이 학 석 사 학 위 논 문

Multidimensional Scaling with Application to Twitter Network Data

거리행렬을 이용한 다차원 척도법과 Twitter 네트워크 자료에의 응용

2013년 2월

서울대학교 대학원

통계학과

이 지 희

Multidimensional Scaling
with Application to Twitter Network Data
거리행렬을 이용한 다차원 척도법과
Twitter 네트워크 자료에의 응용

지도교수 임 요 한

이 논문을 이학석사 학위논문으로 제출함

2012년 10월

서울대학교 대학원

통계학과

이 지 희

이지희의 이학석사 학위논문을 인준함

2012년 12월

위 원 장	이 재 용	(인)
부위원장	임 요 한	(인)
위 원	장 원 철	(인)

Multidimensional Scaling
with Application to Twitter Network Data

by

Jihui Lee

A Thesis
submitted in fulfillment of the requirement
for the degree of
Master of Science
in Statistics

The Department of Statistics
College of Natural Sciences
Seoul National University
February, 2013

Abstract

Jihui Lee
The Department of Statistics
The Graduate School
Seoul National University

Multidimensional scaling is applied in various fields such as marketing, genetics, ecology, molecular biology, psychology and social networks. In majority, multidimensional scaling has its main purpose to verify the relationship between individuals by embedding high-dimensional observations on a sphere to points on a lower-dimensional sphere. Simply put, multidimensional scaling makes it possible to look through the large data by illustrating them with a simple plot. In the process of applying multidimensional scaling to the data, we need to define a dissimilarity matrix, which reflects the distance between each pair of the entities. Under the certain restrictions, there can be a variety of distance measures for constructing the dissimilarity matrix. In this paper, we introduce several different distance measures possibly used for multidimensional scaling and categorize those measures so that they can be used in an appropriate circumstance. An application to the actual data has been done with the network data from Twitter. By implementing different types of measures to the specific data, we would like to show the importance of selecting an appropriate distance measure for the data.

Keywords : *distance measures, dissimilarity matrix, multidimensional scaling, social network, Twitter network data.*

Student Number : 2011-20248

Contents

List of Tables

List of Figures

Chapter 1

Introduction

One way of figuring out the difference between two entities is mapping the two and measure how far they are apart from each other. Even if more than two entities exist, the geographical mapping can be an effective way to display how close a selected number of those individuals are to each other. In order to map the individuals, people often use the concept, proximity. There can be a variety of ways to define the proximity, which is also called closeness or dissimilarity. In case of mapping several cities and towns, for example, airline distances can be used as the proximity between cities and towns. Also, a straight-line distance or the shortest traveling distance can be a good example of defining the proximity. The proximity is defined for each pair of the individuals in a study and plays its role as a measure of association; how alike the two are.

Multidimensional scaling (MDS) is commonly used as a visualizing method given a two-way table of proximities. A primary interest of multidimensional

scaling is to verify the relationship between observations and investigate meaningful clusters or a trend from seemingly complicated data, by plotting the records on a lower-dimensional space. By visualizing the proximity matrix, as known as a dissimilarity matrix as well, multidimensional scaling has its strong advantage that plotting the dissimilarity between observations on a lower-dimensional space facilitates the data interpretation, since it enables us to easily recognize the actual relationship between individuals.

There are a lot of previous studies about multidimensional scaling with different points of view. Big issues in the field of multidimensional scaling include defining the dissimilarity, determining the dimension, and interpreting the dimension. In particular, this paper focuses on the definition of dissimilarity, and thus will introduce previous literatures on the issue. The literatures covered by this paper contain important perspectives of selecting the appropriate distance measure and defining the dissimilarity matrix based on the measure.

In this study, we focus on the way to define a dissimilarity matrix. More specifically, we introduce several distance measures to define the proximity for pair-wise comparison of entities such as the Euclidean distance, the Kulczynski distance, and the Kullback-Leibler divergence. With these measures to formulate the dissimilarity matrix, we would like to apply the multidimensional scaling to the data. In other words, this paper explains multidimensional scaling as a way of visualizing the proximity among individuals and applies it to the Twitter followership network data for an actual application.

The rest of the paper is organized as follows. In chapter 2, some reviews of literature will be presented. Since the main point of this paper is to propose

a new measure of dissimilarity, literatures shown in the chapter focus on the distance measures. Chapter 3 describes the classical multidimensional scaling in general and then suggests several measures for dissimilarity. With the theoretical background from chapter 3, chapter 4 covers a multidimensional scaling application to the Twitter network data. In the final chapter, we make a conclusion based on the application in chapter 4 and suggest a possible research in the future.

Chapter 2

Review of Literature

2.1. Measures of Distance between Partitions

There has been an argument that the tendency to introduce information-theoretic measures without considering their appropriateness prevails in many cases when using multidimensional scaling for a clustering problem (Arabie et al., 1973). It was stated that better understanding of partition distance measures for the use of multidimensional scaling is needed for a more accurate analysis by Arabie et al. (1973). According to Aribie et al. (1973), there are many noteworthy examples that show the importance of the choice of a distance measure. It is remarkable that the analysis of multidimensional stimulus conducted by Lockhead (1970) verified that the sum of the distances between each stimulus and its immediate neighbors and the variance of these distances help to detect the possible set of stimuli most accurately. By locating the multidimensional stimuli in a psychological space based on the

separate components, Lockhead succeeded to conduct a research about those seemingly not analyzable in a physical dimension. Moreover, Arnold (1971) pointed out the importance of non-Euclidean metric, which was applied to the multidimensional space of animal names obtained by Henley (1969).

2.2. The Choice: Distance Functions

For every pair of a fixed set of entities, the proximity is a number that reflects how closely the two entities are related; certainly, a greater degree of proximity implies a smaller difference between the two (Shepard, 1962). Basically, the notion that the proximity − also called nearness or distance − is used for multidimensional scaling suggests that we use the data in a form of special structure through transformation. When it comes to the transformation, it is available to convert the implicit differences into the explicit distances through the transformation as long as some requirements such as monotonicity are guaranteed. Then, the transformation compensates and recovers the spatial structure which is contained only latently in the original data, and enables us to illustrate the data structure even more plainly.

By selecting the proper distance function, we can achieve true reduction of the data since reconstructing the distance of separation for each pair of the entities is of using much smaller set of coordinates for plotting points in the Euclidean space. The choice of the distance function varies; it even can be based on those measures already known. Nonetheless, there is one important requirement that should be satisfied by the distance function: monotonicity.

In order to check whether the distance function is monotonic or not, we need to check if the minimum number of dimensions used for mapping the entities in the Euclidean space through the distance function monotonically coincides with

(i) the initially given proximity(distance)

(ii) the actual set of orthogonal coordinates for the points in this minimum space, and

(iii) a plot that shows the true shape of the initially unknown function relating proximity to distance.

In other words, rank order of distances (including ties) in the original data should be preserved although it is redefined by a new distance function.

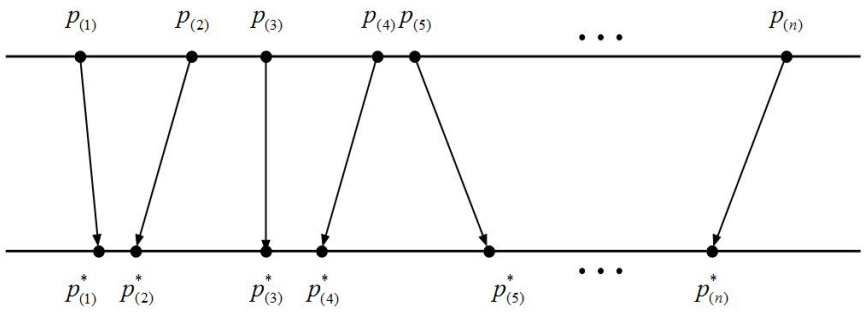

Figure 2.1: Monotonicity; a basic requirement for a distance measure

Figure 2.1 plainly shows the meaning of preserving the rank order. The points $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$ on the upper line indicate the order statistics of the distances (or size) obtained by the original data. That is, they are arranged

by the increasing order. After applying the distance function to the data, we derive the lower line with the points $p^*_{(1)}, p^*_{(2)}, \ldots, p^*_{(n)}$, from $p_{(1)}, p_{(2)}, \ldots, p_{(n)}$ respectively, through the distance function. Obviously, despite the difference in distance between entities, the order has not changed at all.

In addition to the monotonicity requirement, there is another condition to be satisfied for generating a great distance function for multidimensional scaling; the smallest possible dimensionality. Considering that the main purpose of conducting multidimensional scaling is to express the data into a lower-dimensional space so that we can achieve the dimension reduction, an ideal − yet, possibly not unique − distance function leads a configuration with the smallest possible dimensionality.

Chapter 3

Multidimensional Scaling

Multidimensional scaling (MDS) deals with the problem of constructing a meaningful low-dimensional description of n points using information between observations. The information used for multidimensional scaling is about distance, dissimilarity or proximity of each pair of the individuals, and it is defined as a function of the data points. In other words, the aim of multidimensional scaling is to find the points $X_1^*, X_2^*, \ldots, X_n^*$ in k-dimensional space ($k < p$) that successfully represent the trend or characteristics of the original entities $X_1, X_2, \ldots, X_n$ in p-dimensional space with a use of a distance function; if d_{ij} denotes the actual distance between X_i and X_j in p-dimensional space, then d_{ij}^* in k-dimensional space defined by the distance function is similar in some sense to d_{ij}.

3.1. Multidimensional Scaling

For illustrating the classical scaling method of multidimensional scaling, suppose we are given n points $X_1, X_2, \ldots, X_n \in \mathbb{R}^p$. That is,

$$X_1 = (x_{11}, x_{12}, \ldots, x_{1p})^T$$

$$X_2 = (x_{21}, x_{22}, \ldots, x_{2p})^T$$

$$\vdots$$

$$X_n = (x_{n1}, x_{n2}, \ldots, x_{np})^T.$$

For each pair of the entities, we compute the distance between observation i and j, denoted by d_{ij}. Then, with an arbitrary distance measure of d_{ij}, formulate an $(n \times n)$ dissimilarity matrix $\Delta = \delta_{ij}$. Through multidimensional scaling, we identify the points $X_1^*, X_2^*, \ldots, X_n^* \in \mathbb{R}^k$ $(k < p)$, the lower-dimensional representation of the original data. i.e.,

$$X_1^* = (x_{11}^*, x_{12}^*, \ldots, x_{1k}^*)^T$$

$$X_2^* = (x_{21}^*, x_{22}^*, \ldots, x_{2k}^*)^T$$

$$\vdots$$

$$X_n^* = (x_{n1}^*, x_{n2}^*, \ldots, x_{nk}^*)^T.$$

The first step for multidimensional scaling is to convert Δ to a matrix $\mathbf{B}$, the matrix of scalar products by double centering the matrix of which components are δ_{ij}^2. It is to determine the principal coordinates through Singular Value Decomposition (SVM) of $\mathbf{B}$ in the next step. Note that

double centering is subtracting the row and column means of the matrix from its elements, adding the grand mean and multiplying by $-\frac{1}{2}$. i.e., double centering the matrix $X = (x_{ij})$ indicates transforming x_{ij} $(i, j = 1, 2, \ldots, n)$ into

$$-\frac{1}{2}(x_{ij} - \frac{1}{n}\sum_{i=1}^{n} x_{ij} - \frac{1}{n}\sum_{j=1}^{n} x_{ij} + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} x_{ij}).$$

Let $\mathbf{A} = (a_{ij})$ where $a_{ij} = -\frac{1}{2}\delta_{ij}^2$. Then, through double centering, form the $(n \times n)$ matrix $\mathbf{B}$ as below.

$$\mathbf{B} = \mathbf{HAH}$$

where

$$\mathbf{H} = \mathbf{I}_n - n^{-1}\mathbf{J}_n, \quad \mathbf{J}_n = \mathbf{1}_n\mathbf{1}_n^T$$

and $\mathbf{1}_n$ is a n-dimensional vector of ones.

Then, compute the eigenvalues and eigenvectors of the symmetric matrix $\mathbf{B}$. By Singular Value Decomposition (SVD), we derive

$$\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$$

where $\mathbf{\Lambda} = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ is the diagonal matrix of the eigenvalues, and $\mathbf{V} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ is the matrix whose columns are the eigenvectors of $\mathbf{B}$.

If $\mathbf{B}$ is a nonnegative-definite matrix with rank $r(\mathbf{B}) = t(< n)$, the largest t eigenvalues will be positive with the remaining $(n - t)$ eigenvalues zero. Based on the least squares approximation to $\mathbf{B}$ with dimension t,

let $\mathbf{\Lambda}_1 = diag(\lambda_1, \lambda_2, \ldots, \lambda_t)$ denote the $(t \times t)$ diagonal matrix of the positive eigenvalues and $\mathbf{V}_1 = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_t)$ be the corresponding matrix of eigenvectors of $\mathbf{B}$. Then,

$$\mathbf{B} = \mathbf{V}_1\mathbf{\Lambda}_1\mathbf{V}_1^T = (\mathbf{V}_1\mathbf{\Lambda}_1^{1/2})(\mathbf{\Lambda}_1^{1/2}\mathbf{V}_1) = \mathbf{Y}\mathbf{Y}^T$$

where $\mathbf{Y} = \mathbf{V}_1\mathbf{\Lambda}_1^{1/2} = (\sqrt{\lambda_1}\mathbf{v}_1, \sqrt{\lambda_2}\mathbf{v}_2, \ldots, \sqrt{\lambda_t}\mathbf{v}_t) = (\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n)^T$.

Note that $\mathbf{Y}$, the $(n \times t)$ matrix, is made up of $\mathbf{Y}_1^T, \mathbf{Y}_2^T, \ldots, \mathbf{Y}_n^T$, t-dimensional row vectors. These columns are the principal coordinates that enable us to yield the n points into t-dimensional space whose inter-point dissimilarity between the individual i and j is equal to δ_{ij} from the matrix $\Delta = (\delta_{ij})$ for $i, j = 1, 2, \ldots, n$.

In practice, if t is too large, then people often use the largest $t' < t$ positive eigenvalues and its corresponding eigenvectors of $\mathbf{B}$ to construct a reduced set of principal coordinates and their inter-point dissimilarities are also equal to δ_{ij} from Δ. In this situation, the problem of considering the minimum dimensionality criterion for the reasonable choice of t' is brought up.

Actually, the process would come to rest almost immediately. This is because a trivial solution for the case of n points can always be found in $(n-1)$ dimensions by making arbitrarily small adjustments in the regular simplex (Shepard, 1962). However, it does not achieve the real reduction of the original data. Recalling that the key to implementing multidimensional scaling to the data is of using the smallest possible dimensionality, we need to find a reasonable strategy to determine the dimensionality.

In a point of view, determining the dimensionality is a means to the end that we achieve meaningful dimension reduction and express n different

points in p-dimensional space into n points in k-dimensional space ($k < p$) (Torgerson, 1965). Thus, there can be more than one way to determine the dimensionality. One common approach is looking at the eigenvalues of the matrix $\mathbf{B}$. In many cases, by plotting eigenvalues (or a function of them), against dimension, one can determine the dimension at which the eigenvalues get stable. However, since this decision is highly likely to be subjective, choosing the dimension remains to be a problem of multidimensional scaling. In addition that there is no single optimal way to decide the dimensionality of multidimensional scaling, it is hard to figure out its implication.

There are a variety of approaches to defining distances between observations. This study assumes that we can make better use of multidimensional scaling as a tool of understanding relationships between observations by using an appropriate way to define the distances to formulate the dissimilarity matrix. Thus, by introducing several distance measures which suit for certain types of data, this paper suggests a way to make better use of multidimensional scaling.

3.2. Distance Measures

As previously stated, there are various approaches to defining the dissimilarity matrix for multidimensional scaling. In spite of prevailing use of the Euclidean distance in many different fields as a measure of difference, it does not imply that using it is always the optimal way for any type of data. Depending on the specific data structure and pattern, the distance can be

defined accordingly (Cha, 2007). Many studies have shown the necessity to exploit new measures for distance for better implication.

In this section, we would like to introduce several different types of distance measures which can be used to define the dissimilarity. Basically, the choice of a distance measure should be based on the characteristics of the data such as measurement type or representation of the objects. The tables below elucidate a variety of different types of distance measures. According to their pattern and characteristics, each of them belongs to one of the three possible categories; interval, binary(count), and fidelity. The definition of the distance is based on the two partitions A and B, both of which can be expressed by a vector with m elements for the general cases; the vectors are denoted by $\mathbf{A}$ and $\mathbf{B}$. Simply, let the i-th element of vector $\mathbf{A}$ and $\mathbf{B}$ be referred to as a_i and b_i, respectively.

Table 3.1: Distance Measures: 1. Interval

Measure	Definition	
1. Euclidean L_2	$d_{Euc} = \|\|a-b\|\|_2 = \sqrt{\sum_{i=1}^{m} \|a_i - b_i\|^2}$	(1)
2. Squared Euclidean	$d_{SqE} = \|\|a-b\|\|_2^2 = \sum_{i=1}^{m} (a_i - b_i)^2$	(2)
3. City block L_1	$d_{CB} = \|\|a-b\|\|_1 = \sum_{i=1}^{m} \|a_i - b_i\|$	(3)
4. Minkowski L_p	$d_{Min} = \|\|a-b\|\|_p = \sqrt[p]{\sum_{i=1}^{m} \|a_i - b_i\|^p}$	(4)
5. Chebyshev L_∞	$d_{Cheby} = \|\|a-b\|\|_\infty = \max_i \|a_i - b_i\|$	(5)

$*\|a-b\|_p = (\sum_{i=1}^{m} |a_i - b_i|^p)^{(1/p)}$

Table 3.1 describes the most common measures for interval distance; L_p Minkowski family. By varying the value of p in the Minkowski L_p distance, shown in Table 3.1, from 2, 1, to ∞, we can derive the Euclidean distance, the City block distance, and the Chebyshev distance, respectively. Note that the City block distance is also called rectilinear distance, taxicab norm and Manhattan distance. These distance measures are appropriate when the data for analysis are made up of continuous components.

Table 3.2: Distance Measures: 2. Binary (Count)

Measure	Definition	
6. Kulczynski	$d_{Kul} = 1 - \frac{1}{2}(\frac{\sum_{i=1}^{m} \min(a_i, b_i)}{\sum_{i=1}^{m} a_i} + \frac{\sum_{i=1}^{m} \min(a_i, b_i)}{\sum_{i=1}^{m} b_i})$	(6)
7. Jaccard	$d_{Jac} = \frac{\sum_{i=1}^{m} (a_i - b_i)^2}{\sum_{i=1}^{m} a_i^2 + \sum_{i=1}^{m} b_i^2 - \sum_{i=1}^{m} a_i b_i}$	(7)
8. Tanimoto	$d_{Tani} = \frac{\sum_{i=1}^{m} (\max(a_i, b_i) - \min(a_i, b_i))}{\sum_{i=1}^{m} \max(a_i, b_i)}$	(8)
9. Dice	$d_{Dice} = \frac{2 \sum_{i=1}^{m} a_i b_i}{\sum_{i=1}^{m} a_i^2 + \sum_{i=1}^{m} b_i^2}$	(9)
10. Pearson χ^2	$d_{Pear} = \frac{(a_i - b_i)^2}{b_i}$	(10)

Table 3.2 shows the possible options for a distance measure when we have binary or count observations for multidimensional scaling. The choice of a distance measure from the table differs based on the research purpose and its condition. The Kulczynski distance, for example, only concerns the proportion of the difference between observations. According to the specific

definition shown in the table, it bounds between 0 and 1; the higher the Kulczynski value is, the bigger difference there exists between the individuals.

The equations 7 through 10 in Table 3.2 are useful measures of what the entities A and B share with their attributes. Note that they assume that both A and B are with m binary attributes. That is, each attribute of A and B can be either 0 or 1. That is why they are also called the number of matches or the overlap for binary vectors. They are frequently used as similarity measures in the field of biological taxonomy for the binary feature vector comparison. While the Jaccard coefficient imposes the same weight on both difference and similarity, the Dice distance puts double weight on similarity. The Dice distance is also called Sorensen or Czekannowski measure. The Pearson χ^2 is commonly used for count data analysis to measure the similarity.

Table 3.3: Distance Measures: 3. Fidelity

Measure	Definition	
11. Bhattacharyya	$d_{Bha} = -ln \sum_{i=1}^{m} \sqrt{a_i b_i}$	(11)
12. Hellinger	$d_{Hel} = 2\sqrt{1 - \sum_{i=1}^{m} \sqrt{a_i b_i}}$	(12)
13. Kullback-Leibler	$d_{KL} = \sum_{i=1}^{m} a_i ln \frac{a_i}{b_i}$	(13)

Lastly, Table 3.3 explains the distance measures for fidelity similarity. These measures concern about two discrete or continuous probability distributions. The distances are used to determine the relative closeness of the two samples. $\sum_{i=1}^{m} \sqrt{a_i b_i}$ in the Bhattacharyya distance and the Hellinger dis-

tance is called the Bhattacharyya coefficient. When continuous distributions are presented, the summation is simply replaced by integration. However, since we consider the m finite components, there is no need to deal with continuous distributions. If there is no overlap at all, the Bhattacharyya coefficient will be 0 because of the multiplication by zero in every partition. Simultaneously, that leads to the infinitely large value for the Bhattacharyya distance and the upper bound, 2, for the Hellinger distance. The Kullback-Leibler divergence is a non-symmetric measure of the difference between two probability distributions.

Chapter 4

Application to Twitter Network Data

The conjecture that Social Network Services (SNS) such as Twitter, Facebook, and MySpace reflect the political realm of the society has been brought up as more politicians use SNS for their campaign and the public voluntarily and actively express their political preference through SNS; by following specific group of politicians or re-tweeting their comments on Twitter for example. Thus, more and more studies start to have a focus of their analysis on identifying the relationship of individuals on the network. And multidimensional scaling is one of the efficient ways for the network data analysis.

In this paper, we use the Twitter network data based on the number of followers of Korean politicians on Twitter in April, 2012 for the application of multidimensional scaling. By the time of April, 2012, 194 Korean politicians were on Twitter, and they became the targets of the analysis. The type of the

data we have for this analysis is the number of followers for each politician and how many co-followers each pair of politicians has. The statistical software R was used for the application.

The very first step to apply multidimensional scaling to the Twitter network data is to define the dissimilarity matrix. Although the data is obviously a count type, we implement several different types of measures for comparison. Thus, we pick one measure from each category and define the dissimilarity matrix accordingly. The selected measures for the analysis are the Euclidean distance from the interval type, the Kulczynski distance from the count type, and the Kullback-Leibler divergence from the fidelity type, respectively. Denote the sets of followers of the politicians i and j by F_i and F_j. Note that the sign $|A|$ is used to denote the cardinality of the set A, not the absolute value in the arithmetic sense. Then, define the (194×194) dissimilarity matrix $\Delta = \delta_{ij}$, $(i, j = 1, 2, \ldots, 194)$ where

$$\begin{aligned}
\delta_{Euc}(i,j) &= \sqrt{(|F_i| - |F_j|)^2} \\
\delta_{Kul}(i,j) &= 1 - \frac{1}{2}\left(\frac{|F_i \cap F_j|}{|F_i|} + \frac{|F_i \cap F_j|}{|F_j|}\right) \\
\delta_{KL}(i,j) &= |F_i| \, ln \frac{|F_i|}{|F_j|}
\end{aligned}$$

In case of defining the Kulczynski distance, we used the number of co-followers instead of the minimum value of F_i and F_j, because it seems more appropriate to use the intersection of two sets of followers since the analysis is focused on verifying the relationship between politicians.

Based on the definitions above, we can formulate the dissimilarity matrix for each measure. With the matrices, by following the steps described in Section 3.1, we can conduct multidimensional scaling. First, in order to

determine the dimensionality of multidimensional scaling, we plot the eigenvalues against dimension for each distance measure, as shown in Figure 4.1.

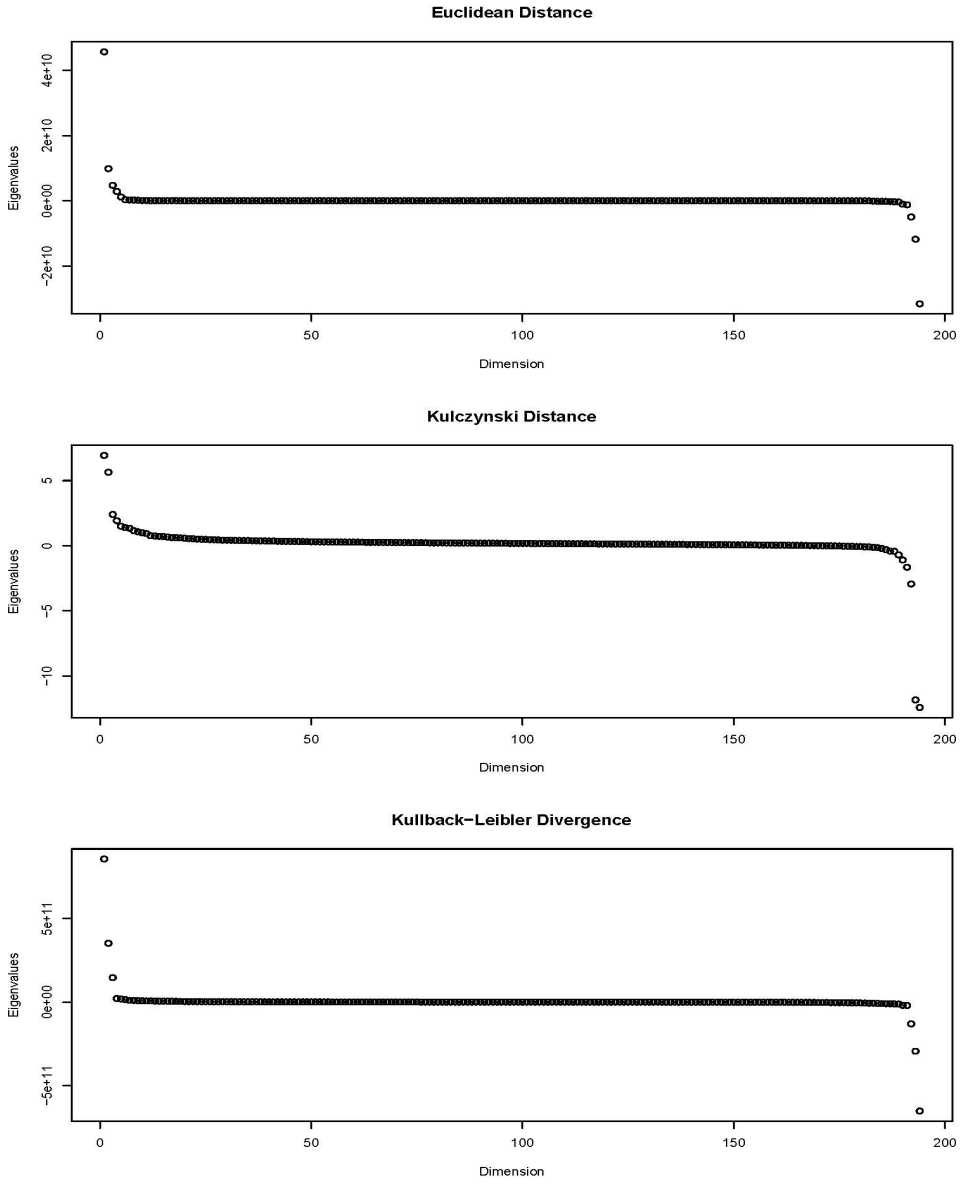

Figure 4.1: Eigenvalues against Dimension

In many cases, as previously stated, the dimension for multidimensional scaling is determined at which the eigenvalues get stable. Based on that perspective, we can take two- or three-dimensional space for the three distance measures. The following figures, Figure 4.2 and 4.3, show the results from multidimensional scaling in two-dimensional space and three-dimensional space, respectively.

In Figure 4.2, the x-axis implies the first principal component from the eigenvector of the biggest eigenvalue, and the y-axis implies the second principal component from that of the next biggest eigenvalue. In case of Figure 4.3, the third biggest eigenvalue and its eigenvector account for the meaning of the z-axis. For each figure, note that the white circle indicates the politicians from the major conservative parties (New Frontier Party, Advancement and Unification Party) while the cross sign illustrates those from the major liberal parties (Democratic United Party, United Progress Party). The rest of them − independent or non-major parties − are marked by the black circle.

It is easy to detect the huge difference in the results from three different distance measures in Figure 4.2. First of all, in case of the result with the Euclidean distance, the values from the first eigenvectors tend to be smaller than zero. In addition, the observations show the linearity slightly, which makes it hard to identify any clustering. Meanwhile, the Kullback-Leibler divergence shows an odd cross-shaped trend without any specific clustering. In fact, applying the Kullback-Leibler divergence seems quite inappropriate to the data since the measure is for probability distribution. And clearly, Twitter network data has its form by count type. Lastly, it is worthwhile to

look through the result from multidimensional scaling with the Kulczynski distance in detail. There exists a strong bond in Twitter followership between politicians from the liberal parties; smaller value of the first principal component. Although there needs extra analysis, estimate or guess to find out what the first principal component implies, it is safe to say the existence of the cluster. In case of those from the conservative parties, though their bond seems weaker than that of the liberal politicians, the conservative politicians tend to have greater values of the first principal component. Clearly, independent politicians or those from the non-major parties do not show any pattern as a cluster.

Figure 4.3 shows the results of multidimensional scaling in three-dimensional space. Based on the result with the Kulczynski distance and the information about the party, we can conclude that the politicians from the liberal parties are inclined to have lower value of all three principal components while those from the conservative parties have totally opposite tendency. Recall that the main purpose of multidimensional scaling is to identify the relationship between entities with the smallest dimensionality. Therefore, there actually is no need to use multidimensional scaling in three-dimensional space, since this trend can be already seen in the result in two-dimensional space.

Along with determining the dimensionality, identifying the meaning of each principal component remains as one of the big problems of multidimensional scaling. Based on the additional information about the parties and other possible sources, one may attempt to find out what each axis implies.

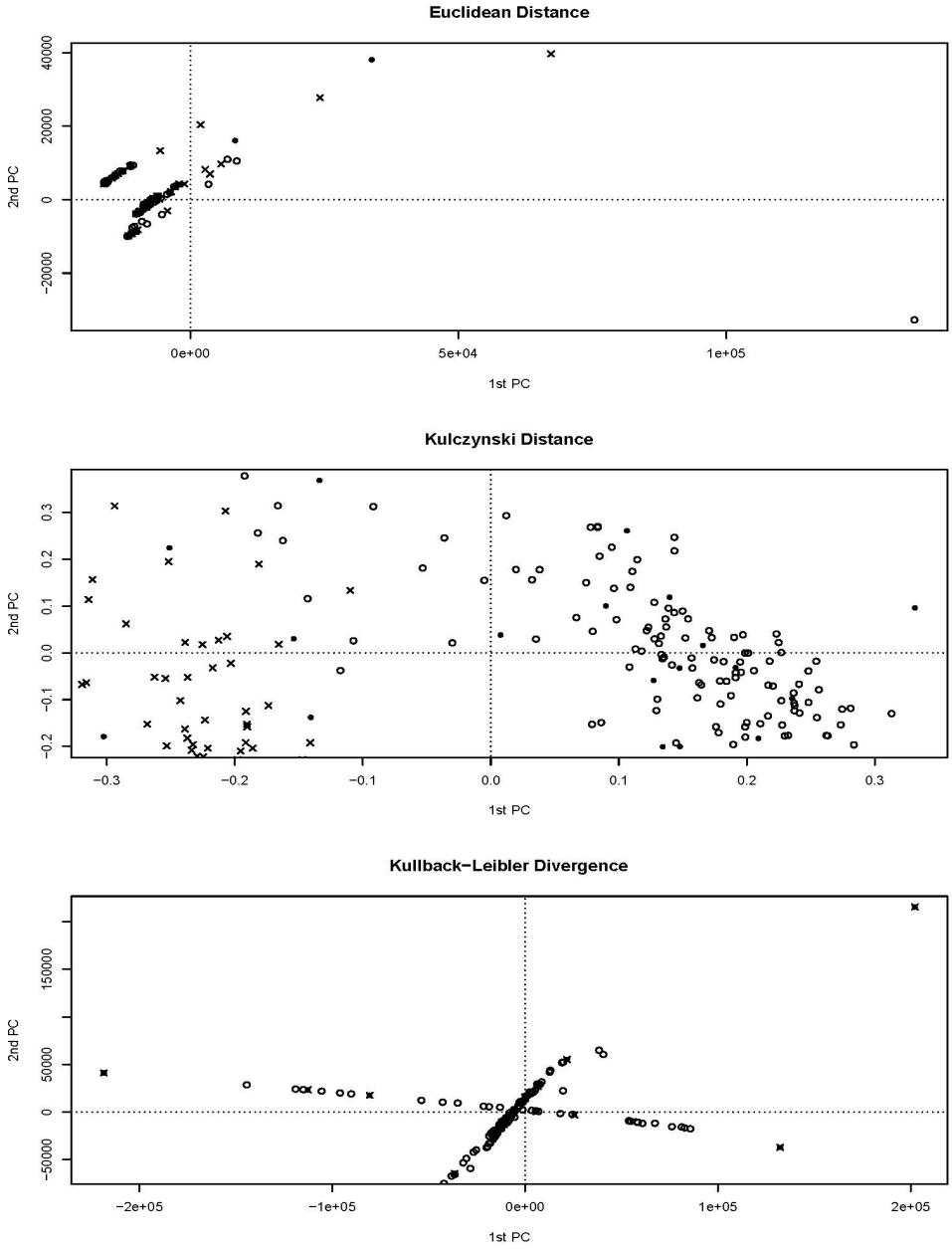

Figure 4.2: Multidimensional Scaling with Dimension=2

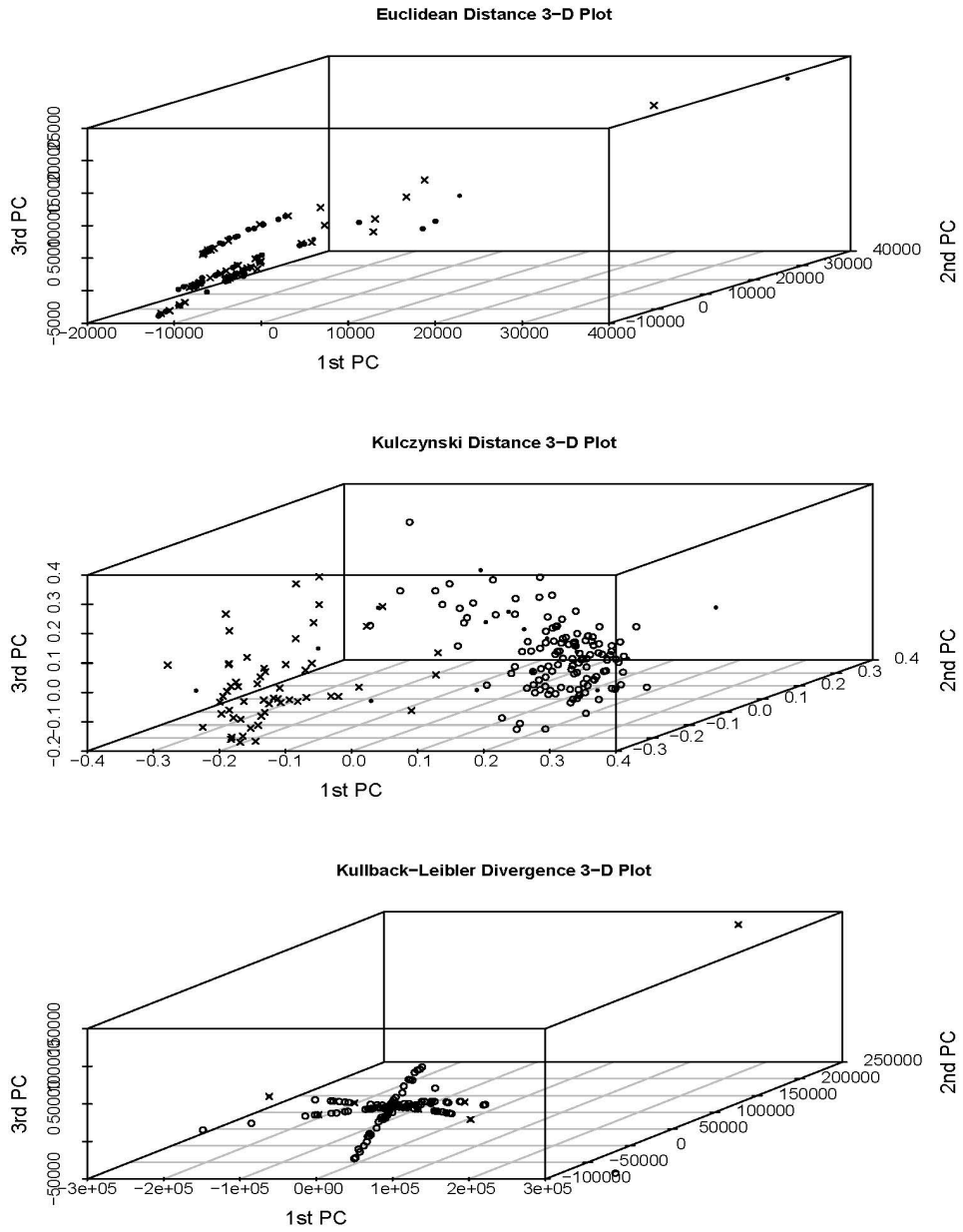

Figure 4.3: Multidimensional Scaling with Dimension=3

Chapter 5

Conclusion

In this paper, we introduced how the classical multidimensional scaling works and applied it to the Twitter network data. In particular, we concentrated on the variety of distance measures that we can use to define the dissimilarity matrix. Although there are several issues still left on multidimensional scaling such as determining the dimensionality and interpreting the meaning of each axis, this paper tried to keep its focus on choosing an appropriate distance measure to construct the dissimilarity matrix reasonably. The choice of measure becomes more important as multidimensional scaling can be applied in various fields such as marketing, genetics, ecology, molecular biology, psychology and social networks (Tzeng et al., 2008). According to their different research aims, the data used for the analysis have different forms with distinctive characteristics. This paper showed that the research can produce a better interpretation of the data, by using a proper measure to define the dissimilarity matrix.

Previously, there have been several studies (Shepard, 1962; Cha, 2007; Arabie et al., 1973) that put an emphasis on the importance of determining the distance measures for the analysis. From the actual application of multidimensional scaling with a variety of distance measures to the Twitter network data, this study also showed how the results differed with different distance measures and implied the importance of choosing the appropriate way to define the distance. Since what we had for multidimensional scaling from Twitter network data is count type, we can make a better inference for the Twitter followership trend through multidimensional scaling with measures such as Kulczynski and Jaccard distance. In other words, simply by considering the pattern and characteristics of the data to define the dissimilarity, we can derive an improved inference from the data through multidimensional scaling.

Bibliography

Agarwal, A., Phillips, J. M., and Venkatasubramanian, S. (2010). Universal multi-dimensional scaling. *16th Annual ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*

Arabie, P., Boorman, S.A. (1973). Multidimensional scaling of measures of distance between partitions. *Journal of Mathematical Psychology* **10**, 148-203.

Arnold J. B. (1971). A multidimensional scaling study of semantic distance. *Journal of Experimental Psychology Monograph* **90**, 2, 349-372.

Cha, S.-H. (2007). Comprehensive survey on distance/similarity measures between probability density function. *International Journal of Mathematical Models and Methods in Applied Sciences*, **4**, Vol 1. 300-307.

Cleeland, C. S., Nakamura, Y., Mendoza, T. R., Edwards, K. R., Douglas, J., and Serlin, R. C. (1996). Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, **9**:179.

Henley N.M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, **8**, 176-184.

Lockhead, G. R. (1970). Identification and the form of multidimensional discrimination space. *Journal of Experimental Psychology*, **85**, 1-10.

Mardia, K. V., Kent J. T., and Bibby J. M. (1979). Multivariate analysis: Probability and mathematical statistics. *Academic Press*, 394-423.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, **27**, No.2, 125-140.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, **27**, No.3, 219-246.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, **17**, 401-419.

Torgerson, W. S. (1965). Multidimensional scaling of similarity. *Psychometrika*, **30**, 379-393.

Tzeng, J., Lu, H. H.-S., and Li, W.-H. (2008). Multidimensional scaling for large genomic data sets. *BMC Bioinformatics*, **9**:179.

국 문 초 록

거리행렬을 이용한 다차원 척도법과
Twitter 네트워크 자료에의 응용
Multidimensional Scaling
with Application to Twitter Network Data

다차원 척도법은 마케팅, 유전학, 생태학, 분자생물학, 심리학, 그리고 네트워크 등 다양한 분야에 적용되고 있다. 대부분의 경우, 다차원 척도법은 고차원 관측치를 낮은 차원의 공간에 표현함으로써 관측치들 간의 관계를 규명하는 데 그 목적을 가지고 있다. 간단히 말해, 다차원 척도법은 복잡한 자료를 간단한 그림으로 표현함으로써 쉽게 그 자료의 특성을 관찰할 수 있도록 한다. 자료에 다차원 척도법을 적용하기 위해서는 모든 쌍에 대하여 거리를 정의, 이를 기반으로 거리 행렬을 생성해야 한다. 거리 행렬을 정의하는 데에는 특정 조건을 만족하는 다양한 종류의 거리 측도가 사용될 수 있다. 본 연구에서는 다차원 척도법에 사용될 수 있는 몇몇의 거리 측도를 소개하고, 그들이 사용되기에 적절한 상황에 맞도록 범주화하고자 하였다. 또한, 다차원 척도법을 Twitter 네트워크 데이터에 적용하는 실제 자료 분석을 실시하였다. 거리 행렬을 정의하기 위해 여러 가지 다른 측도를 적용해봄으로써, 자료의 특성에 적합한 거리 측도를 선택하는 것이 다차원 척도법에 미치는 영향을 밝히고자 하였다.

주요어 : 거리 측도, 거리 행렬, 다차원 척도법, 소셜 네트워크, 트위터 네트워크 데이터.
학 번 : 2011-20248