Advanced methods of statistical learning

Information Criteria 2

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 logistic regression model

$$P_i = P(Y_i = 1) = 1 - P(Y_i = 0); \log\left(\frac{P_i}{1-P_i}\right) = X\beta + \epsilon \ ,$$

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 probabilities of success $||P - \hat{P}||^2$.

   b) Repeat point a) 100 times and report the estimated power, FDR and mean squared error of the estimation of the vector of probabilities P.

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. For $k = 5$ and $\sigma^2 = 16$ generate matrices $F_{50\times k}$ and $C_{k\times 200}$, where elements of F and C are independent $N(0, 1)$ random variables. Then generate the data matrix $X = M + E$, where $M = F \times C$ and the elements of E are independent $N(0, \sigma^2)$ random variables.

   a) Calculate the eigenvalues of the matrix $X^T X$ and use them to visually estimate the rank of the signal matrix M .

   b) Apply PESEL (Sobczyk et al. 2017, "Bayesian dimensionality reduction with PCA using penalized semi-integrated likelihood", JCGS, code available on Piotr Sobczyk github) to estimate the rank of M using the data matrix X.

   c) Repeat the experiment 100 times and draw the histogram of dimensions selected by PESEL.

   d) Repeat point c) for all combinations $k \in \{1, 5, 10, 20\}$ and $\sigma^2 \in \{16, 36\}$.

Malgorzata Bogdan