SDMML HT 2016 - 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? [*hint: what happens when PCA is applied to samples from an isotropic Gaussian?*]

2. We perform PCA on a centred dataset consisting of an i.i.d. sample $\{x_i\}_{i=1}^n$ of a random vector $X = \left[X^{(1)} \dots X^{(p)}\right]^\top$. Denote the projections to principal components by $Z^{(1)}, \dots, Z^{(p)}$. Find the sample variance of $Z^{(j)}$ and show that the sum of the sample variances of individual variables $X^{(1)}, \dots, X^{(p)}$ is equal to the sum of the sample variances of projections $Z^{(1)}, \dots, Z^{(p)}$.

3. Suppose we do PCA, projecting each x_i into $z_i = V_{1:k}^\top x_i$ where $V_{1:k} = [v_1, \dots, v_k]$, i.e., the first k principal components. We can reconstruct x_i from z_i as $\hat{x}_i = V_{1:k} z_i$.

   (a) Show that $\|\hat{x}_i - \hat{x}_j\| = \|z_i - z_j\|$.

   (b) 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 $\lambda_{k+1}, \dots, \lambda_p$ 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. We have a dataset of n vectors $x_1, \dots, x_n \in \mathbb{R}^p$ 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 $z_1, \dots, z_n$ minimizing the reconstruction error.

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

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

   (b) 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.

   (c) 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* $\mathbf{X}$.]

5. 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 $d_{ij} = \|x_i - x_j\|_2$.

6. In lectures we discussed using the Mahalanobis distance to measure distances in K-means:

$$\|x - y\|_M = \sqrt{(x-y)^\top M^{-1}(x-y)}$$

where M is a positive definite matrix. Explain why using this distance is equivalent to applying K-means using the standard Euclidean distance on a transformed data set. What is the choice of the M matrix that leads to an algorithm which is equivalent to first whitening the data?

Optional

7. Download abalone data from `http://archive.ics.uci.edu/ml/datasets/Abalone`, then load it with

```
ab<-read.table( 'abalone.data',sep=',',
      col.names=c('Sex','Length','Diam','Height','Whole',
                  'Shucked','Viscera','Shell','Rings') )
x<-ab[,2:8]
y<-ab[,9]
```

`x` contains seven quantitative attributes, while `y` contains an integer value corresponding to the number of rings of abalone (related to its age). Perform PCA on *correlation* matrix of `x` and look at the biplot. Spot two nasty outliers / likely typos in the data and remove them. Rerun PCA without them. How much variance is explained by the first principal component? Consider the first two principal components - what can you say about the shape of the cloud of points in this 2d space? What can you conclude about the relationship of the first principal component and the number of rings in `y`?