Statistical Inference and Learning- Ex-4

Yaniv Tenzer and Boaz Nadler

Due on Sunday, August 4, 2019,
email submissions: `stat.wisdom@gmail.com`

Q1 In this problem, we will consider developing a Bayesian model for Poisson data; i.e., our observed data consists of i.i.d. observations $y_1, \dots, y_n$, such that $Y_i \sim Poisson(\lambda)$, for all i. Recall, a random variable Y follows a Poisson distribution with mean parameter λ if its pmf is given by

$$P(Y = k|\lambda) = \frac{e^{-\lambda}\lambda^k}{k!},\ k \in \{0, 1, 2, \dots\}$$

The Poisson distribution is a common model to analyze count data.

1. For the Poisson model, identify the conjugate prior. This should be a general class of priors.
2. Under the conjugate prior, derive the posterior distribution of $\lambda|Y = y$. This should be a general expression based on the choice of the hyper-parameters specified in your prior.
3. Find the posterior mean and variance of $\lambda|y$. These should be general expressions based on the choice of the hyper-parameters specified in your prior.
4. Obtain the MLE of λ. Develop and discuss a relationship that exists between the MLE and posterior mean identified in (3).

Q2 Attached to this assignment is the file airquality.csv. The file contains daily air quality measurements in New York, between May 1 to September 30, 1973. In this question we will explore the wind speed, using both frequentist and Bayesian frameworks.

1. load the data set into your workspace.
2. Assume that the wind speed follows a normal distribution, estimate the average wind speed, and its standard deviation using the standard ML estimators.
3. A known fact is that until May 1 1973, the average speed was 12 (mph), and the standard deviation was 2. Under the normality assumption, find the wind speed posterior distribution and use it to estimate the average speed.
4. Build a 95% confidence and a 95% credible set for the average wind speed. Explain your results in words.

Q3 In this problem we continue our analysis of the daily returns of 31 stocks (companies) over the period 2010-2014. The file `A.txt` contains 1258 lines each with 31 columns. Each line contains the price of 31 stocks on a specific day. The list of stock tickers is available in the file `stock_list.txt`. For each stock we will be interested in its daily return in percentage points, defined as

$$R_i(t) = 100 \times (P_i(t) - P_i(t-1))/P_i(t-1).$$

where $P_i(t)$ is the closing price of stock i on trading day t.

This gives a matrix of size 1257x31 that appears in the file `B.txt`. The corresponding trading days [year month day] appear in the file `dates.txt`

i) Compute the sample covariance matrix of the data and the sample correlation matrix of the data (both of size 31x31).

What is the average correlation between all pairs of stocks. Are there negative correlations in this group ?

ii) Find the pair of stocks with highest correlation. Plot their daily returns on an x-y cartesian grid. Which companies correspond to their ticker symbols? Is this surprising ?

iii) Compute the 31 eigenvalues of the sample covariance matrix.

Plot the eigenvalues in descending order.

What do you see? What does this imply about the daily returns of this list of stocks?

Plot the corresponding eigenvector. Note that all of its entries are positive and far from zero. Is this expected / surprising ?

Q4 As you saw in previous question, the daily return covariance matrix has some extremely small eigenvalues. Let's look at their eigenvectors.

Plot the eigenvectors corresponding to the 3 smallest eigenvalues. Note that each of them has many entries close to zero and few large ones. Try to explain this behavior. Look at the daily returns of the specific stocks involved.

Q5 Let X be a p-dimensional random variable with mean zero and covariance matrix Σ. Let $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_p$ be its eigenvalues and $\mathbf{v}_1, \mathbf{v}_2, \ldots \mathbf{v}_p$ its corresponding eigenvectors.

(a) Prove that $Trace(\Sigma) = \sum_j \lambda_j$, and that $\mathbb{E}[\|X\|^2] = Trace(\Sigma)$.

(b) Consider the one-dimensional random variables $Y_i = X^T \mathbf{v}_i$, for $i = 1, \ldots, p$. These are the 1-D projections of the high dimensional X onto the directions $\mathbf{v}_i$. Prove that $\mathbb{E}[Y_i Y_j] = \lambda_i \delta_{i,j}$. Are the random variables Y_i and Y_j independent ? Explain your answer.

(c) Suppose that instead of measuring X, we observe it with additive noise $\tilde{X} = X + \sigma\xi$, where $\sigma > 0$ is the noise level, and $\xi = (\xi_1, \xi_2, \ldots, \xi_p) \in \mathbb{R}^p$ is a random "noise" vector, independent of X. Suppose the entries ξ_i are all random variables with same variance 1 and independent of each other. What is the relation between the eigenvalues and eigenvectors of the population covariance of $\tilde{X}$, denoted $\tilde{\Sigma}$, and those of Σ ?