SDMML HT 2016 - 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? [*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 $\hat{x}_i = V_{1:k} z_i$. 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. 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?

5. 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`?