Leon Wu

CID: 01190736

M3S20 Coursework

February 18, 2020

1 Data

The data for this analysis was collected by myself. The code and commands used for collecting, constructing and analysing this dataset can be found on my Github: https://github.com/leonwu4951/Paper-Configurations

Choice of Data

The aim of this analysis is to reveal configurations of academic papers so that a user can find targeted lists of research materials related to a certain paper or topic, without relying on manually validating papers and using complex searching methods within databases of academic papers. In this way, the user can visualize distances between papers, chose the relevant ones and perhaps observe and explore clusterings of papers that are about similar topics.

This is useful since you can understand better how different topics and individual papers link to one another in a more informative and exploratory way.

The method of collecting the data is as follows:

- Scrape statistics lecture notes to acquire a list of $p \approx 50$ words (ignoring common English words) that will be used to describe each paper. See the wordcloud generated from the lecture notes (**Figure 1**) to get an idea of these words.
- Use keywords and Google Scholar (See Github link for details) to collect papers relating to the subject of statistics
- Count the number of occuruncs of each of the p words in each of the $n \approx 50$ papers.

This result is an n x p matrix. This can easily be expanded to include more words and papers, and a small example ($n, p \approx 50$) is used in order to better illustrate the methods. A heatmap representing the data used can be found in **Figure 2**.

2 Distances / Dissimilarities Between Papers

A distance matrix between the n papers is to be computed. Different metrics to determine these distances can be used to acheive this. In this case, papers with similar (standardized) frequencies of words could be considered "close" to eachother.

A natural choice is to use the Euclidean distance (L2 norm). Other non-euclidean metrics can be used, and may be better for some applications. In this case, the Manhattan distance (L1 norm) could be useful in that if there is a paper which mentions the word "Python" a disproportionate number of times, say, then this paper would have a very large distance between other papers when using the L2 norm. Using the L1 norm reduces this weighting of extreme values which could be desirable.

3 Multi-Dimensional Scaling

Classical Scaling

Classical scaling aims to create a configuration of the data from the distances between points. From **Figure 3** it can be seen that there only 3 large eigenvalues for the Euclidean distance matrix. This means that the data can be projected onto 3 (possibly 2) dimensions without losing too much information. For the Manhattan distance matrix, there are perhaps 3-7 significantly large eigenvalues (**Figure 4**). However, projection onto 2 dimensions would likely yield useful results still. A 2-dimensional plot is also extremely informative to human users and so this is the dimensionality that is chosen. The negative eigenvalues in Figure 4 arise due to the fact the distances between papers are non-euclidean.

The configurations acquired from applying classical scaling are shown in **Figures 5 and 6** for the Euclidean and Manhattan distances respectively. (A reduced number of papers is displayed in order to increase interpretability.)

Ordinal Scaling

Ordinal scaling also aims to create a configuration of the data, but minimizes a penalty function that measures the "stress" of the configuration. Using a random start, the resulting configuration of the ordinal scaling results in a lower stress than that of the classical scaling solution. **Figure 7** shows the configuration from the ordinal scaling of the Manhattan distances matrix.

4 Comparison of Configurations

Procrustes Analysis can be used to compare configurations. **Figure 8** shows the aligned configurations for the classical and ordinal configurations from the Manhattan distances. The configurations have some similarities, especially around the edges, but are distinctly different from eachother. The ordinal scaled configuration will be chosen since its stress was lower as previously discussed.

5 K-Means Clustering and Labelling

A K-Means clusterings is applied to the data to obtain 4 clusters. In order to label these clusters, the 3 nearest neighbours were computed for each centroid. The most common word (standardized) was chosen to be the label for that centroid. The final configuration is plotted using the ordinal scaling with the labelled centroids in **Figure 9**.

6 Conclusion

Using this method, any academic search term can be queried, resulting in a 2D configuration that allows the user to visualize inspect and collect papers that are most similar to search terms in a fast and automated way. In the case of the example configuration, the user can choose to explore papers that mention "inference" often, and find similar papers to the statistics lectures notes (located at position "12") This is extremely useful and time saving when collecting research material!

Extensions

An interactive version of this analysis could be developed, so that general search terms and a larger number of papers and words can be used to construct more complex configurations. The labelling of clusters is also far from optimal; for example the number of clusters was arbitrary and can be decided in a more systematic way using silhouette plots, say.

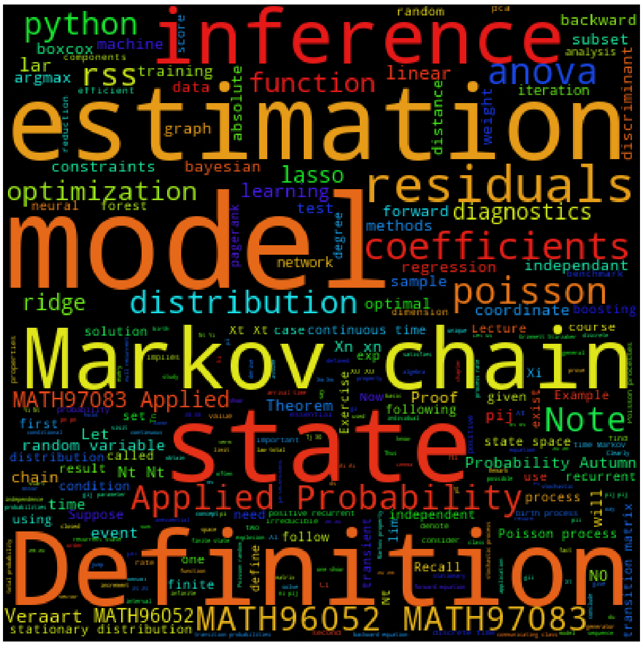

Figure 1: High-frequency words in Statistics Lecture Notes

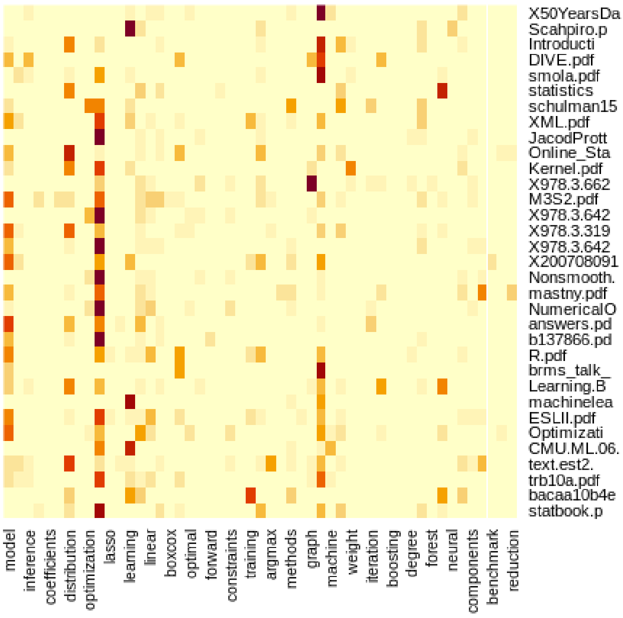

Figure 2: Heatmap word frequencies for each paper

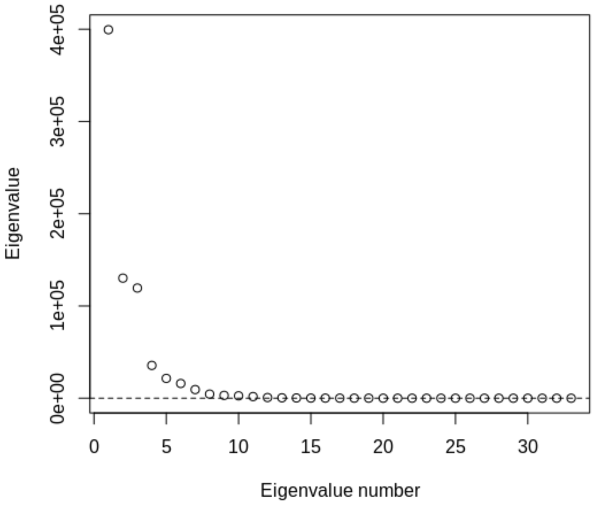

Figure 3: Eigenvalues from the Euclidean Distance Matrix

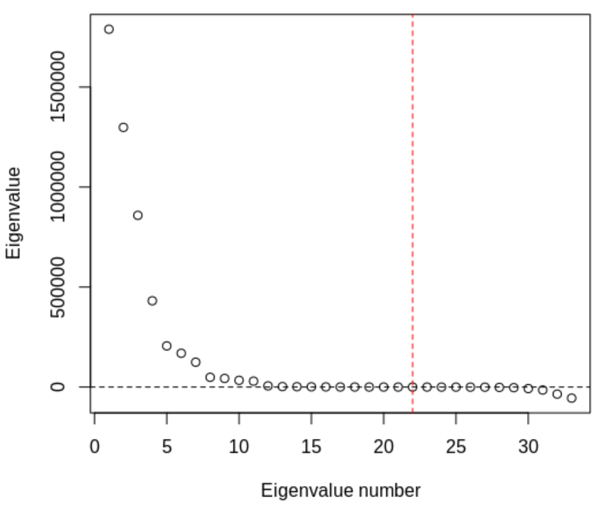

Figure 4: Eigenvalues from the Manhatten Distance Matrix

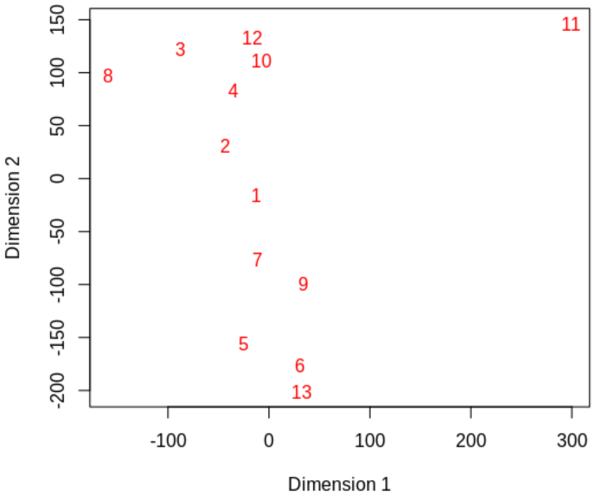

Figure 5: Configuration from Classical scaling (Euclidean)

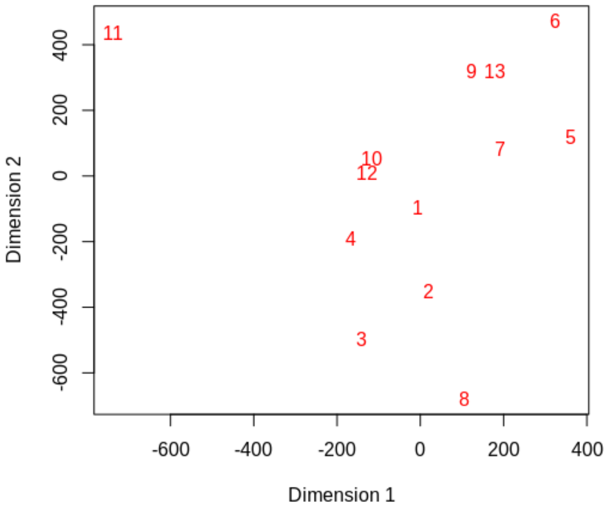

Figure 6: Configuration from Classical scaling (Manhattan)

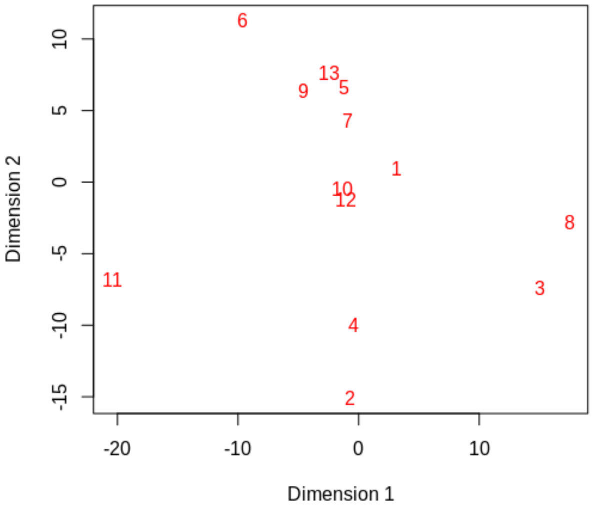

Figure 7: Configuration from Ordinal scaling (Manhattan) with lower stress

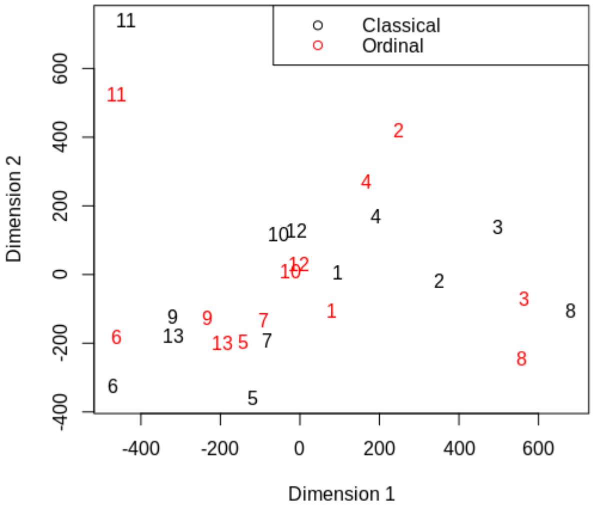

Figure 8: Classical and Ordinal Configurations aligned (Manhatten)

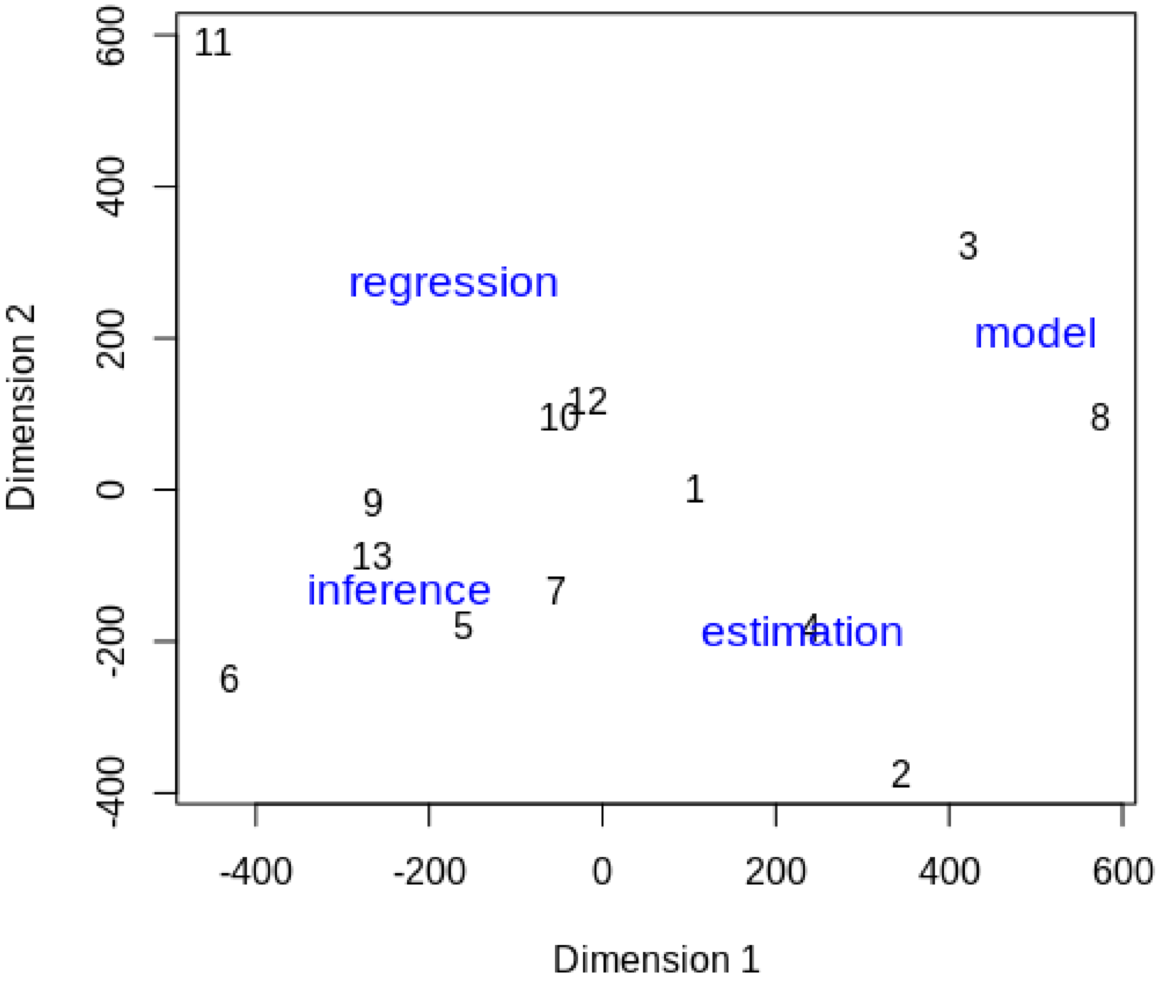

Figure 9: Final Labelled Configurations using K-Means