Regularization and knockoffs

Statistical Learning

Paulina Podgórska

2023-06-16

Ridge regression

Ridge regression is a method of estimating coefficients of multiple-regression models. It is helpful in a situation where data is highly correlated. The ridge regression estimator is given by a formula:

$$\hat{\beta} = argmin_{\beta \in \mathbb{R}^p} \left(||Y - Xb||^2 + \lambda ||b||^2 \right) = (X'X + \lambda I)^{-1} X'Y,$$

where $\lambda > 0$.

Orthonormal design

We will consider a orthonormal design $X'X = I$ and regression model $Y = X\beta + \epsilon$, where $\epsilon \sim N(0, I_{n \times n})$ and the vector of coefficients $\beta_1 = ... = \beta_k = 3.5$ and $\beta_{k+1} = ... = \beta_{950} = 0$ with:

1. $k = 20$,
2. $k = 100$,
3. $k = 200$.

Tuning parameter λ

Our goal is to find a value of tuning parameter λ for ridge regression, so as to minimize the mean square error of the estimator of β:

$$E||\hat{\beta} - \beta||^2 = f(\lambda) = \frac{\lambda^2}{(1+\lambda)^2} ||\beta||^2 + \frac{p\sigma^2}{(1+\lambda)^2}$$

To find the value of λ, the first derivative of the function must be zero.

$$f'(\lambda) = 0 \Leftrightarrow \frac{2(\lambda ||\beta||^2 - p\sigma^2)}{(1+\lambda)^3} = 0$$

from which we can conclude that the value we are looking for is

$$\lambda_0 = \frac{p\sigma^2}{||\beta||^2}.$$

Table 1: Value of λ for each model which minimizes MSE

λ_1	λ_2	λ_3
3.877551	0.7755102	0.3877551

$$MSE = \frac{||\beta||^2\sigma^2 p}{||\beta||^2 + \sigma^2 p}$$

Now, we will calculate the bias and the variance of this optimal estimator.The formulas are as follows.

Bias

$E(\hat{\beta}_i) - \beta_i = E\left((X'X + \lambda I)^{-1}X'Y\right) - \beta_i = E\left(\frac{X'Y}{1+\lambda}\right) - \beta_i = E\left(\frac{\beta_i + X'\epsilon}{1+\lambda}\right) - \beta_i = \frac{\beta_i}{1+\lambda} - \beta_i = -\beta_i\frac{\lambda}{1+\lambda} = \left(\frac{-\sigma^2 p}{||\beta||^2\sigma^2 p}\right)\beta_i$

Variance

$Var(\hat{\beta}_i) = E(\hat{\beta}_i^2) - E(\hat{\beta}_i) = E\left(\left(\frac{\beta_i + X\epsilon}{1+\lambda}\right)^2\right) - \frac{-\beta_i^2}{(1+\lambda)^2} = E\left(\frac{\beta_i^2 + 2\beta_i X'\epsilon + (X'\epsilon)^2}{(1+\lambda)^2}\right) = \frac{\sigma^2}{(1+\lambda)^2} = \frac{\sigma^2||\beta||^4}{(||\beta||^2 + p\sigma^2)^2}$

Table 2: Theoretical values of parameters for our 3 models

	Bias (for $\beta_i = 3.5$)	Variance	MSE
k=20	-2.78	0.04	194.77
k=100	-1.53	0.32	535.06
k=200	-0.98	0.52	684.56

Empirical results

Our next step is to generate 200 replicates of the above model and analyze the data using ridge regression and OLS. We will compare empirical values of above parameters with the theoretical ones as well as with the corresponding parameters of OLS.

Table 3: Empirical results for 3 models

k	var_{RR}	var_{LS}	MSE_{RR}	MSE_{LS}	$bias_{RR}$ (3.5)	$bias_{LS}$ (3.5)	$bias_{RR}$ (0)	$bias_{LS}$ (0)
20	0.201	0.999	194.332	949.576	-2.684	-0.098	0.041	0.027
100	0.565	0.998	553.208	948.117	-1.189	-0.048	-0.254	0.144
200	0.740	1.000	720.180	949.748	-0.525	-0.039	-0.094	0.107

The variance and the bias of ridge regression estimators are very close to the theoretical ones. We observe the biggest difference between the values of variance – our results obtained from generating data are quite higher than the ones calculated by hand. The reason for this might be the influence of noise. Comparing the values of bias for OLS estimators and RR estimators we can draw a conclusion that OLS method tends to overestimate the true value of β whereas RR does the opposite – the estimators have generally lower values than the true betas. We can also observe that number of zero elements in the vector of regression coefficients does not influence the MSE of OLS – in all cases it's higher than RR.

MSE of different estimation methods

We will generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $N(0, \sigma = \frac{1}{\sqrt{n}})$. Then we will generate the vector of the response variable according to the models proposed in previous section.

We will estimate the parameters of those models using:

1. The ridge regression with the tuning parameter λ selected by minimizing the prediction error. We will use the following formula:

$$PE(\lambda) = RSS(\lambda) + 2\sigma^2 Tr(M),$$

   where $M = X(X'X + \lambda I)^{-1}X'$.

2. LASSO with the tuning parameter λ selected by minimizing PE:

$$PE(\lambda) = RSS(\lambda) + 2\sigma^2 k,$$

   where k is the number of variables selected by LASSO.

3. The ridge regression with λ selected by 10 fold CV,
4. LASSO with λ selected by 10 fold CV,
5. OLS,
6. OLS within the model selected by mBIC2,
7. OLS within the model selected by AIC.

After repeating the above experiment 100 times we received below results for MSE of β and $\mu = X\beta$.

Table 4: $||\hat{\beta} - \beta||^2$ for our 7 approaches

	RR_{PE}	$LASSO_{PE}$	RR_{CV}	$LASSO_{CV}$	OLS	OLS_{mBIC2}	OLS_{AIC}
k=20	203.04	112.45	201.13	106.96	18461.11	204.47	203.80
k=100	674.00	455.31	707.13	457.62	19020.38	1217.08	447.78
k=200	1050.26	891.00	1271.14	890.68	18163.53	2489.58	1740.68

Table 5: $||\hat{X(\beta} - \beta)||^2$ for our 7 approaches

	RR_{PE}	$LASSO_{PE}$	RR_{CV}	$LASSO_{CV}$	OLS	OLS_{mBIC2}	OLS_{AIC}
k=20	173.74	105.70	171.78	99.81	938.85	196.01	187.48
k=100	398.48	305.13	426.85	306.21	956.02	1092.87	366.24
k=200	518.29	476.67	704.72	477.33	947.44	2312.28	1373.89

First obvious conclusion is that in cases where there is a lot of coefficients with value of 0, OLS performs significantly worse. Mean square errors are even ten times higher than some other cases. We can also observe a trend where the MSE is higher when value of k is greater. Overall the most satisfying results we achieved using LASSO – 2nd and 4th option. In this case it is intuitive - LASSO produces a sparse model where only a portion of the predictors have non-zero coefficients.

LASSO irrepresentability and identifiability condition

In this section, let's consider the design matrix $X_{100\times 200}$ such that its elements are iid random vectors from $\frac{1}{n}N(0,\Sigma)$, where $\Sigma_{ii}=1$ and for $i\neq j$ $\Sigma_{ij}=0.7$. The vector of regression coefficients is generated in a following way: $\beta_1,...\beta_k=20$ and $\beta_{k+1}=...=\beta_{200}=0$.

Irrepresentability condition

$$||X_{\bar{I}}'X_I(X_I'X_I)^{-1}S(\beta_1)||_\infty \leq 1,$$

where $I=\{i\in\{1,...p\}|\beta_i\neq 0\}$, $X_I=(X_i)_{i\in I}, X_{\bar{I}}=(X_i)_{i\notin I}$ and $S(\beta)$ represents the sign vector of β.

The maximal k for which this condition is satisfied is: $k_{IR}=2$. The obtained value for IR is: 0.9906526. We generated the response variable according to formula

$$Y=X\beta^{k_{IR}}$$

and empirically found such λ that LASSO could recover the sign of β: $\lambda=0.09999052$.

Identifiability condition

The vector β is said to be identifiable with respect to the l_1 norm if the following implication holds

$$X\gamma=X\beta,$$

$$\gamma\neq\beta\Rightarrow||\gamma||_1>||\beta||_1.$$

Maximal k for which the LASSO identifiability condition is satisfied: $k_{ID}=54$. Next, we generated the response variable according to the formula

$$Y=X\beta^{k_{IR}}$$

and tried to find λ such that LASSO can recover the sign of β. The value we achieved is: $\lambda=0.7956665$.

Last step is to generate the response variable according to the formula

$$Y=100X\beta^{k_{ID}+1}$$

and check if there exists λ which allows for separating zero and nonzero elements of β. After generating data again, we got $k_{ID}=60$. The closest value of λ which allows for separating zero an nonzero elements of β we got is 914.7462.

realdata.Rdata

Once again, we will use *realdata* data set to test our prediction methods. Our data set contains the expression levels of 3221 genes for 210 individuals, which we have split into a test and a train set. The test set comprises 30 randomly selected individuals. We then construct models using the ridge regression and LASSO. We select the parameter λ with help of cross-validation. Let's compare the quality of our new models to the model selection criteria from the previous assignment.

Table 6: Number of variables selected by criteria

	LASSO	RR	AIC	BIC	RIC	mBIC	mBIC2
Number of variables	4	3220	193	194	7	6	8

Now let's test the accuracy of our models predictions on the test set using square error.

Table 7: SE

	LASSO	RR	AIC	BIC	RIC	mBIC	mBIC2
SE	40.17365	111.297	31.7597	1.8156	0.0638	0.1101	0.0082

The ridge regression selects all variables. It also got the highest square error among all models. On the other hand, comparing to other models, LASSO behaves the most strictly – it selected only 4 variables. Our methods did not achieve the best results, but we have to keep in mind that the values achieved by criterions from previous assignment were calculated on a different training data, because the division is random.

Now we will take a different approach. We will preselect 300 interesting explanatory variables with the largest marginal correlation with the response variable. Next, we add variables selected with mBIC2. In our case, mBIC2 criterion added 2 more variables. Then, we construct models using RR and LASSO.

Table 8: SE for reduced models

RR	LASSO
3.700647	5.612579

Table 9: Number of variables selected by models on reduced data

RR	LASSO
301	2

k_{ID} with a noisy response

We will consider the same setup as in two sections before but this time we will work on

$$Y = X\beta^{k_{ID}} + \epsilon,$$

where $\epsilon \sim N(0, I)$.

Table 10: Values of parameters for λ for which LASSO MSE is minimal

λ	TD	FD	MSE
0.079822	31	38	142.004

Adaptive LASSO

Now we will run adaptive LASSO with weights $w_i = \frac{1}{|\beta_L|+0.000001}$. Again, we will select the parameter λ so as MSE is minimal.

Table 11: Values of parameters for ad LASSO

λ	TD	FD	MSE
0.079823	31	38	142.0043

Values are the same.

Modification

This time we generate a response

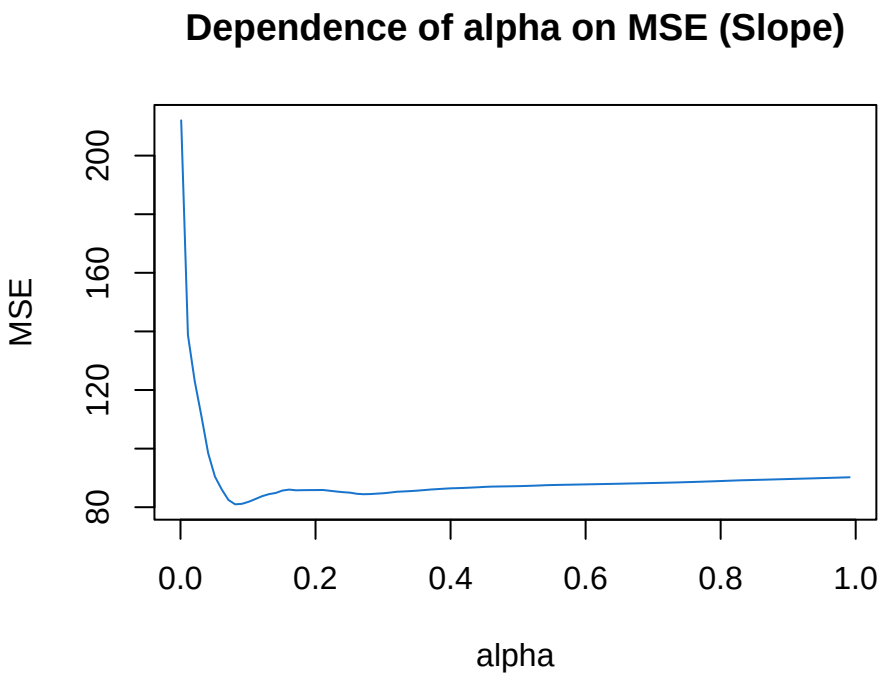
$$Y = X\beta^{k_{ID}} + \epsilon,$$

where $\epsilon \in N(0, I)$.

Table 12: MSE values for LASSO and adaptive LASSO

λ_{LASSO}	MSE_{LASSO}	$\lambda_{adLASSO}$	$MSE_{adLASSO}$
0.001024	210.917	989.673	202.817

SLOPE



We achieve minimal MSE = 80.983 for $\alpha = 0.081$. SLOPE method seems to be the most optimal one. It performs significantly better than LASSO or adLASSO.

Knockoffs

In the last section of the report we will generate the design matrix $X_{100\times200}$ such that its elements are iid random vectors from $\frac{1}{n}N(0, \Sigma)$, where $\Sigma_{ii} = 1$ and for $i \neq j$ $\Sigma_{ij} = 0.7$. The vector or the response variable is generated according to the model

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

where $\epsilon \sim N(0, I)$, $\beta_1 = ... = \beta_k = 30$ and $\beta_{k+1} = ... = \beta_{200} = 0$, $k \in \{5, 20\}$. We will generate 200 replicates of the above model and analyze the data using knockoffs (point a) and multiple knockoffs with RR and LASSO (point b).

Lets compare the power of our four methods.

Table 13: Power of our 4 methods

RR_a	$LASSO_a$	RR_b	$LASSO_b$
0.884	0.208	0.912	0.228

Multiple knockoffs with the ridge regression performs the best with power of 91.2%.