Report 4

Kamil Zaborniak

Contents

Elastic net **2**

Variable selection by LASSO, SLOPE, elastic net and Ridge Regression **3**

The identifiability condition for LASSO **3**

SLOPE and LASSO **4**

Knockoffs **4**

Ridge Regression and the Maximum A Posteriori Bayes rule **5**

Computer Project **5**

Simulation desription 6

Results 6

Elastic net

Let's assume the linear model:

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

where $X^T X = I$ and $\varepsilon \sim N(0, \sigma^2 I)$.

Firstly, we will find the numerical solution for the elastic net in the form:

$$\hat{\beta}_{en} = \underset{b}{\text{argmin}} \tfrac{1}{2}\|Y - X\beta\|_2^2 + \lambda \left(\tfrac{1}{2}(1-\alpha)\|\beta\|_2^2 + \alpha \sum_{i=1}^{p} |\beta_i|) \right).$$

Let's start by considering the function:

$$\begin{aligned} f(b) &= \tfrac{1}{2}(Y - Xb)^T(Y - Xb) + \lambda \left(\tfrac{1}{2}(1-\alpha) b^T b + \alpha \sum_{i=1}^{p} |b_i| \right) \\ &= \tfrac{1}{2}\left(Y^T Y - Y^T X b - b^T X^T Y + b^T X^T X b\right) + \tfrac{1-\alpha}{2}\lambda b^T b + \alpha\lambda \sum_{i=1}^{p} |b_i| \\ &= \tfrac{1}{2} Y^T Y - Y^T X b + \tfrac{1}{2} b^T b + \tfrac{1-\alpha}{2}\lambda b^T b + \alpha\lambda \sum_{i=1}^{p} |b_i| \\ &= \tfrac{1}{2} Y^T Y - Y^T X b + \tfrac{1}{2}\left(1 + \lambda - \alpha\lambda\right) b^T b + \alpha\lambda \sum_{i=1}^{p} |b_i| \\ &= \sum_{i=1}^{p} \left(-\hat{\beta}_i^{(LS)} b_i + \tfrac{1}{2}\left(1 + \lambda - \alpha\lambda\right) b_i^2 + \alpha\lambda |b_i| \right) + \tfrac{1}{2} Y^T Y, \end{aligned}$$

where $Y^T X = (X^T Y)^T = \hat{\beta}_{LS}^T$. Let's denote function:

$$g(b_i) = -\hat{\beta}_i^{(LS)} b_i + \tfrac{1}{2}\left(1 + \lambda - \alpha\lambda\right) b_i^2 + \alpha\lambda b_i \cdot \text{sgn}(b_i).$$

The solution to the following equation:

$$\tfrac{\partial}{\partial b_i} g = 0$$

has form:

$$b_i = \begin{cases} \frac{\hat{\beta}_i^{(LS)} - \alpha\lambda}{1 + \lambda - \alpha\lambda} & \hat{\beta}_i^{(LS)} > b_i > 0; \\ \frac{\hat{\beta}_i^{(LS)} + \alpha\lambda}{1 + \lambda - \alpha\lambda} & \hat{\beta}_i^{(LS)} \leq 0. \end{cases}.$$

What is equivalent to:

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

For $\hat{\beta}^{(LS)} = 3$, $\lambda = 1$, $\alpha = 0.5$:

$$b = \tfrac{1}{1+1-0.5}(3 - 0.5) = \tfrac{5}{3}.$$

The elastic net makes discovery for $\alpha \in [0, 1]$ and $\lambda > 0$, when $\left|\hat{\beta}_i^{(LS)}\right| > \alpha\lambda$. When we set $\alpha = 0$, we use Ridge regression, but when we set $\alpha = 1$, then we use LASSO regression. Generally, the elastic net is the mix of Ridge and LASSO regression. So when the α increase, LASSO works more and Ridge less (Ridge doesn't set any coefficient to 0, but LASSO does), then model starts set to 0 more coefficients.

Let's consider the example with following values: $n = p = 1000$, $p_0 = 950$ (number of important variable), $\sigma = 1$ and $\lambda = 2$. Additionally, let's assume $\beta_1 = 3$. The expected number of false discoveries is equal to:

$$\begin{aligned}
(p-p_0)P_0\left(\left|\hat{\beta}_i^{(LS)}\right| > \alpha\lambda\right) &= (p-p_0)\left(P_0\left(\hat{\beta}_i^{(LS)} < -\alpha\lambda\right) + P_0\left(\hat{\beta}_i^{(LS)} > \alpha\lambda\right)\right) \\
&= (p-p_0)\left(P_0\left(\frac{\hat{\beta}_i^{(LS)}-\beta_i}{\sigma} < -\frac{\alpha\lambda}{\sigma}\right) + P_0\left(\frac{\hat{\beta}_i^{(LS)}-\beta_i}{\sigma} > \frac{\alpha\lambda}{\sigma}\right)\right) \\
&= 2(p-p_0)\left(1-\Phi\left(\frac{\alpha\lambda}{\sigma}\right)\right) \\
&= 2\cdot 50\cdot(1-\Phi(2\alpha)) = 100 - 100\Phi(2\alpha).
\end{aligned}$$

The power of detection X_1 when $\beta_1 = 3$ is equal to:

$$\begin{aligned}
P_1\left(\left|\hat{\beta}_1^{(LS)}\right| > \alpha\lambda\right) &= P_1\left(\hat{\beta}_1^{(LS)} < -\alpha\lambda\right) + P_1\left(\hat{\beta}_1^{(LS)} > \alpha\lambda\right) \\
&= P_1\left(\frac{\hat{\beta}_1^{(LS)}-\beta_1}{\sigma} < -\frac{\alpha\lambda+\beta_1}{\sigma}\right) + P_0\left(\frac{\hat{\beta}_1^{(LS)}-\beta_1}{\sigma} > \frac{\alpha\lambda-\beta_1}{\sigma}\right) \\
&= \Phi\left(-\frac{\alpha\lambda+\beta_1}{\sigma}\right) + 1 - \Phi\left(\frac{\alpha\lambda-\beta_1}{\sigma}\right) \\
&= \Phi\left(-2\alpha-3\right) + 1 - \Phi\left(2\alpha-3\right).
\end{aligned}$$

Variable selection by LASSO, SLOPE, elastic net and Ridge Regression

The LASSO (Least Absolute Shrinkage and Selection Operator), SLOPE (Sorted L-One Penalized Estimation), and elastic net perform variable selection due to the nature of their penalty terms, which induce sparsity in the coefficients. Ridge regression, on the other hand, does not perform variable selection because its penalty term does not induce sparsity. Here's a detailed explanation of the methods behind these models:

- LASSO (Least Absolute Shrinkage and Selection Operator)

  Penalty parameter is equal to sum of absolute values of coefficients. In a result, LASSO sets some coefficients to 0, when the value of penalty parameter is large enough.

- SLOPE (Sorted L-One Penalized Estimation)

  Firstly, the SLOPE sort absolute values of LS estimators of coefficients and then penalizes the coefficients according to their rank after sorting. In a result the SLOPE penalty encourages sparsity similar to LASSO but does so in a manner that takes into account the relative magnitudes of the coefficients. It can provide better control over the false discovery rate, effectively selecting variables by driving some coefficients to zero.

- Elastic net

  As previously, Elastic net is a combination of Ridge regression and LASSO. Because the LASSO sets some coefficients to 0, the Elastic net just does it also with appropriate parameter values.

- Ridge Regression

  The penalty of RR is the sum of squares of coefficients. Ridge Regression shrinks the coefficients but does not force any of them to be exactly zero. It reduces the magnitude of coefficients without performing variable selection.

The identifiability condition for LASSO

The identifiability condition for LASSO makes us sure that the solution to the LASSO is unique. One common identifiability condition is that the design matrix must be full rank, meaning that its columns (predictors) are linearly independent. What is more, $X^T X$ have to be positive definite.

The identifiability condition guarantees that the LASSO optimization has a unique solution, what is important for model selection. It's mean that as size of sample increase, the LASSO estimator will consequently indentify the true non-zeros in the model.

The irrepresentability condition is a more specific requirement for the LASSO to consistently select the correct coefficients. It is directly relates to the ability of LASSO to recover the true sparsity pattern of the coefficient vector. The irrepresentability condition can be stated as follows:

Let S denote the vector: $S(\beta) = (S(\beta_1), \ldots, S(\beta_p)) \in \{-1, 0, 1\}^p$, where for $x \in \mathbb{R}$, $S(x) = \mathbb{I}_{x>0} - \mathbb{I}_{x<0}$. Let $I := \{i \in \{1, \ldots, p\} | \beta_i \neq 0\}$, and let X_I, $X_{\overline{I}}$ be matrices whose columns are respectively $(X_i)_{i \in I}$ and $(X_i)_{i \notin I}$. The irrepresentability condition requires that:

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

When

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

then probability of the support recovery by LASSO is smaller than 0.5.

While the identifiability condition guarantees a unique solution to the LASSO, the irrepresentability condition goes further by ensuring that this unique solution corresponds to the true model, allowing LASSO to perform consistent variable selection.

SLOPE and LASSO

Now, let's try to explain the differences between LASSO and SLOPE.

- LASOO

  $$\hat{\beta}_{LASSO} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2} \|Y - X\beta\|_2^2 + \lambda \sum_{i=1}^{p} |\beta_i| \right),$$

  where λ is a single penalty parameter. Encourages sparsity by shrinking some coefficients to zero, thus performing variable selection. The level of sparsity depends on the single penalty parameter λ. It does not explicitly control the false discovery rate and may struggle with correlated predictors.

- SLOPE

  $$\hat{\beta}_{SLOPE} = \underset{\beta}{\operatorname{argmin}} \left(\frac{1}{2} \|Y - X\beta\|_2^2 + \sum_{i=1}^{p} \lambda_i |\beta_i| \right),$$

  where λ_i values are ordered and decrease, providing a sequence of penalties based on the rank of the absolute values of the coefficients. Also promotes sparsity but does so in a more nuanced way. By using a sequence of decreasing penalties parameters. SLOPE can better control the false discovery rate and adapt to different scales of the coefficients. It offers better performance in selecting true variables, especially in the presence of correlated predictors.

Knockoffs

Knockoffs are a statistical method designed to control the FDR in variable selection problems, especially in high-dimensional settings. The knockoff method creates synthetic variables, called "knockoffs," which mimic the correlation structure of the original variables (predictors) but are constructed in such a way that they do not contain any information about the response variable. These knockoffs are then used as a baseline to identify truly important variables while controlling for false positives. The method is flexible and can be used with various variable selection procedures. Knockoffs are particularly useful in situations with highly correlated predictors, where traditional methods like LASSO might struggle.

The vector of W statistics for the knockoffs procedure is equal to $(8, -4, -2, 2, -1.2, -0.6, 10, 12, 1, 5, 6, 7)$. If we use knockoffs at the false discovery rate level $q = 0.4$, then the indicates of considered variable has form $(1, 4, 7, 8, 9, 10, 11, 12)$.

Ridge Regression and the Maximum A Posteriori Bayes rule

We will show that the Ridge Regression can be viewed as the Maximum A Posteriori Bayes rule with a multivariate normal prior on regression coefficients. In a Bayesian rule, we aim to find the posterior distribution of the regression coefficients β given the data (X, Y). The posterior distribution is proportional to the product of the likelihood and the prior distribution:

$$p(\beta|X, Y) \propto L(Y|X, \beta)p(\beta).$$

We assume the errors are normally distributed, the likelihood $L(Y|X, \beta)p(\beta)$ is given by:

$$Y|X, \beta \sim N(X\beta, \sigma^2 I).$$

The likelihood function is then:

$$L(Y|X, \beta) \propto \exp\left(\tfrac{-1}{2\sigma^2}\|Y - X\beta\|_2^2\right).$$

For Ridge regression, we assume a multivariate normal prior on $\beta \sim N(0, \tau^2 I)$, then the prior distribution is:

$$p(\beta) \propto \exp\left(\tfrac{-1}{2\tau^2}\|\beta\|_2^2\right).$$

Combining the likelihood and the prior, the posterior distribution is proportional to:

$$p(\beta|X, Y) \propto \exp\left(-\tfrac{1}{2\sigma^2}\|Y - X\beta\|_2^2\right)\exp\left(-\tfrac{1}{2\tau^2}\|\beta\|_2^2\right).$$

The Maximum A Posteriori estimator has form:

$$\hat{\beta}_{MAP} = \underset{\beta}{\operatorname{argmin}}\left(\tfrac{1}{2}\|Y - X\beta\|_2^2 + \tfrac{1}{2} \cdot \tfrac{\sigma^2}{\tau^2}\|\beta\|_2^2\right).$$

Comparing this to the ridge regression objective function, we see that they are equivalent if we set the regularization parameter λ in ridge regression to be the ratio of the noise variance to the prior variance:

$$\lambda = \tfrac{\sigma^2}{\tau^2}.$$

Thus, ridge regression can be interpreted as MAP estimation with a multivariate normal prior on the regression coefficients, where the regularization parameter λ corresponds to the ratio of the error variance to the prior variance.

Computer Project

Now, it's time to check the theory using computer simulation. We will check how work the LASSO and Ridge Regression with cross-validation and Knockoffs. We will estimate the FDR and the power of these methods and we will also estimate MSE of estimators of coefficients vector (β) and $\mu = X\beta$.

Simulation desription

With the single replication, we will generate the design matrix $X_{500\times 450}$ such that its elements are independent and identically distributed random variables form $N\left(0, \frac{1}{500}\right)$. Next, we will generate the vector of the response variable according to the model:

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

where $\varepsilon \sim 2N(0, I)$, $\beta_i = 10$ for $i \in \{1, \ldots, k\}$ and $\beta_i = 0$ for $i \in \{k+1, \ldots, 450\}$. For this, we will estimate the regression coefficients and identify the important variable using Least Squares, Ridge regression and Lasso with tuning parameters selected by cross-validation and Knockoffs with Ridge and Lasso at the nominal FDR equal to 0.2. We will repeat this replication 100 times for all 3 cases corresponding to $k \in \{5, 20, 50\}$.

Results

As a result of simulation we obtained the following measures.

		Methods		
		LASSO (CV)	knockoff	
			LASSO	Ridge R.
k	5	0.829	0.108	0.040
	20	0.761	0.243	0.185
	50	0.663	0.173	0.222

Table 1: Table of FDR of methods.

We can see that knockoffs controls FDR, better for the LASSO. Cross-validated LASSO make more more false discoveries. We also can see that FDR increase when k decrease.

		Methods		
		LASSO (CV)	knockoff	
			LASSO	Ridge R.
k	5	1	0.800	0.360
	20	1	0.990	0.550
	50	1	0.995	0.702

Table 2: Table of Powers of methods.

On the above table we can see, that the cross-validated LASSO has the greatest power in comparison to knockoffs. The 2nd. place has the LASSO with knockoffs. The worst, independent on k, is the Ridge Regression with knockoffs.

		Methods of β estimation		
		Least Squares	Cross-validation	
			LASSO	Ridge R.
k	5	49.304	0.283	0.910
	20	43.935	1.027	2.744
	50	37.355	1.978	4.652

Table 3: Table of MSE of β estimators.

		Methods of β estimation		
		Least Squares	Cross-validation	
			LASSO	Ridge R.
k	5	3.488	0.231	0.664
	20	3.576	0.691	1.494
	50	3.492	1.227	2.066

Table 4: Table of $\mu = \frac{1}{n}\|X\beta - X\hat{\beta}\|_2^2$ for method of coefficients estimation.

We can see that the cross-validation method (especially LASSO) estimate the coefficients vector the best. The worst method is the Least Squares.

To sum up, the best method among those considered is the LASSO with knockoffs and the 2nd. is the Ridge regression with knockoffs.