of square matrix, we can distinctly characterize the square matrix in terms of two unique things. So, what are they? They are the eigen pairs. So, for example, here we have the data matrix M and this E and lambda are the eigen values and eigen vectors respectively. So, this is your eigen vector, this is your eigen value.



So, how do you define them? If we multiply the data matrix with the eigen vector which it will become equal to eigen value into multiplied by the same eigen vector. So, here M is the square matrix connecting it to the last thing that I told you actually in case we are dealing with the data and characterizing the data our covariance matrix can be one of the such square matrix. So, M is a square matrix, lambda is a constant and E is a non-zero column vector. Now, this lambda we call as the eigen value of M and E is the corresponding eigen vector. Now, if E is an eigen vector of M and C is any constant, then it is also true that C E is an eigen vector of M as well and that too with the same eigen value.

Now, to avoid ambiguity regarding the length what we do? We may try to normalize it, so that if we try squaring and summing up the value becomes 1. So, finding this eigen vectors is fairly easy if you remember you have to construct something called characteristics equation. So, how do you get this characteristics equation? If we go back to the previous slide, if we this is basically M E

minus lambda E equal to 0. So, from this we can get the characteristics equation by multiplying this with E. So, from this characteristics equation where I is the identity matrix, we can get a quadratic equation in case we have a 2 dimensional matrix, but as the size increase we may get certain higher order polynomial.

Finding Eigen Pairs

Example

$M = \begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix}$ ⇒ $M - \lambda I = 0$ ⇒ $\lambda^2 - 9\lambda + 14 = 0$ ⇒ $\lambda^2 - 9\lambda + 14 = 0$

⇒ $\lambda = 7$ and $\lambda = 2$ ⇒ $\begin{bmatrix} 3 & 2 \\ 2 & 6 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = 7 \begin{bmatrix} x \\ y \end{bmatrix}$ ⇒ $\begin{matrix} 3x+2y & = & 7x \\ 2x+6y & = & 7y \end{matrix}$ ⇒ $y = 2x$

⇒ A possible random vector is $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$ ⇒ After normalizing the principal eigen vector is $\begin{bmatrix} 1/\sqrt{5} \\ 2/\sqrt{5} \end{bmatrix}$

For the second eigenpair, we repeat the above with eigenvalue 2 in place of 7. The equation involving the components of e is x = −2y, and the second eigenvector is $\begin{bmatrix} 2/\sqrt{5} \\ -1/\sqrt{5} \end{bmatrix}$

For higher dimension the procedure is analogous

$$M\mathbf{e} = \lambda\mathbf{e}$$

$$\Rightarrow \quad (M - \lambda I)\mathbf{e} = \mathbf{0}$$

$\mathbf{e} \neq \mathbf{0}$ => determinant of $M - \lambda I$ must be 0

The determinant of $(M - \lambda I)$ is an nth-degree polynomial in λ

So, because this is a very simple problem we can solve it and find out the corresponding eigen values and eigen vector. So, here this x this vector x y and this vector x y when multiplied with 7 we can get 2 equations and solving them we can get solving and after normalizing we can get the eigen vector. So, now, this same concept can be now because see from this we were getting 2 values of lambda for 1 value of lambda we got this is our eigen vector and when we change it to 2 because lambda equal to 2 was the second value we got this as our eigen vector. Now, for higher dimensional we can similarly find it out. Now, the question is why are we talking about this eigen vectors because they have some very nice properties which is going to help us in finding out the principal components which in turn will help us in dimensionality reduction.

Now, eigen values of the real symmetric matrices are always real if m is an symmetric matrix then it has exactly n eigen vectors. Now, when we talk about the symmetric matrix try remembering our covariance matrix is a symmetric matrix. Now, any 2 eigen vectors that came from that come from distinct eigen values are orthogonal. So, orthogonal in the sense if we try multiplying those 2 vectors and add them together then the value becomes I mean the if component wise we multiply and the value becomes 0 and each of them are standardized and they have they can be scaled to have length 1. Now, we can stack up all these eigen vectors to represent it in a matrix form where E is the set of all eigen vector and this lambda is a matrix where the diagonal elements are eigen values and rest of the values are 0.

Now, if all eigen vectors are linearly independent then E is invertible that is we get this relationship. Now, suppose M is a symmetric matrix then this E transpose is inverse. Therefore, we have the relationship where M can be decomposed into E lambda and E transpose. Now, connect this because this is because M is a symmetric matrix we can represent it in this form. Now, with this background on eigen values and eigen vectors let us look at what exactly is principal component analysis and what is its purpose.

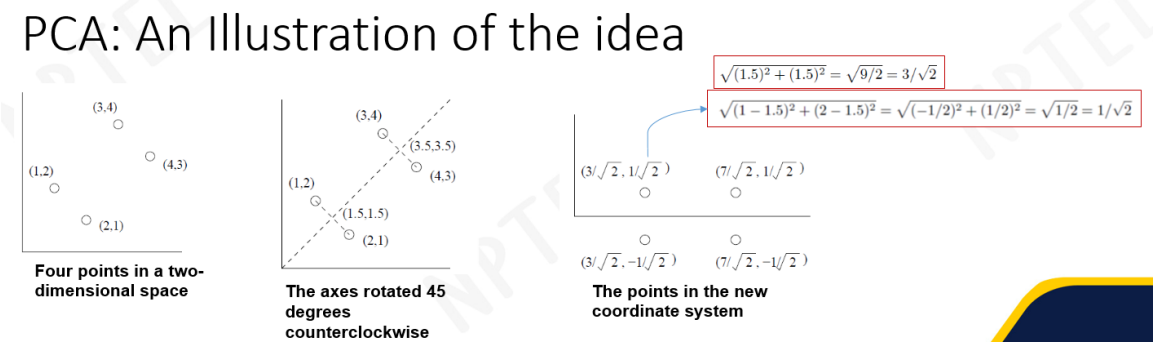
- We stack up all the eigen vectors to get $ME = E\Lambda$

  Where $E = [e_1 \ e_2 \ \ldots . e_n]$; $\Lambda = diag(\lambda_1 \quad \lambda_2 \quad \ldots \quad \lambda_n)$

- If all eigenvectors are linearly independent , E is invertible. That is, $M = E\Lambda E^{-1}$

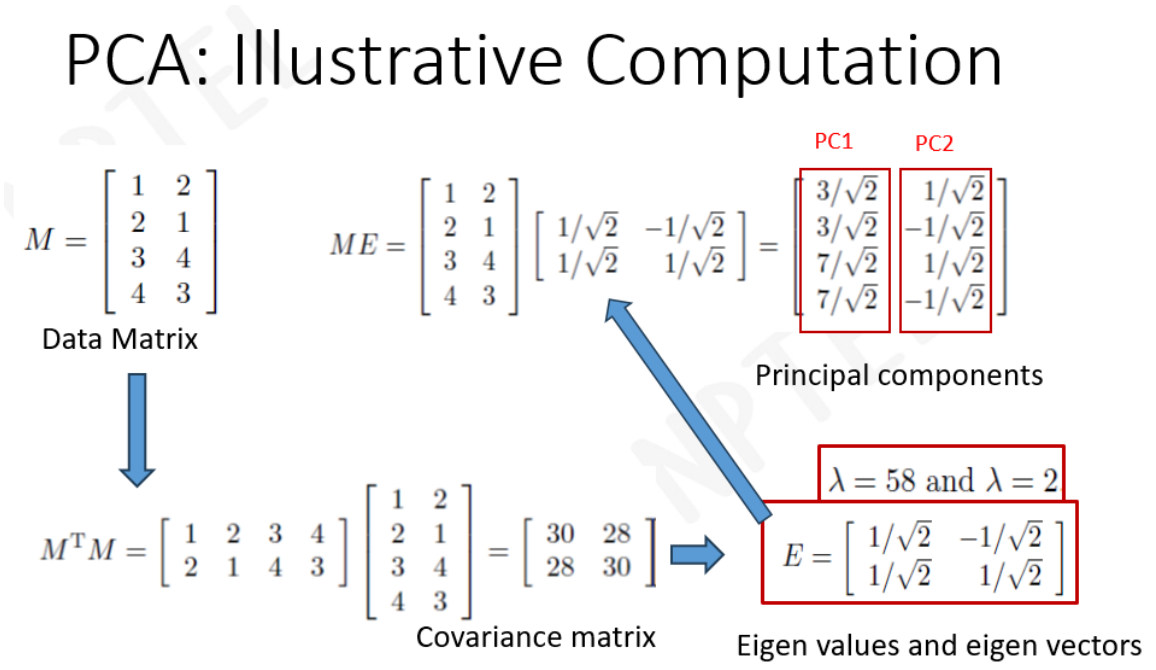
- Suppose M is symmetric matrix, then $E^T = E^{-1}$, Therefore $M = E\Lambda E^T$

Now, the purpose of PCA is to represent a d dimensional data in a low dimensional space without much loss of its content. So, this is the task of dimensionality reduction which is also called feature extraction. So, which means out of let us say d dimensions we can extract some p dimensions where p is less than d. The aim is to discover some new axis which are of course, orthonormal vectors for this data. Now, this eigen vectors of M M transpose and N transpose M turn out to be this orthonormal vectors.

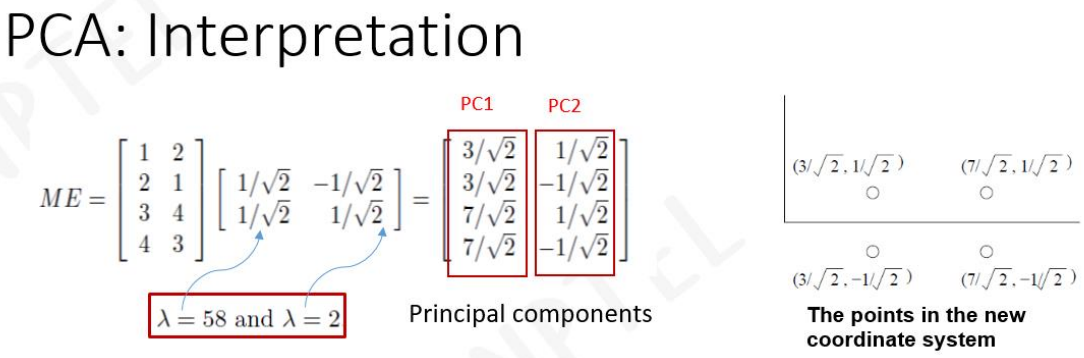


Which orthonormal vectors? The new axis that we are searching for. Now, what is M? M is the data matrix and M transpose is the transpose of that data matrix. Now, the axis corresponding to the eigen vector is the one along which the variance of the data is maximized. So, therefore, if we choose top few eigen vectors then they will be characterizing most of the variance that is discovered in the data. Now, why do we really consider the components with maximum variance? Because variance in the data represents uniqueness of this data. For example, let us say consider some points and mean of all those points is let us say 50. If all the N data are 50 then all the data will be here itself. We have total N numbers and if all the N numbers are 50 they will be here.

So, there is nothing no surprise component in this. So, if you would like to if you have different numbers then all of them will have spread around this 50.

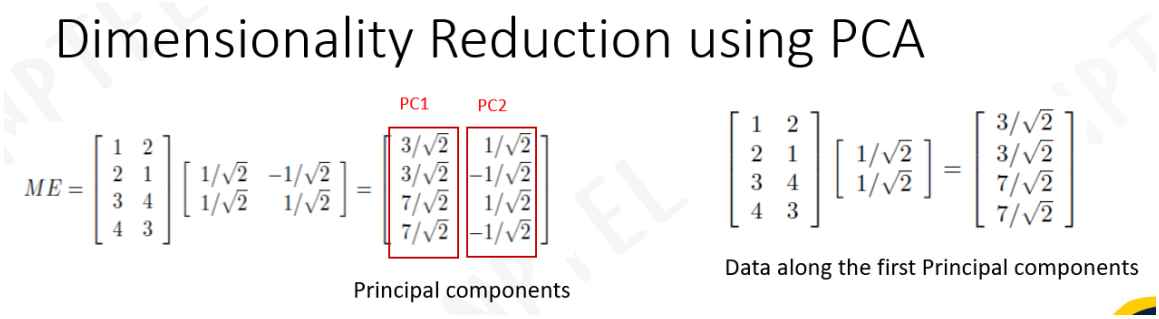


So, each of them will be unique and distinct from 50. So, variability describes uniqueness of the individual elements in a data set. So, this is a case of one dimensional data, but when you have multiple dimensions then variability in all those dimensions are to be considered. Now, the axis corresponding to the second eigen vector the first eigen vector captures maximum variability. And the second eigen vector is the next feature which captures little bit less than less variability than the first eigen vector, but still it is higher than the others.

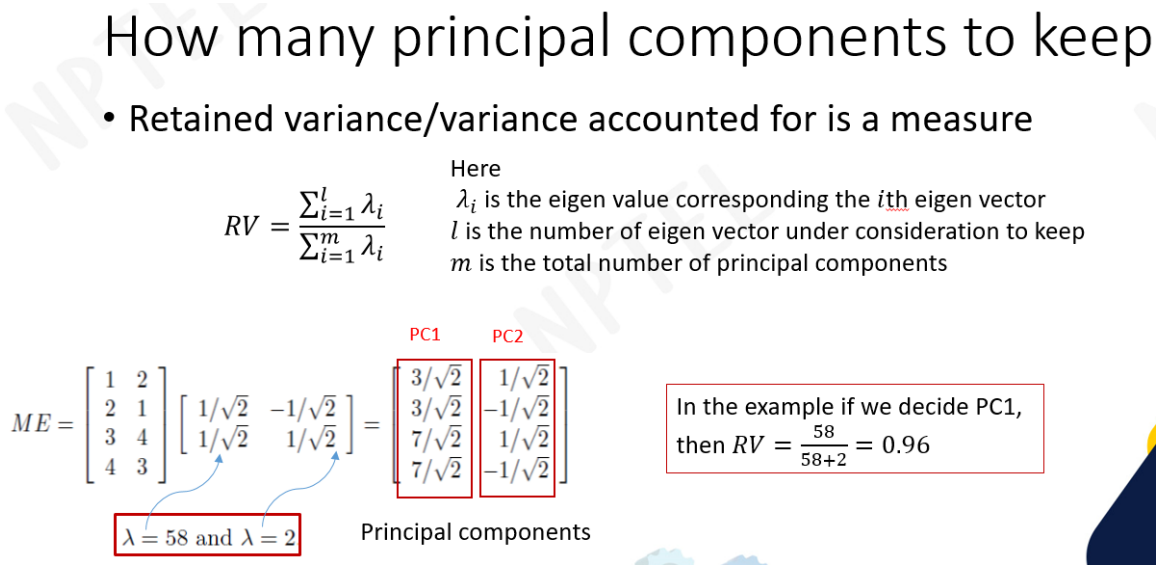


So, which means if you have total N number of total d dimensional vectors which is represented in a new axis using eigen d dimensions with the feature that all these are new axis are also orthonormal to each other and this is just the you have just made a rotation of the original data. In that case if you capture the variability in those principal components which again is same as that of the original dimension of the data, but starting from the first one till the last one the you capture first in the first component you capture maximum variability and in next little bit less

variability and so on. In the last component the least variability. So, as a result if you take top few then the maximum variability that is there in the data which makes the data unique can be represented by top few components. So, this is an illustration look at this here in this in this space in this 2 dimensional space we have 4 points.



So, these are the 4 points 1 2 3 4. Now, what is our aim? Our aim is to discover a new coordinate system. So, that in this new coordinate system sorry new axis maximum variability can be captured in first few components. So, now look at this. Now, suppose this is we consider this axis is rotated by 45 degrees and this is our new axis. So, now, in this new space these are the 4 points again. Now, look at when they were here this was their x component, this was y component, this was x component, this was for this point this was the y component, for this one this is x, this is y, this is y, for this one this is this is x, this is x and this is y. Now, look at the variability of this on this here the variability and here the variability on y axis they are almost looking same not much difference, but now you look at this. This is the x component on this new axis, this is the y component, this is the x component, this is the y component, this is the x component, this is again y component. So, maximum variability is captured on this line this new x.

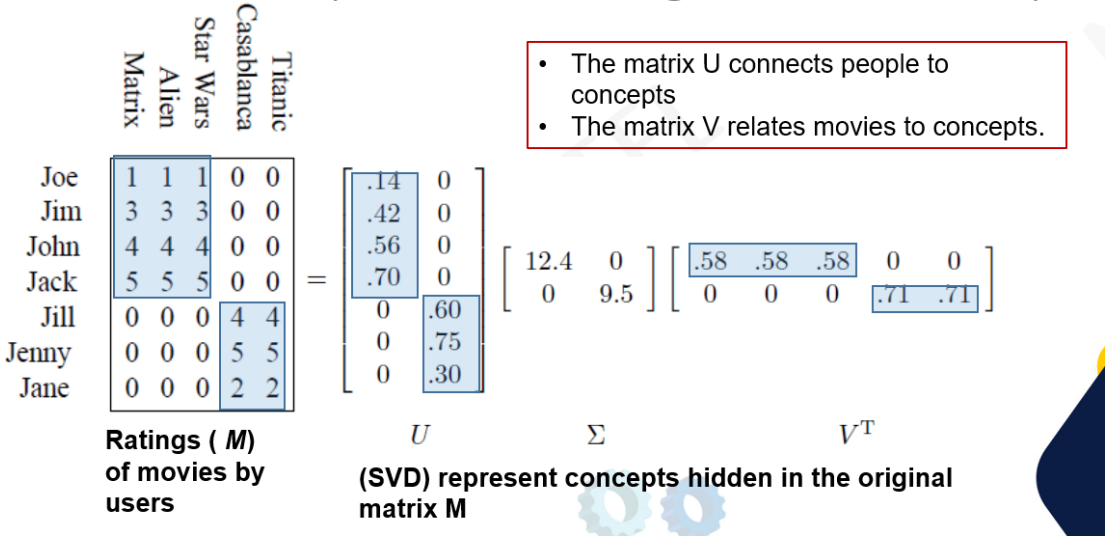


So, if we represent this in this terms of this new rotated axis the points are now in this now look at this is 1.5. So, the point the first point becomes 1.5 square plus 1.2 root of I mean the distance from this. So, this is in 3 by root 2 and similarly this one this one this one is the distance between this point that is this point is 1.5 1.5. So, distance from 1 minus 1.5 and 2 minus 1.5 square. So,

that makes it 1 by root 2. So, now, this one which was originally 1.1 comma 2 this is now this point and as I told as I have seen maximum variability is captured on this axis. So, in this new coordinate system the first axis the one corresponding to the largest eigenvalue is the most significant formally the variance of the points along that axis is the largest. The second axis corresponds to the second eigen pair in the which is the next most significant in the same sense and the pattern continues for each of the eigen pairs.

Now, you have to select top few to capture the maximum variability. So, this is an illustrative computation of PCA that we saw that whatever we saw geometrically that we can now compute. So, these are the 4 points we saw 1 2 2 1 then 3 4 4 3. So, for them this is how you compute. So, this computation how do you get these eigenvectors that is why we discussed little bit in the beginning about how to get this eigenvectors and eigenvalues.

Significance of SVD in the context of recommender system: Deriving Hidden Concepts



So, you can now in this following the same procedure that we discussed you can find out the eigenvectors and once you find out the eigenvectors then you get the PC 1 principal component 1 and principal component 2 principal component 1 and principal component 2. Now here look how did we get this eigenvectors and eigenvalues look M transpose M that is your original transpose of the original data matrix and this is the data matrix and this we get the covariance matrix. Typically when we get the covariance matrix this data has to be mean centered I mean it has to be normalized and standardized and because of this assuming that they are already standardized this M transpose M gives you the covariance matrix and this is a symmetric matrix and these are the corresponding eigenvectors and eigenvalues and this eigenvectors how do you get it few slides before first I have shown you how to get the eigenvalues and eigenvectors, but once you get it then you can multiply it with the data matrix ok. What is the data matrix? Data matrix has 4 elements and 2 dimensions this is first dimension this is second dimension. So, naturally your corresponding covariance matrix was a 4 cross 4 matrix and corresponding eigenvectors you also had 2 eigenvectors.

So, you multiply and you got it. Now try remembering the characteristics of this new access system. Now if you have this is these 2 are our 2 principal components then you can see the maximum values are occurring here. So, maximum sorry maximum variability is occurring here ok. So, this is what I told you now because of this new principal components and this is my new access system. So, because of this in new principal components my data is represent in this form.

Now this is the first data point, this is the second data point, this is the third data point and this is fourth data point ok. And the same observation you can make the first access corresponds to the largest eigenvectors and variability is the most ok. So, now, if we want to retain one of the components which one will contain retain the first one that is PC 1 which contains maximum variability. Now for this the lambda was 58 for this lambda was 2 ok. So, for whatever the eigenvalue was highest that one we have to retain.

Significance of SVD in the context of recommender system: Deriving Hidden Concepts

- The matrix U connects people to concepts
- The matrix V relates movies to concepts.

	Matrix	Alien	Star Wars	Casablanca	Titanic
Joe	1	1	1	0	0
Jim	3	3	3	0	0
John	4	4	4	0	0
Jack	5	5	5	0	0
Jill	0	2	0	4	4
Jenny	0	0	0	5	5
Jane	0	1	0	2	2

$=$

$$\begin{bmatrix} .13 & .02 & -.01 \\ .41 & .07 & -.03 \\ .55 & .09 & -.04 \\ .68 & .11 & -.05 \\ .15 & -.59 & .65 \\ .07 & -.73 & -.67 \\ .07 & -.29 & .32 \end{bmatrix} \begin{bmatrix} 12.4 & 0 & 0 \\ 0 & 9.5 & 0 \\ 0 & 0 & 1.3 \end{bmatrix} \begin{bmatrix} .56 & .59 & .56 & .09 & .09 \\ .12 & -.02 & .12 & -.69 & -.69 \\ .40 & -.80 & .40 & .09 & .09 \end{bmatrix}$$

Ratings (*M*) of movies by users

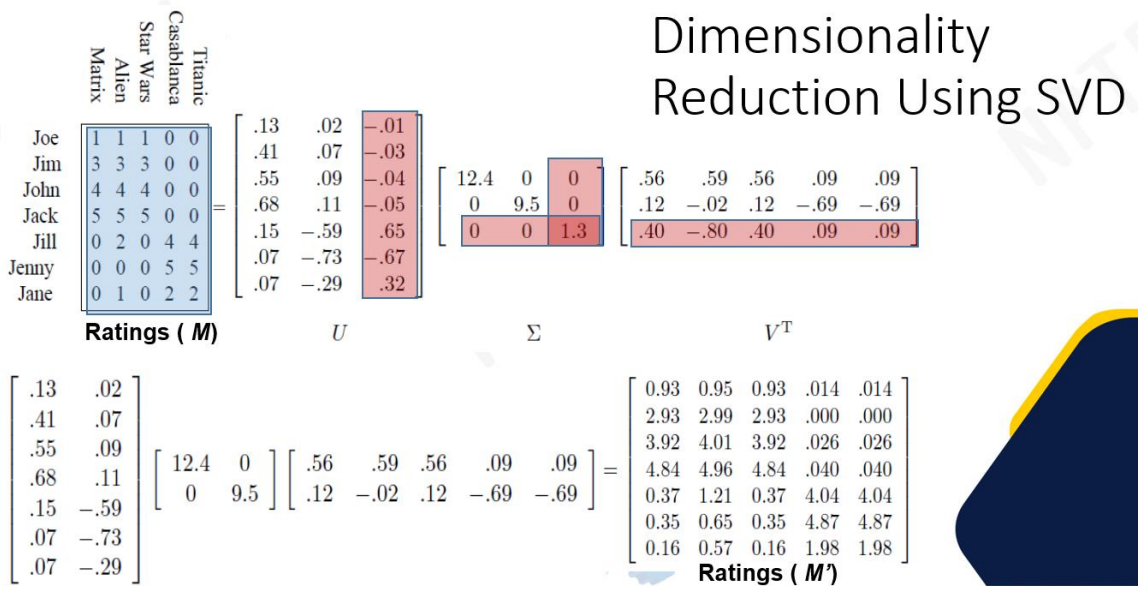
U Σ V^{T}

(SVD) represent concepts hidden in the original matrix M

So, now, if we retain the first component then how much variability is captured that is measured in terms of retained variance and this or the variance accounted for. So, this is basically is represented by this formula. If we have total m number of eigen values, then the denominator represents sum of all the eigenvalues and if we want to retain l of those l principal components then this is the sum of that l. So, in this particular example if we try retaining the first principal component then the corresponding eigenvalue is 58 and sum of these two is 60. So, total we have 96 percent variability is retained.

Next comes our principal component analysis I mean in after principal component analysis let us look at the second methodology that is singular value decomposition. In case of singular value decomposition as the name indicates we have to decompose the data matrix into 3 distinct matrices. So, m is represented by the product of one matrix called U which is m cross where a U is an m cross r column orthonormal matrix. So, when we say it is column orthonormal what do we mean? Each of the component U is a matrix.

So, let us say this is the matrix. So, these are the columns. So, if these column elements are multiplied with each other that is you make the dot product then that dot product turns out to be 0. So, take any 2 columns these 2 columns or these 2 columns or consider these 2 columns. So, if you multiply individual elements and take the dot product then it becomes 0. So, U is an m cross r column orthonormal matrix, V is an n cross r column orthonormal matrix. So, if we consider V transpose then it is a row orthonormal. So, now, we have sigma which is a diagonal matrix. So, here in this U sigma V transpose sigma is an diagonal matrix, diagonal matrix sigma is an diagonal matrix and all the elements which means all the elements except on the main diagonal are 0. So, these elements of this diagonal matrix are called the singular values. Now, let us look at one example and that too in the context of recommender system. So, now, if you look at in this particular matrix this is one rating matrix. In this rating matrix there are 4 people who are rating distinctly 3 movies and there are 3 people who are distinctively rating these 2. If you look at some are action movies, some are some other kind of some romantic movies or something. Now, if we make the singular value decomposition of this and represents it in terms of this U sigma and V transpose we get this U matrix which here what is r this was your this matrix was your 7 rows and 5 columns. So, this is 7 rows and 2 columns this is 2 cross 2 and this is 2 cross 5.



So, m is 7 r is 2 n is 5. So, m cross r r cross r and this is r cross n this is V transpose. So, here we are supposed to note couple of things. What is to be noted? Look these 4 people rated movies of one category and the second 3 rated these movies in one category. So, if you look at this U matrix this actually connects these 7 peoples would with 2 distinct concepts. So, these 3 let us say represents the concept of some category of movie and these 2 represents concept of another category movie.

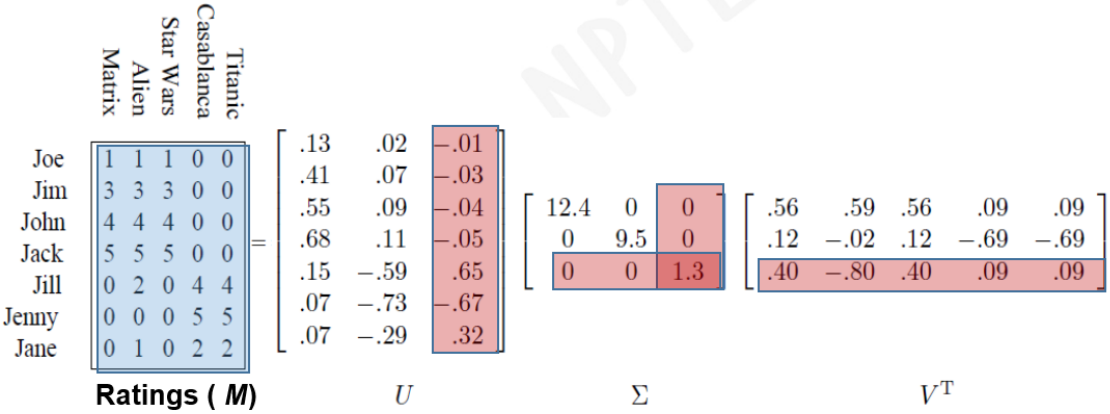
So, this is my concept 1 and this is my concept 2. So, which means instead of representing all the movies which are action type I can now represent in terms of a latent feature through concept 1. Yes, I forgot to tell you in case of PCA also when we consider the first principal component we call it the first latent features the that is one latent feature, second principal component second latent feature because they will be actually making a linear combination of the original features.

So, here we in continuing in the similar manner here also in SVD we have conceptualized these movies into 2 categories and this we are representing by these 2 column vectors. So, also the movies the movies the first 3 movies belong to one category. So, here we have non zero values and the second category it is 0.

$$M = U\Sigma V^T$$

$$MM^{\mathrm{T}}U = U\Sigma^2 \qquad M^{\mathrm{T}}MV = V\Sigma^2$$

So, looking at from this rating matrix through SVD it is possible to represent the data in certain conceptual manner instead of the entire data we represent this as first concept and this is as the second concept. So, these are some of the movies of one genre this is another set of movies I mean these people like movies from another genre. Now, moving ahead suppose there are some people I mean there we distinctly categorized distinctly categorized into 2 groups of people. Now, there is let us say there is some overlap. So, as a result of this overlap we got the value of r which was then 2 now we got this is 7 cross 3 this is 3 cross 3 and this is 3 cross 5.



But still each of the concept which was liked look at this the first 3 people look at the original one the first 4 people were liking this concept. Now, this first 4 people are liking the concept, but now this is little jumble up they are not distinct because there is some rating some movie which is liked by both category of people. So, still the matrix U continuing in the same manner still we can say matrix U connects the people to concepts. So, which means now instead of using all these 3 we can use any 2 of this which maximally describes the concepts belonging to which are liked by this set of people. So, similarly the movies also we can decrease the dimension by considering the first few vectors.

So, this is what I was telling you we can now think of reducing the dimensions by dropping this one and dropping this one. So, if we drop this and try multiplying then what do we get? We get this matrix which originally was having many 0s look here there are so many 0s. Now, all these 0s are most of these 0s are now filled up. However, we may note that if we have a value 1 we are still getting some value which is closer to 1 and if we have value let us say here 5 we are getting

some value which is closer to 5. And additionally what was our original problem in recommender system? Our original problem is if some person has not given an rating let us say in this people are giving rating in a scale of 1 to 5, but these 0s are some of the places where ratings were not given.

Now, we are able to generate some ratings for this 0 this was 0 now this is 0.014 this is also 0.014. So, at least some value we are able to find out. Now, how do we compute SBT? This SBT is again closely relates to the eigenvalues and eigenvectors of M M transpose and M transpose M. The same thing we saw in case of PCA as well. So, this is again connected to the concept of covariance matrix. So, this when we talk about covariance M transpose M was the covariance matrix. So, if we look at this u and sigma M M transpose which was a symmetric matrix into u was u sigma transpose which means u is the eigenvector and sigma square were the eigenvalues. So, if we take the eigenvalues of M M transpose and take the root over then we get sigma.

So, also the case here. Now, the question is what is the dimension of M M transpose M? What was the dimension of M? Dimension of M was in the last example it was 7 cross 5 and what was the dimension of dimension of M M transpose? M M transpose was 5 cross 7. So, what was the dimension of M M transpose? It is a 7 cross 7 matrix and this is a 5 cross 5 matrix. So, of course, 5 cross 5 because this is actually the covariance matrix and you had 5 found distinct dimensions. So, in that case in both the cases if you can see the sigma is actually same. So, which means some of the components in this sigma was a diagonal matrix.

So, some of the diagonal elements towards the end are actually 0 when the matrix size will be more. So, now, the question is how many singular values to retain? Just like in case of our principal component we were trying to find out how much we were trying to retain maximum variability possible. So, here also we try retaining as much as much energy as possible. So, now, here how do you define energy? So, here the total energy is defined to be square of all these values.

So, 12.4 this value 9.5, 1.3 this is the total energy. Now, if you drop this lowest possible value then this is the my new energy level. So, which means in this particular example by dropping the lowest singular value which in turn will be dropping this dimension and this dimension of the user concept matrix and item concept matrix we will be actually retaining 99 percent of the energy and that is a good number. So, all the examples that we have considered here are derived from this massive mining massive data sets and these are some of the other references which I have used. So, this is now we summarize. In this we considered about we talked about the data reduction which involves many tasks along with dimensionality reduction and specifically we considered SVD, APCA and SVD.

As the two approaches for discovering the latent features we saw that through PCA we can transform the data to a new axis where the maximum variability it is retained in the first principal component and when we consider the second component it decreases a bit and so on. So, therefore, by keeping the first few principal components we can actually retain maximum variability. So, the data which was originally let us say in d dimension can now be represented in p dimension where p is less than d. This SVD can also transform the data by connecting the rows and columns to latent features and the rows in the in the example that we considered the rows in

the context of recommender system the rows where are actually representing the users and columns were representing the movies. And we saw that the the people the can be connected to various movie concepts such as genre etcetera and items can also be categorized based on the genre. And we also made it little bit generic and understood that even if we cannot physically ay that they are actually belong to different genre and etcetera, but that example through that example you understood if we consider the first few components of that u and v matrix then those latent features can be used to used for dimensionality reduction. Thank you. With this we wind up this lecture.