PhUSE US Connect 2018

Paper AB06

Shaken but Not Stirred: An Example of Subject Classification Using Multidimensional Scaling

Manuel Sandoval Garcia, Bayer, Whippany, U.S.

ABSTRACT

A fully randomized clinical trial is necessary to conduct unbiased and objective research. However, randomization can also hide some relations between study subjects that could be interesting to the investigator. Classification and similarity techniques are a powerful tool to discover these lost relations. Among these, Multidimensional scaling creates a graphical display of subjects (or characteristics) based on their differences. This paper provides an example of multidimensional scaling usefulness, in which safety and efficacy values will be displayed using different sets of identifiers to create the corresponding distance function.

INTRODUCTION

As part of the analyses conducted in a clinical trial, several classifications of the data are usually created. The classifications most frequently used are subject and treatment, but we are not limited to them. One of the questions that can arise is how similar can the points on one of these classifications be, regarding the results. A graphical technique to represent these similarities can be helpful to explore them and discover relationships between them at a glance.

MULTIDIMENSIONAL SCALING

Cox and Cox [Cox and Cox 3] define multidimensional scaling as “the search for a low dimensional space, usually Euclidean, in which points in the space represent the objects […], one point representing one object, and such that the distances between the points in the space match, as well as possible, the original dissimilarities”. In a sense, what we are looking for is a low-dimensional representation of the differences in a set of points.

There is no limit to the number of dimensions that can be used to create the resulting representation. However, if the intention is to generate a visual output, we should limit ourselves to two dimensions. Three dimensions are feasible, but the results can be confusing. If more dimensions are used in the analysis, projections into two dimensions can be created, and a formal analysis can still be performed. That, however, is out of scope of this paper.

For the analysis used in the examples that follow, I used SAS® PROC MDS. However, libraries in R®, Python® and other specialized programs can also be used to create the desired result. The bibliography has examples and guidance on how to use them.

The general code used in the examples is shown below. Some of the statements (such as “BY” or “VAR”) and options (“DATA =”, “OUT =”) are self-explanatory. During the analysis that follows the choice of some of the options will be discussed in more detail. However, an exhaustive description of all the possible options and sentences that can be used in the SAS reference documents is available online [SAS Institute 4].

```
ods graphics on;
proc mds data =  <datain> out = <dataout> pdata pfit similar;
    by  <byvar>;
    var <matrix_vars>;
RUN;
ods graphics off;
```

Turning ODS GRAPHICS ON automatically generates a graphical representation of the results and the goodness of fit of our model. The creation of this additional graph is not strictly necessary, since we can replicate these figures easily with the dataset output created by SAS. One of the examples below is a case when the creation of a separate graph is required.

DISSIMILARITIES

It should be noted that the variables to include in the VAR statement need to be those that contain the distances, and not the measurements. Multidimensional scaling (and, more specifically, PROC MDS in this case), provides us with a representation of the distances between any two points of a set, in fewer dimensions. However, we need to create the distances ourselves, based on previous knowledge of our topic and what we want to analyze.

We should not limit ourselves just to using the Euclidean distance function to evaluate the differences between elements. Any dissimilarity function (or similarity, which SAS will automatically convert to its inverse) can be used. The function that one actually chooses will depend on the nature of the data and on the interpretation of this data. For example, we could use city-block or Minkowski distance functions if we know they represent our data better.

In general, any function that satisfies the following four properties can be considered a dissimilarity:

1. For any given point, the distance to itself is 0.
2. Given any two points, a and b, the distance from a to b is greater or equal to 0.
3. Given any two points, a and b, the distance from a to b is equal to the distance from b to a.
4. Given three points a, b and c, the distance from a to c is less than or equal to the sum of the distances from a to b and b to c.

Sometimes, one or more of the properties listed above will not be met. This can happen particularly when the subjective perception of the difference between two points is considered. For example, we can ask a group of people how different two sounds are perceived by them. They could consider two sounds to be different, even when they were actually the same one. Likewise, they could have perceived sound a more different from sound b than b from a.

In cases like the one described above, we need to transform our differences so that we will meet the four conditions above. Following the example of the perceived sounds, we could disregard the differences of a to a, and consider the average of the distance between a and b, with b and a.

For the examples studied below, we will use the differences between two points and correlations. As we will see, correlations will need a particular transformation.

EXAMPLE 1: SAFETY DATA – ADVERSE EVENTS.

Consider a clinical trial that has adverse events of special interest. These adverse events can be grouped into classes according to their interest for the study. What we want to know is how close any two of them are. In order to do so, I calculated the correlations between the number of events every study subject had, and planned to run multidimensional scaling on them.

However, correlation coefficients can take values between -1 and 1, and so they violate the second of the distance principles stated above. One possible transformation was to add 1 to all values, forcing them all to be greater than 0. Additionally, if we want to keep the same range than the original values for cosmetic reasons, we could divide them by two to keep all values below one. Adding one constant and multiplying by another do not modify the proportion of similarities. But the new results were misleading. A correlation of -1 would imply that two observations were highly correlated, *in a negative way*. However, in the new scale, they would be shown as having no similarity at all.

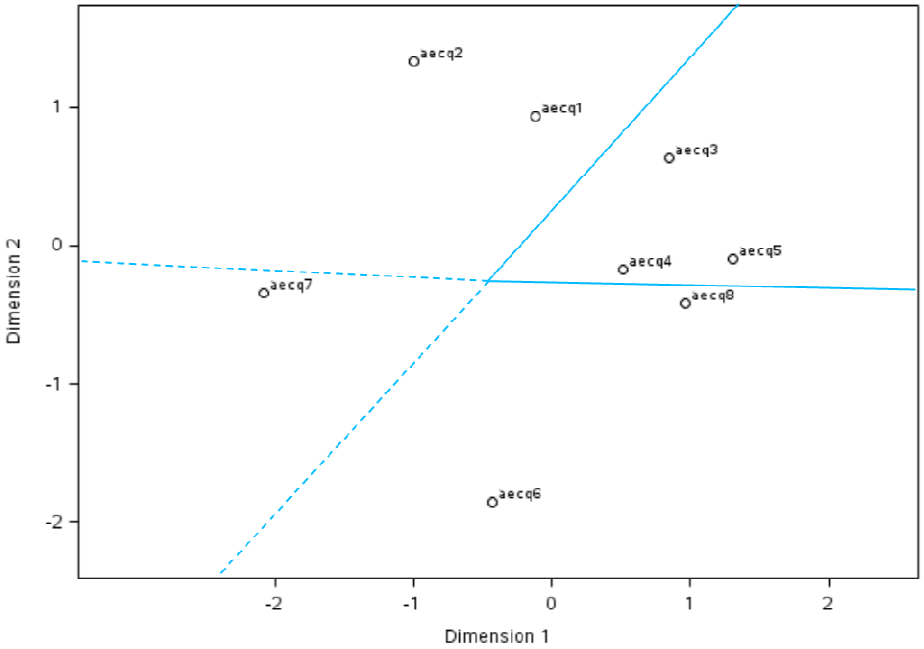
We could also transform the data by squaring the correlation coefficients (an absolute value function would also work). This way, two elements with a high correlation, either positive or negative, would remain highly similar, while those closer to 0 would also remain so.

The price to be paid is that we would no longer know if two elements were positively or negatively related. Considering that the interest is just how similar two events were, and not necessarily in what direction, the sacrifice could be done. However, we need to ensure that the direction of the correlation is not part of what we want to study.

ORDINAL SCALING

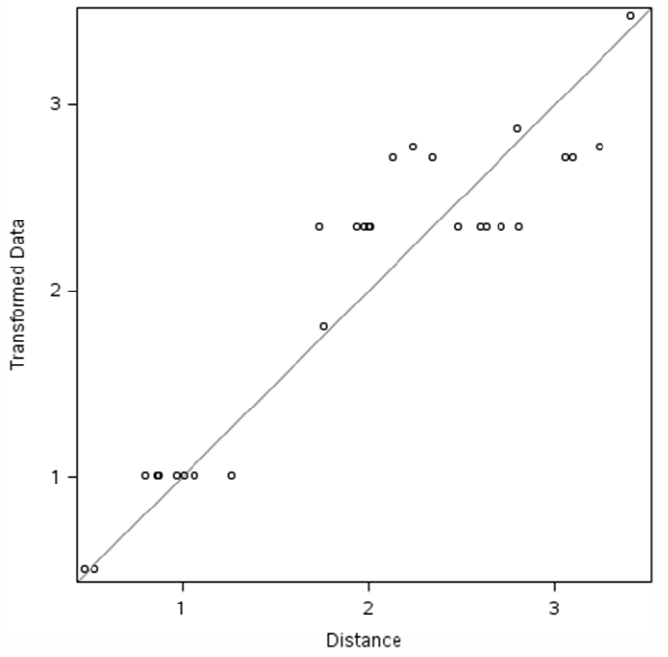
There are several ways in which the map of dissimilarities can be created in a multidimensional scaling process. If we care about the actual distance in the graph matches the actual distance between two points, a metric scaling needs to be done. In SAS, we can specify this using the “LEVEL =” option within the PROC MDS statement. The option LEVEL = ORDINAL (which is the default) will create the map according to a non-metric function, which will keep the respective distances order, but in which the magnitudes are not accounted. This approach was taken with the transformed correlations between AE counts, since it will keep the differences as they are, without the actual

magnitudes, which we had already lost. So, if an AE (say aecq7) is twice as far from aecq1 than aecq2, we can know that it is more different, but not that it doubles the distance. The resulting map is shown below:



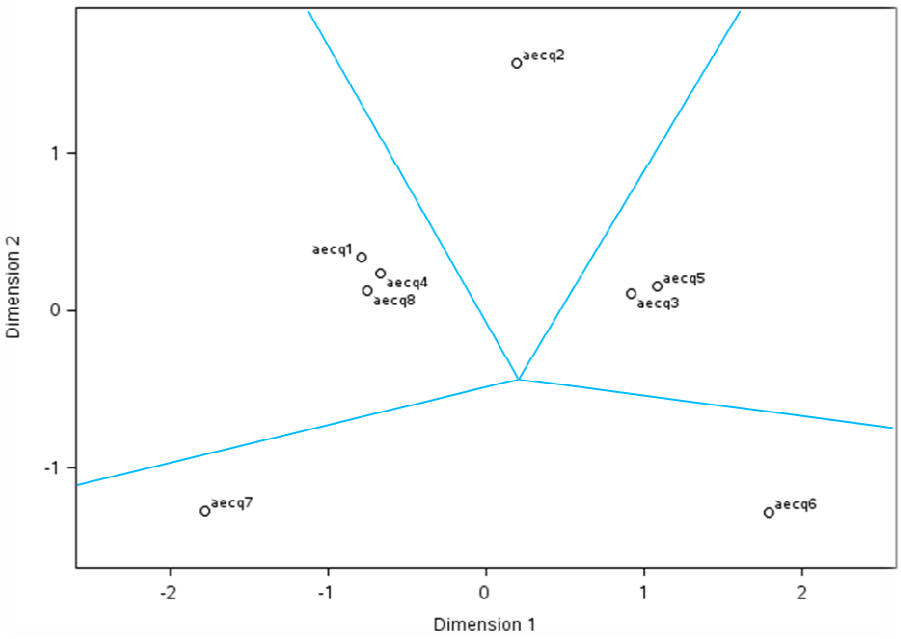
The AE groups aecq1 and aecq2 depict adverse events related to the initial condition, so it is not surprising that they ended up close to one another. In a similar way, groups 3, 4 and 5 are related to the treatment and were expected to be closely related. Groups 6, 7 and 8 were expected conditions that needed special care, and so the fact that group 8 is close to 4 and 5 merits further investigation, as is that groups 6 and 7 seem isolated from the rest. The lines drawn inside the graph create a partition that helps investigation, but no particular method was used to create them. The process of creating these partitions is called unfolding and is out of the scope of the present paper. They do not have to create full axis to be helpful, and that is why I am extending some as dotted lines, considering that I am not sure if they would have any meaning.

We have now a graphical representation of the similarities. How good (or bad) is it? Remember that we know how far the points really are. It might be impossible to build a representation in two dimensions that keeps all the original distances. The second figure that SAS outputs by default shows us those differences by each estimate calculated, so we can see, graphically, how far we are. For the model built above, the discrepancies are the following:



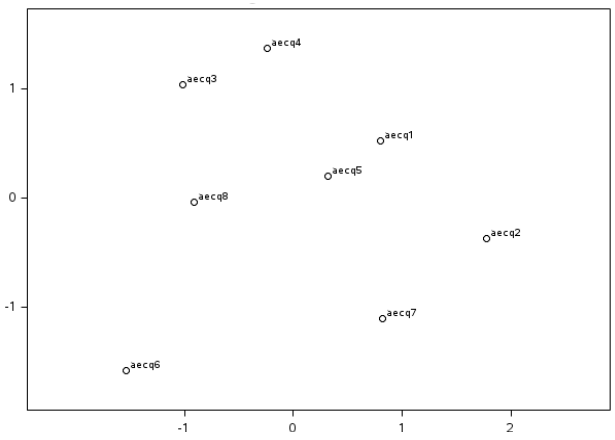
It is also possible to generate a single statistic to measure how good the overall model is. The stress function, or Badness-of-Fit Criterion, is part of the PROC MDS output and measures how far the total residuals are from the real values. In the example above, the aggregate value is 0.15, which is not particularly good, but it can still be used. Criteria for acceptance vary, but usually less than 0.15 is considered good, and less than 0.20 is a poor fit. [Borg and Groenen 2]

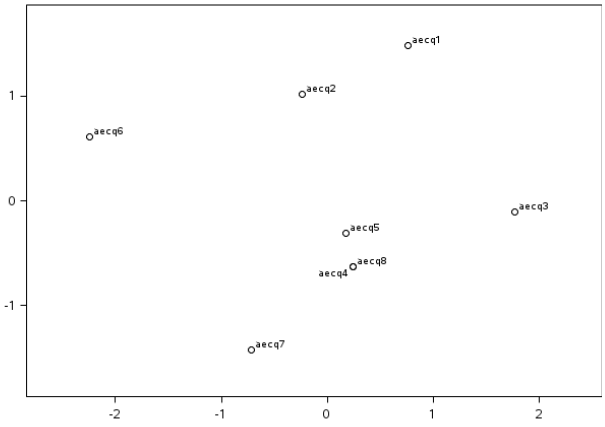
The information that we can get through the analysis above is limited, though. We could not really expect that many subjects will have adverse events in more than one of the categories. In that case, correlation will be leaning heavily towards a small group of subjects. However, we can group the AEs in many different ways. For example, if, instead of summarizing by subject, we do it by a degree of sickness severity at baseline (a category that only has 6 different classes), the correlations might be more meaningful. So, the analysis becomes the following:



Groups 6 and 7 are isolated from the other groups, just like before, but group 2 became isolated too. With the extra information, we can now see a clear association between groups 3 and 5, as well as between 1, 4 and 8. All these would be legitimate questions to further research. The badness of fit for this particular model is of just 0.04, which means that the distances between the points are really similar to those from the correlation coefficients.

A third analysis that could be done with the same approach is to count the adverse events by treatment. The resulting numbers, however, are very similar for all treatments, and so the squared correlations are almost 1 for all 8 AE classes. Instead of creating that analysis, a more interesting method is to analyze by treatment (using the BY statement in SAS) counting for subject as in the original approach. The results are shown below, in the two contiguous figures.





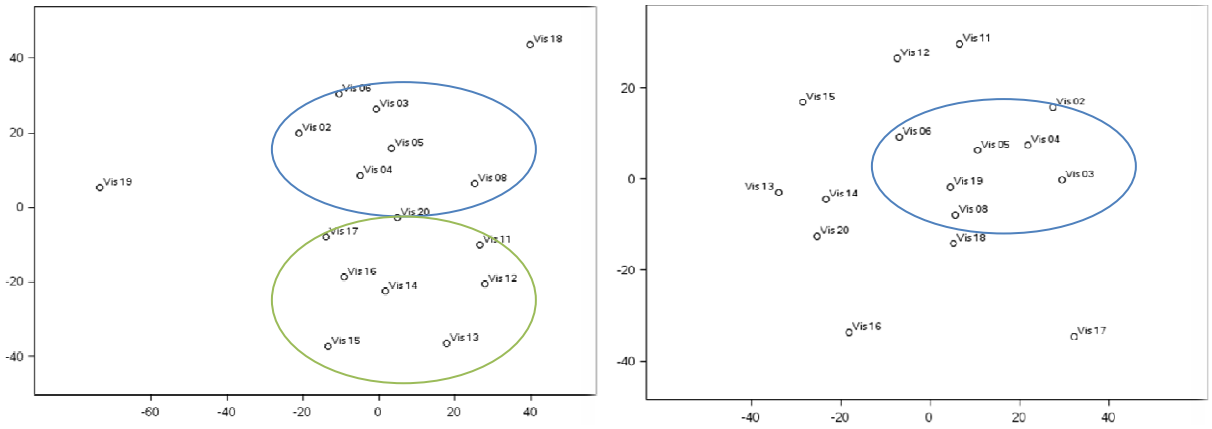
Unfortunately, there is not much that we can read from this new analysis. For the first treatment, all points look dispersed in an even cloud. The second is somehow similar with a 90 degree rotation to the right. Groups 4 and 8 are extremely similar in the second treatment (with group 5 somehow near), so that might be worth reviewing.

EXAMPLE 2: EFFICACY DATA – VISITS.

For the second example, let us consider an efficacy result. Suppose that we have a clinical trial with 20 visits, and a particular endpoint that we measure on all visits except 1, 9 and 10. We know that the ten first visits correspond to the first phase of the study, while the following 10 to a second. Multidimensional scaling was used as a first approach to see graphically if there are any similarities in the improvement between each visit. In order to do so, we considered the difference between the analysis values in a visit with the result in all the other visits for every single subject. The mean values of these differences were then used in the VAR statement of PROC MDS.

Since we are interested in the real distances between the visits and not just in the relative position according to the distances, the option LEVEL = ABSOLUTE was used in the PROC MDS statement instead of the default LEVEL = ORDINAL, used in the example above.

The result, separated by treatment, is the following:

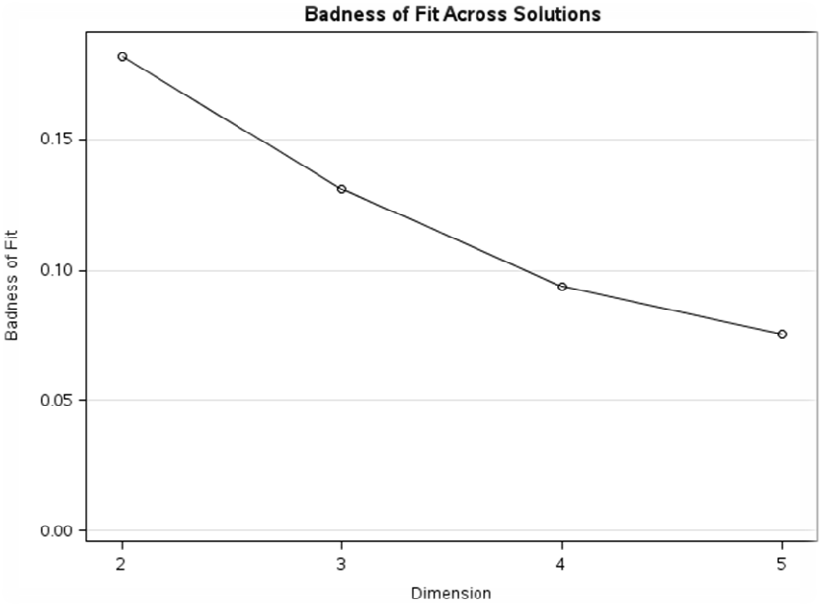


The first thing to note is that we have to create two separate graphs for each treatment (as in the previous example, because the dissimilarity function was defined between visits, but not between visits and treatment. There might be a way to define the difference so that both visits and treatment are taken into account, but that would need to include a distance between every possible pair of combinations visit/treatment. We also cannot use the same graph and overlap the figures because, even if the ranges of the dimensions were the same (they are not), there is no guarantee that the dimensions are equal. We could be comparing different things. The only reasonable comparison that we can do, at least with this approach, is to compare them side by side, and only as a rough indicator of the possible similarities.

Using this interpretation method, we can see that the visits for the first phase are all located near the upper right section of each graph (enclosed in a blue oval). For the second phase, the first treatment has them localized below those of the first, with the exception of visits 18 and 19, which are isolated from the others. The second treatment, however, has the visits from the second phase all over the figure, with visit 19 even inside the area of the first phase. Some differences in the design of the two phases could have caused this discrepancy.

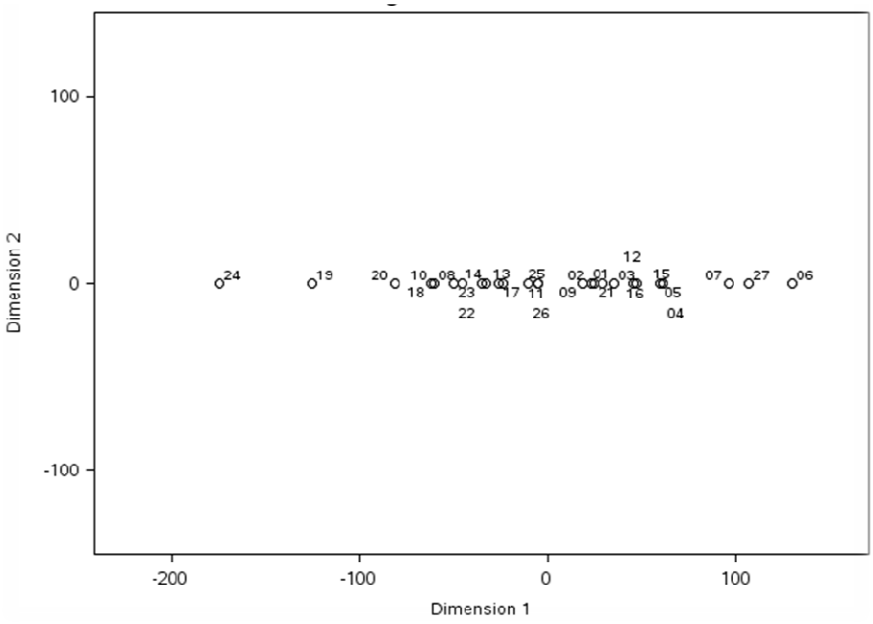
Unfortunately, the badness-of-fit for this analysis is not that good. For the first treatment, the value is 0.14, but for the second it is 0.18. This might account for the differences in the graph, and not necessarily a difference in treatment for the second phase. Would the badness-of-fit have been better if we had used more dimensions to create the multidimensional representation? Using the option “DIM =” in the PROC MDS statement the exact dimensions to be run can be specified, and SAS includes a figure to compare the badness-of-fit statistic.

The badness-of-fit figure of the second treatment for the example, using “DIM = 2 to 5 by = 1”, is shown below.



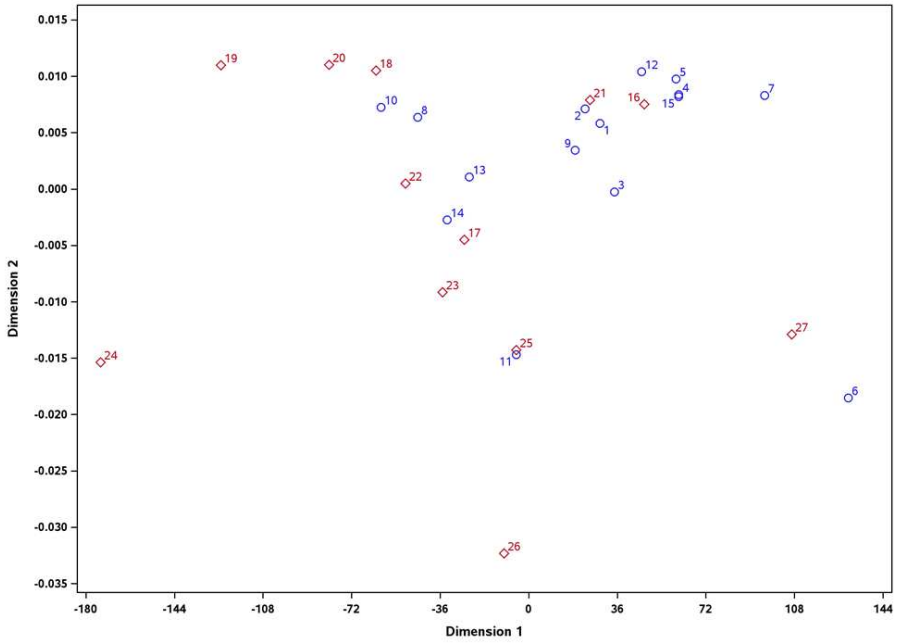
EXAMPLE 3: EFFICACY DATA – SUBJECTS.

For the final example, let us consider the same study as in the example above. However, instead of grouping by visit, the data will be grouped by subject. Let us take a group of subjects from the study, and a statistic of the result under study. In this case, the maximum value of each subject was used, but any other value such as the mean, last value or any statistic that might have more meaning can be used. As a dissimilarity function, the absolute value of the difference between the results was considered again and processed using PROC MDS. The resulting graph is shown below:



Two difficulties are immediate. Even if we had a quick way to determine which treatment each subject belonged to, it would be better if the figure showed it to us in the graph. Also, the points seem to be all allocated within a very flat line, almost as an unidimensional scale. SAS uses the same scale for the graphical representation of both dimensions, which in this case is not helpful.

As had already been mentioned above, we can use the coordinates given in the output dataset and recreate the figure by ourselves, using any method we fancy. The recreated version, using red squares for one treatment and blue circles for the second, is the following:



As we can see, the result is not unidimensional at all! And there is a very clear pattern, with the red squares coalescing in the left side while the blue circles grouping to the right. It might also be worthy to investigate why subjects 6, 24, 26 and 27 are isolated from the other subjects.

CONCLUSION

As we saw, Multidimensional scaling can be a useful technique for screening possible differences in the data. Different groupings and the use of different dissimilarity functions can generate different results, so it should be useful to experiment with different configurations. Finally, as usual with any statistical technique, Multidimensional scaling will never be a substitute of the knowledge about the data being researched, so it is good to have the experts at hand for final interpretation.

REFERENCES

1. Borg, Ingwer, and Patrick J.F. Groenen, *Modern Multidimensional Scaling,* 2nd Edition, Springer, 2005
2. Borg, Ingwer, Patrick J.F. Groenen, and Patrick Mair, *Applied Multidimensional Scaling*, Springer, 2013
3. Cox, Trevor F., and Michael A.A. Cox, *Multidimensional Scaling,* Chapman and Hall, 2000
4. "The MDS Procedure", *SAS/STAT® 13.1 User's Guide*, SAS Institute, Inc, 2013, *support.sas.com/documentation/onlinedoc/stat/131/mds.pdf*

ACKNOWLEDGMENTS

I would like to thank Sascha Ahrweiler, Michael Badlani and Karnika Dalal for their support (and pushing me) to create this paper; Yoriko de Sanctis and Miriam Tamm for their encouragement and support and to my wife, Claudia, for listening to all my complaints and help me to review it.