
TMA4270 Multivariate Analysis H2008 Computer Exercise 4

Exercise report due Wednesday, November 26, 2008, 12hrs
Instituttkontoret, 7.etg Sentralbygg 2, or to H.Rue or A.Ottavi
STUDENTNR not NAME!

Groupsize ≤ 2

Number of pages handed in need not be more than 5, excluding figures and R code

Portraits and eigenfaces

- The report should start with a short summary (one or two paragraphs) explaining your conclusions and interpreting your model in a language suitable for a nonspecialist.
- Code, output and plots from R should be included.
- The report may be handwritten.
- The report may be written in Norwegian.
- You may work alone, or in groups of maximal group size two (and hand in one report).

Introductory information

In this exercise we will study portraits of young males or females. We will look at so-called “eigenfaces” and use a factor model to represent the portraits.

Data: Your first step is to make the data available in R. Do the following:

- Create a directory for this exercise, and go there. (Large data...)
- Start R.
- Read the female or male portraits into R by:

```
faces <- dget("http://www.math.ntnu.no/~hrue/TMA4270-2008/faces.dd")
```

Now the portraits are available in a list structure.
- Make the function `display.matrix` available

```
source("http://www.math.ntnu.no/~hrue/TMA4270-2008/display.matrix.r")
```

About the data: The data set is a collection of portraits of young medical students from at Stanford. To see the first 18 of each gender, do

```
par(ask=T)
par(mfrow=c(3,3))
for(k in 1:18) display.matrix(faces$matrix[,k])
for(k in 101:118) display.matrix(faces$matrix[,k])
```

There are 100 portraits (grey level images) of each gender; males from $k=1$ to 100 and womans from 101 to 200. Each grey level image is 100×100 . You can think of each image \mathbf{m} as a 100×100 matrix

$$\mathbf{m} = \begin{pmatrix} m_{11} & \cdots & m_{1,100} \\ & & \\ & m_{ij} & \\ & & \\ m_{100,1} & \cdots & m_{100,100} \end{pmatrix}$$

To display this matrix, we simply map each m_{ij} to a grey value and display it as an image.

The function `display.matrix`, which is included in `display.matrix.r`, does this. The argument can be either a matrix or a vector.

(Images of size 200×200 are originally available, but this is rather inconvenient due to the large memory requirements.)

Data-layout: The male and female portraits are stored as vectors

`faces$matrix[,1]` to `faces$matrix[,200]`

Each column `faces$matrix[,1]` is a vector of length $100^2 = 10\,000$ representing a 100×100 matrix stored column by column.

In “the normal notation”, we have that

$$\mathbf{x}_1, \dots, \mathbf{x}_{200}$$

are iid observations from both the faces and male (sub-)population, where each \mathbf{x}_i is a vector of length 10 000. You may convert back and forth between a vector representation of an image and a matrix representation of an image, using

```
par(mfrow=c(1,3))
x = faces$matrix[,1] ## vector
display.matrix(x)    ## display it
x = matrix(x,100,100) ## convert it to a matrix
display.matrix(x)    ## display it, you see the same...
x = as.vector(x)     ## convert it back to a vector
display.matrix(x)    ## display it, you see the same...
```

Note that the function `display.matrix`, accepts both formats.

Observe that `faces$matrix` is a (number of variables) \times (number of observations) matrix.

Study the different list elements of the `faces` list.

Creating files of images: A quick way to generate a postscript/PDF/PNG file of the images is to use

```
dev.copy2eps(file="image.eps")
dev.copy2pdf(file="image.pdf")
```

which will create the file `image.eps` or `image.pdf`, with the contents of the graphical window.

Alternatively you can use

```
postscript("image.eps")
##pdf("image.pdf")
##png("image.png",width=960,height=960)
display.matrix(faces$matrix[,1])
dev.off()
```

to make a postscript file with the image for the first portrait.

The task

1. **Mean portrait.** Estimate the mean of the portraits, **one for each gender**. Does the mean vary across the image? Which features are represented in the estimated mean portraits?
2. **Eigenvalues and eigenvectors of the covariance matrix.** In this task we will compute the eigenvalues and eigenvectors of the estimated covariance matrix of the portraits \mathbf{x} , $\hat{\Sigma}$. **Each gender has its own covariance matrix.**

What is the dimension of the covariance matrix $\hat{\Sigma}$? Is this so large that you will expect problems

- just to store it?
- What about computing the eigenvectors and eigenvalues?

Prove the following result and explain, *carefully*, why this result solves the above problems. Write out the relationship between the eigenvalues and eigenvectors of $\hat{\Sigma}$ and the below λ_i 's and \mathbf{v}_i 's.

Theorem 1 Let \mathbf{X} be a $r \times m$ matrix where $r \geq m$ with rank $k \leq m$. Denote by $\lambda_1, \dots, \lambda_k$ the k non-zero eigenvalues of $\mathbf{X}\mathbf{X}^T$ and let $\mathbf{v}_1, \dots, \mathbf{v}_k$ be the corresponding eigenvectors. Denote by $\gamma_1, \dots, \gamma_k$ the k non-zero eigenvalues of $\mathbf{X}^T\mathbf{X}$ and let $\mathbf{w}_1, \dots, \mathbf{w}_k$ denote the corresponding eigenvectors. Then $\lambda_i = \gamma_i$ and $\mathbf{v}_i \propto \mathbf{X}\mathbf{w}_i$ for $i = 1, \dots, k$.

Hint: Start with $\mathbf{X}^T\mathbf{X}\mathbf{w}_k = \gamma_k\mathbf{w}_k$ and premultiply by \mathbf{X} .

3. **Eigenfaces** Sort the eigenvalues, λ_i , of $\hat{\Sigma}$ so that

$$\lambda_1 \geq \lambda_2 \geq \dots$$

and let $\mathbf{v}_1, \mathbf{v}_2, \dots$ denote the corresponding eigenvectors. **Do this for each gender.**

The eigenvector \mathbf{v}_1 is called the *first eigenface*, \mathbf{v}_2 is called the *second eigenface*, and so on. Explain why this is a reasonable terminology.

Display the first eigenfaces (for each gender) and discuss what features they extract.

Discuss the connection between principal components and eigenfaces.

4. **Factor model**

We will now construct a factor model representing the faces, **one for each gender**.

Explain why

$$\mathbf{x} = \boldsymbol{\mu} + \sum_{i=1}^k s_i \sqrt{\lambda_i} \mathbf{v}_i + \boldsymbol{\epsilon} \quad (1)$$

where $\mathbf{s} \sim N(\mathbf{0}, \mathbf{I})$, $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \boldsymbol{\Psi})$ and $\text{Cov}(\mathbf{s}, \boldsymbol{\epsilon}) = \mathbf{0}$, is a reasonable choice.

Discuss the relation with the factor model as presented in Johnson & Wichern book.

Is (1) a reasonable model for the faces? How can such a model be verified?

Choose k so that about 80% of the estimated total population variance is explained (for each gender), and estimate the factor scores for some of the portraits. Try to verify if the assumption $\mathbf{s} \sim N(\mathbf{0}, \mathbf{I})$ is "reasonable".

Compare the values of the factor-scores for some of the portraits. Does specific features in the portraits also show up in some of the factor-scores?

Which gender seems best represented by the factor model?

5. Simulated portraits

Use the estimated factor model (1) to *simulate* portraits. Use the R-function `rnorm` to sample normal random variables.

Why are the simulated portraits looking the way they do? What assumptions are violated in the factor model? Are there any differences in the “performance” due to gender?