TMA4268 Statistical Learning

Chapter 10: Unsupervised Learning - NYT Stories

Thiago G. Martins, Department of Mathematical Sciences, NTNU

Spring 2022

Contents

PCA on New York Times stories

This example (code and data) are based on the lecture of Brian Junker and Cosma Shalizi about Principal components and factor analysis.

You can download the `pca-examples.Rdata` on our course website.

Data Exploration

New stories randomly selected from the New York Times Annotated Corpus. There are 57 stories about art and 45 about music on the dataset available.

```
# Modify the location based on your filesystem
load("../10_unsupervised_learning/datasets/pca-examples.Rdata")

# We will work with nyt.frame
nyt_data = nyt.frame
```

The `nyt_data` is a dataset containing 102 observations (102 new stories) and columns containing class labels of the stories (art and music) and the count of each word in a given story (weight by the inverse document frequency and normalized by vector length).

```
str(nyt_data)
```

```
## 'data.frame':    102 obs. of  4432 variables:
##  $ class.labels    : Factor w/ 2 levels "art","music": 1 1 1 1 1 1 1 1 1 1 ...
##  $ X.              : num  0.00871 0.00585 0.01604 0.02641 0.00729 ...
##  $ X.d             : num  0 0 0 0 0 ...
##  $ X.nd            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ X.s             : num  0 0 0.0114 0 0.011 ...
##  $ X.th            : num  0.00925 0 0 0 0 ...
##  $ X.this          : num  0 0 0 0 0 ...
##  $ a               : num  0.00756 0.00142 0.01006 0.00868 0.00839 ...
##  $ abandoned       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ abc             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ability         : num  0 0 0 0 0 0 0 0 0 0 ...
```

```
##  $ able            : num  0 0.0399 0 0 0 ...
##  $ about           : num  0.0533 0 0 0.0125 0 ...
##  $ above           : num  0 0 0.0536 0 0 ...
##  $ abroad          : num  0 0 0 0.041 0 ...
##  $ absorbed        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ absorbing       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ abstract        : num  0.0216 0.0435 0 0 0 ...
##  $ abstraction     : num  0 0 0 0 0 ...
##  $ abstractions    : num  0 0 0 0 0 ...
##  $ abundance       : num  0.0349 0 0 0 0 ...
##  $ academic        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ academy         : num  0 0.273 0 0 0 ...
##  $ accents         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accept          : num  0 0 0 0 0 ...
##  $ access          : num  0 0 0 0 0 ...
##  $ accessible      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acclaimed       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accommodate     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accompanied     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accompanying    : num  0.0268 0 0 0 0 ...
##  $ according       : num  0 0 0 0.0736 0 ...
##  $ accordingly     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ account         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accounted       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ accused         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ achieved        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ achievement     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acknowledge     : num  0 0 0 0.041 0.0438 ...
##  $ acknowledged    : num  0 0 0 0.0315 0 ...
##  $ acquired        : num  0 0 0.0348 0 0 ...
##  $ acquisition     : num  0 0 0 0 0 ...
##  $ acquisitions    : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acre            : num  0 0 0.0454 0 0 ...
##  $ across          : num  0 0 0.02 0 0 ...
##  $ acrylics        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ act             : num  0.0198 0 0 0 0 ...
##  $ acted           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acting          : num  0 0 0 0 0 ...
##  $ action          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actions         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ active          : num  0.0349 0 0 0 0 ...
##  $ activities      : num  0 0 0 0 0 ...
##  $ actor           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actors          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actress         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ acts            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ actually        : num  0.0198 0 0 0 0.0248 ...
##  $ adam            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adams           : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adamss          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adaptation      : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ add             : num  0 0 0 0.0736 0 ...
##  $ added           : num  0 0 0 0 0.0443 ...
##  $ adding          : num  0 0 0 0 0 ...
```

```
##  $ addition        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ additional      : num  0 0 0 0 0 ...
##  $ address         : num  0.0349 0 0 0 0 ...
##  $ addresses       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adds            : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adhering        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adjacent        : num  0 0 0.0454 0 0 ...
##  $ administration  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ admired         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ admission       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ admits          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adopted         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ ads             : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adults          : num  0 0 0 0 0.0361 ...
##  $ advance         : num  0 0 0 0 0 ...
##  $ advanced        : num  0 0 0 0 0.0438 ...
##  $ advantage       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adventure       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adventurous     : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advertisements  : num  0 0 0 0 0.0438 ...
##  $ advertising     : num  0 0 0 0 0.118 ...
##  $ advice          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advised         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ adviser         : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advising        : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ advocates       : num  0 0 0 0.041 0 ...
##  $ aesthetic       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ affair          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ affairs         : num  0 0 0 0.101 0 ...
##  $ affect          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ affected        : num  0.0627 0 0 0 0 ...
##  $ affection       : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ afford          : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ afraid          : num  0 0 0 0 0 0 0 0 0 0 ...
##   [list output truncated]
```

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

```
##   art music
##    57    45
```

Let's check some word samples:

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

```
##  [1] "penchant"  "brought"   "structure" "willing"   "yielding"
##  [6] "bare"      "school"    "halls"     "challenge" "step"
## [11] "largest"   "lovers"    "intense"   "borders"   "mall"
## [16] "classic"   "conducted" "mirrors"   "hole"      "location"
## [21] "desperate" "published" "head"      "paints"    "another"
## [26] "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
```

PCA

We will now perform PCA. We exclude the first column since it is related to the response of the data. We did not set `scale=TRUE` because the data has already been normalized.

```
nyt_pca = prcomp(nyt_data[,-1])
```

Loadings

Note that we can have at most 102 PCs in this case.

```
nyt_loading = nyt_pca$rotation
dim(nyt_loading)
```

```
## [1] 4431  102
```

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

```
##                 PC57     PC86     PC70      PC30     PC90     PC53
## display     -0.01220  0.00697 -0.00888 -0.011200  0.01630 -0.00246
## feminist     0.00365  0.01830  0.00131 -0.021800 -0.01170 -0.03150
## posturing   -0.00165  0.02020  0.00473  0.000924 -0.00708 -0.01140
## emotionally  0.00194  0.02900  0.00470  0.002250 -0.00552 -0.03400
## finally      0.00973 -0.00408  0.00160  0.008220  0.00737 -0.01060
##                  PC27     PC32      PC88     PC59
## display     -0.001150  0.00257 -0.004390 -0.00447
## feminist     0.016800  0.00121 -0.005450 -0.00646
## posturing    0.000349  0.00834  0.000553 -0.00466
## emotionally -0.004020 -0.00187 -0.007260 -0.00595
## finally     -0.009170  0.00992  0.007140  0.01370
```

Show the 30 words with the biggest positive loading on PC1:

```
signif(sort(nyt_loading[,1],decreasing=TRUE)[1:30],2)
```

```
##       music        trio     theater   orchestra   composers       opera
##       0.110       0.084       0.083       0.067       0.059       0.058
##    theaters           m    festival        east     program           y
##       0.055       0.054       0.051       0.049       0.048       0.048
##      jersey     players   committee      sunday        june     concert
##       0.047       0.047       0.046       0.045       0.045       0.045
##    symphony       organ     matinee   misstated instruments           p
##       0.044       0.044       0.043       0.042       0.041       0.041
##         X.d       april      samuel        jazz     pianist     society
##       0.041       0.040       0.040       0.039       0.038       0.038
```

Show the 30 words with the biggest negative loading on PC1:

```
signif(sort(nyt_loading[,1],decreasing=FALSE)[1:30],2)
```

```
##       she       her        ms         i      said    mother    cooper
##    -0.260    -0.240    -0.200    -0.150    -0.130    -0.110    -0.100
##        my  painting   process paintings        im        he       mrs
##    -0.094    -0.088    -0.071    -0.070    -0.068    -0.065    -0.065
##        me  gagosian       was   picasso     image sculpture      baby
##    -0.063    -0.062    -0.058    -0.057    -0.056    -0.056    -0.055
##   artists      work    photos       you    nature    studio       out
##    -0.055    -0.054    -0.051    -0.051    -0.050    -0.050    -0.050
##      says      like
##    -0.050    -0.049
```

Show the 30 words with the biggest positive loading on PC2:

```
signif(sort(nyt_loading[,2],decreasing=TRUE)[1:30],2)
```

```
##         art      museum      images     artists   donations     museums
##       0.150       0.120       0.095       0.092       0.075       0.073
##    painting         tax   paintings   sculpture     gallery  sculptures
##       0.073       0.070       0.065       0.060       0.055       0.051
##     painted       white    patterns      artist      nature     service
##       0.050       0.050       0.047       0.047       0.046       0.046
##  decorative        feet     digital      statue       color    computer
##       0.043       0.043       0.043       0.042       0.042       0.041
##       paris         war collections     diamond       stone     dealers
##       0.041       0.041       0.041       0.041       0.041       0.040
```

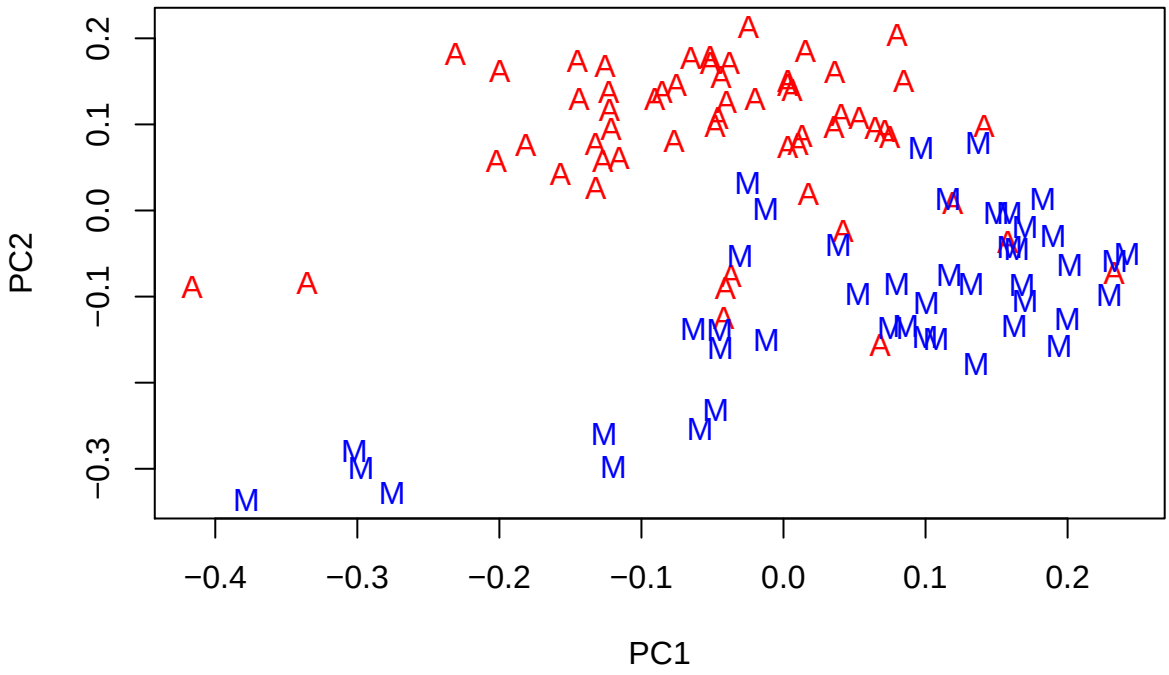
Show the 30 words with the biggest negative loading on PC1:

```
signif(sort(nyt_loading[,2],decreasing=FALSE)[1:30],2)
```

```
##          her          she      theater        opera           ms
##       -0.220       -0.220       -0.160       -0.130       -0.130
##            i         hour   production         sang     festival
##       -0.083       -0.081       -0.075       -0.075       -0.074
##        music      musical        songs        vocal    orchestra
##       -0.070       -0.070       -0.068       -0.067       -0.067
##           la      singing      matinee  performance         band
##       -0.065       -0.065       -0.061       -0.061       -0.060
##       awards    composers         says           my           im
##       -0.058       -0.058       -0.058       -0.056       -0.056
##         play     broadway       singer       cooper performances
##       -0.056       -0.055       -0.052       -0.051       -0.051
```

Plot the projection of the stories on to the first 2 components. Arts stories with red As and music stories with blue Ms. The separation is very good, even with only two components.

```
plot(nyt_pca$x[,1:2],type="n")
points(nyt_pca$x[nyt_data[,"class.labels"]=="art",1:2],pch="A",col="red")
points(nyt_pca$x[nyt_data[,"class.labels"]=="music",1:2],pch="M",col="blue")
```


PC2
0.2
0.1
0.0
−0.1
−0.3
−0.4
−0.3
−0.2
−0.1
0.0
0.1
0.2
PC1