Statistical Machine Learning HT 2020 - Problem Sheet 1

Please send any comments or corrections to Pier Palamara (email on course website).

1. **(PCA identifiability).** 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. **(PCA variance).** 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. **(PCA using k PCs).** 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. **(PCA compression).** 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. **(PCA and MDS).** 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. **(*Coding*: PCA and biplots).** 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?