Contents

Chapter 1

Distance Measures and Standardization

1.1 Cluster Analysis : Proximity Matrices

- A **proximity** is a measurement of the **similarity** or **dissimilarity**, broadly defined, of a pair of objects. If measured for all pairs of objects in a set (e.g. driving distances among a set of U.S. cities), the proximities are represented by an object-by-object proximity matrix

- The joining or tree clustering method uses the dissimilarities (similarities) or distances between objects when forming the clusters. Similarities are a set of rules that serve as criteria for grouping or separating items.

- A proximity is thought of as a similarity if the larger the value for a pair of objects, the closer or more alike we think they are. Examples of similarities are co-occurrences, interactions, statistical correlations and associations, social relations, and reciprocals of distances. A proximity is a dissimilarity if the smaller the value for a pair of objects, the closer or more alike we think of them. Examples are distances, differences, and reciprocals of similarities.

- Proximities are normally symmetric, so that the proximity of object a to object b is the same as the proximity of object b to object a. For example, the distance from Boston to NY is 206 miles, and the distance from NY to Boston is also 206 miles.
  However, in the case of one-way streets, it is possible for distances to be non-symmetric.

- For n items - the full proximity matrix is a symmetic square matrix in which the entry in cell (j, k) is some measure of the similarity (or distance) between the items to which row j and column k correspond.

- The main diagonal contains zeroes. There are $\left[\frac{n}{2}\left(n-1\right)\right]$ distinct distance calculations required.

  For example, for ten cases, 45 distinct measures must be calculated.

1.2 Using Proximity Matrices for Hierarchical Clustering

Using ***nearest neighbour*** linkage, describe how the agglomeration schedule based on the following proximity matrix. With nearest neighbour, a case is assigned to the cluster of the case with which it has the shortest distance. Cluster are also joined on this basis.

Case	1	2	3	4	5	6	7	8	9	10
1	0.00	**4.82**	89.39	85.97	46.26	71.87	56.42	23.75	31.57	11.70
2	**4.82**	0.00	94.24	38.96	**5.55**	35.07	74.52	71.27	61.84	**4.84**
3	89.39	94.24	0.00	57.65	27.27	25.31	20.89	**2.84**	63.50	89.39
4	85.97	38.96	57.65	0.00	**22.94**	**7.13**	70.49	23.09	**12.75**	85.97
5	46.26	**5.55**	27.27	**22.94**	0.00	39.44	17.43	79.22	14.47	46.26
6	71.87	35.07	25.31	**7.13**	39.44	0.00	27.50	30.65	13.34	71.87
7	56.42	74.52	20.89	70.49	17.43	27.50	0.00	91.16	44.92	**6.42**
8	23.75	71.27	**2.84**	23.09	79.22	30.65	91.16	0.00	**3.18**	23.75
9	31.57	61.84	63.50	**12.75**	14.47	13.34	44.92	**3.18**	0.00	31.57
10	11.70	**4.84**	89.39	85.97	46.26	71.87	**6.42**	23.75	31.57	0.00

- The closest pair in terms of distance (2.84) are cases 3 and 8. So this is the first linkage.
- The next closest pair (3.18) are 8 and 9. The next linkage joins case 9 to 3 and 8.
- The next closest pair (4.82) are 1 and 2. So this is the next linkage. [So far (3,8,9) and (2,10)]
- The next closest pair (4.84) are 2 and 10. The next linkage joins case 1 to 2 and 10.
- The next closest pair (5.55) are 2 and 5. The next linkage joins case 5 to 1, 2 and 10. [So far (3,8,9) and (1,2,5,10)]
- The next closest pair (6.42) are 7 and 10. The next linkage joins case 7 to 1, 2, 5 and 10.
- The next closest pair (7.13) are 4 and 6. The next linkage joins case 4 to 6. [So far (3,8,9), (4,6) and (1,2,5,10) All cases are in clusters. This is a 3 cluster solution.]
- The next closest pair (11.70) are 1 and 10. Disregard, because they are already clustered together.
- The next closest pair (19.44) are 4 and 9. This joins cluster (4,6) to cluster (3,8,9) [So far (3,4,6,8,9) and (1,2,5,10). This is a 2 cluster solution.]
- The next closest pairing is 4 and 5. This linkage joins all cases together in one cluster.

1.3 Distance measures

- Distance can be measured in a variety of ways. There are distances that are Euclidean (can be measured with a ruler) and there are other distances based on similarity.
- For example, in terms of geographical distance (i.e. Euclidean distance) Perth, Australia is closer to Jakarta, Indonesia, than it is to Sydney, Australia.
- However, if distance is measured in terms of the cities characteristics, Perth is closer to Sydney (e.g. both on a big river estuary, straddling both sides of the river, with surfing beaches, and both English speaking, etc).
- A number of distance measures are available within SPSS. The ***squared Euclidean distance*** is also a most widely used measure.
- There are various measures to express (dis)similarity between pairs of objects. A straightforward way to assess two objects proximity is by drawing a straight line between them.This type of distance is also referred to as ***Euclidean distance*** (or straight-line distance) and is the most commonly used type when it comes to analyzing ratio or interval-scaled data.

$$d_{Euclidean}(B,C) = \sqrt{(x_B - x_C)^2 + (y_B - y_C)^2}$$

The Euclidean distance is the square root of the sum of the squared differences in the variables values. Suppose B and C were positioned as $(7,6)$ and $(6,5)$ respectively.

$$d_{Euclidean}(B,C) = \sqrt{(6-5)^2 + (7-6)^2} = \sqrt{2} = 1.414$$

This distance corresponds to the length of the line that connects objects B and C. In this case, we only used two variables but we can easily add more under the root sign in the formula. However, each additional variable will add a dimension to our research problem (e.g., with ten clustering variables, we have to deal with ten dimensions), making it impossible to represent the solution graphically.

- The ***Squared Euclidean distance*** uses the same equation as the Euclidean distance metric, but does not take the square root. In the previous example, the squared Euclidean distance between B and C is 2. As a result, clustering with the Squared Euclidean distance is computationally faster than clustering with the regular Euclidean distance.
- We can compute the distance between all other pairs of objects. All these distances are usually expressed by means of a ***distance matrix***. In this distance matrix, the non-diagonal elements express the distances between pairs of objects and zeros on the

diagonal (the distance from each object to itself is, of course, 0). In our example, the distance matrix is an 8×8 table with the lines and rows representing the objects under consideration.

Objects	A	B	C	D	E	F	G
A	0						
B	3	0					
C	2.236	1.414	0				
D	2	3.606	2.236	0			
E	3.606	2	1.414	3	0		
F	4.123	4.472	3.162	2.236	2.828	0	
G	5.385	7.071	5.657	3.606	5.831	3.162	0

- There are also alternative distance measures: The ***Manhattan distance*** or city-block distance uses the sum of the variables absolute differences. This is often called the Manhattan metric as it is akin to the walking distance between two points in a city like New Yorks Manhattan district, where the distance equals the number of blocks in the directions North-South and East-West. Using the points B and C that we used previously, the manhattan distance is computed as follows:

$$d_{City-block}(B,C) = |x_B - x_C| + |y_B - y_C| = |6-5| + |7-6| = 2$$

1.4 Euclidean Distance

The Euclidean distance between two points, x and y, with k dimensions is calculated as:

$$\sqrt{\sum_{j=1}^{k}(x_j - y_j)^2}$$

The Euclidean distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.

1.4.1 Example

Compute the Euclidean Distance between the following points: $X = \{1, 5, 4, 3\}$ and $Y = \{2, 1, 8, 7\}$

x_j	y_j	$x_j - y_j$	$(x_j - y_j)^2$
1	2	-1	1
5	1	4	16
4	8	-4	16
3	7	-4	16
			49

The Euclidean Distance between the two points is $\sqrt{49}$ i.e. 7.

1.4.2 Euclidean Distance

- The most straightforward and generally accepted way of computing distances between objects in a multi-dimensional space is to compute Euclidean distances, an extension of Pythagoras's theorem.
- If we had a two- or three-dimensional space this measure is the actual geometric distance between objects in the space (i.e. as if measured with a ruler).
- The Euclidean distance is probably the most commonly chosen type of distance. It simply is the geometric distance in the multidimensional space. The Euclidean distance between two points, x and y, with k dimensions is calculated as:

$$\sqrt{\sum_{j=1}^{k}(x_j - y_j)^2}$$

- In a univariate example, the Euclidean distance between two values is the arithmetic difference, i.e. ***value1 - value2***. In the bivariate case, the minimum distance is the hypotenuse of a triangle formed from the points, as in Pythagoras's theorem.
- Although difficult to visualize, an extension of the Pythagoras's theorem will give the distance between two points in n-dimensional space.
- The Euclidean distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.
- The Squared Euclidean distance may be preferred to the Euclidean distance as it is slightly less computational complex, without loss of any information.

Euclidean Distance : Worked Example Compute the Euclidean Distance between the following points: $X = \{1, 5, 4, 3\}$ and $Y = \{2, 1, 8, 7\}$

x_j	y_j	$x_j - y_j$	$(x_j - y_j)^2$
1	2	-1	1
5	1	4	16
4	8	-4	16
3	7	-4	16
			49

The Euclidean Distance between the two points is $\sqrt{49}$ i.e. 7.

1.4.3 Squared Euclidean distance

- The squared Euclidean distance is used more often than the simple Euclidean distance in order to place progressively greater weight on objects that are further apart.
- The Squared Euclidean distance between two points, x and y, with k dimensions is calculated as:

$$\sum_{j=1}^{k}(x_j - y_j)^2$$

- The Squared Euclidean distance may be preferred to the Euclidean distance as it is slightly less computational complex, without loss of any information.

1.4.4 Manhattan (City Block) Distance

- The **City-block (Manhattan) distance** is simply the aggregrate difference across dimensions.
- In most cases, this distance measure yields results similar to the simple Euclidean distance. However, note that in this measure, the effect of single large differences (outliers) is dampened (since they are not squared).
- The City block distance between two points, x and y, with k dimensions is calculated as:

$$\sum_{j=1}^{k} |x_j - y_j|$$

- The City block distance is always greater than or equal to zero. The measurement would be zero for identical points and high for points that show little similarity.
- **Example**
  Compute the Manhattan Distance between the following points: $X = \{1, 3, 4, 2\}$ and $Y = \{5, 2, 5, 2\}$, based on four numeric variables.

	Case X	Case Y	Difference	|Diff|
Variable 1	1	5	-4	4
Variable 2	3	2	1	1
Variable 3	4	5	-1	1
Variable 4	2	2	0	0
				6

- The Manhattan Distance between the two points is 6.

1.4.5 Other Measures

- When working with metric (or ordinal) data, researchers frequently use the ***Chebychev distance***, which is the maximum of the absolute difference in the clustering variables values. This distance measure may be appropriate in cases when we want to define two objects as "different" if they are different on any one of the dimensions. The Chebychev distance is computed as:

$$\text{distance(x,y)} = \text{Maximum}|x_i - y_i|$$

For B and C, this result is:

$$d_{Chebychec}(B,C) = \max(|x_B - x_C|, |y_B - y_C|) = \max(|6-5|, |7-6|) = 1$$

- **Power distance.** Sometimes we may want to increase or decrease the progressive weight that is placed on dimensions on which the respective objects are very different. This can be accomplished via the power distance. The power distance is computed as:

$$\text{distance(x,y)} = (\sum_i |x_i - y_i|^p)^{1/r}$$

Parameter p controls the progressive weight that is placed on differences on individual dimensions, parameter r controls the progressive weight that is placed on larger differences between objects. If r and p are equal to 2, then this distance is equal to the Euclidean distance.

A few example calculations may demonstrate how this measure "behaves."

  * Parameter p controls the progressive weight that is placed on differences on individual dimensions
  * parameter r controls the progressive weight that is placed on larger differences between objects
  * If r and p are equal to 2, then this distance is equal to the Euclidean distance.

- **Percent disagreement.** This measure is particularly useful if the data for the dimensions included in the analysis are categorical in nature. This distance is computed as: distance(x,y) = (Number of xi yi)/ i

- There are other distance measures such as the Angular, Canberra or Mahalanobis distance. In many situations, the ***Mahalanobis distance*** is desirable as this measure compensates for ***multi-collinearity*** between the clustering variables. However, it is unfortunately not menu-accessible in SPSS.

1.5 Standardizing the Variables

Generally, it is good practice to transform the dimensions of all variables so they have similar scales.

- Note that Euclidean (and squared Euclidean) distances are usually computed from raw data, and not from standardized data. This method has certain advantages (e.g., the distance between any two objects is not affected by the addition of new objects to the analysis, which may be outliers).
- However, the distances can be greatly affected by differences in scale among the dimensions from which the distances are computed.
- For example, if one of the dimensions denotes a measured length in centimeters, and you then convert it to millimeters (by multiplying the values by 10), the resulting Euclidean or squared Euclidean distances (computed from multiple dimensions) can be greatly affected (i.e., biased by those dimensions which have a larger scale), and consequently, the results of cluster analyses may be very different.
- If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values.
- In this example, both variables are measured on the same scale, so thats not much of a problem, assuming the judges use the scales similarly. But if you were looking at the distance between two people based on their IQs and incomes in dollars, you would probably find that the differences in incomes would dominate any distance measures.
- Variables that are measured in large numbers will contribute to the distance more than variables recorded in smaller numbers.
- In the hierarchical clustering procedure in SPSS, you can standardize variables in different ways. You can compute standardized scores or divide by just the standard deviation, range, mean, or maximum. This results in all variables contributing more equally to the distance measurement. Thats not necessarily always the best strategy, since variability of a measure can provide useful information.

1.6 Example: Motivation for Standardized Distance

Let us consider measuring the distances between between two points using the three continuous variables pollution, depth and temperature. Let us suppose that a difference of 4.1 in terms of pollution is considered quite large and unusual, while a difference of 48 in terms of depth is large, but not particularly unusual. What would happen if we applied the Euclidean distance formula to measure distance between two cases.

Variables	case 1	case 2
Pollution	6.0	1.9
Depth	51	99
Temp	3.0	2.9

Here is the calculation for Euclidean Distance:

$$d = \sqrt{(6.0 - 1.9)^2 + (51 - 99)^2 + (3.0 - 2.9)^2}$$

$$d = \sqrt{16.81 + 2304 + 0.01} = \sqrt{2320.82} = 48.17$$

The contribution of the second variable depth to this calculation is huge ,therefore one could say that the distance is practically just the absolute difference in the depth values (equal to $|51 - 99| = 48$) with only tiny additional contributions from pollution and temperature. These three variables are on completely different scales of measurement and the larger depth values have larger differences, so they will dominate in the calculation of Euclidean distances.
The approach to take here is **standardization**, which is is necessary to balance out the contributions, and the conventional way to do this is to transform the variables so they all have the same variance of 1. At the same time we ***center*** the variables at their means, this centering is not necessary for calculating distance, but it makes the variables all have mean zero and thus easier to compare.

The transformation commonly called standardization is thus as follows (using artificial data):

$$\text{standardized value} = \frac{\text{observed value - mean}}{\text{standard deviation}}$$

Variables	Case 1	Case 2	Mean	Std. Dev	Case 1 (std)	Case 2 (std)
Pollution	6.0	1.9	4.517	2.141	0.693	-1.222
Depth	51	99	74.433	15.615	-1.501	1.573
Temp	3.0	2.9	3.057	0.281	-0.201	-0.557

$$d_{std} = \sqrt{(0.693 - (-1.222))^2 + (-1.501 - 1.573)^2 + (-0.201 - (-0.557))^2}$$

$$d_{std} = \sqrt{3.667 + 9.449 + 0.127} = \sqrt{13.243} = 3.639$$

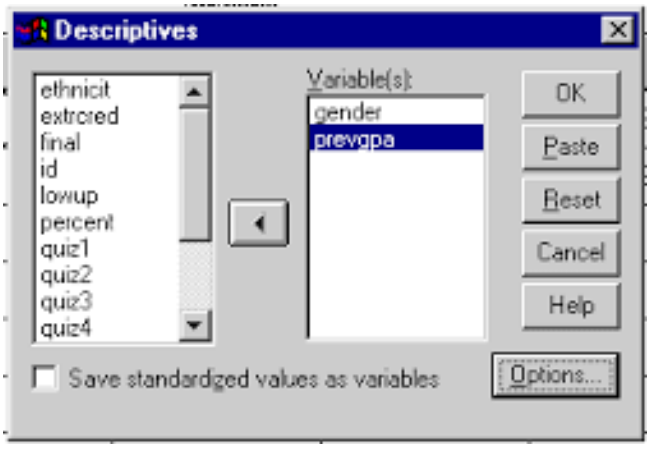
Pollution and temperature have higher contributions than before but depth still plays the largest role in this particular example, even after standardization. But this contribution is justified now, since it does show the biggest standardized difference between the samples.

1.6.1 Logarithmic Transformation

As as alternative to scaling or standardization, the user may opt to use the logarithm of a value, rather than the value itself.

1.7 Standardizing the Variables: SPSS Implementation

- If variables are measured on different scales, variables with large values contribute more to the distance measure than variables with small values.
- Variables that are measured in large numbers will contribute to the distance more than variables recorded in smaller numbers.
- In the hierarchical clustering procedure in SPSS, you can standardize variables in different ways.
- You can compute standardized scores or divide by just the standard deviation, range, mean, or maximum.
- This results in all variables contributing more equally to the distance measurement.
- Thats not necessarily always the best strategy, since variability of a measure can provide useful information.



1.8 R Distance Measures supported by the `dist` Function

- This is not core material for MA4128.
- The `dist()` function computes and returns the distance matrix computed by using the specified distance measure to compute the distances between the rows of a data matrix.
- The distance measures supported by `dist()` are
  - `euclidean` (but not squared euclidean directly)
  - `maximum`
  - `manhattan`
  - `canberra`
  - `binary`
  - `minkowski`.