Advanced methods of statistical learning

Shrinkage methods and knockoffs

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 $\epsilon \sim N(0, I)$, $\beta_i = 3$ for $i \in \{1, \ldots, k\}$ and $\beta_i = 0$ for $i \in \{k+1, \ldots, 950\}$ and $k \in \{5, 20, 50, 100, 200\}$.

   a) For each of the above settings use one replication of the experiment and cross-validation in *glmnet* package to select the optimal tuning parameter for ridge regression and LASSO.

   b) For 500 replications of the above experiments estimate the regression coefficients using

      i) least squares

      ii) ridge regression and LASSO with the tuning parameters selected in point a)

      iii) adaptive version of SLOPE at FDR level 0.2

      Additionally, estimate the prediction errors of the methods provided in i) and ii) using their relationship with the residual sum of squares and calculate FDR and power of LASSO and adaptive SLOPE.

      Compare the mean square errors of the estimators of β for all the methods and use 500 of independent replicates to evaluate the accuracy of the estimation of the prediction error for all the methods apart from SLOPE.

   c) Use knockoffs with the estimates provided by the above methods so as to control FDR at the level of 0.2. For each of the methods described in i)-iii) estimate the FDR and power of the knock-off procedures.

Malgorzata Bogdan