# Applied Multivariate Statistical Analysis: Homework 1

Homework format: all homework must be written in latex. You must turn in both your tex and pdf files. Attach your code and computer output if there is any programming.

1. (a) Consider an $N \times N$ invertible matrix $A$ and a vector $a \in \Re^N$, find explicit expressions for $|A + aa^T|$ and $(A + aa^T)^{-1}$, where $|A|$ denotes the determinant of a square matrix $A$.

   (b) For two non-singular $p \times p$ matrices $A$ and $B$, find an expression of $(A + B)^{-1}$ that contains only $A^{-1}$ and $B^{-1}$.

2. Suppose symmetric matrices $A$ and $B$ are both $(J \times J)$. Denote eigenvalues of $A$ and $B$ as $\{\lambda_j(A)\}$ and $\{\lambda_j(B)\}$, respectively. Please show that:

$$\sum_{j=1}^{J} \{\lambda_j(A) - \lambda_j(B)\}^2 \leq trace\{(A - B)(A - B)^T\}$$

3. $A$ is a $J \times J$ symmetric matrix, $U$ is a $(J \times K)$ matrix where $K \leq J$. Suppose that $U^T U = I_K$. Please show that

$$\lambda_j(U^T A U) \leq \lambda_j(A)$$

   and the equality holds if the columns of $U$ are the first $K$ eigen-vectors of $A$.

4. Consider matrix $X = \begin{pmatrix} 1 & 1 & 1 \\ 2 & 1 & 1 \\ 1 & 2 & 1 \\ 2 & 2 & 1 \end{pmatrix}$,

   (a) Find the QR and singular value decomposition of $X$. What are the two sets of basis vectors of the column space $C(X)$?

   (b) Use the SVD of $X$ to find the eigen-decomposition of $X^T X$. What are the eigenvalues and eigenvectors?

5. Consider a random sample $X_1, \ldots, X_N$ that are uniformly distributed in a unit ball in $\Re^p$, i.e., $\{x \in \Re^p : \|x\| \leq 1\}$.

   (a) Derive the median distance $M$ from the origin to the closest data point. What are the median distances for a sample of size $10^6$ and $p = 1, \ldots, 15$, respectively.

   (b) Derive the mean distance $D$ from the origin to the closest data point. What are the mean distances for a sample of size $10^6$ and $p = 1, \ldots, 15$, respectively.

6. Let $X_1$ be $N(0,1)$ and $X_2 = \begin{cases} -X_1, & -c \leq X_1 \leq c \\ X_1, & \text{otherwise} \end{cases}$ for some constant $c > 0$.

   (a) Show that $X_2$ also has a $N(0,1)$ distribution for any fixed $c$, but $(X_1, X_2)^T$ does not have a bivariate normal distribution.

   (b) Show that there exists $c > 0$ such that $\text{cov}(X_1, X_2) = 0$, but the resulting $X_2$ is not independent of $X_1$.

7. Consider a linear model with $p$ parameters, fit by ordinary least squares to a set of training data $(x_1, y_1), \ldots, (x_n, y_n)$ with the OLS estimate $\hat{\beta}_{OLS}$. Suppose we have some test data $(\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)$ drawn at random from the same population as the training data. Denote $R_{tr}(\beta) = n^{-1} \sum_{i=1}^{n} (y_i - \beta^T x_i)^2$ and $R_{te}(\beta) = m^{-1} \sum_{i=1}^{m} (\tilde{y}_i - \beta^T \tilde{x}_i)^2$, show that $E\{R_{tr}(\hat{\beta}_{OLS})\} \leq E\{R_{te}(\hat{\beta}_{OLS})\}$, where the expectations are taken over all random quantities.

8. Consider the linear model $y = X\beta + \epsilon$, where $X$ is $n \times p$ and $y \in \Re^n$, and of interest is

$$\hat{\beta} = \text{argmin}_{\{\beta \in \Re^p : A\beta = a\}} (y - X\beta)^T (y - X\beta),$$

   where the $q \times p$ matrix $A$ is of rank $q$, $q \leq p$, and $a \in \Re^q$. Show that

$$\hat{\beta} = \hat{\beta}_{OLS} - (X^T X)^{-1} A^T \{A (X^T X)^{-1} A^T\}^{-1} (A \hat{\beta}_{OLS} - a),$$

   where $\hat{\beta}_{OLS} = (X^T X)^{-1} X^T y$ is the ordinary least square estimator.