Statistical Learning
Assignment 4

Exercises

1. Assume the linear model:
$$Y = X\beta + \epsilon,$$
where $X'X = I$ and $\epsilon \sim N(0, \sigma^2 I)$.

   Find the numerical solution for the elastic net in the form:
$$\hat{\beta}_{\text{en}} = \text{argmin}_b \frac{1}{2}\|Y - Xb\|_2^2 + \lambda \left(\frac{1}{2}(1-\alpha)\|b\|_2^2 + \alpha \sum_{i=1}^{p} |b_i| \right)$$

   - What would be the value of the elastic net estimator with $\lambda = 1$ and $\alpha = 0.5$ if $\hat{\beta}_{\text{OLS}} = 3$?
   - How does the number of discoveries depend on the parameter α?
   - Provide the numerical value for the expected number of false discoveries when $n = p = 1000$, $p_0 = 950$, $\sigma = 1$, and $\lambda = 2$, and the power of detection of X_1 when $\beta_1 = 3$.

2. Why do the LASSO, SLOPE, and elastic net perform variable selection, while ridge regression does not?

3. Formulate the identifiability condition for LASSO. What does it guarantee in terms of model selection? How does it compare to the irrepresentability condition?

4. Define SLOPE. How is it different from LASSO in terms of formulations and properties?

5. What are knockoffs?

6. The vector of W statistics for the knockoffs procedure is equal to:
$$W = (8, -4, -2, 2, -1.2, -0.6, 10, 12, 1, 5, 6, 7).$$
Which variables would be considered important if we use knockoffs at the false discovery rate (FDR) level $q = 0.4$?

7. Show that ridge regression can be viewed as the Maximum A Posteriori (MAP) Bayes rule with a multivariate normal prior on regression coefficients.

Computer project

Generate the design matrix $X_{500\times 450}$ such that its elements are independent and identically distributed (iid) random variables from $\mathcal{N}(0, \sigma = \sqrt{\frac{1}{n}})$. Then generate the vector of the response variable according to the model:

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

where $\epsilon \sim 2\mathcal{N}(0, I)$, $\beta_i = 10$ for $i \in \{1, \ldots, k\}$, $\beta_i = 0$ for $i \in \{k+1, \ldots, 450\}$, and $k \in \{5, 20, 50\}$.

For 100 replications of the above experiments, estimate the regression coefficients and/or identify important variables using:

i) Least squares.

ii) Ridge regression and LASSO with the tuning parameters selected by cross-validation.

iii) Knockoffs with ridge and LASSO at the nominal false discovery rate (FDR) equal to 0.2.

Perform the following analyses:

a) Estimate the false discovery rate (FDR) and the power of the cross-validated LASSO and the knockoffs with ridge and LASSO.

b) For all three methods in i) and ii), estimate the mean square errors of the estimators of β and $\mu = X\beta$.