3-Point Pseudo-Scale Distance Measure for Measuring Indirect Proximity

Norliza Adnan¹, Norhaiza Ahmad² and Maizah Hura Ahmad³

Department of Mathematical Sciences, Faculty of Science
Universiti Teknologi Malaysia, 81310 UTM Skudai, Johor, Malaysia
¹ orlyy2liza@gmail.com, ² norhaiza@utm.my, ³ maizahutm@gmail.com

Copyright © 2013 Norliza Adnan et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

A typical proximity measure employed to find the distance between two points in multidimensional scaling techniques (MDS) is the Euclidean distance. This direct measurement approach however lacks the ability to extract more information from other points in the data. In this paper, we introduce a 3-point pseudo-scale distance measure based on Kendall’s tertiary treatment of ties (TTT). This distance is coupled with ANOVA-Tukey’s test approach to reveal the hidden structure where the Euclidean distance used in multidimensional scaling technique is not able to cater. To demonstrate the proposed technique, a set of rainfall data recorded between the years 1968 and 2003 over nine rain-gauge stations in Peninsular Malaysia are used.

Keywords: Euclidean-distance, Pseudo-distance, Multidimensional Scaling, ANOVA-Tukey

1 Introduction

The purpose of multidimensional scaling (MDS) is to allow insight into the underlying structure of relations between variables by providing a geometrical representation of these relations. MDS is a statistical technique whose characteristics can be classified as: (i) the type of observed relations which can be used in data analysis, (ii) the type of geometrical representation of these relations. The graphical display of the correlations amongst variables provided by MDS
enables the data analyst to literally look at the data and to explore their structures visually [1]. This visual approach could show the regularities that remain hidden when studying the data in the form of arrays of numbers. The MDS representations are also useful as a representational basis for various mathematical models of categorization, identification and recognition memory and generalization [2].

Visualization of points in a two-dimensional plot in MDS shows the closeness amongst points and its pattern. The attributes or samples that are studied are represented by the points on the plot. The possible input for MDS involves a kind of relation between pairs of variables which often indicate dissimilarities or similarities that can be translated by a proximity matrix or distance matrix. The Euclidean distance is a popular distance measure used to find pairwise distances among samples in the MDS algorithm [3]. Unfortunately, this type of distance measure lacks the ability to extract the relation between other variables in the data.

The purpose of the current paper is to enhance the visualization of MDS geometrical representation by using pseudo distance measure. This is aided by the ANOVA-Tukey test approach in order to obtain a proximity data matrix that exhibits the relation between more than two variables at the same time.

This paper will be organized as follows. We start by presenting the methodology used in the study. We then present the results when the proposed procedure is applied to rainfall data. Finally we give our conclusions based on the evaluation method presented in this study.

2 Methodology

Multidimensional scaling with Euclidean distance

In this section, a brief description of MDS based on Euclidean distance measure is presented. The MDS models are defined by specifying the similarity or dissimilarity data that becomes the proximity. Through a set of algorithm, these proximity measures will be mapped onto distances on an $n$-dimensional MDS configuration. The distance between any two of its points, $i$ and $j$ are computed using various distance measure and the most natural distance function is the Euclidean distance.

The MDS algorithm determines coordinates for each item in an $n$-dimensional space such that the Euclidean distances among the variables best approximates the input proximities. The optimal scaling of the classical non-metric MDS can be seen through the distance between two variables, $i$ and $j$ in an $n$-dimensional Euclidean space. The formula used is
3-Point pseudo-scale distance measure

\[ d_{ij} = \sqrt{\sum_{a=1}^{n} (x_{ia} - x_{ja})^2} \]  

The measurement using Euclidean distance in (1) shows the direct connection between variables in the distance matrix. Based on that matrix, a 2-dimensional plot in MDS will be constructed. Thus, the similarities or dissimilarities among variables depend on the distance values and the patterns that can be seen through the 2-dimensional plot. For example, if the distance between A and B is 23.0, while between A and C is 30.0, then B is said to be closer to A than C is to A.

In the distance measurement process, only the direct measurement between two variables is obtained. In other words, only the nearness between variables will be identified and no other information from other variables can be extracted. Therefore, the ability to get more information that can be extracted from the similarity measurement process will improve the quality of geometrical representation in MDS.

ANOVA-Tukey test

To construct the distance matrix, \(d_{ij}\), we use Tukey’s test in the procedure. The aim of using Tukey’s test is to determine the significance difference between variables which represented by points. Here, the difference between any two mean scores is compared against HSD (honestly significant difference). A mean difference is statistically significant only if it exceeds HSD.

The multiple comparison of means among any variables is investigated by using Tukey’s HSD. The Tukey’s test is a statistical test that is generally used in conjunction with ANOVA (Analysis of Variance) to find which means are significantly different.

The Tukey’s test starts by performing the ANOVA test to select the appropriate means in order to calculate Tukey's test for each mean comparison. A Tukey's score is then checked for statistical significance against Tukey's probability or critical values by taking into account the appropriate degrees of freedom (within) and number of treatments.

The results from a Tukey’s test will show the significant difference between the variables and the purpose of using this method is to identify which pairs have no differences between them. The results from a Tukey’s test will be used to set up the distance matrix, \(d_{ij}\).

Pseudo-distance

To set up the distance matrix, \(d_{ij}\), a 3-point pseudo-scale analysis will be performed. A pseudo distance is represented as \(n\)-point pseudo-scale that consist
of a penalty score which is based on the Kendall’s Primary (PTT), Secondary (STT) and Tertiary (TTT) treatment of ties [4]. Given that A, B and C represent the variables, the distance measures will be as follows:

\[ \delta_{AB} = \begin{cases} 1 : & \text{if } A \approx B \\ 2 : & \text{if } A \approx C \text{ and } C \approx B \text{ with } A \neq B \\ 3 : & \text{for all other cases} \end{cases} \]

where \( \delta_{AB} = 1 \) indicates that A and B are similar to each other, \( \delta_{AB} = 2 \) indicates that A is different to B but A is similar to C and C is similar to B, and \( \delta_{AB} = 3 \) indicates the conditions for all other cases.

Assessing the reliability of MDS

In order to assess the reliability of the proposed approach, a specific evaluation will be used to measure the deviation from monotonicity between distances, \( d_{ij} \) and the observed dissimilarities. The evaluation is done by using STRESS evaluation.

STRESS(S), is the square root of a normalized residual sum of squares that measures the mismatch between the rank order of distances in the data and the rank order of distances in the ordinations. As \( S \to 0 \), the configuration is said to be approaching a perfect fit to the observed dissimilarity. Therefore, the aims of STRESS is to find the configuration which minimizes the STRESS value.

The formula for calculating STRESS evaluation is as follows:

\[
S = \sqrt{\frac{\sum_{i \neq j} (d_{ij} - \hat{d}_{ij})^2}{\sum_{i \neq j} d_{ij}^2}}
\]

The interpretation of the STRESS measure follows the suggestion made by [6] as listed in Table 1:

<table>
<thead>
<tr>
<th>Stress (in %)</th>
<th>Goodness of fit</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>Poor</td>
</tr>
<tr>
<td>10% or 0.1</td>
<td>Fair</td>
</tr>
<tr>
<td>5% or 0.05</td>
<td>Good</td>
</tr>
<tr>
<td>2.5% or 0.025</td>
<td>Excellent</td>
</tr>
<tr>
<td>0</td>
<td>Perfect</td>
</tr>
</tbody>
</table>

Based on Table 1, a STRESS value of 0.20 means that 80% of the variance of the \( d_{ij} \) is explained by the distances. Therefore, it is considered as poor fit compared to STRESS values 0.1, 0.05, 0.025 and 0.
3 Numerical Results

In order to illustrate our new distance measure, we use MDS technique for visualizing rainfall data over nine rain-gauge stations in Peninsular Malaysia for duration of 36 years from 1968 to 2003. The rain-gauge stations are listed in Table 2:

<table>
<thead>
<tr>
<th>Notations</th>
<th>Stations</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOS</td>
<td>Alor Setar</td>
</tr>
<tr>
<td>BYN</td>
<td>Bayan Lepas</td>
</tr>
<tr>
<td>IPH</td>
<td>Ipoh</td>
</tr>
<tr>
<td>MLK</td>
<td>Malacca</td>
</tr>
<tr>
<td>SWN</td>
<td>Sitiawan</td>
</tr>
<tr>
<td>SBG</td>
<td>Subang</td>
</tr>
<tr>
<td>KBH</td>
<td>Kota Bharu</td>
</tr>
<tr>
<td>KTN</td>
<td>Kuantan</td>
</tr>
<tr>
<td>MSG</td>
<td>Mersing</td>
</tr>
</tbody>
</table>

We start the analysis by setting the nine stations as variables that will be considered as variables in the analysis and perform the Tukey’s test. Then, the similarity (or dissimilarity) between pairs of stations is identified as a base to determine the scale in pseudo-distance. The MDS algorithm in computing the coordinates will be performed based on the distance matrix, and the results will be shown in a 2-dimensional configuration.

As a result, a lower triangle matrix (Table 3) is constructed. A lower triangle matrix is produced because elements $a_{ij}$ and $a_{ji}$ have no difference in meaning since the values are the same. The matrix indicates the significant difference among rain-gauge stations based on the results from Tukey’s test for rainfall data. For those pairs in the lower triangle matrix that have no sign at all, this means that the stations have significant difference. The ‘≈’ shows the similarity between stations based on the Tukey’s test or in other words, no significant difference between rain-gauge stations. The difference is significant at $p < 0.05$ value which means that the station was dissimilar to each other. The values in Table 3 highlight that the difference between rain-gauge stations was not significant (or the rain-gauge stations were similar to one another).
Table 3: Tukey’s test significant difference results for rainfall data

<table>
<thead>
<tr>
<th></th>
<th>AOS</th>
<th>BYN</th>
<th>IPH</th>
<th>MLK</th>
<th>SWN</th>
<th>SBG</th>
<th>KBH</th>
<th>KTN</th>
<th>MSG</th>
</tr>
</thead>
<tbody>
<tr>
<td>AOS</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BYN</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPH</td>
<td>-</td>
<td></td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLK</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SWN</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SBG</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KBH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KTN</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>≈</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>MSG</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>≈</td>
<td>-</td>
</tr>
</tbody>
</table>

Based on the Tukey’s hypotheses testing, there are ten pairs that shows no difference (in Table 3) and is shown using “≈” sign. For example, there is no significance difference between MALACCA rain-gauge station with IPOH, SITIAWAN, SUBANG and MERSING rain-gauge stations. The “-” entries correspond to no meaning.

The similarity matrix, $d_{ij}$ is shown in Table 4.

Table 4: The similarity matrix, $d_{ij}$ using Euclidean-distance

\[
d_{ij} = \begin{bmatrix}
2870.478 & 2961.379 & 2961.379 & 0 & 3363.941 & 3237.927 & 2653.590 & 2783.549 & 2958.499 \\
2848.449 & 2884.959 & 2884.959 & 3237.927 & 2821.691 & 0 & 1900.715 & 2696.037 & 2854.808 \\
2727.521 & 2755.077 & 2755.077 & 2653.590 & 2653.590 & 1900.715 & 0 & 2568.321 & 2711.109 \\
1881.863 & 2822.736 & 2822.736 & 2783.549 & 2783.549 & 2696.037 & 2568.321 & 0 & 1913.074 \\
2076.899 & 2028.759 & 2032.612 & 2958.499 & 1983.051 & 2854.808 & 2711.109 & 1913.074 & 0
\end{bmatrix}
\]

The similarity matrix in Table 4 shows the distance measurement based on Euclidean distance. The values in the entries shows the distance measures amongst the stations. For example, the distance between Bayan Lepas and Alor Setar is 1977.138 units while the distance between Alor Setar and Ipoh is 2008.230. These means that Bayan Lepas is similar to Alor Setar compared to Ipoh.

The results in Table 4 are used to the construct a similarity matrix using the proposed pseudo-distance as shown in Table 5.
3-Point pseudo-scale distance measure

Table 5: The similarity matrix, $d_{ij}$ using pseudo-distance

$$
\begin{bmatrix}
0 & 3 & 3 & 3 & 3 & 3 \\
3 & 0 & 3 & 3 & 3 & 3 \\
3 & 3 & 0 & 3 & 3 & 3 \\
3 & 3 & 1 & 0 & 3 & 3 \\
3 & 3 & 1 & 1 & 0 & 3 \\
3 & 3 & 1 & 1 & 1 & 0 \\
3 & 3 & 1 & 1 & 1 & 0 \\
3 & 3 & 3 & 3 & 3 & 3 \\
3 & 3 & 3 & 3 & 3 & 3 \\
3 & 3 & 3 & 3 & 3 & 0 \\
3 & 3 & 3 & 3 & 3 & 0 \\
3 & 2 & 3 & 1 & 1 & 3 \\
3 & 3 & 3 & 3 & 3 & 0 
\end{bmatrix}
$$

Table 5 shows the similarity matrix based on pseudo-scale as the measurements. The values in the entries shows the level of scales of the distances amongst the stations. In contrast to the distance measurement using Euclidean distance, the entries are levels of scales that shows the similarity amongst stations. Therefore, scale 0 corresponds to no meaning. Scale 1 corresponds to “similar” (for example, Malacca rain-gauge station is similar to Ipoh rain-gauge station) as defined in the pseudo-scales. Scale 2 corresponds to “A is similar to C and C is similar to B, but A has no connection (dissimilar) to B at all” (for example, Bayan Lepas rain-gauge station is similar to Malacca rain-gauge station. Mersing rain-gauge station is similar to Malacca rain-gauge station). But, Bayan Lepas rain-gauge station is dissimilar to Malacca and Mersing rain-gauge stations). Scale 3 shows other cases which are not considered in Scale 1 and Scale 2. However, since only three cases are considered as scales, many similarity measures have equal scale, that is, the pairs with Scale 3 [5], [6].

In addition to meaningful dimensions, we should also look for clusters of points or particular patterns and configurations such as circles and manifolds. Figure 1 shows the visual plot of MDS using Euclidean distance. In terms of quadrant plot, SITIAWAN, IPOH, BAYAN LEPAS and ALOR SETAR are placed in the first quadrant. SUBANG is the only station in second quadrant. While MERSING and KUANTAN are in third quadrant, MALACCA and KOTA BHARU are in the fourth quadrant.
In Figure 2, a visual plot of MDS using Pseudo distance is plotted. Figure 2 shows that MERSING, MALACCA and SITIAWAN are in the first quadrant. ALOR SETAR and KUANTAN are in the second quadrant. While KOTA BHARU and BAYAN LEPAS are in the third quadrant, IPOH and SUBANG are in the last quadrant.

The MDS configurations is satisfied at 2-dimension because MDS plots are more difficult to interpret and impracticable to represent in print, as dimensionality increases beyond three [3].
The validation of the results are carried out by using STRESS evaluation. The STRESS value for MDS using Euclidean distance is 0.3287, while the STRESS value for MDS using Pseudo distance is 0.1826 which means that 81.74% of the variance of the $d_{ij}$ is explained by the distances.

**Conclusion**

Comparing the performance of the proposed approach and the well-known Euclidean distance in the MDS algorithm, we found that the ANOVA-Tukey coupled with the 3-points pseudo-scale gives the smallest STRESS value which means that the mismatched between the distance rank order in the data and the rank order of distances in the ordinations is reduced when compared to Euclidean distance measurement. In the other words, the mismatched between the distance rank order in the data and the rank order of distances in the ordinations is small compared to the Euclidean distance approach. Therefore, the coarse in MDS plot is refined using the proposed approach.

**Acknowledgement**

This study was supported by Universiti Teknologi Malaysia and the Ministry of Higher Education (MOHE), Malaysia under the Fundamental Research Grant (FRGS) Vot 78487.

**References**