Statistical Inference and Learning- Ex-5

Yiqing Wang and Boaz Nadler

Due by Thursday Feb. 4, 2016

Q1. Let X be a $\mathbb{R}^p$-valued random column vector. Let us denote its mean $\mu = \mathbb{E}[X]$ and its covariance matrix $\Sigma = \mathbb{E}[(X-\mu)(X-\mu)^T]$. We assume for simplicity that the covariance matrix's eigenvalues $(\lambda_i)_{1\le i\le p}$ are distinct, that is, $\lambda_i \neq \lambda_j$ for all $i \neq j$. Their associated column eigenvectors are denoted by $(v_i)_{1\le i\le p}$. In the following questions, we assume $\mu = 0$ until stated otherwise.

1. Show that the covariance matrix is positive semidefinite, that is, $\forall y \in \mathbb{R}^p$, $y^t \Sigma y \ge 0$.

2. Show that for all $1 \le i, j \le p$, $\mathbb{E}[(X^T v_i)(X^T v_j)] = \lambda_i 1_{i=j}$.

3. Show that $\mathbb{E}\|X\|^2 = \mathrm{Tr}(\Sigma)$ where $\|X\|^2 = \sum_{i=1}^p X_i^2$ and $\mathrm{Tr}(\Sigma) = \sum_{i=1}^p \Sigma_{i,i}$ (Hint: Show that $\mathrm{Tr}(\Sigma) = \sum \lambda_i$).

4. Show that for all $1 \le k \le p-1$, $\mathbb{E}\|X - \sum_{i=1}^k (X^T v_i) v_i\|^2 = \sum_{i=k+1}^p \lambda_i$.

5. Let $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ be an estimate of Σ from n samples $x_1, \ldots, x_n \in \mathbb{R}^p$. Suppose that Σ is invertible with smallest eigenvalue $\lambda_p > 0$. Suppose that $n < p$ (namely we have fewer observations that the dimension). Show that $\hat{\Sigma}_n$ is not invertible and that $\|\hat{\Sigma}_n - \Sigma\| \ge \lambda_p$, namely when $n < p$, $\hat{\Sigma}_n$ is in general far from Σ (Hint: what is the rank of $\hat{\Sigma}_n$?).

6. We now drop the assumption $\mu = 0$ and seek to estimate the covariance matrix from its n i.i.d. samples $(x_1, \ldots, x_n) \in \mathbb{R}^{p\times n}$. Show that with their sample mean defined as $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^n x_i$, the following estimator

$$\hat{\Sigma}_n = \frac{1}{n-1}\sum_{i=1}^n (x_i - \hat{\mu}_n)(x_i - \hat{\mu}_n)^t$$

is unbiased, namely,

$$\mathbb{E}[\hat{\Sigma}_n] = \Sigma.$$

Q2. In this exercise you will apply principal component analysis on actual data. The file `stock_data.mat` contains, in matlab format, the following matrices: `A` of size 1258×31, with the daily stock prices of 31 different companies traded in the United States. The list of stocks appears in the variable `stock_list`. The matrix `B` of size 31×1257 where $B_{i,t} = 100 \times (A_{t+1,i} - A_{t,i})/A_{t,i}$. Namely, it contains the daily returns of the 31 different stocks. The matrix `B_mc` is the mean-centered version of `B`.

To load the data into the matlab workspace simply type `load stock_data.mat`.

The files A.txt, B.txt and stock_list.txt are text files with the same data (in case you are using some other software to analyze it).

(a) Compute the sample covariance matrix S_n of the 31 companies, and the sample correlation matrix R_n of the 31 companies. The latter can be computed as DS_nD where $D = diag(1./sqrt(S_n))$.

Plot the two matrices, in matlab this can be done via `figure(1); imagesc(S_n)` and similarly for R_n. Can you see interesting structure in one or both of these matrices ? Is there structure apparent in one that is less apparent in the other ? Explain your answer.

(b) Let us first look at only 2 specific companies, indices 4 and 5 in the stock list. These are `DUK` and `ED`. `DUK` is the ticker symbol of Duke company - operates as an energy company in the United States and Latin America. `ED` stands for Consolidated Edison. It engages in regulated electric, gas, and steam delivery businesses in the United States. It offers electric services to approximately 3.4 million customers in New York City and Westchester County.

Plot the daily return prices of these two companies one against the other, namely `plot(B(4,:),B(5,:),'bo')`. What do you see ? How correlated are these two different companies ?

(c) Compute the eigenvalues and eigenvectors of S_n, namely the principal components of the data matrix B containing all 31 stocks. Note that in matlab this can be directly computed from B via

```
[coeff, score, latent] = pca(B);
```

where `latent` are the eigenvalues, and `coeff` is a 31×31 matrix whose columns are the principal components.

Plot the eigenvalues. What do you see? How much of the overall variability is explained by the first principal component ?

(c) Typically in PCA, people look at the top principal components. However, sometimes one may find interesting things in the smallest ones instead. Plot the eigenvector that corresponds to the *smallest* eigenvalue. What do you see? Look at the corresponding companies where the entries of this eigenvector are significantly different from zero. Which are they and how does it explain the structure of this eigenvector?

Q3. Let $f = (f_1, f_2, \ldots, f_m)$ be a vector whose entries are all strictly positive. Consider the following Markov chain $(X_n)_{n\geq 1}$ on the finite states $\{1, \cdots, m\}$ whose transition matrix is

$$p(i,j) = \begin{cases} \frac{1}{m} \min\left(1, \frac{f_j}{f_i}\right), & i \neq j \\ 1 - \frac{1}{m} \sum_{j \neq i} \min\left(1, \frac{f_j}{f_i}\right), & i = j. \end{cases}$$

1. Show that the Markov chain that corresponds to the matrix P is irreducible and aperiodic.

2. Show that its stationary distribution π is proportional to f. Hint: prove that the condition $\pi(i)p(i,j) = \pi(j)p(j,i)$ holds.

3. Let $m = 50$ and $f_j = 2 + \sin(\pi j/100)$. Simulate the Markov chain for $10,000$ consecutive steps and plot the histogram of the observed chain in the m states.