Advanced methods of statistical learning

Information Criteria

1. Generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $N(0, \sigma = 0.1)$. Then generate the vector of the response variable according to the model

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

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

   a) Use BIC, AIC, RIC, mBIC i mBIC2 (you can use *bigstep* library in R) to identify important covariates when the search is performed over the data base date consisting of

      i) 20 first variables
      ii) 100 first variables
      iii) 500 first variables
      iv) all 950 variables.

      Report the number of false and true discoveries and the square error of the estimation of the vector of expected values of Y : $||X\beta - \hat{Y}||^2$.

   b) Repeat point a) 100 times and report the estimated power, FDR and mean squared error of the estimation of expected values of Y. Compare results of different criteria between themselves as well as to the resuts obtained in Problem 2 from list 1.

2. Compare RIC, mBIC and mBIC2 using example iv) of Problem 1 when the vector of true regression coefficients contains 50 nonzero entries, i.e. $\beta_i = 3$ for $i = 1, \ldots, 50$ and $\beta_i = 0$ for $i = 51, \ldots, 950$.

3. Generate the vector of the response variable according to the model

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

, where $\beta_1 = \ldots = \beta_{30} = 10$, $\beta_{31} = \ldots = \beta_{950} = 0$ and $\epsilon_1, \ldots, \epsilon_n$ are iid from a

   a) shifted exponential distribution with $\lambda = 1$
   b) Cauchy distribution.

   i) Use mBIC, mBIC2, rBIC and rBIC2 to identify important covariates. Report the number of true and false discoveries.
   ii) Use variables selected by rBIC2 and estimate the corresponding regression coefficients using least squares as well as the robust regression based on the Huber and Bi-square objective functions. Report the mean square error of estimation of regression coefficients.
   iii) Repeat this experiment 100 times and report estimated FDR and Power. For this part you do not need to estimate regression coefficients (i.e. you do not need to perform step ii).

Malgorzata Bogdan