# Statistical Learning

## Midterm 1

1. The three-dimensional multivariate normal vector $\mathbf{X}$ has the mean equal to $\mu = (2, 4, 1)'$ and the covariance matrix

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0.5 \\ 0 & 0.5 & 1 \end{bmatrix}.$$

   a) What is the distribution of the first coordinate $X_1$ ? What would be the best prediction for $X_1$ ?

   b) What is the conditional distribution of $X_1$ given that $X_2 = 2$ ? What is the best prediction for $X_1$ given that $X_2 = 2$ ?

   c) What is the conditional distribution of $X_1$ given that $X_2 = 2$ and $X_3 = 0$ ? What is the best prediction for $X_1$ given that $X_2 = 2$ and $X_3 = 0$ ?

2. Consider a three dimensional multivariate normal distribution $N(\mu, \Sigma)$. The parameters' estimates based on the sample of $n = 5$ observations from this distribution are equal to

$$\hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \hat{\mu}_3)' = (\bar{X}_1, \bar{X}_2, \bar{X}_3)' = (10, 12, 11)'$$

$$\hat{\Sigma} = S = \begin{bmatrix} 3 & 1 & 1 \\ 1 & 3 & 1 \\ 1 & 1 & 3 \end{bmatrix}$$

.

   a) Calculate the value of the Hotelling T-test statistic for the hypothesis $\mathcal{H}_0 : \ \mu = (10, 10, 10)$.

   b) Provide the formula for the critical value of this statistic and the decision rule at the significance level $\alpha = 0.01$.

3 Let $X_1, \ldots, X_n$ be iid random vectors from the five dimensional multivariate normal distribution with independent coordinates $N(\mu, \sigma^2 I)$.

   - Provide the formula for $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_5)$, where $\hat{\mu}$ is the maximum likelihood estimate of $\mu$.
   - What is the distribution of $\hat{\mu}$ ?
   - What is the bias and the variance of $\hat{\mu}_1$ ?
   - What is the value of the mean squared error $MSE(\hat{\mu})$ ?
   - What is the bias and the variance of $0.5\hat{\mu}_1$ ?
   - What is the value of the mean squared error of $\tilde{\mu} = 0.5\hat{\mu}$ ?
   - Under which scenarios $\tilde{\mu}$ has the smaller MSE than $\hat{\mu}$.

4 Define the James-Stein estimator. What are its statistical properties ? Is it unbiased ? How does its variance and MSE compare to the corresponding characteristics of the maximum likelihood estimator.

5 P-values for 10 tests are equal to $0.005, 0.013, 0.16, 0.18, 0.47, 0.81, 0.90, 0.91, 0.92, 0.98$.

a) Which null hypothesis are rejected by the Benjamini-Hochberg procedure at the nominal FDR level $q = 0.05$ ?

b) Assume that the indicators of the null hypothesis are $(1, 1, 2, 1, 1, 2, 2, 2, 2, 2, 2)$. What is the False Discovery Proportion of the Benjamini-Hochberg procedure on this data set ?

6. Your data contains 10 variables. You fit 10 regression models including the first variable, the first two variables, etc. The residual sums of squares for these 10 consecutive models are equal to (1731, 730, 49, 48.9, 42, 39.5, 39.2,1 38.8, 37.6, 37.6). The sample size is equal to 50.

   a) Assuming that the standard deviation of the error term is known; $\sigma = 1$, find the estimate of the prediction errors for the first four models.

   b) Which of these 10 models is selected by AIC ? And which model is selected by BIC ?