UML computer project 1

- Handed out: April 1 (Fri)
- Hand in: April 17 (So) midnight the latest,
  by email to michael.gutmann@helsinki.fi
- You can do the exercises in pairs. Submit in that case only one report. Please write in the report the names and the student numbers of both of you.
- Submit your report as a single pdf containing figures and **discussion**. Don't put figures in the report without explaining the result they show. The report should be enjoyable to read; remember that the grading will be based on the report.
- Submit the source code as well. The code needs to be such that running it for every exercise will produce the figures in the report.
- Each exercise in the project gives you points. They will, at the end, determine your grade for the computer project. Each exercise in all the assignments gives you an equal amount of points.

Exercise 1 : Basic PCA

In this exercise, you will need to compute eigenvalue decompositions. In matlab, this is done with the command eig*, see* help eig*. Note that in matlab, the smallest eigenvalue and its corresponding eigenvector is returned first.*

1. Create artificial data as shown in Figure 1. The data needs to be just qualitatively similar. In your report, make a scatter plot as in the figure, and explain how you created the data. Especially, say how you control the "direction" of the data, and its spread.

2. Do PCA for two of the four scatter plots. Plot the principal component (directions) on the top of the scatter plots. For the report, create figures similar to Fig 4.1, p.30, in the lecture handout (but show both principal component directions).

3. For one of the four scatter plots, project the data on each of the two PC directions, and make a histogram (20 bins worked fine for me). Compute the variance of the projections (=principal components="PCs") and say how the variances relate to the covariance matrix of the data.

4. Create artificial data which has the vectors

$$\mathbf{v}_1 = \frac{1}{\sqrt{2}}[1\ 1]^T \quad \mathbf{v}_2 = \frac{1}{\sqrt{2}}[-1\ 1]^T$$

   as principal component directions with PCs having variance 1 and 3, respectively (take 1000 sample points). Make a scatter plot of the data and compute its covariance matrix. What are the eigenvectors and eigenvalues of the covariance matrix?

5. Reduce the dimension of this data to 1 so that the reconstruction error is minimized (see p.31 section 4.2 or ex.16 and ex17). For the 50 first data points, make a scatter plot which shows both the original points and their approximation (like in Figure 2). How large is the average reconstruction error (take the average over all data point, not only the 50)? How large is the proportion of variance explained? What is the relation between average reconstruction error and proportion of variance explained?

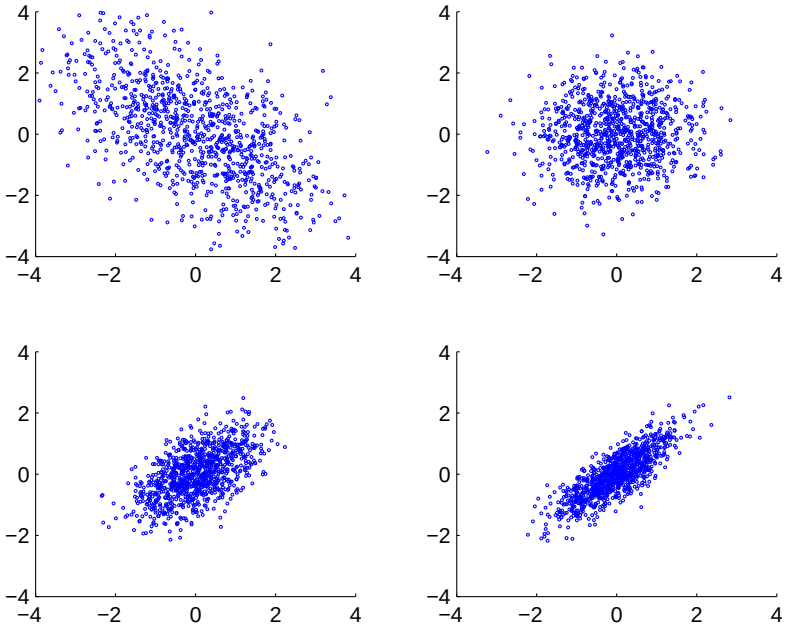

Figure 1: Scatter plots for exercise 1. They are of the same type as the scatter plot on p.30, section 4.1 in the lecture handout. The data is Gaussian with different kinds of covariance matrices. Every scatter plot shows 1000 data points.

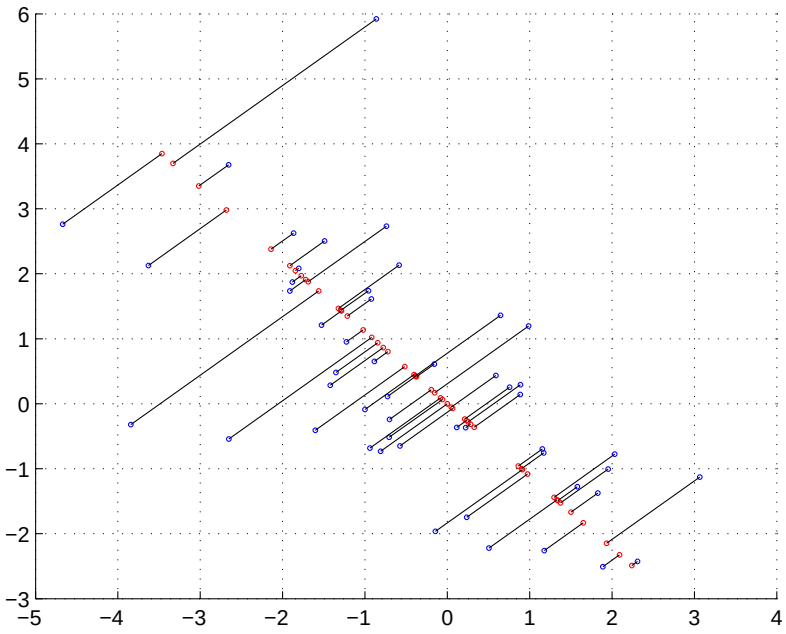

Figure 2: Sample solution for exercise 1, question 5. Original data in blue, approximation in red. The black lines show the correspondance.

Exercise 2: PCA and FA

1. Reproduce Figure4.2 on p35 in the lecture handout (You can drop the numbers in the plot, which gives you Figure 3). Explain the figure in your own words. What decides the length of the red lines? (this exercise is related to math Ex. 20 in the lecture handout).

2. Make a plot which shows the proportion of variance explained in function of the number of principal components (see section 4.3.5 p 34).

3. Implement the quartimax rotation for factor analysis (FA), which is explained in section 5.5, p. 39 of the lecture handout and the topic of math Ex. 23. *Note: There is a typo in the solution to math ex 23: The gradient should be* $4A^T(AU)^3$ *where the power is taken elementwise* Test your solution: does the algorithm bring you from Eq (5.12) to Eq (5.13) in the lecture handout? (you might get the sign-inverted matrix, that's also ok)

4. Now you are ready to reproduce Fig. 5.1, p.40 in the lecture handout. Again, explain the figure in your own words. (note: the signs of the new coordinates you get might be flipped compared to those in the lecture handout, that's ok)

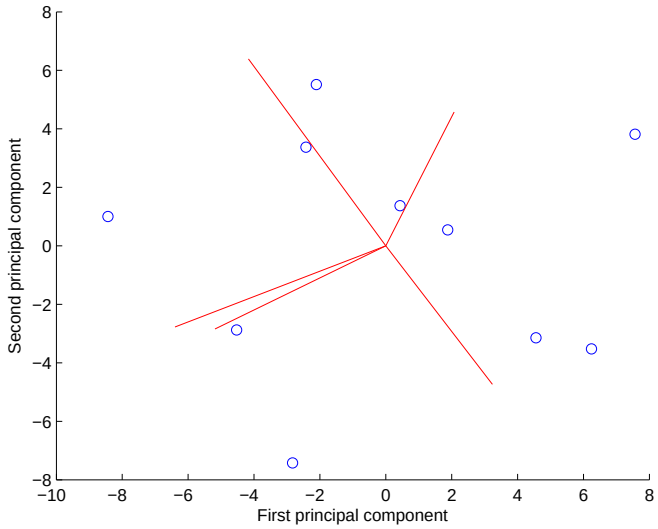

Figure 3: For exercise 2: Reproduction of Fig4.2 on p35 in the lecture handout.

Exercise 3: Denoising and compression by PCA and FA

For this exercise, you will use images of handwritten digits from the MNIST database. You can download the database from the homepage. The file `digits.txt` contains the data matrix X of size 784×1000 where the 28×28 handwritten digit images are rearranged as the columns of the data matrix.

1. Load the data into matlab and, for preprocessing, remove the average from the images, and normalize each image to unit norm. Then, center the data by removing the mean of each of the 784 random variables. Visualize some digits before the preprocessing, the mean, and the same digits after preprocessing. If you use matlab, you can use the provided `visual.m`.(`visual(X(:,1:100), 2, 10)` will show the first 100 numbers, magnified by a factor 2, arranged in 10 columns)

2. Perform PCA on the data. Show the proportion of variance explained in function of the number of principal components. Show the first 20 principal component directions, how much variance do they explain?

3. Choose 10 digits you like. Project each digit on a 1,2,4,8,16,32,64, and 128 dimensional subspace spanned by the first principal component directions (see section 5.6 in lecture handout and math Ex17). Show the approximations and the original digits. Discuss the result. For each dimension, what is the average reconstruction error? Explain how projection on subspaces can be used for compression.

4. Use the computed mean and principal component directions to clean up the 100 numbers in the file `noisyDigits.txt`.