

Module 10: Recommended Exercises

TMA4268 Statistical Learning V2024

Daesoo Lee, Kenneth Aase, Stefanie Muff, Sara Martino
Department of Mathematical Sciences, NTNU

April 4, 2024

PCA on New York Times stories

This exercise is based on the New York Time stories example (code and data) on the [lecture](#) of Brian Junker and Cosma Shalizi about Principal components and factor analysis.

Data Exploration

New stories were randomly selected from the [New York Times Annotated Corpus](#). There are 57 stories about art and 45 about music on the dataset available.

The `nyt_data` is a dataset containing these 102 stories, which represent the 102 observations. The first column contains class labels of the stories (art and music) and, and the other stories contain the counts of each word in a given story (weight by the inverse document frequency and normalized by vector length).

The New York Times stories dataset are contained in the file `pca-examples.rdata` which is already included in the directory. You can still download it from google drive (<https://drive.google.com/open?id=1vaK9GDvMw4Hsuv0T1jHeq9ZyrqLhJ6MR>) and store in the directory of your Rmd file. The `pca-examples.rdata` can be loaded with the following code.

```
load("pca-examples.rdata") # use this when knitting
# load("Exercise10\\pca-examples.rdata") # use this when running the code in Rmd

# We will work with nyt.frame
nyt_data <- nyt.frame # dimension: (102, 4432)
```

```
str(nyt_data)
summary(nyt_data$class.labels)
```

Let's check some word samples:

```
colnames(nyt_data)[sample(ncol(nyt_data),30)]
```

```
## [1] "pendant" "brought" "structure" "willing" "yielding" "bare"
## [7] "school" "halls" "challenge" "step" "largest" "lovers"
## [13] "intense" "borders" "mall" "classic" "conducted" "mirrors"
## [19] "hole" "location" "desperate" "published" "head" "paints"
## [25] "another" "starts" "familiar" "window" "thats" "broker"
```

Let's check some values in the dataset. We have many zeroes, as most words do not appear in most stories.

```
signif(nyt_data[sample(nrow(nyt_data),5),sample(ncol(nyt_data),10)],3)
```

```
##      jacket patch tapes   want   ford failed condemn intentional confined
## 24         0     0     0 0.0000 0.0000 0.0000         0         0         0
## 2          0     0     0 0.0275 0.0704 0.0000         0         0         0
## 85         0     0     0 0.0482 0.0000 0.0000         0         0         0
## 59         0     0     0 0.0000 0.0000 0.0000         0         0         0
## 76         0     0     0 0.0000 0.0000 0.0215         0         0         0
##      destroyed
## 24             0
## 2              0
## 85             0
## 59             0
## 76             0
```

Problem 1

- For the New York Times stories (`nyt_data`) dataset:
 - Create a biplot and explain the type of information that you can extract from the plot.
 - Create plots for the proportion of variance explained (PVE) and cumulative PVE. Describe what type of information you can extract from the plots.

Problem 2

Show that the algorithm below is guaranteed to decrease the value of the objective

$$\underset{C_1, \dots, C_K}{\text{minimize}} \left\{ \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i, i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2 \right\}$$

at each step.

Algorithm 10.1 *K-Means Clustering*

1. Randomly assign a number, from 1 to K , to each of the observations. These serve as initial cluster assignments for the observations.
 2. Iterate until the cluster assignments stop changing:
 - (a) For each of the K clusters, compute the cluster *centroid*. The k th cluster centroid is the vector of the p feature means for the observations in the k th cluster.
 - (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).
-

Problem 3

Perform k -means clustering in the New York Times stories dataset.

Problem 4

Perform hierarchical clustering in the New York Times stories dataset.