MULTIDIMENSIONAL SCALING, CLUSTERING, AND NETWORK MODELS

Yi Chen

Columbia University

Midterm Review

MULTIDIMENSIONAL SCALING, CLUSTERING, AND NETWORK MODELS

# Problem 1

Minkowski distance with equals to 1, 2, and 3 correspond to three different models defined by Cross (1965). In a two-dimensional stimulus space:

* The city-block metric (=1) refers to the *discrimination model*, where the strength of reacting to a stimulus different from the correct stimulus on both dimensions corresponds to the sum of the generalization of on both dimensions. Consequently, when the psychological dimensions are **separable (dimensions are “obvious, compelling and clearly distinguishable”)**, city-block metric can be used. For example, all the students in the class give their score about the similarity between a pair of countries. If the similarity is defined based on the dimensions as Carbon emissions per capita and Proportion of population with the education equal or above undergraduate level. Every student strictly follows this rule to measure the similarity. Then, the overall psychological distance between to country can be calculated based on the sum of distance on each of two dimensions since they are all
* The Euclidean distance metric (=2) refers to the *excitation model*, where the generalization gradient decreases evenly around Sr into all directions of the psychological space. Consequently, when the psychological dimensions are **interacting (dimensions are “fuzzy and not distinguished”)**, Euclidean distance metric can be used. For example, all the students in the class give their score about the similarity between a pair of countries. However, this time there is no limitation about what dimension every student should pick. For a given number of dimension (e.g., dimension equals to 2), the definition of each dimension is usually fuzzy and unclear. One possible situation could be: on dimension can be defined as a combination of the cultural and geographic factors, while the other one can be defined as a combination of economics, politics, and technology factors. In the situation like this, using the Euclidean metric is a safe choice since there may have a certain level of interacting effects among different dimensions, and we are not able to assign any unequal weight towards any dimension.
* The maximum metric (=) refers to the *dominance model*, where the strength of reacting to is determined by only that dimension on which and differ most. For example, all the students in the class give their score about the similarity between a pair of countries. This time, three main dimensions are already given: economics factor, politics factor, and geographic factor. When measuring the similarity, students are required to use only **one** out of three dimensions which they believe will be the most obvious factor to distinguish two countries. Thus, when calculating the distance between two countries, we may only need to pick the dimension with the biggest distance to represent the overall distance since only one dimension in reality is dominate the location of the stimuli in the psychological space.

# Problem 2

The goal of this analysis is to identify “the common patterns of test performance among subjects”. The first approach is to use the PCA to generate the low dimension representation of subjects’ response correlations. The second approach is to use MDS algorithms to generate the low dimension representation of Euclidean distance based psychological proximity space using the subject’s responses.   
 For the first approach, the correlation is calculated based on the agreements of pairs of subjects in response. If all the item response of two subjects are the same then the correlation will be 1. If all the item response of two subjects are different then the correlation will be -1. Since the correlation (instead of variance-covariance) matrix is used, PCA is based on the standardized item response for each subject. This is actually not necessary since the item response are scaled dichotomously. Then, the principal components generated represents dimensions which capture the biggest variances in the original item response space through rotating the base. We may could reduce the dimension of 50 into a smaller number of principal components. However, there are two possible issues: (1) since the item response are scored dichotomously, the scale difference is limited in 50-dimension space. I could expect the number of dimensions that could be reduced will be small and the interpretation will be hard to find the unique patterns; (2) Principal components are the linear combinations of all the original item response of the 50 subjects. For each principal component, we can tell which subject contribute most positively or negatively. But it is still hard to clearly identify the pattern.

For the second approach, the Euclidean distance measure the number of items that a pair of subjects have different response. Bigger the distance it is means more different item response two subjects has. The Euclidean distance in this case range from 0 to . Consequently, this data is a dissimilarity metric. Based on the MDS algorithms, we are able to find a low dimension representation of the psychological space through which we can identify the location of each subject. Thus, it will be convenient to find the possible subject pattern groups through identifying the potential cluster in the visualization. The problem similar to the PCA approach is to determine the number of dimensions. Bigger dimension leads to less loss of information and smaller stress but lead to higher difficulty in interpretation.

In summary, we could expect that MDS approach will give an empirically better result since MDS can capture this dichotomous scored item response through transfer it into the dissimilarity distance metric. PCA approach, on the other hand, has the difficulty to handle this “nearly circle” parameter space which does not have some dominant dimensions.

# Problem 3

The data set provided is the co-occurrence data which indicates the number of cases when product and are bought together.

* **Question 1**

The main diagonal of this matrix represents the frequency of the product will be brought more than once in a purchase at the same time. For example, the diagonal element of “expensive tie” is 0 means no subjects in the data buy more than one example tie. Similarly, the diagonal element of “cheap tie” is 13 means 13 subjects in the data buy at least two cheap ties together.

In general, the data is the similarity data. Bigger the number is more frequently people will buy the corresponding pair of products together, which mean these two products are “close”. Smaller the number is less frequently people will buy the corresponding pair of products together, which mean these two products are “far away”.

* **Question 2**

The frequency of purchase is the based on which co-purchase is calculated. If the frequency of purchase of one product is small, the co-purchase of this product with other products cannot be big. Consequently, a better approach could be normalizing the co-occurrence data by calculating the proportion of co-occurrence within all purchase one product has.

Using this data set directly will lead to the dominant influences from the products with high purchase frequency in the MDS algorithms (e.g., Cheap knitwear) since much more obvious distance information could be obtained by these products compared with the products with low purchase frequency. Consequently, the dimension generated by the MDS will be largely defined by the characteristics in the high purchase frequency products.

Consequently, we can image that the high purchase frequency products will be more likely to appear in the outside area of the visualization and has more extreme value on each dimension. One the other hand, the low purchase frequency products will be in the middle area with less distinguishedness between each other.