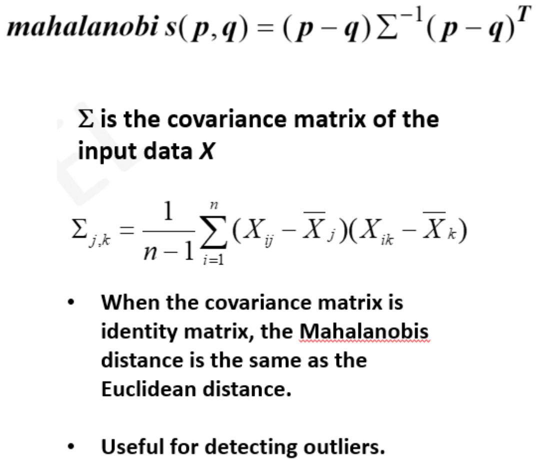
Course Name - Recommender Systems
Professor Name - Prof. Mamata Jenamani
Department Name - Industrial and Systems Engineering
Institute Name - Indian Institute of Technology Kharagpur
Week - 03
Lecture - 12

Lecture 12: Distance and Similarity (Continued)

Hello everyone! We are going to continue on the concept of distance and similarity in continuation with my last lecture. To start with, let me tell you we are talking it in the context of collaborative filtering that too neighborhood based collaborative filtering. However, as I told you in the last class, we need not think that this distance and similarity concept is related only to collaborative filtering. This is a very widely adopted concept applied almost in every area of machine learning. In fact, when we talked about k nearest neighbors and clustering methods, we also talked about distance and similarity concepts. Now coming to the gist of the last lecture that we covered, we tried talking about the concept of distance function.

Then we looked at the distance in binary setting in case of the values which are not numeric and they are binary. And then we tried talking about the binary based, then we tried talking about the quantitative distance concepts and we saw different kinds of norms that we can be using for that purpose. Now we continue with our idea of distance and similarity. So now, we will be talking about our distance measure which is for multivariate numeric data.



$$mahalanobi\,s(\boldsymbol{p},\boldsymbol{q}) = (\boldsymbol{p}-\boldsymbol{q})\Sigma^{-1}(\boldsymbol{p}-\boldsymbol{q})^{\boldsymbol{T}}$$

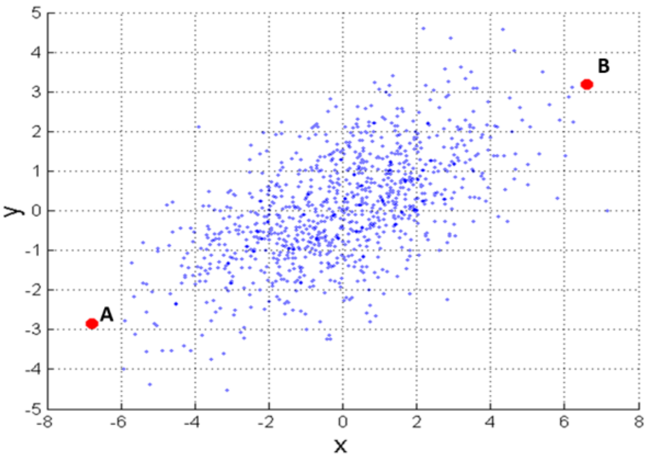
Σ is the covariance matrix of the input data *X*

$$\Sigma_{j,k} = \frac{1}{n-1}\sum_{i=1}^{n}(X_{ij}-\overline{X}_j)(X_{ik}-\overline{X}_k)$$

- **When the covariance matrix is identity matrix, the Mahalanobis distance is the same as the Euclidean distance.**
- **Useful for detecting outliers.**

This distance measure is called Mahalanobis distance. Mahalanobis is a Indian statistician and this particular distance measure is proposed by him. Now in this distance measure which also considers numeric data, the main idea is we have to also consider the distribution from which the data point is drawn. So, this distance is a generalization of finding how many standard deviation away a point P is from the mean of the multivariate distribution D. Now this distance is 0 for a point P at the mean of the distribution and as we go far away from P, this value increase and in the direction of the principal components.

Now looking at this, this can be defined as Mahalanobis distance can be defined as P minus Q where P and Q are two vectors and I mean the two points in the multidimensional plane and all of them have some P number of components. So basically, P is a point with let us say points P 1 to P P and Q is another point Q 1 to Q P and this P dimensional vector, these two points are part of many points from which some distribution is coming up. Now that how do you get the distribution? If you remember, we had the data matrix in which we had n number of observations. So, this n number of observations each were having let us say P dimensions. So, P elements they were having.

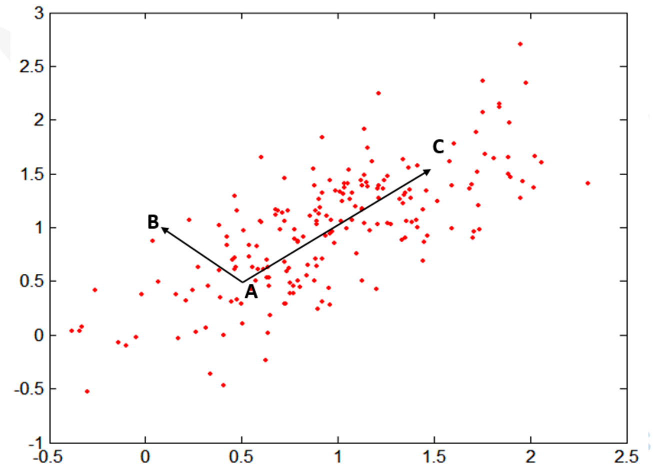


So, from this n observations, we can get its covariance matrix which is a P cross P matrix. P is the number of dimensions here. So, there is because I am also using P as the points, please do not confuse with it. So, P is the point here and in this P that I am talking about is the dimension. So now, rather I will be changing this to let us say this point to x and this point to y and this is x y, this is x y and there are many such vectors x number of vectors.

Now this sigma is the covariance matrix of this input data x and when we compute this covariance matrix because we are drawing it from a sample, we can divide this by 1 by n. This is of course, P. So now, when the covariance matrix is identity matrix, this Mahalanobis distance is same as that of the Euclidean distance. Now, if we look at the points point A and B, the Euclidean distance between these two points is 4 points 14.7 whereas the Mahalanobis distance is 6.

Why so? If we look at this as I told you in the last slide, as I told you in the last slide, actually in this data shows positive correlation among the data points in this two dimension. So therefore, if we try drawing the major axis on which maximum variability lies, we will be getting there, we will be getting a rotation and get this principal component 1 and principal component 2. Now, if we look at these points from the perspective of these principal components, their maximum variability will be captured on this PC 1. As a result, the points will have different representations, different values in terms of this new axis. So, as a result, this distance is less because there will be the mean will be somewhere here.

So, naturally they will be more on this axis, the distance will be they will be more close in terms of this axis. Now, moving ahead, moving ahead, look at this, we have these three points A, B and C. Let us see how this computation is happening. Let us say for from this data, this is the covariance matrix. How do we get this covariance matrix? As I told you from whatever data set you already have with you, from that sample you will be making your covariance matrix.



Now, once you have this covariance matrix which is in this case given to you, take this inverse and this p minus p here the two points when we take A and B, the first point is this that is your p and second point is this that is your q. So, p minus q sigma inverse, then p minus q transpose will be giving you my logis distance values. So, if we look at this point, the distance between A to C is less than that of A to B which if you look in terms of your Euclidean distance, this one looks longer, whereas this one is short. Whereas, because this data is actually lies I mean they are I mean the if you look at the PC 1 and PC 2, if this data is because of its new rotation, it is probably the your PC 1 will be somewhere here and PC 2 will be here. So, in terms of that this new values emerge.

$$mahalanobis(p,q) = (p-q)\Sigma^{-1}(p-q)^{T}$$

Covariance Matrix: $\Sigma = \begin{bmatrix} 0.3 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$

$$\Sigma^{-1} = \begin{bmatrix} 6 & -4 \\ -4 & 6 \end{bmatrix}$$

A: (0.5, 0.5)

B: (0, 1)

C: (1.5, 1.5)

Mahal(A,B) = 5

Mahal(A,C) = 4

So, there are many vector based similarity measures. So, in some situations the distance measures provide a skewed view of the data. Skewed view in the sense the data may be very sparse and sometimes the zeros that are there part of the data are actually not significant. Significant in the sense zeros are not meaningful. In case of asymmetric binary variable this is the case.

Now, in such cases vector based similarity measures are found to be very useful. Most common such vector based similarity measure is your cosine similarity. Let there be two vectors, the first one x with n component, second one y again with n component. So, if we take the dot product then element wise we multiply. Now, what is cosine similarity? Cosine similarity is normalized dot product.

So, this when we normalize this similarity this dot product we are supposed to divide it by the norms of both the values. So, both this x and y this is of course, capital Y capital Y. So, norm of x individual element squared and taken root over individual element squared and took a root over and this gives you the vector based cosine similarity. This vector based cosine similarity also have another variation adjusted cosine similarity. So, this is one example here we have two vectors d 1 and d 2 with elements only zeros and ones.

Two vectors: $X = \langle x_1, x_2, \cdots, x_n \rangle$ $Y = \langle y_1, y_2, \cdots, y_n \rangle$

Dot product of two vectors: $sim(X,Y) = X \bullet Y = \sum_i x_i \times y_i$

Cosine Similarity = normalized dot product

The norm of a vector $\|X\| = \sqrt{\sum_i x_i^2}$

The cosine similarity

$$sim(X,Y) = \frac{X \bullet Y}{\|X\| \times \|y\|} = \frac{\sum_i (x_i \times y_i)}{\sqrt{\sum_i x_i^2} \times \sqrt{\sum_i y_i^2}}$$

Now, element wise these two these two these two these two are getting multiplied and this is the norm. Norm individual squares of each element taken the norm and you divide. So, this is your cosine similarity, this is your dot product and this is your cosine similarity. Correlation can also be used as a similarity measure. So, about correlation we have already studied when we talked about the statistical foundation.

Correlation basically relates two different attributes. Now, when we talk about correlation as a similarity measure we are now not talking about the attributes we are talking about two vectors let us say x and y x 1 to x n y 1 to y n two vectors and this will have some mean x bar and this will have mean y bar. So, given this mean individual elements of this will be a mean will be subtracted from individual elements and they will be getting multiplied and you take the root over. This is the computation of Pearson correlation coefficient. This is the first vector x second vector y these are the mean in the sum of all this divided by n sum of all this divided by n these are the values where x is getting subtracted from mean and this is y is getting subtracted from mean.

Now, this value at individual component level is getting multiplied and this sum. So, this makes the first part of this the numerator. Now, coming to the denominator what was the denominator? Denominator was taking the square summing is taking the square of the individual from where mean is getting subtracted and summing them up and finally, taking the root over. So, these are the squares of individual elements when you add them up you are getting these values and your correlation coefficient is basically root over of this value divided by this into this root over. So, it turns out to be this much.

$$D1 = [1, 1, 1, 1, 1, 0, 0]$$
$$D2 = [0, 0, 1, 1, 0, 1, 1]$$

$$D1 \cdot D2 = 1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 1 + 1 \times 0 + 0 \times 1 + 0 \times 1 = 2$$

$$\|D1\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2} = \sqrt{5}$$

$$\|D2\| = \sqrt{0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 1^2} = \sqrt{4}$$

$$similarity(D1, D2) = \frac{D1 \cdot D2}{\|D1\|\|D2\|} = \frac{2}{\sqrt{5}\sqrt{4}} = \frac{2}{\sqrt{20}} = 0.44721$$

Now, come to the second approach of correlation based coefficient. So, this is called Spearman's rank order correlation coefficient. The benefit of Spearman's rank correlation coefficient over Pearson's correlation coefficient over the Pearson's correlation coefficient is it is specifically suitable for ordinal scale. Whereas, Pearson though it is also used in case of ordinal scale, but that is most suitable when the data is in some kind of continuous scale. But if the order matters then the numeric value can also be represented in terms of order and you can find out.

And moreover one very important thing is that here in case of Pearson's when you take the correlation basically you talk about a linear relationship. Whereas, in case of Spearman's nonlinearity can be captured pretty well. So, if the variable is not in terms of rank then you have to first convert them into rank. So, this is basically once you rank the variables this is rank of xi that is individual data points minus the average rank. Individual rank minus the average rank for both the variables.

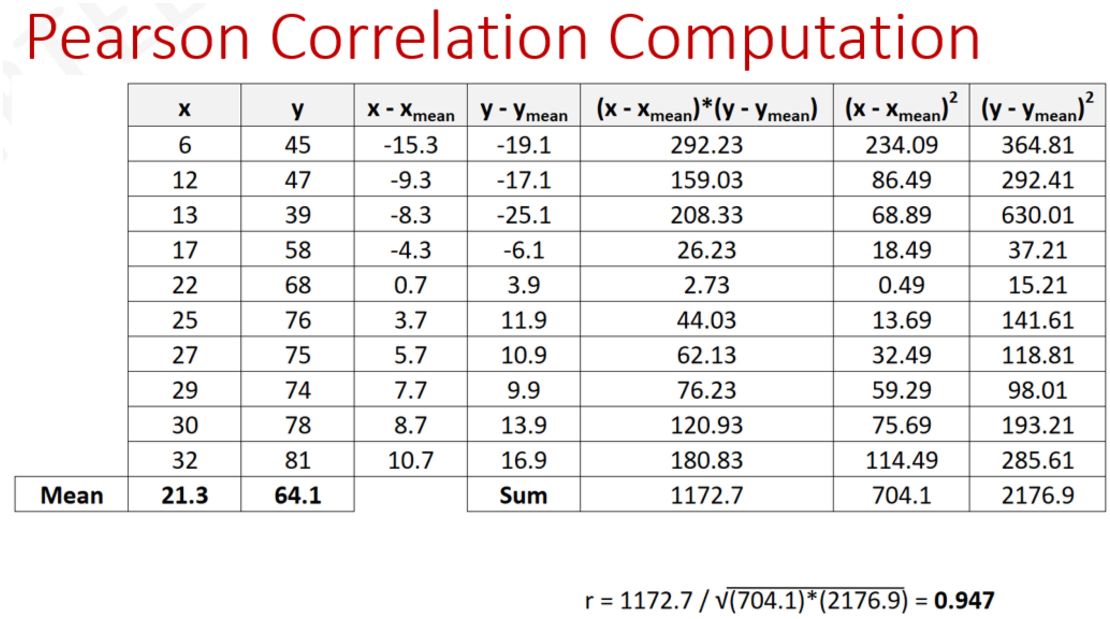
Pearson Correlation $$r = \frac{\sum_{i=1}^{n}\left((x_i - \bar{x})(y_i - \bar{y})\right)}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

So, x is one but x sorry both the both the observations x is one observation y is another observation. And just like your Pearson here also individual rank minus average rank getting squared for both the variables getting added these squares are getting added up and then multiplied and you take the root over. And it can it is also observed that this Spearman's row can also be represented in terms of this formula. Where n is the number of attributes and di is the square of the rank. Now look at this di so di square is the square of the rank.

Now look at this here we have two objects x and y both of them are not in ordinal scale but they are numeric values. So, had they been in ordinal scale this rank would have directly computed but now we have to do little bit sorting. So, if you sort this variable x what is the order what is the lowest what is the highest value highest value is 80. So, rank is 1. What is the second highest value 76 rank is 2.



Pearson Correlation Computation

	x	y	x - x_{mean}	y - y_{mean}	(x - x_{mean})*(y - y_{mean})	(x - x_{mean})^2	(y - y_{mean})^2
	6	45	-15.3	-19.1	292.23	234.09	364.81
	12	47	-9.3	-17.1	159.03	86.49	292.41
	13	39	-8.3	-25.1	208.33	68.89	630.01
	17	58	-4.3	-6.1	26.23	18.49	37.21
	22	68	0.7	3.9	2.73	0.49	15.21
	25	76	3.7	11.9	44.03	13.69	141.61
	27	75	5.7	10.9	62.13	32.49	118.81
	29	74	7.7	9.9	76.23	59.29	98.01
	30	78	8.7	13.9	120.93	75.69	193.21
	32	81	10.7	16.9	180.83	114.49	285.61
Mean	**21.3**	**64.1**		**Sum**	1172.7	704.1	2176.9

r = 1172.7 / √(704.1)*(2176.9) = **0.947**

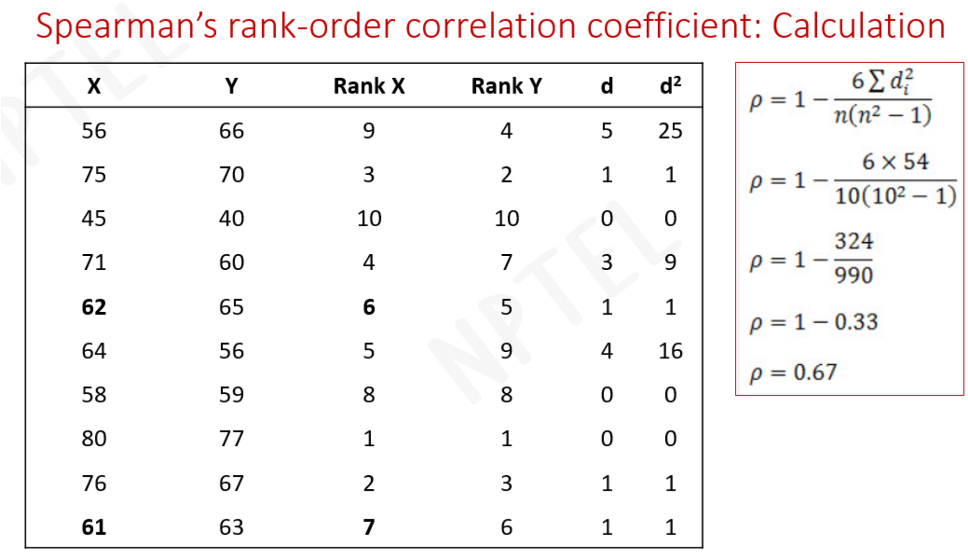
So, in an in an decreasing order start from the highest and keep on sorting first position second position third position and so on. Now look at the second variable in the second variable again 77 is the highest then next one is 70. So, both the variables with respect to their individual data points are ranked. So, while ranking it may so happen that let us say value 77 occurs twice. So, in that case the first position which the first position which was 77 and it was occurring twice then 1 will be divided by 2 and both will have rank 0.

5.5. Similarly, here is another example suppose there would have been two 61s two 61s two 61s the position number of each would have been now the position number of 61 is 6 and 62 is 7 but now both 61 and 61 would have occupied position 6 and 7. So, the individual rank would be 6 plus 7 divided by 2. So, that will be 6.5 6.5, but the next element that is after 7 this is this was the next element 8 next element 8 that will have the value 8 as such.

$$\rho = \frac{\sum_{i=1}^{n}\left(\left(rank(x_i) - \overline{rank(x)}\right)\left(rank(y_i) - \overline{rank(y)}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(rank(x_i) - \overline{rank(x)}\right)^2 \sum_{i=1}^{n}\left(rank(y_i) - \overline{rank(y)}\right)^2}} \qquad \rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

So, once you decide these ranks you have to find out this difference in the ranks sorry I told you that d is the difference between the ranks. So, in the last lecture I told that d is the rank now d is the difference between the ranks. So, now this 9 minus 4 5 3 minus 1 2 3 minus 2 1 that is how we calculated this difference between the ranks d is the difference and you calculate the d square and this Spearman's row is calculated using this formula. So, now let us go to the next correlation based similarity this is called Kendall's correlation this is again specifically designed to capture the association between two ordinal variables. In case of Spearman's even if it is not ordinal we were trying to rank them and order them.

Now here it was specifically for two ordinal variables. Now when it is between two ordinal variables let us say x and y again are two vectors x and y are two vectors let us go back to the example one example. So, this here this in that context we are calling it as x calling it as y. So, they were two vectors. So, what we were doing this Kendall's tau this n was the total number of pairs which how I mean that this we when we come take these combinations we have to make a number of pairs.



Spearman’s rank-order correlation coefficient: Calculation

X	Y	Rank X	Rank Y	d	d²
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	**6**	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	**7**	6	1	1

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2-1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

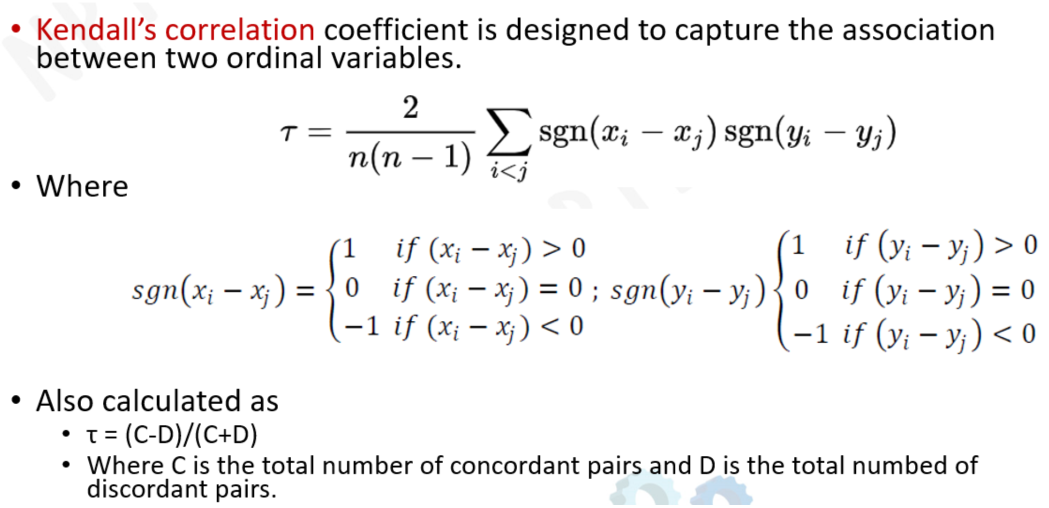
So, there are four elements here and four elements here. So, we take n to be 4. So, now here what we do you take the sign of xi minus xj what is xi xj they are the elements of the vector x and i is less than j. So, which means how many pairs you will be getting from this example there are four elements.

So, this is x1 x2 x3 x4. So, you will be taking pairs and the relationship is i should be less than j. So, x1 is 1 x2 is 3 x3 is 2 and x4 is 4. So, you take the combinations 1 with 3 2 and 4 3 2 and 4 this first 3 then 2 with other 2 that is these 2 then 3 with the last one. So, that makes it total 6 number of combinations. So, similarly for the second vector you are taking all the combinations.

Now, after you take all the combinations we are supposed to see are both these vectors provide. Look here in this particular example we are specifically talking about the four item ratings given by two customers. So, customers when the customers are giving rating this is for the first item there are there are four items here and two customers let us say customer R1 and R2 are giving and there are four items R1 and R2 and these are the ratings. So, item 1 item 2 item 3 item 4.

So, 1 1 3 4 2 2 4 3. So, same customer has given item 1 rating 1 to item 3 it is giving rating 3 and second customer given same 1 to item 1, but 4 to item 2. So, is there any difference in their rating behavior when this person is giving 1 giving 1 and 3 while comparing two items this person is giving 1 and 4 comparing while comparing two items. So, we take the difference in which there is I mean if there is any difference in rating. So, if this one is higher than this we give it minus 1.

So, that is defined by this sine function. So, if xi minus xj is greater than 0 then it is 1 if it is less than 0 it is minus 1 if they are equal they are 0. So, when two pairs over here. So, here this minus this is negative. So, this is minus this minus this is negative.



- Kendall's correlation coefficient is designed to capture the association between two ordinal variables.

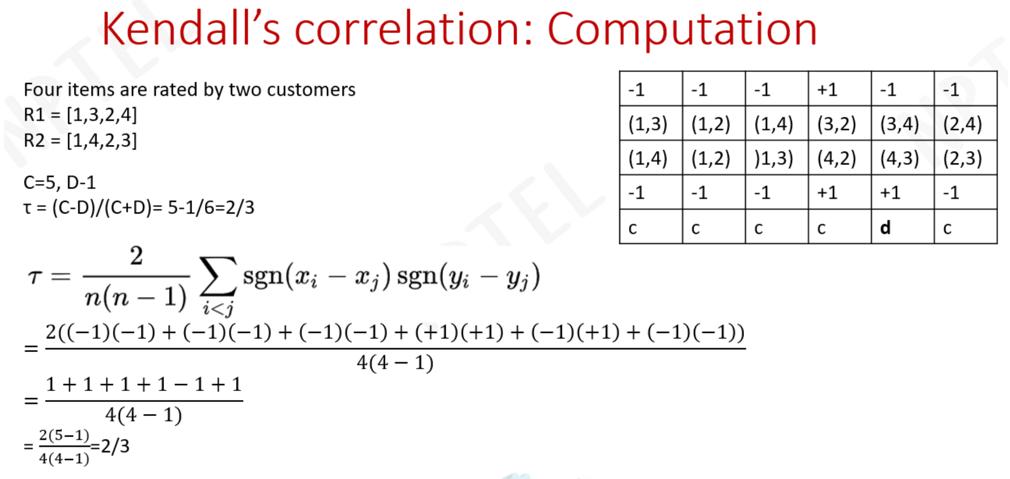
$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$$

- Where

$$sgn(x_i - x_j) = \begin{cases} 1 & if\ (x_i - x_j) > 0 \\ 0 & if\ (x_i - x_j) = 0 \\ -1 & if\ (x_i - x_j) < 0 \end{cases} ;\ sgn(y_i - y_j) \begin{cases} 1 & if\ (y_i - y_j) > 0 \\ 0 & if\ (y_i - y_j) = 0 \\ -1 & if\ (y_i - y_j) < 0 \end{cases}$$

- Also calculated as
  - τ = (C-D)/(C+D)
  - Where C is the total number of concordant pairs and D is the total numbed of discordant pairs.

So, this is minus 1 and so on. So, now, when we multiply this signs of each now look at this here we have a concept of concordance concordant pairs and discordant pairs concordant pairs and discordant pairs. So, what are concordant pairs? We what are we comparing? We are comparing rating behavior of user 1 and user 2 R1 and R2. So, both of them are saying item 3 is better than item 1. So, both of them agree.



Kendall's correlation: Computation

Four items are rated by two customers
R1 = [1,3,2,4]
R2 = [1,4,2,3]

C=5, D-1
τ = (C-D)/(C+D)= 5-1/6=2/3

-1	-1	-1	+1	-1	-1
(1,3)	(1,2)	(1,4)	(3,2)	(3,4)	(2,4)
(1,4)	(1,2)	)1,3)	(4,2)	(4,3)	(2,3)
-1	-1	-1	+1	+1	-1
c	c	c	c	**d**	c

$$\tau = \frac{2}{n(n-1)} \sum_{i<j} \operatorname{sgn}(x_i - x_j) \operatorname{sgn}(y_i - y_j)$$

$$= \frac{2((-1)(-1) + (-1)(-1) + (-1)(-1) + (+1)(+1) + (-1)(+1) + (-1)(-1))}{4(4-1)}$$

$$= \frac{1+1+1+1-1+1}{4(4-1)}$$

$$= \frac{2(5-1)}{4(4-1)} = 2/3$$

So, there is concordance in their opinion. Here also both of them agree that item 2 item 1 sorry item 2 is better than item 1 both of them agree both of them agree both of them also agree here and here they do not agree because here this person is saying this item 2 is better than item 4 because to item 2 it has given the rating 3 to item 4 it has given 4, but here it is just opposite. So, both of them disagree and here also both of them agree. So, using either the original formula by multiplying the signs of all this like minus 1, minus 1, minus 1, minus 1 getting multiplied individually multiply and then divide by n into 4 minus 1. What where the what the 4 is? 4 is coming from the number of what I mean the the dimension of each vector.

So, here the number of item in this particular example. So, that divided by this will be giving you 2 by 3. Similarly, if we consider the concordant and discordant pairs how many concordant pairs are there? 5 concordant pairs where they agree. So, here they agree here they agree here. So, there are 5 places they agree and 1 place they do not agree.

So, c minus d by c plus d is again 2 by 3. So, whether you go by this formula or this formula they lead to the same observation I mean they in fact are same that tau value of tau in this particular example is 2 by 3. So, with this we complete our discussion on distance and similarity. Now, let me tell you one thing even we limited ourselves to the discussion of few distance and similarity measures there are many more and as the nature of the data changes then the type of similarity or distance function we use will also change. For example, in case of let us say time series data probably you have to take a different distance measures. In case the data is some kind of text data probably the distance measure will be different, but anyway right now we have limited to this ourselves to only this kind of measures and in subsequent lecture if we come across a situation where a new distance measure is required we may be discussing on that.

So, to conclude this Mahalanobis distance we started with the consider Mahalanobis distance and we understood that it is a multivariate distance measure that considers the correlation among the variables and while doing so, it tries finding out some kind of centroid of the data and tries to find out the distance from that centroid. Now, when it comes to centroid of the data that centroid is basically the centroid is is measured considering the correlation that exists among the variables. In other words, we represent the data in that new rotated scale in terms of our we we put this in case of we rotate it and in terms we present it in terms of principal components and along this principal components if we measure the distance that turns out to be the Mahalanobis distance. Now, if there is no relationship among the data then this Mahalanobis distance turns out to be the Euclidean distance. So, when there is no relationship in the data means all the variables are not correlated which means your covariance matrix will be a diagonal will have all the ones in the diagonals and rest of the elements non diagonal elements are 0.

In that case this Mahalanobis distance will become the Euclidean distance only. Now, when the data is very sparse and there are number of zeros in the vector there are number of zeros and the zeros in the vector are not significant then vector by similarity measures are used. Now, in case there is high variance across data objects correlation coefficient is accepted to be a good option. In case we are not interested in the linear relationship which is represented by correlation coefficient and we are interested in the rank. So, we can use Spearman's rank order correlation coefficient it is a good measure it is a ranked version of Pearson's correlation coefficient and it is a good measure when there is a non non-linear relationship in the data as well.

And Kendall's correlation coefficient is also a very good measure when both the variables both the data have ordinal variables and in this context we specifically saw one example where we have rating given by two users. Thank you very much.