Statistical Learning

Regularization and knockoffs

1. Generate the orthonormal ($X^T X = I$) matrix od dimension 1000×950. Consider the regression model

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

with $\epsilon \sim N(0, I_{n \times n})$ and the vector of regression coefficients $\beta_1 = \ldots = \beta_k = 3.5$ and $\beta_{k+1} = \ldots = \beta_{950} = 0$ with

   a) k = 20,

   b) k = 100,

   c) k = 200.

   For each of these cases

   i) To be done by hand: Calculate the value of the tuning parameter λ for the ridge regression, so as to minimize the mean square error of the estimation of β.

   ii) To be done by hand: Calculate the bias, the variance and the mean squared error of this optimal estimator.

   iii) Generate 200 replicates of the above model and analyze the data using ridge regression and OLS. Compare empirical bias, variance, mse of the ridge regression with the theoretical values of these parameters, calculated above, and with the corresponding parameters of OLS.

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 models proposed in Task 1, above.

   Estimate the parameters of this model using the ridge regression and LASSO and the tuning parameter λ selected by

   a) minimizing the prediction error criterion PE (assume you know σ)

   b) 10 fold cross-validation
   and with

   c) OLS

   d) OLS within the model selected by mBIC2 and AIC.

   Compare the estimation errors $||\hat{\beta} - \beta||^2$ and $||X(\hat{\beta} - \beta)||^2$ for these 8 approaches.

   Repeat the above experiment 100 times and compare the mean square errors of estimation of β and $\mu = X\beta$ for the above approaches.

3. Generate the design matrix $X_{100 \times 200}$ such that its elements are iid random vectors from $\frac{1}{n} N(0, \Sigma)$, where $\Sigma_{ii} = 1$ and for $i \neq j$ $\Sigma_{ij} = 0.7$. Now, consider the vector $\beta^k \in R^{200}$, such that $\beta_1 = \ldots = \beta_k = 20$ and $\beta_{k+1} = \ldots = \beta_{200} = 0$.

- Find the maximal k for which the LASSO irrepresentability condition is satisfied and call it k_{IR}. Then generate the response variable according to the formula

$$Y = X\beta^{k_{IR}} \ ,$$

(noisless case) and empirically find λ such that LASSO can recover the sign of β. If this turns out not to be possible, increase the magnitude of the nonzero elements of β.

- Find the maximal k for which the LASSO identifiability condition is satisfied and call it k_{ID}. Then generate the response variable according to the formula

$$Y = X\beta^{k_{ID}}$$

and empirically find λ such that LASSO can properly separate zero and nonzero elements of β. If this turns out not to be possible, increase the magnitude of the nonzero elements of β.

- Generate the response variable according to the formula

$$Y = 100 * X\beta^{k_{ID}+1} \ ,$$

and check if there exists λ which allows for separating zero and nonzero elements of β.

4. For this problem use the set realdata.Rdata from List 2 and the same split of the data into the training and the test set as the one you used for the previous assignment.

   a) Use the training set (180 individuals) to construct the regression model explaining the expression level of gene 1 (first column in the data set) as the function of expression levels of other genes. Use Ridge regression and LASSO and apply crossvalidation to select the tuning paramter (verify that cv.glmnet indeed identifies the minimum of the prediction error). Use the test set to verify the predictive accuracy of considered models. Compare to the predictive performance of model selection criteria from the previous assignment. Compare also the number of variables selected by different methods.

   b) Preselect interesting explanatory variables. Select 300 variables with the largest marginal correlation with the response variable and add variables selected by mBIC2. Then apply regularization methods (ridge and LASSO) to build a predictive model on such reduced set of variables. Use the test set to verify the predictive performance of the obtained models and compare to the predictive properties of models obtained in earlier experiments.

5. Consider again the setup from Problem 3 but now generate a noisy response

$$Y = X\beta^{k_{ID}} + \epsilon \ ,$$

where $\epsilon \sim N(0, I)$.

   a) Find the value of the tuning parameter λ for which LASSO MSE is minimal and use $\hat{\beta}^L$ to denote the corresponding LASSO estimate of β.

   b) Calculate the number of true and false discoveries for this selection of λ.

   c) Run the adaptive LASSO with weights $w_i = \frac{1}{|\beta_i^L| + 0.000001}$. Select the tuning parameter so as MSE is minimal and compare this value of MSE with the optimal value of MSE for LASSO (see point a)).

d) Calculate the number of true and false discoveries for the above selection of λ for adaptive LASSO and compare to the values obtained in point b).

6. Consider again the setup from Problem 5 but now generate a response

$$Y = X\beta^{k_{ID}+10} + \epsilon \ ,$$

where $\epsilon \sim N(0, I)$.

a) Find the value of the tuning parameter λ for which LASSO MSE is minimal and use $\hat{\beta}^L$ to denote the corresponding LASSO estimate of β.↙ *Runtheadaptive LASSO with weights* $\mathrm{w}_i = \frac{1}{|\beta_i^L|+0.000001}$. Select the tuning parameter so as MSE is minimal and compare this value of MSE with the optimal value of MSE for LASSO (see point a)).

b) Analyze this data set with SLOPE with BH sequence at the FDR level 0.2. Adjust the parameter α so as to obtain the minimal MSE. Compare this MSE to the MSE obtained by LASSO and adaptive LASSO.

7. Generate the design matrix $X_{100\times 200}$ such that its elements are iid random vectors from $\frac{1}{n}N(0, \Sigma)$, where $\Sigma_{ii} = 1$ and for $i \neq j$ $\Sigma_{ij} = 0.7$. Then generate the vector of the response variable according to the model

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

where $\epsilon \sim N(0, I)$, $\beta_i = 30$ for $i \in \{1, \ldots, k\}$ and $\beta_i = 0$ for $i \in \{k+1, \ldots, 200\}$ and $k \in \{5, 20\}$.

For 100 replications of the above experiment:

- use knockoffs with ridge and LASSO to identify important variables while keeping FDR equal to 0.1,
- use multiple knockoffs (with 5 knockoffs copies) with LASSO and ridge at the FDR level 0.1.

Summarize the results.

Malgorzata Bogdan