Statistical Learning

List 1 - Multiple Regression and Multiple Testing

1. Generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $N(0, \sigma = \frac{1}{\sqrt{1000}})$. Then generate the vector of values of the response variable

$$Y = X\beta + \epsilon \ ,$$

where $\beta = (3, 3, 3, 3, 3, 0, ..., 0)^T$ and $\epsilon \sim N(0, I)$.

Perform the following analysis using reduced models containing only

i) first 10 columns of X

ii) first 100 columns of X

iii) first 500 columns of X

iv) all 950 columns.

a) For each of the above models find the least squares estimator of β and perform tests for significance of individual regression coefficients at the significance level $\alpha = 0.1$.

b) For each of the above models find the average standard deviation of the estimators of individual regression coefficients and the average length of the respective 90% confidence intervals. Compare these values for models of different sizes.

c) Calculate and compare the number of true and false discoveries for different models

i) without adjusting for multiple testing

ii) using Bonferroni correction

iii) using Benjamini-Hochberg correction

2. Repeat the above experiments 500 hundred times and calculate

a) the average variance of the estimators of individual regression coefficients - compare with the theoretical value (use the inverse Wishart distribution)

b) the average length of the 90% confidence interval (compare with the theoretical estimate)

c) the average number of true and false discoveries and the estimators of the Family Wise Error Rate (FWER) and the False Discovery Rate (FDR) for the procedures used in point c) of Problem 1. For the procedures without the correction and with the Bonferroni correction find the respective theoretical values.

Malgorzata Bogdan