MTAT.03.227 Machine Learning
Spring 2016 / Exercise session X
Nominal score: 10p
Maximum score: 15p
Deadline: 12th of April 16:15 EET

1. This exercise explores the close connection between principal component analysis and multivariate normal distribution. Your first task is to generate four data distributions starting form white Gaussian noise $x_1 \sim \mathcal{N}(0,1)$ and $x_2 \sim \mathcal{N}(0,1)$. The remaining distributions are given by the following affine transformations:

$$\begin{cases} y_1 &= 0.5 \cdot x_1 + 1 \ , \\ y_2 &= x_2 - 0.5 \ , \end{cases}$$

$$\begin{cases} z_1 &= \cos(30°) \cdot y_1 + \sin(30°) \cdot y_2 \ , \\ z_2 &= \sin(30°) \cdot y_1 - \cos(30°) \cdot y_2 \ , \end{cases}$$

$$\begin{cases} u_1 &= \frac{y1+y2}{\sqrt{2}} \ , \\ u_2 &= \frac{y1-y2}{\sqrt{2}} \ . \end{cases}$$

   (a) Express distributions (z_1, z_2) and (u_1, u_2) through a direct affine transformation from (x_1, x_2). Verify experimentally that two ways to generate the data leads to the same distribution. (**1p**)

   (b) Implement the the PCA algorithm for two-dimensional case. For that do the following steps. Centre the data and find the covariance matrix Σ. Use `eigen` function find eigen vectors of Σ. Use `dataEllipse` function to plot the data. Add origin of the ellipse and both axes of the ellipse based on the PCA analysis. Apply the method on the distribution (z_1, z_2) and verify that the axes and the centre found by you indeed matches the visualisation given by `dataEllipse`. (**1p**)

   (c) Use the results form the PCA analysis to invert the affine transformation for (z_1, z_2). That is, find the coordinates of the data given the new origin and new directions of the axes computed by the PCA. Visualise the reconstruction quality by visualising 30 original data points (x_1, x_2) compared to reconstructed data points $(\hat{x}_1, \hat{x}_2)$. How is the result related with impossibility of full reconstruction? (**1p**)

2. This exercise explores how principal component analysis can be used for dimensionality reduction. Assume that the actual data $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ is located on the low-dimensional space or curve but our observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m$ are shifted due to white Gaussian noise.

   (a) Generate the data located on the straight line specified by parametrised

equation

$$\begin{cases} x_1 &= t \ , \\ x_2 &= 2t \ , \\ x_3 &= 3 - t \ . \end{cases}$$

For instance sample t value uniformly 500 times from the range $[0, 1]$ and add white Gaussian noise $\varepsilon_1, \varepsilon_2, \varepsilon_3 \sim \mathcal{N}(0, 0.1)$. Visualise the original data and disturbed data using `plot3d` and `rgl.points` form the `rgl` package. You can use functions `rgl.postscript` and `rgl.snapshot` to take snapshots of the `rgl` visualisation tool. (**0.5p**)

(b) Use the function `prcomp` to find the principal components of the data. Since the original data is one-dimensional, use the first coordinate of the PCA and the first principal component to reduce the data as a one-dimensional object. Compare the reconstructed line with the original line. Are they well aligned? Visualise also the distance between the original data points and reconstructed data points. Compare it with the distance between the original data points and observed data points. Use histograms for that. What is the main reason that the data is not completely reconstructed and how precise the reconstruction is on average? (**1p**)

(c) Generate the data located on the plane specified by parametrised equation

$$\boldsymbol{x} = t_1 \boldsymbol{p}_1 + t_2 \boldsymbol{p}_2$$

where $\boldsymbol{p}_1 = (1, 1, 1)^T$ and $\boldsymbol{p}_2 = (0, 1, -1)^T$. Again, generate 500 data points on the plain and add white Gaussian noise $\varepsilon_1, \varepsilon_2, \varepsilon_3 \sim \mathcal{N}(0, 0.1)$. Visualise the original data and disturbed data using `plot3d` and `rgl.points`. Reconstruct the plane using PCA. Again, estimate the reconstruction precision by comparing the distance distributions between original, reconstructed and observed data. (**1p**)

(d) The data from a file `mysterious-data.Rdata` seems to form a 3 dimensional surface. However, it is corrupted by additive white Gaussian noise. Your task in this exercise is to reconstruct the surface and estimate the size of the noise component. For that you can do the following steps. First, come up with a surface description. Second, build a data distribution model and find a feature of the data that allows you to find the parameters fixing the surface. Use either simulations or maximum likelihood estimate to find how model parameters through defined features. You get an extra point if your solution is analytical. (**2p**)

3. In this exercise, you have to apply the principal component analysis in order to analyse the socioeconomic data of the biggest and richest countries in the world. Understanding the richest and biggest countries in the world

is important for many reasons. Often they are responsible for economic recessions; most of the political and economic global decisions are made by them; they spread their values to the rest of the world etc. Your goal is to apply PCA to the data of the richest and the biggest countries in the world and write a nice report explaining your analysis of the data. The data is stored in the file `countries.Rdata` and the detailed instructions are given in the last years exercise `PCA-analysis-task.pdf`. You can score up to **10 points** form this exercise.

4. One of the most striking applications of principal component analysis and independent component analysis is image processing. Standard way to approach this problem is to split the original image into small sub-images of size 16×16 or 32×32 and flatten them back to one-dimensional vectors. This yields many data-points in the high dimensional space. Now if we apply PCA or ICA to reduce the dimensionality, we obtain principal or independent components that correspond to repetative patterns in the image. Further steps depend on the exact application.

   (a) Use the package `jpeg` to read in few images of Jackson Pollock paintings. As these images are coloured you get individual signals for red, green and blue. For simplicity, you can work with one colour channel or consider intensity of the image. Split the image into small 16×16 or 32×32 squares. There are several ways how to do it. First, you can just split the image into sub-images. Second, you can just sample enough squares from random locations. Third, you can sample all possible squares. Choose one option and justify your choice (**1p**)

   (b) After you have extracted the squares you need to flatten them into vectors. Define a method for that and discuss whether the choice of flattening method influences the further analysis (**1p**)

   (c) Perform PCA analysis on the data and show the first ten corresponding principal components as coloured squares, i.e., unflatten them. Interpret the results (**2p**)

   (d) Pack the original image using few principal components. What can be said about the image quality. In particular, how does the mean square error depend on the number of used components. Find the number of components that are needed to make the picture recognisable and visually indistinguishable from the original (**3p**)

   (e) Repeat the same analysis with `fastICA` package and report the major differences in the reconstruction quality and interpretability of components. (**3p**)

   (f) Discuss whether it is possible to extract rotation invariant components from the image? How one should sample the sub-images and how one should flatten the sub-image into the vector. (**3p**)

5. Similarly to fitting a Gaussian distribution on the data, we can modify the PCA so that it can handle weighted data. Test the behaviour of

the algorithm on the data where the most of the data is generated by two-dimensional normal distribution. To make it non-trivial, assure that the the first principal component is rotated 45° and the second principal component covers up to 25% of variance. Additionally add uniform noise that has is significantly larger variance than the normal distribution.

(a) Implement a principal component analysis with integer weights, i.e., a data point $\boldsymbol{x}_i$ has a multiplicity c_i. Define a reasonable weighting scheme for data points, which has marginal weights for the outlier points. Draw plots that show the original PCA fit and new fit based on weights. Discuss whether the weighting is useful. (**2 p**)

(b) Generalise the previous approach for the fractional weights. For that you have to reimplement the entire principal component analysis from scratch, since you need to form a correlation matrix for weighted data. Consult `em-cookbook.pdf` from the previous exercise session and the PCA derivation material from the lecture. Repeat the same experiment as in the previous subtask. (**3p**)