ECMA 31330 - Problem Set 4

Bruno Aravena M.

Due date: February 20th 3.30pm

Instructions: For every assignment, you will be asked to submit a document where you write up your answers and insert the figures/ tables requested for the assignment.You will also be required to submit the code where you code your simulations and produce your answers, figures, tables, etc. You can write your code in any programming language you want, but I will be submitting solutions and presenting examples in TA sessions in R.

This means you will submit a .pdf where you answer the questions and include your final figures, and also a script of code (e.g. .R) where I can see your work. Even if you work with a document that integrates code and answers (e.g. R Markdown or Jupyter Notebook I would insist in you also sending a .pdf version of it.

1 Principal Component Analysis

For this problem set we will be using datasets from the Statistical Learning textbook. In R these can be accessed through the ISLR2 package, details and explanation for each dataset can be found in the link. If you are not using R, the .csv files can be found here.

1. Consider the projection interpretation of PCA (section 12.2.1 of the textbook): Explain why it is fundamental to standarize your variables before running PCA. (Hint: think on how having different variances in your x's would influence the optimization problem).

2. Consider the projection interpretation of PCA (section 12.2.1 of the textbook): Explain what are the loadings and scores in this interpretation.

3. For the rest of the question we will revisit the College dataset used in TA Session. Load this dataset and convert the Private variable into a binary. You will be left with only numeric variables.

4. Run PCA. Plot the loadings of all variables in the first two principal components. Comment on which 5 variables load more into the first factor, and which 5 variables load more on the second.

5. Consider now the scores of each observation. List the 5 observations with highest score for the first principal component, and the 5 observations with the lowest. Repeat this for the second principal component.

6. Considering your answers to the previous two questions can you offer a raw description of what each principal component is more or less capturing?

7. Report the variance for the scores of each principal component. What relationship do you observe and how does this relate to PCA?

2 Matrix Completion

For this question, we will be going through an example of Matrix Completion. I would suggest going over section 12.3 of the textbook.

1. Suppose we have a matrix with missing entries. A simple "baseline" approach is to replace a missing entry x_{ij} by the average of the observed entries for that column $\bar{x}_j$. Suppose instead you are asked to come up with vectors $\{a\}_{i=1}^N$ and $\{b\}_{j=1}^p$ such that the missing entry x_{ij} is now replaced by $a_i b_j$. Describe a choice of $\{a\}_{i=1}^N$ and $\{b\}_{j=1}^p$ that still implements the "baseline" imputation? In what sense is this new approach better than the "baseline"?

2. Fix your $\{b\}_{j=1}^p$ from your answer above. To simplify things further, consider that there is only one column which has some missing entries. You now want to consider an imputation that does not treat each unit of observation i equally. Can you provide an intuition that explains which units should get larger values and which units should get lower values?

3. How does this intuition connect to PCA?

4. Read the Boston dataset from the textbook. This dataset gives information about different neighborhoods in this city. More detail can be found here. Standardize the columns and save this version of the data. From this data, create a different version where the columns `crim`, `indus`, `age`, `dis`, `tax` and `medv` have 20% chances of being missing. This probability should be independent across variables i.e. You should **NOT** end up with the same 20% of rows missing all these columns.

5. Implement matrix completion with all principal components and with only 2. There are different packages to do this. You should end up with 2 different "completed" matrices.

6. For each variable that had missing data create a plot where you overlay: the real missing values, the values predicted by matrix completion with all principal components and the values predicted by matrix completion with only 2 principal components.

7. Report a mean squared error (comparing real value and imputed) for each of these two approaches and each of the 6 missing columns. Which approach works better ?

8. *(Optional, feel free to respond a subset)* Are there variables for which the matrix completion works better than for others? Can you propose a pattern that explains this difference? Why does PCA might perform different in this case? Are there fixes we could have implemented?