Exercise 2 report :

Data preprocessing

Tilman Marquart

Baptiste Piofret

Maxime Gatineau

University of Southern Denmark

*Spring Semester 2020*

Table of contents

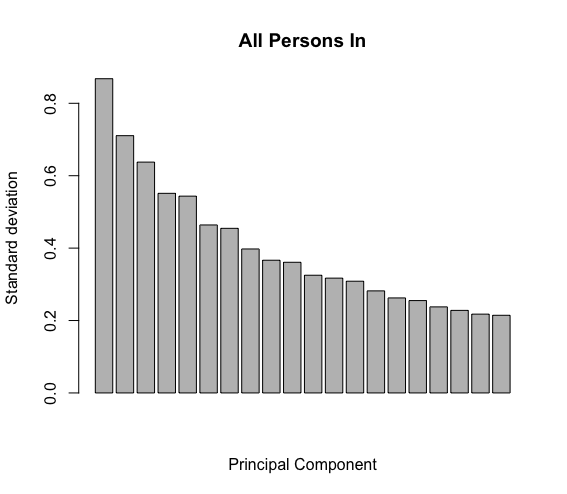
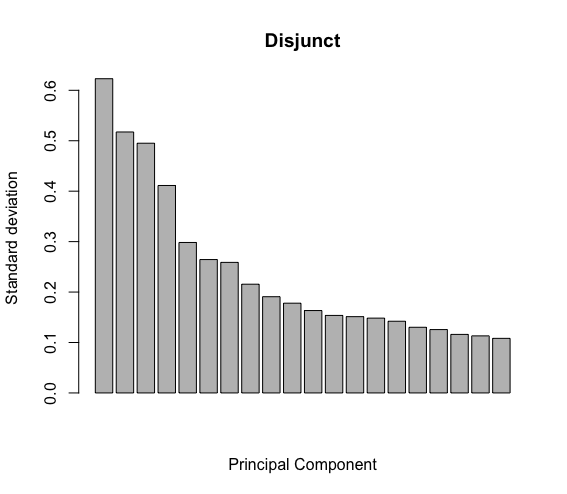
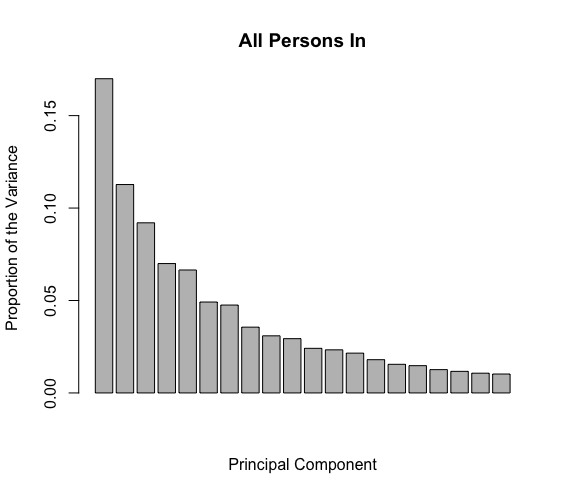
[Exercise 2.1 : Principal Component Analysis (PCA) 3](#_Toc36321891)

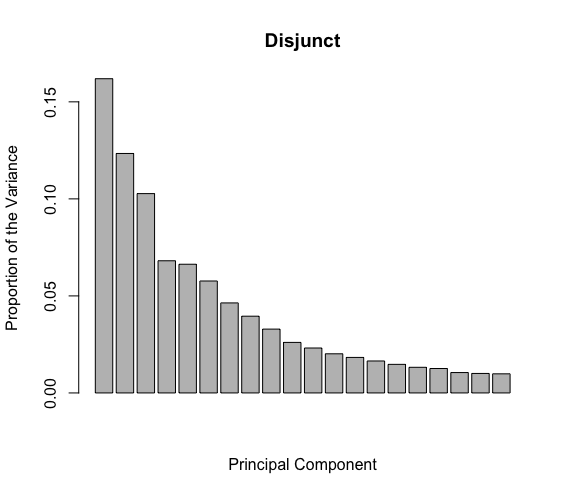
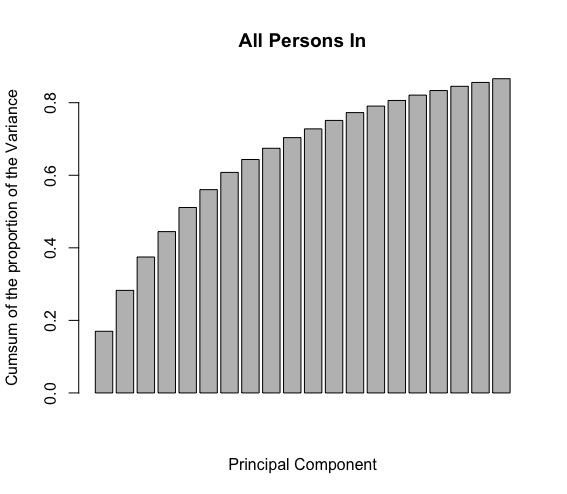
[Exercise 2.2: Normalization 6](#_Toc36321892)

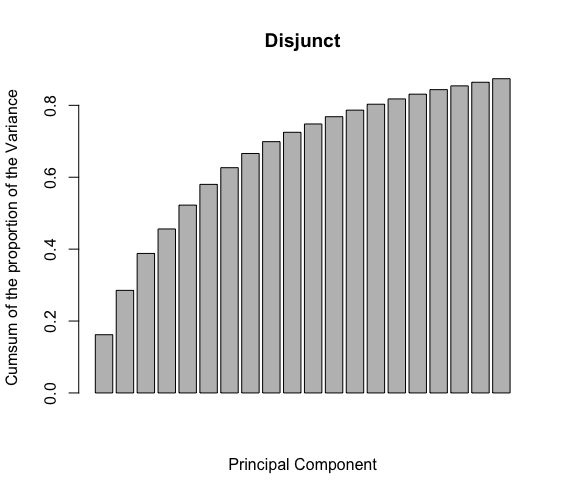
[Exercise 2.3: Preprocessing 7](#_Toc36321893)

[Exercise 2.4: Reconstruction with PCA 8](#_Toc36321894)

# Exercise 2.1 : Principal Component Analysis (PCA)







The standard deviation decreases exponentially with the importance of the parameter. We can say the same of the proportion of the variance with a squared factor to get the variance from the standard deviation. We then see the cumulative sum of the proportion of the variance follows a logarithmic shape and approaches the 100% with the number of parameters increasing.

Dataset split: All Persons In

Variance threshold 80 % needs 15 PCA components. Proportional cumulative sum of the variance from all of them is: 0.8061658

K: 3 Accuracy: 97.01 Runtime: 6.316975

K: 7 Accuracy: 96.775 Runtime: 6.525208

K: 12 Accuracy: 96.49 Runtime: 6.130448

Variance threshold 90 % needs 25 PCA components. Proportional cumulative sum of the variance from all of them is: 0.905796

K: 3 Accuracy: 97.72 Runtime: 10.35009

K: 7 Accuracy: 97.43 Runtime: 10.10074

K: 12 Accuracy: 97.125 Runtime: 10.59192

Variance threshold 95 % needs 36 PCA components. Proportional cumulative sum of the variance from all of them is: 0.9508849

K: 3 Accuracy: 97.915 Runtime: 24.41177

K: 7 Accuracy: 97.545 Runtime: 24.39653

K: 12 Accuracy: 97.225 Runtime: 21.28739

Variance threshold 99 % needs 74 PCA components. Proportional cumulative sum of the variance from all of them is: 0.9901393

K: 3 Accuracy: 97.955 Runtime: 53.18392

K: 7 Accuracy: 97.53 Runtime: 51.94326

K: 12 Accuracy: 97.165 Runtime: 52.11162

Dataset split: Disjunct

Variance threshold 80 % needs 14 PCA components. Proportional cumulative sum of the variance from all of them is: 0.8030988

K: 3 Accuracy: 78.27 Runtime: 5.582443

K: 7 Accuracy: 78.565 Runtime: 5.95058

K: 12 Accuracy: 78.48 Runtime: 5.657591

Variance threshold 90 % needs 24 PCA components. Proportional cumulative sum of the variance from all of them is: 0.9041168

K: 3 Accuracy: 82.665 Runtime: 10.27751

K: 7 Accuracy: 82.57 Runtime: 10.73692

K: 12 Accuracy: 82.23 Runtime: 9.624222

Variance threshold 95 % needs 35 PCA components. Proportional cumulative sum of the variance from all of them is: 0.9507284

K: 3 Accuracy: 83.335 Runtime: 22.57845

K: 7 Accuracy: 83.22 Runtime: 21.62002

K: 12 Accuracy: 82.995 Runtime: 19.5195

Variance threshold 99 % needs 75 PCA components. Proportional cumulative sum of the variance from all of them is: 0.9903329

K: 3 Accuracy: 83.77 Runtime: 62.75142

K: 7 Accuracy: 83.595 Runtime: 57.43834

K: 12 Accuracy: 83.095 Runtime: 1.049605

For both All persons in and Disjunct, with a greater K, the accuracy decreases but the runtime as well. And to cover a higher variance %, the runtime doubles but not the accuracy. It might sometimes be more efficient to cover a smaller percentage of the variance but save time running the algorithm.

# Exercise 2.2: Normalization

We used the min-max normalization. We chose to work with a 90% variance threshold because the runtime is much smaller, and the accuracy doesn’t vary too much.

Normalization before PCA:

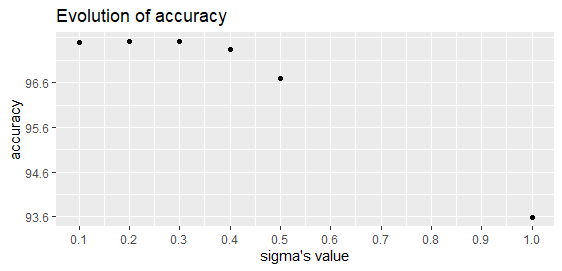
We get a mean accuracy of 98.095% and a mean KNN of 4.2602 sec.

After PCA:

We get a mean accuracy mean of 98.0675% and a mean KNN Runtime of 4.227885 sec.

The normalization is faster when performed after PCA.

# Exercise 2.3: Preprocessing

Une image contenant eau

Description générée automatiquement

We notice that the accuracy decreases when sigma (the amount of smoothing) increases. The runtimes do not seem to be depending on the value of sigma.

# Exercise 2.4: Reconstruction with PCA

2.4.1

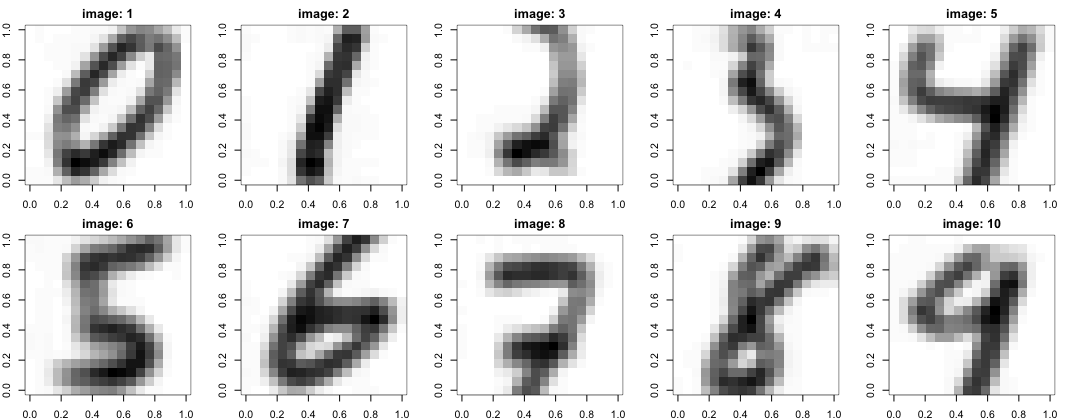


Image of each cipher that was created according to the data

2.4.2

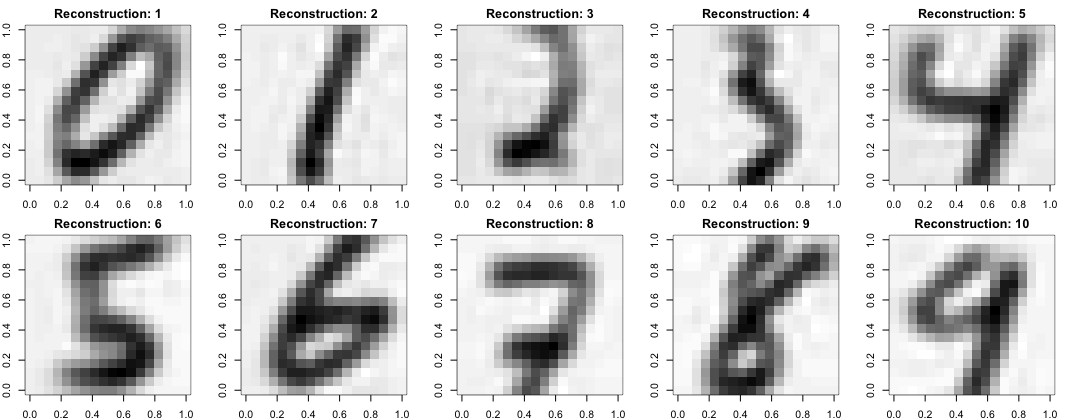
Une image contenant différent, afficher, horloge, cité

Description générée automatiquement

The first 10 eigenvectors as an image

Each eigenvector is a group of pixels in our output. When using the full parameters, our output is described with a vector of dimension 324 and each number in that vector describes the value of one pixel. What we can do instead of using the whole dataset is to find “patterns” and describe our digits with this set of patterns. In PCA these patterns are the eigenvectors (which also have the same dimensions) and by giving each eigenvector a weight and summing them up we can also describe our image.

2.4.3



Recreated image using all the principal components

2.4.4

Une image contenant différent, horloge, afficher

Description générée automatiquement

Recreated image using 80% of variance

Une image contenant horloge

Description générée automatiquement

Recreated image using 90% of variance

Une image contenant horloge

Description générée automatiquement

Recreated image using 95% of variance

The higher the percentage of cumulated variance gets, the more eigenvectors we use to recreate the image. Therefore, the quality of the image gets better because we have more “patterns” available to describe the image.

We can see the amount of data saved in the following table (that uses the assumption that one pixel needs one byte)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| % of variance | Number of Eigenvectors | Bytes for EV | Bytes for scores for one image | Total for N images |
| 80% | 14 | 14 bytes | N\*14+324 | (N\*324)/(N\*14+324) |
| 90% | 24 | 24 bytes | N\*24+324 | (N\*324)/(N\*24+324) |
| 95% | 35 | 35 bytes | N\*35+324 | (N\*324)/(N\*35+324) |

2.4.5

Scores for 43:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
| 0.89881150 | 0.13218921 | -0.26592992 | 0.02620752 | 0.32310330 | 0.35373291 |

|  |  |  |  |
| --- | --- | --- | --- |
| PC7 | PC8 | PC9 | PC10 |
| 0.06616662 | -0.31915394 | 0.35917030 | -0.36144824 |

Scores for 456:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| PC1 | PC2 | PC3 | PC4 | PC5 | PC6 |
| 1.507370730 | -0.445392431 | 0.055150857 | 0.004585787 | 0.064170593 | 0.125690786 |

|  |  |  |  |
| --- | --- | --- | --- |
| PC7 | PC8 | PC9 | PC10 |
| 0.028805782 | 0.065597007 | -0.184172389 | -0.039411853 |