Assignment 4

Emre Beray Boztepe

07 06 2024

Exercises

1-) Assume the linear model:

Y = Xβ + ϵ

where $X'X = I$ and $e \sim N(0, \sigma^2 I)$ Find the numerical solution for the elastic net in the form:

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

a-) What would be the value of the elastic net estimator with λ = 1 and α =0.5 if $\hat{\beta}_{OLS} = 3$?

We can use this formula:

$$\hat{\beta}_{en} = \frac{\hat{\beta}_{OLS}}{1 + \lambda(1-\alpha)}$$

When we substitute with what we have in the question:

$\frac{3}{1+1(1-0.5)} = \frac{3}{1.5} = 2$

b-) How does the number of discoveries depend on the parameter α?

As α increases, lasso component increases (L1 penalty), potentially reducing the number of non-zero coefficients.

As α decreases, the ridge component (L2 penalty) becomes dominant, generally leading to more non-zero coefficients though they may be small in magnitude.

c-) Provide the numerical value for the expected number of false discoveries when n = p = 1000, p0 = 950, σ = 1, and λ = 2, and the power of detection of X1 when β1 = 3.

p - p0 = 1000 - 50 = 50, which is the number of non-null predictors.

Proportion of FD = $\frac{FD}{TD}$

- TD: Total discoveries which is ≥ 1

$$Power = P(RejectH_0|\beta_1 \neq 0)$$

- H_0: null hypothesis states that there is no effect of X_1

```
## Number of False Discoveries:  1

## Total Discoveries:  51

## Proportion of False Discoveries:  0.01960784

## Power of detection for X1:  1
```

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

LASSO uses an L1 penalty ($\lambda\sum|\beta_i|$) and this promotes sparsity in the coefficient estimates. When λ incsreases, L1 penalty can drive some coefficients exactly to 0.

SLOPE applies a sequence of penalties to the sorted absolute values of coefficients. These sequence of penalties often becomes increasingly stringent and may help control the false discovery rate in variable selection. It also can drive certain coefficients to zero (it depends on their size relative to their assigned penalty)

Elastic net, combines L1 and L2 penalties. L1 component of the penalty induces sparsity by enabling coefficients to shrink to zero. L2 component helps handle situations where there are highly correlated variables or more variables than observations

Ridge uses L2 penalty $\lambda\sum\beta_i^2$. L2 minimizes the square of coefficients but ensures that none of the coefficients can actually zero.

So, as a summary, methods incorporating an L1 component are capable of variable selection, making them useful in scenarios where feature selection is crucial.

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

The identifiability condition refers to the conditions under which the LASSO solution is guaranteed to correctly identify the true model (the correct set of non-zero coefficients)

Formula:

$$\left|X_{null}^T X_{signal}\left(X_{signal}^T X_{signal}\right)^{-1} sgn\left(\beta_{signal}\right)\right| \leq \alpha$$

- X_{signal}: columns of X corresponding to the non-zero coefficients.
- β_{signal}: non-zero coefficients.
- X_{null}: columns of X corresponding to the zero coefficients.
- $sgn\left(\beta_{signal}\right)$: sign vector of the non-zero coefficients.
- α: some constant which $\alpha \in (0,1]$

When the irrepresentable condition is met, LASSO is guaranteed to consistently select the correct model as the sample size goes to infinity. It can be said that it will correctly identify

all and only the relevant predictors as having non-zero coefficients. This guarantee is under the assumption of sufficient regularization and certain conditions on the design matrix X (ex. having more observations than predictors)

The identifiability in a more general sense refers to the unique determination of model parameters from the observed data. For LASSO, this goes beyond mere consistency and touches on the uniqueness of the solution and this is not always guaranteed like where the predictors are highly correlated.

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

SLOPE (Sorted L-One Penalized Estimation) is a regularization technique designed for linear regression that adapts and extends the ideas from LASSO. It aims to control FDR in variable selection and makes it particularly suitable for scenarios with high-dimensional data where multiple hypothesis testing becomes an issue.

$$\hat{\beta}_{SLOPE} = argmin_b \left(\frac{1}{2} |Y - Xb|_2^2 + \sum_{i=1}^{p} \lambda_i |\beta|_{(i)} \right)$$

- $|\beta|_{(i)}$: i-th largest absolute value of the coefficient vector β
- λ_i: non-negative constants sorted in non-increasing order ($\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$)

LASSO applies a uniform penalty ($\lambda \sum |\beta_i|$) across all coefficients, promoting overall sparsity by potentially shrinking all coefficients to zero. Thus, emphasizes regularization and simplicity. SLOPE assigns a sequence of penalties based on the rank of the absolute values of coefficients, targeting a structured sparsity by imposing stronger penalties on larger coefficients.

LASSO is well-documented in its consistency and sparsity (under conditions like the irrepresentable condition), SLOPE outclasses in controlling the FDR under a wider range of conditions and adapts to unknown sparsity patterns of coefficients. LASSO is favored for its efficiency through algorithms like coordinate descent, suitable for large datasets. SLOPE, due to its sorted penalty structure, requires more sophisticated algorithms to manage the complexity introduced by this ordering. IT makes it essential in applications where controlling false discoveries is critical.

5-) What are knockoffs?

Knockoffs method used for variable selection that controls the false discovery rate (FDR). This technique was developed to address the challenges of multiple hypothesis testing in scenarios where there are many potential predictors by creating fake or “knockoff” versions of each variable. These knockoff variables serve as a control group to test which of the original variables have a true association with the response, as opposed to correlations arising by chance.

Knockoff methods are particularly useful in settings where variables are highly correlated. They can be applied in conjunction with any feature selection method that provides

importance measures (LASSO, ridge etc.) and the method provides theoretical guarantees for FDR control under fairly general conditions.

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?

First, we solve W in descending order:

W_{sorted}: (12,10,8,7,6,5,2,1,−0.6,−1.2,−2,−4)

Then, we can use this formula:

$$T = min\{t \in W: \frac{1 + \#\{j: W_j \leq -t\}}{\#\{j: W_j \geq t\} \vee 1} \leq q\}$$

Simulation:

```
## Threshold T: 4

## Important variables: 8 10 12 5 6 7
```

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

Ridge can be interpreted within Bayesian frameworks where the prior distribution over the regression coefficients is assumed to be multivariate normal.

Ridge Regression:

$$\hat{\beta}_{ridge} = argmin_{\beta}\left(\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 + \lambda|\beta|_2^2\right)$$

- y: observed outputs
- x_i: feature vectors
- β: coefficients to be estimated
- λ: regularization parameter

Bayesian Interpretation:

Assuming a linear model with Gaussian noise

$$p(y|X,\beta) \propto exp\left(-\frac{1}{2\sigma^2}|y - X\beta|^2\right)$$

- the likelihood of observing y given X

- X: design matrix
- β: coefficient vector
- σ^2: variance of the Gaussian noise

In the Bayesian context, ridge regression assumes a multivariate normal prior on the regression coefficients β

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

- τ^2: variance of the prior distribution (smaller τ^2 leads to greater shrinkage)

According to Bayes' rule, the posterior distribution of β given the observed data y and X is proportional to the product of the likelihood

$$p(\beta|y,X) \propto p(y|X,\beta)p(\beta)$$

Substituting the expressions for the likelihood and the prior

$$p(\beta|y,X) \propto exp\left(-\frac{1}{2\sigma^2}|y-X\beta|^2 - -\frac{1}{2\tau^2}|\beta|^2\right)$$

To find the MAP estimate, we can minimize the negative log of the posterior:

$$\hat{\beta}_{MAP} = argmin_\beta\left(\frac{1}{2\sigma^2}|y-X\beta|^2 + \frac{1}{2\tau^2}|\beta|^2\right)$$

When we set $\lambda = \frac{\sigma^2}{\tau^2}$:

$$\hat{\beta}_{ridge} = argmin_\beta(|y-X\beta|^2 + \lambda|\beta|^2)$$

So, we show that ridge regression is equivalent to the MAP estimation under a Bayesian framework where the prior distribution over the coefficients is a multivariate normal distribution.

Computer project

Generate the design matrix $X_{500\times450}$ such that its elements are independent and identically distributed (iid) random variables from $N\left(0, \sigma = \sqrt{\frac{1}{n}}\right)$. Then generate the vector of the response variable according to the model:

Y = Xβ + ϵ

where $\epsilon \sim 2N(0,I)$, $\beta_i = 10$ 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 $\mu = X\beta$.

	Lasso		**Knockoff with Lasso**		**Knockoff with Ridge**	
	FDR	**Power**	**FDR.1**	**Power.1**	**FDR.2**	**Power.2**
5	0.90	1	0.17	1	0.14	1.00
20	0.77	1	0.24	1	0.28	1.00
50	0.67	1	0.20	1	0.16	0.94

	MSE			**MSE MEAN**		
	MSE_OLS	**MSE_ridge**	**MSE_lasso**	**MSE_mean_OLS**	**MSE_mean_ridge**	**MSE_mean_lasso**
5	750	161	4.9	1743	1216	102
20	846	374	19.2	1824	2008	355
50	1058	760	40.2	1822	2673	655

Comments:

Lasso:

FDR: The False Discovery Rate is quite high across all scenarios, ranging from 0.90 to 0.67 as k increases. This indicates that Lasso, while powerful, tends to include a significant number of false positives, especially when fewer predictors are truly non-null (k=5).

Power: Remains constant at 1 across all k values, suggesting that Lasso is consistently effective at identifying all true non-null predictors.

Knockoff with Lasso:

FDR: Shows a lower FDR compared to traditional Lasso, ranging from 0.17 to 0.24. This substantial reduction underscores the effectiveness of the Knockoff method in controlling false discoveries.

Power: Maintains a power of 1 across all settings, indicating that the Knockoff method does not sacrifice the ability to detect true effects when improving control over false discoveries.

Knockoff with Ridge:

FDR: Exhibits a relatively low FDR, although generally higher than Knockoff with Lasso. It varies from 0.14 to 0.28, increasing with the number of non-null predictors (k).

Power: The power is very high, at or near 1 for all k settings, with a slight decrease to 0.94 at k = 50. This indicates strong effectiveness in identifying true predictors, though there's a slight reduction in power as the complexity (number of true predictors) increases.

While traditional Lasso is very powerful, its high FDR can be problematic in settings where the cost of false positives is significant. In contrast, Knockoff techniques, especially with Lasso, provide much better control over false discoveries without compromising on the ability to detect true effects.

Both Knockoff with Lasso and Knockoff with Ridge demonstrate their effectiveness in managing the FDR more stringently than Lasso alone. The choice between using Knockoff with Lasso or Ridge should consider other factors such as the model assumptions, the nature of the data, and specific analytical needs.

Means

OLS:

MSE increases with the number of non-null predictors, suggesting that as the complexity of the model increases, OLS's performance degrades, likely due to overfitting or inability to manage higher dimensional data effectively. The mean MSE significantly increases with k, highlighting potential variability and stability issues in higher-dimensional settings.

Ridge:

Shows a less steep increase in MSE compared to OLS as k increases, indicating better handling of more complex models, likely due to its regularization nature which helps manage multicollinearity and overfitting. The mean MSE also increases with k, but the pattern suggests that while Ridge manages complexity better than OLS, it still struggles as the number of true predictors grows.

Lasso:

Exhibits the lowest MSE across all k values, significantly outperforming OLS and Ridge. This demonstrates Lasso's effectiveness in not only handling sparsity but also in feature selection, effectively zeroing out many irrelevant features. The mean MSE remains much lower than the other two methods across all k values, suggesting that Lasso maintains good performance and stability even as the complexity of the model increases.