SMLDM HT 2014 - MSc Problem Sheet 1

1. Suppose a p-dimensional random vector X has a covariance matrix Σ. Under what condition will the first principal component direction be identifiable? (It is not identifiable if there are more than one direction satisfying the defining criterion). Supposing it is not identifiable, can you describe the behaviour of the first principal component computed using a dataset, when the dataset is perturbed by adding small amounts of noise?

2. In lectures we defined the total sample variance to be

$$\sum_{i=1}^{n} S_{ii} = \lambda_1 + \cdots + \lambda_p$$

where S is the sample covariance and $\lambda_1, \ldots, \lambda_p$ are its eigenvalues. Show that the total sample variance is equal to the sum of the sample variances of each individual variable, $X_1, \ldots, X_p$.

3. Suppose we do PCA, projecting each x_i into $z_i = V_{1:k}^{\top} x_i$ the first k principal components. We can reconstruct x_i from z_i by inverting the process, $\hat{x}_i = V_{1:k} z_i$. Show that the error in the reconstruction equals:

$$\sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 = (n-1) \sum_{j=k+1}^{p} \lambda_j$$

where $\lambda_{k+1}, \ldots, \lambda_p$ are the $p-k$ smallest eigenvalues.
Thus the more principal components we use for the reconstruction, the more accurate it is. Further, using the top k principal components is optimal in the sense of least reconstruction error.

4. As in the lectures, suppose we have a dataset of n vectors $x_1, \ldots, x_n \in \mathbb{R}^p$ with zero mean. We wish to "compress" the dataset by representing each vector x_i using a lower dimensional vector $z_i \in \mathbb{R}^k$ with $k < p$. We assume a linear model for reconstructing x_i from z_i. That is, there is a matrix $M \in \mathbb{R}^{p \times k}$ such that Mz_i is close to x_i. We measure the reconstruction error using Euclidean distance, so that the total error is:

$$\sum_{i=1}^{n} \|x_i - Mz_i\|_2^2$$

We wish to find a reconstruction model M and representations $z_1, \ldots, z_n$ minimizing the reconstruction error.

   (a) Suppose M is given and that it is full rank. Show that the representations $z_1, \ldots, z_n$ minimizing the reconstruction error is given by:

$$z_i = (M^{\top} M)^{-1} M^{\top} x_i.$$

   (b) Show that PCA projection gives an optimal M. [Hint: there are a few ways to show this. One way is to recall the property that SVD of $\mathbf{X}$ gives the best rank k approximation to $\mathbf{X}$.]

   (c) If M is a solution minimizing the total reconstruction error, explain why MQ is also a solution, where Q is any $k \times k$ invertible matrix.

5. In this question, you will use biplots to interpret a data set consisting of US census information for the 50 states. The dataset can be obtained using the R commands:

```
data(state)
state <- state.x77[, 2:7]
row.names(state)<-state.abb
```

The data consists of estimates (in 1975) of population, per capita income, illiteracy rate, life expectancy, murder rate, high school graduate proportion, mean number of days below freezing, and area. We will not look at population level and area.

(a) Give the R commands to apply PCA to the correlation matrix and to show the biplot. Include a printout of the biplot. You can produce a pdf printout by using the command

```
pdf("statebiplot.pdf",onefile=TRUE)
```

before the biplot command, and

```
dev.off()
```

afterwards.

(b) According to the plot, what variables are positively correlated with graduating high school *HS Grad*? Which are negatively correlated? In each case, give a possible explanation.

(c) Run the `summary` command on output of the PCA routine. What is the proportion of variance explained by the first two principal components? How trustworthy are your observations?

6. Let $x_1, \ldots, x_n$ be a dataset of p-dimensional vectors and $C = \{C_1, C_2, \ldots, C_K\}$ a partition of $\{1, \ldots, n\}$.

For each cluster C_k, define

$$\begin{aligned} \bar{x}_k &= \frac{1}{n_k} \sum_{i \in C_k} x_i \text{ be the within-cluster mean} \\ \bar{x} &= \frac{1}{n} \sum_{k=1}^{K} n_k \bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_i \text{ be the overall mean} \end{aligned}$$

and let

$$\begin{aligned} T &= \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x})(x_i - \bar{x})^\top \text{ be the total deviance to the overall mean} \\ W &= \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - \bar{x}_k)(x_i - \bar{x}_k) \text{ be the within-cluster deviance to the cluster mean} \\ B &= \sum_{k=1}^{K} \sum_{i \in C_k} (\bar{x}_k - \bar{x})(\bar{x}_k - \bar{x})^\top \text{ be the between cluster deviance} \end{aligned}$$

T, W and B are $p \times p$ matrices.

(a) Verify that $T = W + B$.

(b) Explain how the K-means objective is related to W.

(c) How does T change during the course of the K-means algorithm? How does B change?

7. (Optional) Under the assumption that your data are centred, show that you can compute the $n \times n$ Gram matrix B such that $b_{ij} = x_i^\top x_j$ using the dissimilarity matrix D where $d_{ij} = \|x_i - x_j\|_2$.