Statistical learning

Assignment 4

Jakub Skalski

October 6, 2024

1. Elastic Net

Elastic net is a regularization technique used in linear regression models to improve their predictive accuracy and interpretability. It combines the penalties of two popular methods, Ridge Regression (L2 penalty) and Lasso Regression (L1 penalty).

1.1 Formulation

$$\hat{\beta}_{\text{en}} = \arg\min_b \left(\frac{1}{2}\|Y - Xb\|_2^2 + \lambda \left(\frac{1}{2}(1-\alpha)\|b\|_2^2 + \alpha \sum_{i=1}^{p} |b_i| \right) \right)$$

1.2 Closed form solution

Assuming the linear model:

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

where $X'X = I$ and $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. In orthogonal case we can obtain at the solution analytically:

$$\nabla_b \left(\frac{1}{2}\|Y - Xb\|_2^2 + \lambda \left(\frac{1}{2}(1-\alpha)\|b\|_2^2 + \alpha \sum_{i=1}^{p} |b_i| \right) \right) =$$

$$\frac{1}{2} \left(\nabla_b \|Y - Xb\|_2^2 + \lambda \left(\frac{1}{2}\nabla_b (1-\alpha)\|b\|_2^2 + \alpha \nabla_b \sum_{i=1}^{p} |b_i| \right) \right)$$

The resulting gradient is depends on the OLS solution $\hat{b}$ and a sign vector $S(b)$:

$$b - \hat{b} + \lambda(1-\alpha)b + \lambda\alpha S(b)$$

From now on let us consider each coefficient separately and solve for its zero derivative:

$$b_i - \hat{b}_i + \lambda(1-\alpha)b_i + \lambda\alpha \operatorname{sgn}(b_i) = 0$$

Adjusting for the general case we arrive at the closed form solution:

$$\hat{\beta}_i = b_i = \text{sgn}(\hat{b}_i)\left(\frac{|\hat{b}_i| - \lambda\alpha}{1 + \lambda(1-\alpha)}\right)^+$$

1.3 Ordinary least squares relation

Knowing the OLS solution beforehand we can compute elastic net estimator. Suppose $\lambda = 1$, $\alpha = 0.5$, $\hat{\beta}_{OLS} = 3$. We can simply plug those into our elastic net formula:

$$\hat{\beta}_i = 1 \cdot \left(\frac{3 - \frac{1}{2}}{1 + \frac{1}{2}}\right)^+ = \frac{5}{3}$$

1.4 Parameterization

The α parameter dictates of the raltion between ridge and LASSO regularization terms. LASSO promotes sparsity, while ridge discovers everything, so then the expected number of discoveries drops as alpha increases.

2. Variable Selection in Regression

In LASSO, SLOPE and elastic net optimization L1 norm constraint creates corners in the feasible region where the optimization can be performed. When the optimization hits these corners, some coefficients become zero, leading to variable selection.

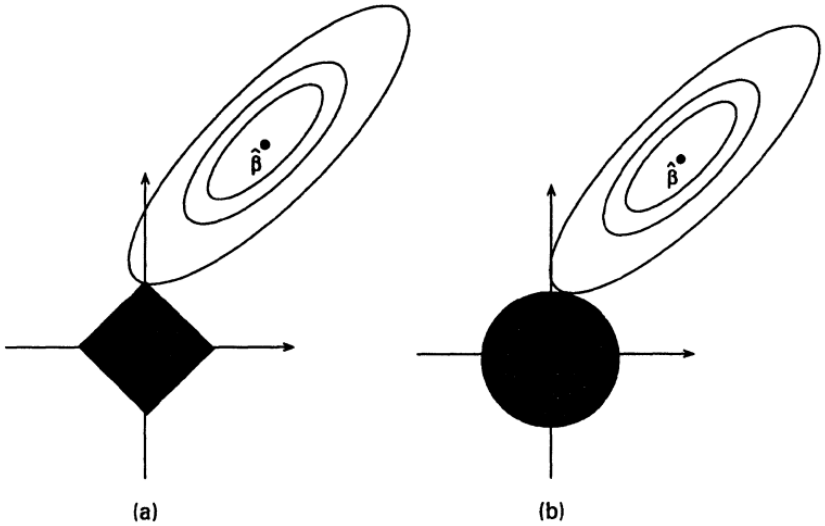

Figure 1: Estimation for (a) the lasso and (b) ridge regression

3. Identifiability and irrepresentability conditions

3.1 Identifiability

A model is considered identifiable if, in theory, it is possible to determine the true values of its underlying parameters given an infinite number of observations. Mathematically, this means that distinct parameter values must produce different probability distributions for the observable variables.

Theorem (Tardivel, Bogdan, 2019)

For any $\lambda > 0$, LASSO can separate well the causal and null features if and only if vector β is identifiable with respect to ℓ_1 norm and $\min_{i \in I} |\beta_i|$ is sufficiently large.

Corollary

Appropriately thresholded LASSO can properly identify the sign of sufficiently large β if and only if β is identifiable with respect to ℓ_1 norm.

3.2 Irrepresentability

The sign vector of β is defined as

$$S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p,$$

where for $x \in \mathbb{R}$, $S(x) = 1_{x>0} - 1_{x<0}$. Let $I := \{i \in \{1, \ldots, p\} \mid \beta_i \neq 0\}$, and let $X_I, X_{\bar{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$. Irrepresentable condition:

$$\|X_{\bar{I}}^T X_I (X_I^T X_I)^{-1} S(\beta_I)\|_\infty \leq 1$$

When

$$\|X_{\bar{I}}^T X_I (X_I^T X_I)^{-1} S(\beta_I)\|_\infty > 1$$

then the probability of the support recovery by LASSO is smaller than 0.5 (Wainwright, 2009). Should the irrepresentability condition be violated, Lasso may fail to select the correct model even if identifiability conditions are met.

4. SLOPE relation to LASSO

4.1 Adaptive LASSO

Adaptive LASSO introduces weights to the ℓ_1 penalty, allowing for differential shrinkage of coefficients based on initial estimates. The formulation of adaptive LASSO is:

$$\text{Objective:} \quad \min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \lambda \sum_{j=1}^{p} w_j |\beta_j| \right\}$$

Here, w_j are the weights applied to the j-th coefficient. These weights are typically chosen as the inverse of the absolute values of initial estimates of the coefficients (e.g., from an ordinary least squares (OLS) fit or a preliminary LASSO fit).

4.2 Sorted L-One Penalized Estimation (SLOPE)

SLOPE penalizes the sorted absolute values of the coefficients and is formulated as follows:

$$\text{Objective:} \quad \min_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - X_i\beta)^2 + \sum_{j=1}^{p} \lambda_j |\beta_{(j)}| \right\}$$

Here, $\beta_{(j)}$ denotes the j-th largest coefficient in absolute value, and λ_j are a sequence of tuning parameters with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$.
This method tends to group variables with similar effect sizes due to the sorted nature of its penalties, making it more effective in selecting groups of correlated variables and can also be tuned to control FDR.

5. Knockoffs

Knockoffs are synthetic variables created to provide a means of controlling the false discovery rate (FDR) in variable selection procedures. The idea is to create ”knockoff” versions of the original variables that mirror the correlation structure of the original variables but are conditionally independent of the response. By comparing the importance of original variables to their knockoff counterparts, one can identify truly significant variables while controlling FDR.

5.1 Necessary requirement

The real variables X and the knockoff variables $\tilde{X}$ need to satisfy the following:

$$\Sigma_X = \Sigma_{\tilde{X}} \text{ and for } i \neq j \, \text{Cov}(X_i, \tilde{X}_j) = \text{Cov}(X_i, X_j).$$

5.2 Knockoff filter

Define a random threshold as

$$\hat{t}(\lambda) = \min \left\{ t > 0 : \frac{1 + \#\{j : W_j(\lambda) \leq -t\}}{\#\{j : W_j(\lambda) \geq t\}} \leq q \right\}$$

where $W_j = |\hat{\beta}_j| - |\hat{\beta}_{p+j}|$ and select

$$\hat{S}(\lambda) = \{j : W_j(\lambda) \geq \hat{t}(\lambda)\}.$$

5.3 Identifying important variables

Suppose we have computed a knockoff statistic $W = (8, -4, -2, 2, -1.2, -0.6, 10, 12, 1, 5, 6, 7)$ and we want to control FDR at level 0.4. We start by first computing the random threshold which in this case is $t = 4$. Variables above the threshold are therefore at the indices 1, 7, 8, 10, 11, 12.

6. Computer project

In this project we evaluate the performance of different regression methods — least squares, ridge regression, LASSO, and knockoffs with ridge and LASSO — in terms of false discovery rate (FDR), power, and mean square error of the estimators β and $\mu = X\beta$ through repeated simulations.

6.1 Project setting

Data is a single-time generated design matrix $X_{500\times450}$ such that its elements are independent and identically distributed (iid) random variables from $\mathcal{N}(0, \sigma = \sqrt{\frac{1}{n}})$. Then the each-time generated vector of the response variable is the following:

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

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

6.2 FDR and power estimations

Table 1: False discovery rates

k	LASSO	LASSO + knockoffs	ridge + knockoffs
5	0.205	0.158	0.123
20	0.501	0.174	0.178
50	0.520	0.186	0.191

Table 2: Powers

k	LASSO	LASSO + knockoffs	ridge + knockoffs
5	0.914	0.782	0.406
20	0.993	0.968	0.744
50	0.993	0.957	0.612

We can see that the knockoffs excel at larger values of k. We achieve a comparable power to pure LASSO while keeping FDR below 0.2.

6.3 Mean squared estimation errors

Table 3: Mean squared difference of coefficients

k	OLS	LASSO	ridge
5	16577	249	497
20	16709	555	1454
50	15558	1124	2376

Table 4: Mean squared difference of predictions

k	OLS	LASSO	ridge
5	1790	234	478
20	1816	474	995
50	1779	780	1287