Statistical learning

Report 4

Maciej Szczutko

2024-06-09

Elastic-net regression

The elastic net is another extension proposed by Hui Zou and Trevor Hastie in 2005 in “Regularization and variable selection via the elastic net”. The main motivation for this method was handle the data with highly correlated data. The LOSS functon is defined as

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

Assume orthonormal design we can derived explicit formula for $\hat{\beta}_{\text{en}}$.

$$\begin{aligned}
\beta^{\text{EN}} &= \arg\min \sum \frac{1}{2}\|Y - X\beta\|_2^2 + \lambda\left(\frac{1-\alpha}{2}\|b\|_2^2 + \alpha\sum_{i=1}^{p}|b_i|\right)\\
f(b) &= \frac{1}{2}(Y - Xb)^T(Y - Xb) + \lambda\left(\frac{1-\alpha}{2}b^T b + \alpha\sum_{i=1}^{p}|b_i|\right) =\\
&= \frac{1}{2}\left(Y^T Y - 2Y^T Xb + b^T X^T Xb\right) + \frac{\lambda(1-\alpha)}{2}b^T b + \lambda\alpha\sum_{i=1}^{p}|b_i| =\\
&= b^T b\left(\frac{1}{2} + \frac{\lambda(1-\alpha)}{2}\right) + \lambda\alpha\sum_{i=1}^{p}|b_i| - Y^T Xb + \frac{1}{2}Y^T Y =\\
&= \sum_{i=1}^{p}\left[b_i^2\left(\frac{1}{2}(1 + \lambda(1-\alpha))\right) + \lambda\alpha\,|b_i| - \hat{\beta}_i^{\text{OLS}} b_i\right] + \text{ const}\\
\frac{\partial f\left(b_i\right)}{\partial b_i} &= b_i(1 + \lambda(1-\alpha)) + \lambda\,\alpha\operatorname{sgn}\left(b_i\right) - \hat{\beta}_i^{\text{OLS}} = 0
\end{aligned}$$

When we analyse the above equation considering $sgn(\beta_i)$ we end with explicit formula

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

Number of discoveries

Recall that elastic net make discovery if $\left|\widehat{\beta}_i^{OLS}\right| > \lambda\alpha$. Then if

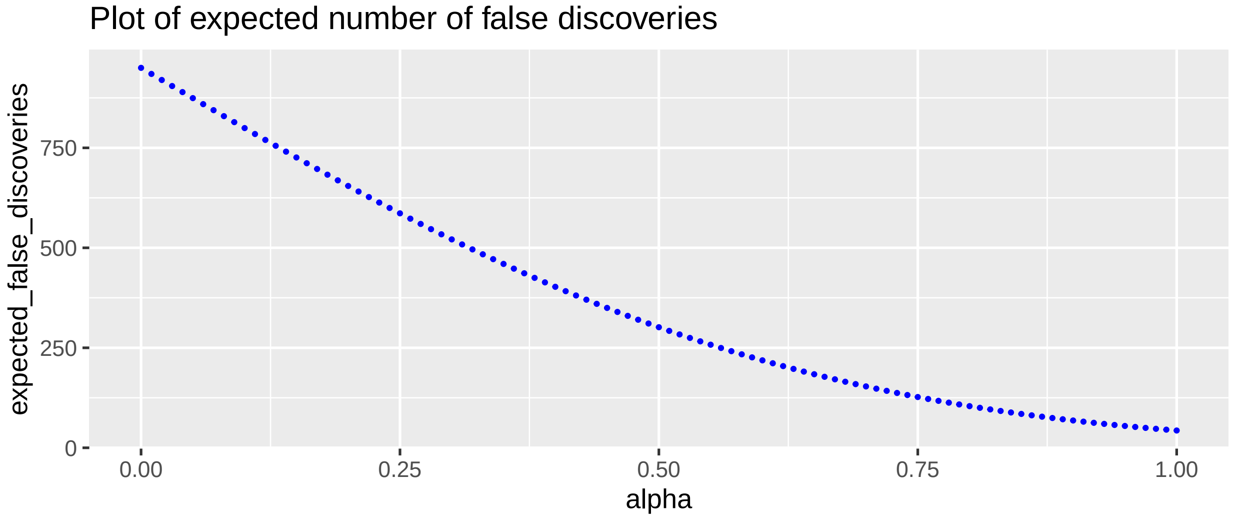
$$\begin{aligned}
\alpha = 0, \left|\widehat{\beta}_i^{OLS}\right| > 0 \equiv \text{ RIDGE },\\
\alpha = 1, \left|\widehat{\beta}_i^{OLS}\right| > \lambda = \text{ LASSO.}
\end{aligned}$$

In case when EN reduce to ridge we don't perform any selection (0 shrinkage only), but if $\alpha = 1$ we do. From this we conclude number of discoveries should decrease as α grows.

Now we can derived formula for single false discovery (similar way as in report 3.).

P (I type error) $= P\left(\left|\hat{\beta}_i^{OLS}\right| > \alpha\lambda|\beta_i = 0\right) = P\left(\frac{|\hat{\beta}_i^{OLS}|}{\sigma} > \frac{\alpha\lambda}{\sigma}|\beta_i = 0\right) = 2\left(1 - \Phi\left(\frac{\alpha\lambda}{\sigma}\right)\right).$

With square design matrix $n = 1000$ and if we assume $p_0 = 950$ and $\lambda = 2, \sigma^2 = 1$ the average number of false discoveries is $950 \cdot 2(1 - \Phi(2\alpha)$.



The above plot confirm previous conclusion. α grows, EN make less discoveries and intuitively less the false ones.

We can perform similar calculation under assumptions that $\beta_i \neq 0$. As OLS is unbiased and normally distributed we need to just subtract true β_i value. Probability of single correct decision is:

$$1 - \Phi\left(\frac{\lambda\alpha - \beta_i}{\sigma}\right) + \Phi\left(\frac{-\lambda\alpha - \beta_i}{\sigma}\right).$$

So for $\beta_i = 3$ we have $1 - \Phi(2\alpha - 3) + \Phi(-2\alpha - 3)$.

What would be the value of the elastic net estimator with $\lambda = 1$ and $\alpha = 0.5$ if $\hat{\beta}_{\text{OLS}} = 3$?

According to above formula we got $\hat{B}_{en} = \frac{1}{1+1(1-0.5)} max(3 - 0.5, 0) = \frac{2}{3} \cdot \frac{5}{2} = \frac{5}{3}$.

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

Because the LASSO, SLOPE and elastic net include L1 penalty when Ridge regression has additional penalty in terms of L2 norm. Ridge shrink some less important coefficient to 0 but not set them to 0. The other one do. This is because of the obtained estimator which perform some threshold selection.

Identifiability condition for LASSO

As we already know the LASSO selects the variables for which $|\widehat{\beta}_i^{OLS}| > \lambda$. The identifiability condition is criterium that says when LASSO select correct model (reject all null predictors, and do not reject any true predictor). I know there exist theorem that says the model is correctly identified for any $\lambda > 0$ only if

$\min_{i\in I}|\beta_i|$ is sufficiently large. The I represent indices for true predictors within data. I have not been able to get direct formulas based on β and statistical properties of X.

Irrepresentability conditon give use direct probability bound for correct model identification in terms of X, β and with respect to supremum norm. When

$$\left\|X_I' X_I \left(X_I' X_I\right)^{-1} S\left(\beta_l\right)\right\|_\infty > 1$$

then probability of correct model selection by LASSO is less than 0.5. Demanding to above term be lower or equal 1 is needed to correct model idendification by LASSO with high probability. Bogdan preprint 1.2

SLOPE

SLOPE which is "Sorted L-One Penalized Estimation" is regularization with new norm penalty.

$$\hat{\beta} = \text{argmin}_{b\in\mathbb{R}^p} \frac{1}{2}\|y - Xb\|_{\ell_2}^2 + \sum_{i=1}^{p} \lambda_i |b|_{(i)}$$

As we can see the main difference is the sorting coefficients in order to their magnitudes (in descending order). Also each coordinate has associated weight λ_i. The sequence $\{\lambda_i\}_1^p$ must satisfy

$$\lambda_1 \geq \ldots \geq \lambda_p \geq 0.$$

The interpretation is as follow. We can adjust penalty according to feature importance and we assume the importance is greater if coefficient is greater. Greater importance should also result in a greater penalty. If we set all $\lambda_i = c$ then we end up with ordinary LASSO.

Knockoff

The Knockoff method is designed to handle case when some variables is highly correlated with other within data and estimate coefficient might be large. Then Lasso or Slope might not detect is not important and shouldn't be use to predict response. In nutshell idea behind knockoff is as follow. We create copy of all features. Let say we have design matrix X and copy will be $\tilde{X}$. We want to preserve covariance between features but the covariance between exact copy of given feature should be as low as possible. Formally $\forall i \neq j \quad \text{cov}\left(X^i, \widetilde{X}^{(j)}\right) = \text{cov}\left(X^i, X^{(j)}\right)$.

Then we crate matrix $D = (X, \widetilde{X})_{nx2p}$. We using D as design matrix in LASSO model.

Then we need to deduce which feature are important on predicting response. We measure variable importance computing:

$$Z_j = \left|\hat{\beta}_j(\lambda)\right|, \quad \tilde{Z}_j = \left|\hat{\beta}_{j+p}(\lambda)\right|, \quad j = 1, \ldots, p$$

Note: Method description from Candes.

Then we compute

$$W_j = h\left(Z_j, \tilde{Z}_j\right) = -h\left(\tilde{Z}_j, Z_j\right), \quad j = 1, \ldots, p$$

where the h mus be anti-symmetric. h is called symmetrized knockoff statistics.

Finally, the knockoff procedure selects predictors with large and positive values of W_j, according to the adaptive threshold defined as

$$T = \min\left\{t : \frac{1 + \#\{j : W_j \leq -t\}}{\#\{j : W_j > t\}} \leq \alpha\right\},$$

where α is the (desired) target FDR level.

For further work we use implementation *knockoff* library developed by Stanford scientist.

Example of selection based on calculated W statistic

Assume we applied knockoff procedure and obtain vector $W = (8, -4, -2, 2, -1.2, -0.6, 10, 12, 1, 5, 6, 7)$. We want to control FDR on level $q = 0.4$. To find the proper threshold we can just evaluate expression $\frac{1+\#\{j:W_j(\lambda)\leq -t\}}{\#\{j:W_j(\lambda)\geq t\}}$ for $t = W'$ where W' is sorted absolute values of W.

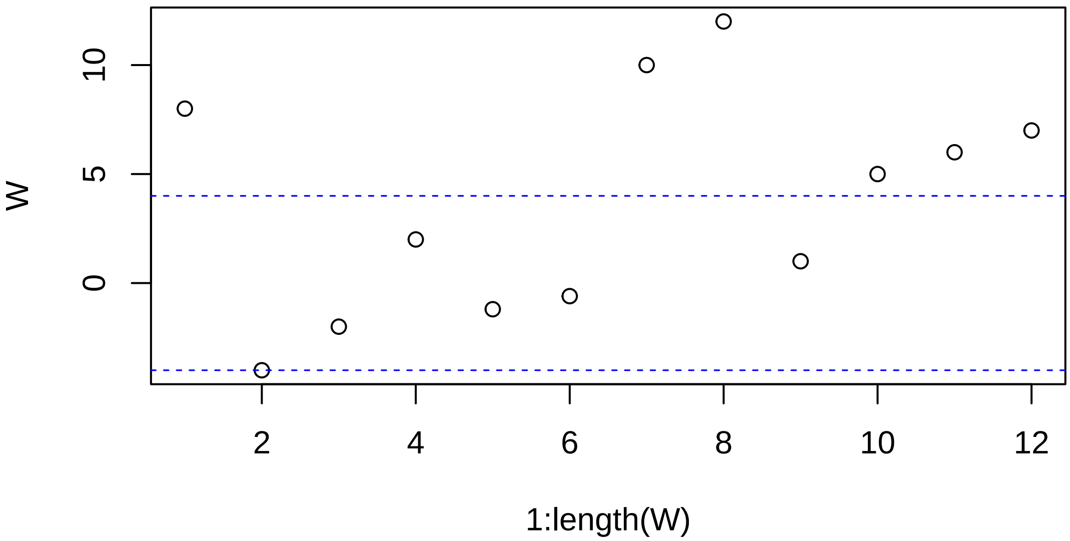


Figure 1: Threshold selection for Knockoff

Iterating over W' we find that setting $\hat{t}(\lambda) = 4$* give us proportion $\frac{1+1}{6} = \frac{1}{3} < 0.4$. Now we select variable only above (positive) threshold which is $\widehat{\mathcal{S}(\lambda)} = \left\{j : W_j(\lambda) \geq \hat{t}(\lambda)\right\} = \{1, 7, 8, 10, 11, 12\}$.

Note * The threshold should be lower by definition, but this simplify procedure and the results are equivalent.

Simulation part

In this simulation we generate response vector from model

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

where $\epsilon \sim 2\mathcal{N}(0, I), \beta_i = 10$ for $i \in \{1, \dots, k\}, \beta_i = 0$ for $i \in \{k+1, \dots, 450\}$, and $k \in \{5, 20, 50\}$. The design matrix $X_{500x450}$ is orthonormal.

We fit OLS, LASSO, Ridge and perform Knockoff procedure based on coefficients from LASSO and from Ridge (FDR=0.2). We check the following statistic $E|\beta - \hat{\beta}|_2^2$, $E|X(\beta - \hat{\beta})|_2^2$. Also we want to compare FDR for ordinary LASSO and both Knockoff version.

The all models use lambda.1se instead lambda.min.

Table 1: Estimated FDR and power based on 100 replication.

	Lasso		Knockoff with Lasso		Knockoff with Ridge	
	FDR	Power	FDR.1	Power.1	FDR.2	Power.2
5	0.13	1	0.14	1	0.20	1.00
20	0.47	1	0.18	1	0.22	1.00
50	0.57	1	0.19	1	0.19	0.94

Table 2: Estimated mean based on 100 replication.

	$E\|\beta - \hat{\beta}\|_2^2$			$E\|X(\beta - \hat{\beta})\|_2^2$		
	MSE_OLS	MSE_ridge	MSE_lasso	MSE_mean_OLS	MSE_mean_ridge	MSE_mean_lasso
5	780	167	10	1775	1476	234
20	779	430	24	1795	2048	485
50	806	837	53	1810	3092	770

Results

We can see that LASSO always have full power, but the FDR increase when we decrease the sparsity of data. For $k = 50$ over half of discoveries are false. Same time both version of Knockoff almost always correctly identify all true predictors. They control FDR on given $q = 0.2$. The small deviation for case with $k = 20$ in knockoff based on Ridge is acceptable from my point of view.