Duże bazy danych

LASSO and Ridge regression

1. For this problem use the same orthonormal matrix of dimension 1000×950, which you generated for List 2. Consider the regression model

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

with $\epsilon \sim N(0, I_{n \times n})$ and the vector of regression coefficients $\beta_1 = \ldots = \beta_k = 3$ and $\beta_{k+1}, \ldots, \beta_{950} = 0$ with

a) $k = 20$,

b) $k = 100$,

c) $k = 200$.

For each of these cases

i) For each $i \in \{1, \ldots, 950\}$ calculate the bias, the variance and mse of the LASSO estimator with the tuning parameter λ.

ii) Calculate the value of the tuning parameter λ for LASSO, so as to minimize the mean square error of the estimation of the vector β, $MSE = E||\hat{\beta} - \beta||^2$.

iii) Compare the minimal MSE provided by LASSO to MSE provided by OLS and the minimal value of MSE provided by the ridge regression.

iv) Find the critical value and calculate the power of the statistical test based on the LASSO estimator and controlling FWER at the level 0.1.

iv) Generate 200 replicates of the above model and analyse the data using LASSO. Compare empirical bias, variance, mse and power of the test based on the LASSO estimator with the theoretical values of these parameters, calculated above and with the corresponding parameters of OLS and ridge, calculated previously.

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

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

with $\beta = (\beta_1, \ldots, \beta_p)$, where $\beta_1 = \ldots = \beta_k = 3$, $\beta_{k+1} = \ldots, \beta_p = 0$ with $k = 10$ and $\epsilon \sim N(0, I)$.

a) Estimate the parameters of this model using LASSO and ridge regression with the tuning parameter selected by *crossvalidation* and compare $SE = ||\hat{\beta} - \beta||^2$ of both methods.

b) For both methods use model-free knockoffs to identify important predictors at FDR level 0.2 and compare numbers of true and false positives.

c) Repeat points a)-b) 100 times and compare average SE for both methods with cross-validarion and the power and FDR of the knockoffs based on both methods.

3. Repeat 2 with $k = 100$ and $\beta_1 = \ldots = \beta_k = 2$.

4. Repeat 2 when rows of X are iid random vectors from $\frac{1}{n}N(0, \Sigma)$, where $\Sigma_{ii} = 1$ and for $i \neq j$ $\Sigma_{ij} = 0.5$, and with $\beta_1 = \ldots = \beta_{10} = 5$.

Malgorzata Bogdan