Introduction to Statistical Learning Coursework

Alix Vermeulen (CID: 01564896)

1. Data Selection

My data is a contraceptive method choice data set based on the demographic and socio-economic characteristics of women in Indonesia in 1987 found on the UCI Machine Learning Repository. I found this dataset interesting to investigate because family planning & contraception is an important issue in emerging countries such as Indonesia where contraception prevalence is low. This issue is one that is tackled under the Sustainable Development Goals hence I thought it was worth investigating. As the population keeps increasing and is predicted to reach 9 billion by 2050, birth control is an important topic to tackle, especially in emerging countries. Besides, I also chose this dataset to investigate which factors have the most influence on the choice of contraception for women. The choice of contraception here is between 3 categories: 1. No use of contraception, 2. Short-term use and 3. Long-term use.

My data contains entries from 1473 individuals (women) and 9 attributes. These attributes correspond to: age, education level, standard-of-living index, religion, employment etc. (see Appendix). Most of the variables are ranked from 1 to 4 where 1=low and 4=high, some are ranked by either 1=Yes and 0=No, only the age is not on a scale. Therefore, I first scaled my data so that all variables would be on the same scale. My data was then ready for a distance-based method analysis.

My reason for using distance-based methods here is to be able to locate dissimilarities between social-economic factors and explore how these affect women's choice of contraception. My hypothesis is that women in Indonesia with a high economic and social status (for example, high level of education, currently employed, high standard of living index etc.) will benefit most likely from long-term use of contraception because they should be able to afford it and/or have access to it. Besides, they most likely have less time to be having babies because they are working.

I will firstly conduct Classical Multidimensional Scaling on my data and then a Self-Organising Map analysis.

2. Classical Multidimensional Scaling

I firstly performed Classical Multidimensional Scaling (MDS) on my data to see how my data was spread (see Figure 2 in Appendix). I decided to use a non-Euclidean distance, the Manhattan distance as this gave me the best results and best clusters. I initially picked a large number of dimensions (9) and then reduced to 2 dimensions by the help of the eigenvalue plot. My eigenvalue plot (Figure 3) only seems to show one eigenvalue that stands out from the rest which would imply that only one dimension is needed for MDS. However, my aim is to visualise the data to see patterns hence I will pick the first two biggest eigenvalues and hence project the data in 2 dimensions for my analysis.

From the MDS plot, there appears to be some darker spots towards the right-hand side where more data is clustered meaning that women who have a similar socio-economic background tend to use similar contraceptive methods. Perhaps with less data, this would be slightly easier to observe.

3. SOM Analysis

Through the use of self-organising maps (Figure 5), we can observe which attributes have the most influence on the data. It appears that there is some clustering in the bottom right corner of the SOM where we can see that the most influential variables are the women's religion and whether she is employed or not. And on the other side, the left-hand side of the SOM, the level of the wife's education and her husband's education seem to be the most influential factors. This corresponds to my hypothesis in that the more educated you are, the more likely you are to have a job and earn money to afford contraception. Moreover, media exposure and the number of children the women already has don't seem to have a big influence on the data.

4. Conclusion

To conclude, distance-based methods have been helpful in examining the dissimilarities in socio-economic factors that lead to different uses of contraception in women in Indonesia.

The data set I chose is from 1987 and I believe that there has been many technological advancements and social progress since then meaning that more women have access to contraception. The data for Indonesia probably looks very different now. Nevertheless, I still think it is worth investigating because less developed countries today (compared to Indonesia) can learn and understand patterns that were previously observed in emerging countries.

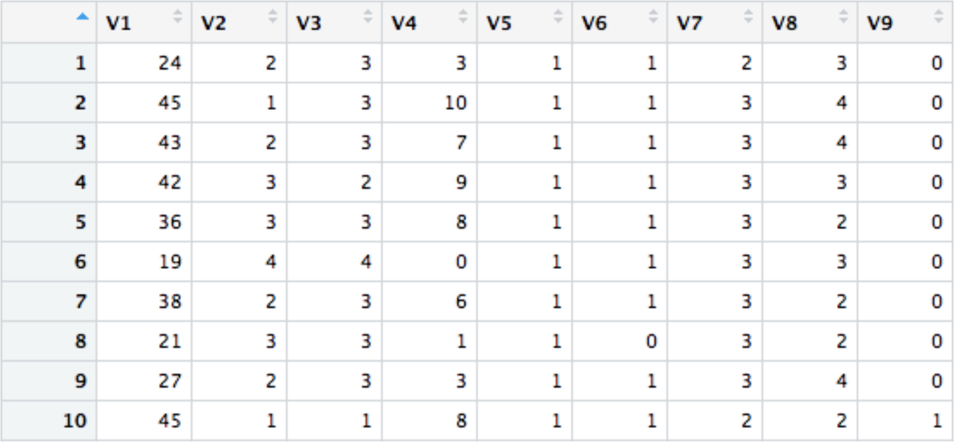
Finally, I think the limitations of this data was the fact that it was only gathered data from 1473 women. With a larger sample size, we would get more precise results. I think this data set was interesting however because it also highlights the distinct inequalities and separation within one country. An extension of this task would be to do this analysis on a world scale as we could see a very big contrast between wealthy/developed countries and emerging/developing countries.

5. References

My data set was obtained from UCI Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/Contraceptive+Method+Choice

6. Appendix

The 9 attributes correspond to:
V1. Wife's age (numerical)
V2. Wife's education (categorical) 1=low, 2, 3, 4=high
V3. Husband's education (categorical) 1=low, 2, 3, 4=high
V4. Number of children ever born (numerical)
V5. Wife's religion (binary) 0=Non-Islam, 1=Islam
V6. Wife's now working? (binary) 0=Yes, 1=No
V7. Husband's occupation (categorical) 1, 2, 3, 4
V8. Standard-of-living index (categorical) 1=low, 2, 3, 4=high
V9. Media exposure (binary) 0=Good, 1=Not good



	V1	V2	V3	V4	V5	V6	V7	V8	V9
1	24	2	3	3	1	1	2	3	0
2	45	1	3	10	1	1	3	4	0
3	43	2	3	7	1	1	3	4	0
4	42	3	2	9	1	1	3	3	0
5	36	3	3	8	1	1	3	2	0
6	19	4	4	0	1	1	3	3	0
7	38	2	3	6	1	1	3	2	0
8	21	3	3	1	1	0	3	2	0
9	27	2	3	3	1	1	3	4	0
10	45	1	1	8	1	1	2	2	1

Figure 1: First 10 rows of the data

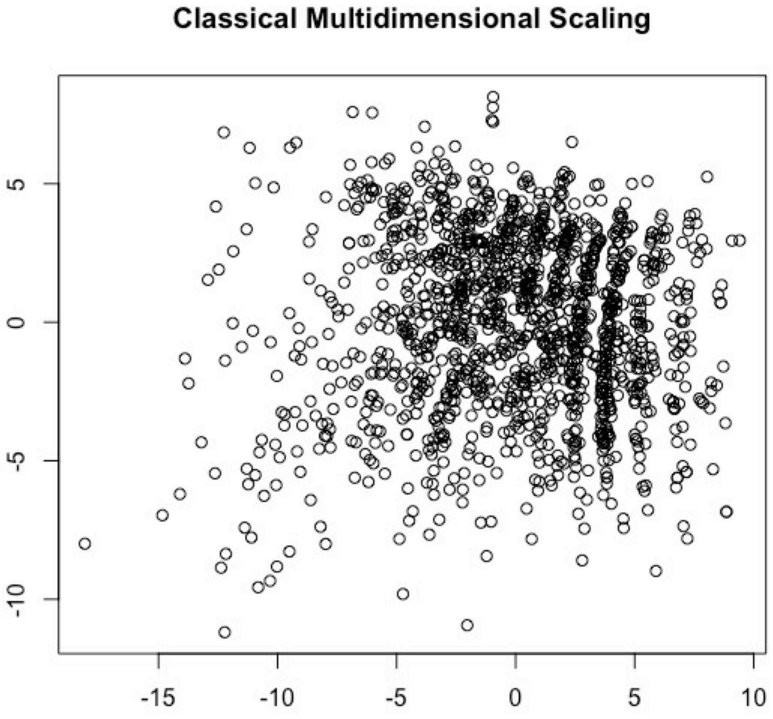


Figure 2: CMS of the data using Manhattan distance

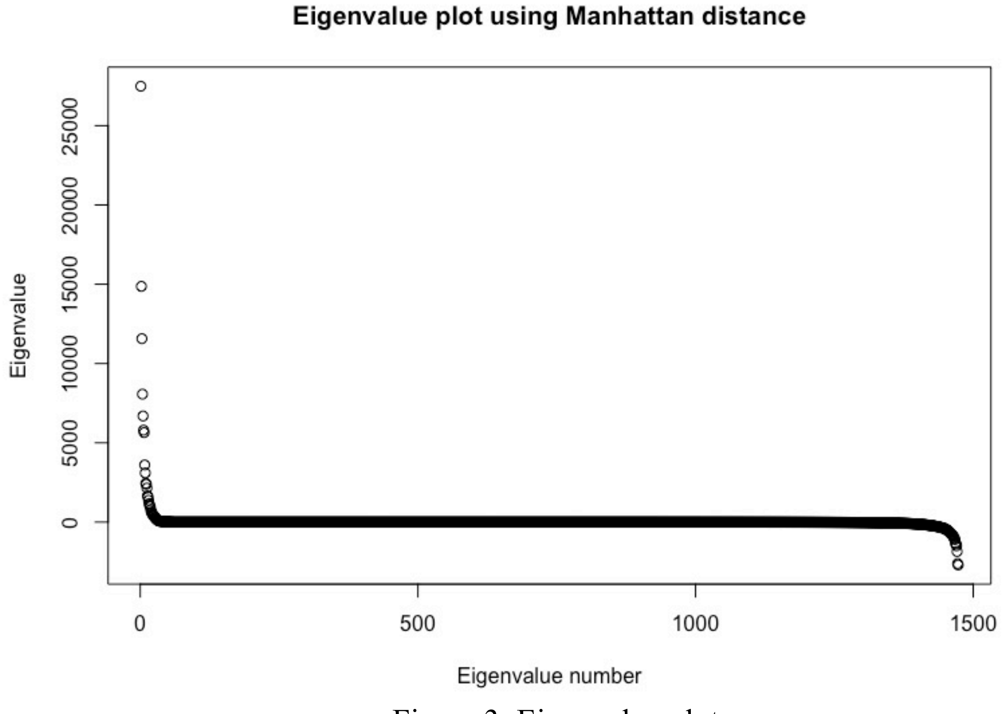


Figure 3: Eigenvalue plot

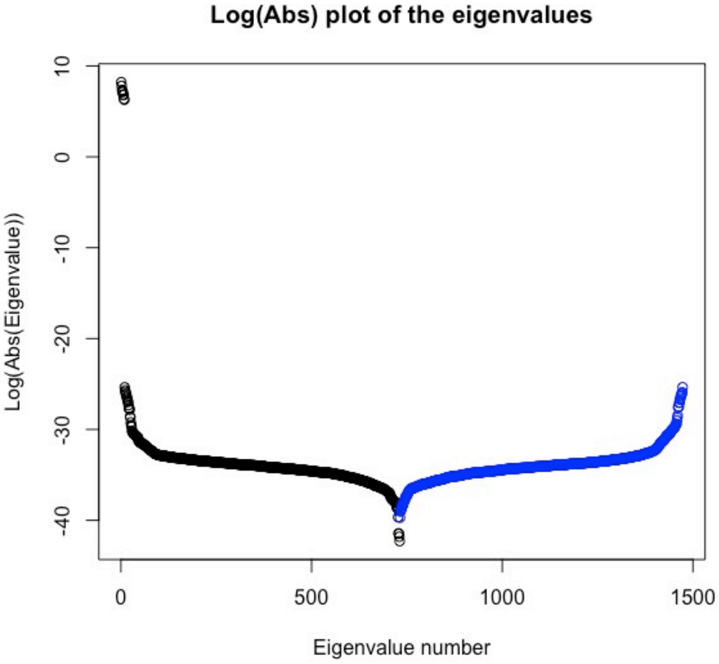


Figure 4: Log(Abs) plot of eigenvalues

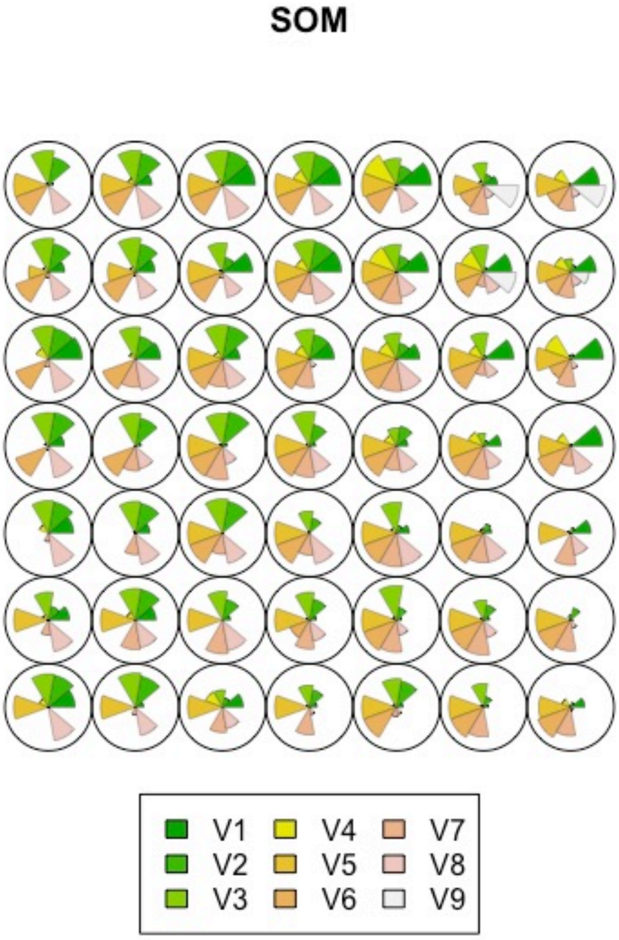


Figure 5: Self-organising map