SMLDM HT 2014 - Part C 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?

   Answer: If the largest eigenvalue λ_1 of Σ has multiplicity 1 (equivalently, that λ_1 is strictly larger than the second largest eigenvalue λ_2 in the EVD), then the first principal component direction is identifiable since its eigenspace is one-dimensional. Otherwise, it is not identifiable, as any of the infinitely many directions within the eigenspace will satisfy the maximum variance criterion.

   The computed first PC will not be stable under perturbations of the dataset, since small perturbations will cause the direction of largest variance to vary significantly.

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, $X_1, \ldots, X_p$.

   Answer: The sample variance of X_i is S_{ii}. So the sum of the sample variances is $\sum_{i=1}^{p} S_{ii} = Tr[S]$. The trace operator is preserved under similarity transforms, so $Tr[S] = Tr[V\Lambda V^{-1}] = Tr[\Lambda] = \sum_i \lambda_i$.

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.

   Answer: Suppose V is the full $p \times p$ matrix containing all p eigenvectors. Denote by $\tilde{V} = V_{k+1:p}$ the matrix consisting of all eigenvectors except the first k. Since V is orthogonal, $VV^\top = I = V_{1:k}V_{1:k}^\top + \tilde{V}\tilde{V}^\top$.

$$\begin{aligned}
\sum_{i=1}^{n} \|x_i - \hat{x}_i\|_2^2 &= \sum_{i=1}^{n} \|x_i - V_{1:k}V_{1:k}^\top x_i\|_2^2 = \sum_{i=1}^{n} \|(I - V_{1:k}V_{1:k}^\top)x_i\|_2^2 = \sum_{i=1}^{n} \|\tilde{V}\tilde{V}^\top x_i\|_2^2 \\
&= \sum_{i=1}^{n} x_i^\top \tilde{V}\tilde{V}^\top \tilde{V}\tilde{V}^\top x_i = \sum_{i=1}^{n} x_i^\top \tilde{V}\tilde{V}^\top x_i \\
&= \sum_{i=1}^{n} Tr[x_i^\top \tilde{V}\tilde{V}^\top x_i] = \sum_{i=1}^{n} Tr[\tilde{V}^\top x_i x_i^\top \tilde{V}] = (n-1) Tr\left[\tilde{V}^\top S \tilde{V}\right] \\
&= (n-1) \sum_{j=k+1}^{p} \lambda_j
\end{aligned}$$

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:

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

Answer: The derivative of the total error loss function with respect to z_i is

$$2M^\top (x_i - Mz_i)$$

Setting the derivative to zero gives $M^\top M z_i = M^\top x_i$. Since M is full rank, we can invert $M^\top M$ so that $z_i = (M^\top M)^{-1} M^\top x_i$. The second derivative is negative definite so this is the optimal z_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 $\mathbf{X}$.]

Answer: Substituting the solved for value for z_i, the cost function is now

$$\sum_{i=1}^{n} \|x_i - M(M^\top M)^{-1} M^\top x_i\|_2^2$$
$$= \left\| \mathbf{X} - \mathbf{X} M (M^\top M)^{-1} M^\top \right\|^2$$

Note that $\mathbf{X}M(M^\top M)^{-1}M^\top$ has rank at most k. Consider the SVD of $\mathbf{X} = UDV^\top$. If we keep only the top k singular values and associated vectors, then $\tilde{\mathbf{X}} = U_{1:k} D_{1:k,1:k} V_{1:k}^\top$ minimizes

$$\left\| \mathbf{X} - \tilde{\mathbf{X}} \right\|^2$$

where $D_{1:k,1:k}$ is the diagonal matrix of top k singular values, $U_{1:k} \in \mathbb{R}^{n \times k}$ the associated columns of U, and similarly $V_{1:k}$. Letting $M = V_k$, we see that

$$\mathbf{X}M(M^\top M)^{-1}M^\top = \tilde{X}$$

so must be a minimizing M.

(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.

Answer: We have that $MQ(Q^\top M^\top MQ)^{-1} QM^\top = M(M^\top M)^{-1} M^\top$ so the reconstruction error is the same.

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)
```

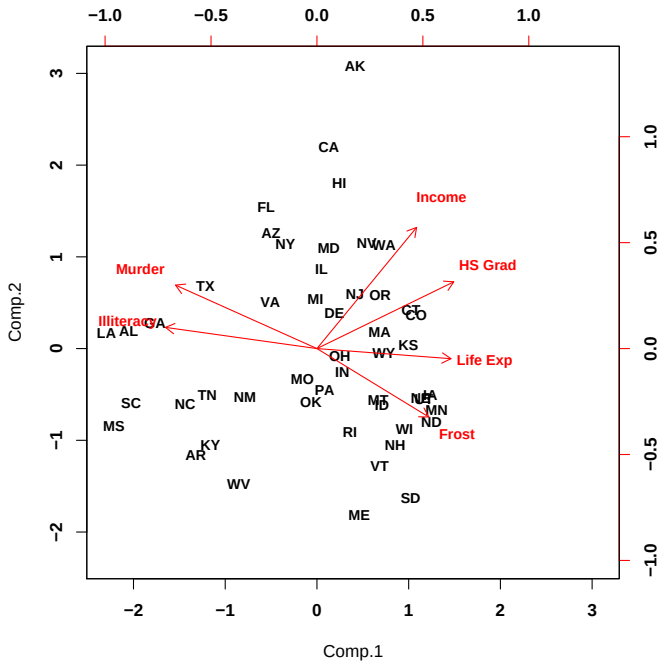
before the biplot command, and

```
dev.off()
```

afterwards.

Answer:

```
state.pca <- princomp(state,cor=TRUE)
biplot(state.pca)
```



(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.

Answer: *Income* and *Life Exp* are positively correlated to *HS Grad*, *Illiteracy* and *Murder* are negatively correlated.

Higher income implies more successful parents which in turn implies greater expectations

of children. Greater life expectancy implies higher standard of living which implies happier family environments. Or, better education may lead to higher income, which leads to better healthcare and better living standards which lead to higher life expectancy.

Greater illiteracy directly implies lower educational levels, i.e. high school graduation. Higher murder rates suggest poor backgrounds which suggest lower educational levels.

Of course these are just suggested "explanations" and others may be sensible too. E.g. better educated people may be more successful in life and have better healthcare so live longer (US does not have universal healthcare).

(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?

Answer: The first two PCs explained about 76% of the variance. The observations are relatively accurate.

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

Answer: We have

$$\|x_i - x_j\|_2^2 = \|x_i\|_2^2 + \|x_j\|_2^2 - 2x_i^T x_j$$

so

$$\begin{aligned}
\sum_{j=1}^n \|x_i - x_j\|_2^2 &= n\|x_i\|_2^2 + \sum_{k=1}^n \|x_k\|_2^2 \Leftrightarrow \sum_{j=1}^n d_{ij}^2 = nd_{ii}^2 + \sum_{k=1}^n d_{kk}^2 \\
\sum_{i=1}^n \|x_i - x_j\|_2^2 &= \sum_{k=1}^n \|x_k\|_2^2 + n\|x_j\|_2^2 \Leftrightarrow \sum_{i=1}^n d_{ij}^2 = nd_{jj}^2 + \sum_{k=1}^n d_{kk}^2 \\
\sum_{i,j=1}^n \|x_i - x_j\|_2^2 &= 2n\left(\sum_{k=1}^n \|x_k\|_2^2\right) \Leftrightarrow \sum_{i,j=1}^n d_{ij}^2 = 2n\sum_{k=1}^n d_{kk}^2
\end{aligned}$$

We have

$$b_{ij} = \frac{1}{2}d_{ii}^2 + \frac{1}{2}d_{jj}^2 - \frac{1}{2}d_{ij}^2$$

where

$$d_{ii}^2 = \frac{1}{n}\left(\sum_{j=1}^n d_{ij}^2 - \sum_{k=1}^n d_{kk}^2\right) = \frac{1}{n}\left(\sum_{j=1}^n d_{ij}^2 - \frac{1}{2n}\sum_{i,j=1}^n d_{ij}^2\right)$$

so

$$\begin{aligned}
b_{ij} &= \frac{1}{2n}\left(\sum_{j=1}^n d_{ij}^2 - \frac{1}{2n}\sum_{i,j=1}^n d_{ij}^2\right) + \frac{1}{2n}\left(\sum_{i=1}^n d_{ij}^2 - \frac{1}{2n}\sum_{i,j=1}^n d_{ij}^2\right) - \frac{1}{2}d_{ij}^2 \\
&= \frac{1}{2n}\left(\sum_{i=1}^n d_{ij}^2 + \sum_{j=1}^n d_{ij}^2\right) - \frac{1}{2n^2}\sum_{i,j=1}^n d_{ij}^2 - \frac{1}{2}d_{ij}^2
\end{aligned}$$

Denoting $A_{ij} = -\frac{1}{2}d_{ij}^2$ and $J_{ij} = 1$ one can check that

$$B = \left(I - n^{-1}J\right) A \left(I - n^{-1}J\right).$$