Statistical Learning

Dawid Dieu

Assignment 3

Ridge Regression and LASSO

tags: **SL**

Task 1

1. Generate orthonormal ($X^TX = I$) matrix od dimension 1000×950. Consider the regression model

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

with $\epsilon \sim N(0, I_{n\times n})$ and the vector of regression coefficients $\beta_1 = \ldots = \beta_k = 3.5$ and $\beta_{k+1} = \ldots = \beta_{950} = 0$ with

   a) k = 20,

   b) k = 100,

   c) k = 200.

   For each of these cases

   i) To be done by hand: Calculate the value of the tuning parameter λ for the ridge regression, so as to minimize the mean square error of the estimation of β.

   ii) To be done by hand: Calculate the bias, the variance and the mean squared error of this optimal estimator.

   iii) Generate 200 replicates of the above model and analyze the data using ridge regression and OLS. Compare empirical bias, variance, mse of the ridge regression with the theoretical values of these parameters, calculated above, and with the corresponding parameters of OLS.

First for each `k` we generated all needed parameters.

k	lambda	Bias for nonzero Beta	Variance	MSE
20	3.877551	-2.7824268	0.04203358	194.7699
100	0.7755102	-1.5287356	0.31721496	535.0575
200	0.387755	-0.9779412	0.5192474	684.5588

Now for each k we do 200 steps of simulation to compare theoretical values with the empirical ones.

- MSE

k	theoretical	ridge estimator	OLS estimator
20	194.7699	194.7489	946.549
100	535.0575	536.115	946.549
200	684.5588	684.5665	946.549

- Variances

k	theoretical	ridge estimator	OLS estimator
20	-2.7824268	-2.830794	-0.2359125
100	-1.5287356	-1.486266	0.07540591
200	-0.9779412	-0.9516881	0.03643289

- Biases for zero Beta:

k	theoretical	ridge estimator	OLS estimator
20	0	-0.001011	-0.001297
100	0	-0.001524	-0.002243
200	0	-0.002389	-0.002731

- Biases for nonzero Beta:

k	theoretical	ridge estimator	OLS estimator
20	-2.7824268	-2.830794	-0.2359125
100	-1.5287356	-1.486266	0.07540591
200	-0.9779412	-0.9516881	0.03643289

As we can see, the Ridge estimator performed much better in all cases; resuls are very close to the theoretical values.

Task 2

2. Generate the design matrix $X_{1000\times 950}$ such that its elements are iid random variables from $N(0, \sigma = 1/\sqrt{n})$. Then generate the vector of the response variable according to the models proposed in Task 1, above.

   Estimate the parameters of this model using the ridge regression, LASSO and elastic-net with $\alpha = 0.5$ and the tuning parameter λ selected by

   a) minimizing the SURE criterion
   b) 10 fold cross-validation
      and with

   c) OLS
   d) OLS within the model selected by mBIC2 and AIC.

   Compare the estimation errors $||\hat{\beta} - \beta||^2$ and $||X(\hat{\beta} - \beta)||^2$ for these 8 approaches.

   Repeat the above experiment 100 times and compare the mean square errors of estimation of β and $\mu = X\beta$ for the above approaches.

After estimating the parameters we will validate those models using two metrics.
\[SSE = ||\hat{\beta} - \beta ||^2\]
\[XSSE = ||X(\hat{\beta} - \beta) ||^2\]

- k = 20

estimation method	SSE	XSSE
Ridge	200.114	168.1079
Ridge with CV	199.8085	168.1367
LASSO	245	256.5255
LASSO with CV	94.53028	93.03195
ElasticNet	245	256.5255
ElasticNet with CV	150.249	141.109
OLS	19571.29	902.6226
OLS with mBIC2	201.0817	211.8074
OLS with AIC	189.1572	176.279

- k = 100

estimation method	SSE	XSSE
Ridge	683.3758	394.5984
Ridge with CV	668.127	377.3932
LASSO	1225	1219.411
LASSO with CV	382.2562	277.7989
ElasticNet	1225	1219.411
ElasticNet with CV	490.2368	316.9383
OLS	19707.03	941.783
OLS with mBIC2	1230.186	1099.447
OLS with AIC	459.4462	378.2465

- k = 200

estimation method	SSE	XSSE
Ridge	1127.882	523.0009
Ridge with CV	1073.185	497.7537
LASSO	2450	2455.473
LASSO with CV	926.991	446.5228
ElasticNet	2450	2455.473
ElasticNet with CV	928.0622	454.1362
OLS	15594.33	887.1635
OLS with mBIC2	2472.888	2304.903
OLS with AIC	1751.143	1348.727

Now we want to repeat the above experiment 100 times and compare the mean square errors.

- k = 20

estimation method	SSE	XSSE
Ridge	201.1945	172.402
Ridge with CV	201.4302	173.0229
LASSO	245	256.5255
LASSO with CV	107.1974	99.333
ElasticNet	245	256.5255
ElasticNet with CV	146.2316	132.0562
OLS	16899.99	937.6378
OLS with mBIC2	189.8139	185.9191
OLS with AIC	215.5512	190.9234

- k = 100

estimation method	SSE	XSSE
Ridge	747.0916	436.3712
Ridge with CV	724.087	420.0539
LASSO	1225	1219.411
LASSO with CV	493.806	327.5307
ElasticNet	1225	1219.411
ElasticNet with CV	579.9695	363.8726
OLS	16927.76	964.7021
OLS with mBIC2	1208.939	1129.874
OLS with AIC	445.8001	372.164

- k = 200

estimation method	SSE	XSSE
Ridge	747.0916	436.3712
Ridge with CV	724.087	420.0539
LASSO	1225	1219.411
LASSO with CV	493.806	327.5307
ElasticNet	1225	1219.411
ElasticNet with CV	579.9695	363.8726
OLS	16927.76	964.7021
OLS with mBIC2	1208.939	1129.874
OLS with AIC	445.8001	372.164

As we can observe, the more significant the `k` is, the bigger the errors are. It's worth noting that CV is almost always beneficial, and the errors are much more minor in many cases.
As one could predict, the simple OLS performed the worst in all cases.

Task 3

In this task we repeat the calculations from above but with:

$$\beta_1 = ... = \beta_k = 5$$

For each `k` we simulate 100 experiments.

- k = 20

estimation method	SSE	XSSE
Ridge	370.08	275.3401
Ridge with CV	366.8801	274.3001
LASSO	500	523.0645
LASSO with CV	134.034	124.383
ElasticNet	500	523.0645
ElasticNet with CV	235.2017	193.7925
OLS	22904.16	997.6643
OLS with mBIC2	168.9514	152.9448
OLS with AIC	218.686	194.3768

- k = 100

estimation method	SSE	XSSE
Ridge	1146.44	570.9534
Ridge with CV	1088.325	540.7069
LASSO	2500	2643.077
LASSO with CV	476.711	332.7864
ElasticNet	2500	2643.077
ElasticNet with CV	735.9433	438.7563
OLS	17959.47	958.7895
OLS with mBIC2	2223.85	1981.496
OLS with AIC	869.2478	741.2573

- k = 200

estimation method	SSE	XSSE
Ridge	1908.856	654.3468
Ridge with CV	1727.1	617.5927
LASSO	5000	5213.801
LASSO with CV	1039.317	490.1576
ElasticNet	5000	5213.801
ElasticNet with CV	1322.157	554.8434
OLS	16562.42	916.5706
OLS with mBIC2	4997.1	4664.591
OLS with AIC	3429.092	2851.967

In this task, we use a stronger signal. Both loss functions return now bigger values. CV still significantly improves the quality of the model.

Task 4

4. Repeat 2 and 3 when rows of X are iid random vectors from $\frac{1}{n}N(0,\Sigma)$, where $\Sigma_{ii}=1$ and for $i \neq j$ $\Sigma_{ij}=0.5$.

We run 100 experiments again.
First we use \(\beta=3.5\) from task 2.

- k = 20

estimation method	SSE	XSSE
Ridge	445.29536	227.42275
Ridge with CV	120.5263732	73.741772
LASSO	359.2206668	203.753585
LASSO with CV	141.49714	65.43056
ElasticNet	285.119773	163.544058
ElasticNet with CV	137.327924	64.876841
OLS	38010.854	952.909033
OLS with mBIC2	248.7183	113.76565
OLS with AIC	432.326	188.87178

- k = 100

estimation method	**SSE**	**XSSE**
Ridge	950.1566814	363.6333
Ridge with CV	725.7073	273.65686
LASSO	1587.86784	790.13938
LASSO with CV	735.402253	267.3908
ElasticNet	1412.60039	720.2052
ElasticNet with CV	694.492714	252.5433
OLS	37879.9928	960.13311
OLS with mBIC2	1345.27248	612.83898
OLS with AIC	503.801647	210.318714

- k = 200

estimation method	SSE	XSSE
Ridge	1433.05391	433.2023
Ridge with CV	1286.6696	407.367759
LASSO	2938.09128	1465.47057
LASSO with CV	1477.00699	463.400821
ElasticNet	2719.92623	1359.0562
ElasticNet with CV	1363.0692	423.77349
OLS	38647.8999	958.5379
OLS with mBIC2	2722.70378	1229.95198
OLS with AIC	1805.4559	725.924591

Now we use \(\beta=5\) from task 3.

- k = 20

estimation method	SSE	XSSE
Ridge	645.5601409	297.4956066
Ridge with CV	128.4152988	66.35780845
LASSO	357.1922033	197.1430232
LASSO with CV	146.1026887	80.43129619
ElasticNet	287.0214765	170.4786317
ElasticNet with CV	141.2333257	71.92879677
OLS	37765.694	943.8831959
OLS with mBIC2	250.7961621	123.4175934
OLS with AIC	430.1188163	199.5527017

- k = 100

estimation method	SSE	XSSE
Ridge	1463.051278	449.4469049
Ridge with CV	710.6091703	265.6142953
LASSO	1584.91296	815.8979346
LASSO with CV	718.8060363	275.7655605
ElasticNet	1417.035613	736.4243208
ElasticNet with CV	681.5163971	251.9197325
OLS	38789.81018	953.3254101
OLS with mBIC2	1335.275561	599.6270596
OLS with AIC	496.8045085	214.7270858

- k = 200

estimation method	SSE	XSSE
Ridge	2121.795055	537.8004469
Ridge with CV	1297.801884	400.4248966
LASSO	2938.874996	1494.589493
LASSO with CV	1494.806636	455.5511071
ElasticNet	2721.218099	1390.477981
ElasticNet with CV	1358.10604	409.733123
OLS	39751.83141	952.0101199
OLS with mBIC2	2732.804951	1225.377811
OLS with AIC	1817.030296	736.4988962

Not much changed in those experiments. CV was still the best way to train a model. Plain OLS was significantly worst than other methods.

Task 5

5. Generate the design matrix $X_{100\times 200}$ such that its elements are iid random variables from $N(0, \sigma = 0.1)$. Now, consider the vector $\beta^k \in R^{200}$, such that $\beta_1 = \ldots = \beta_k = 20$ and $\beta_{k+1} = \ldots = \beta_{200} = 0$.

- Find the maximal k for which the LASSO irrepresentability condition is satisfied and call it k_{IR}. Then generate the response variable according to the formula

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

  where $\epsilon \sim N(0, I)$ and empirically find the minimal λ such that LASSO can recover the sign of β. If this turns out not to be possible, increase the magnitude of the nonzero elements of β.

- Find the maximal k for which the LASSO identifiability condition is satisfied and call it k_{ID}. Then generate the response variable according to the formula

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

  where $\epsilon \sim N(0, I)$ and empirically find the minimal λ such that LASSO can properly separate zero and nonzero elements of β. If this turns out not to be possible, increase the magnitude of the nonzero elements of β.

- Generate the response variable according to the formula

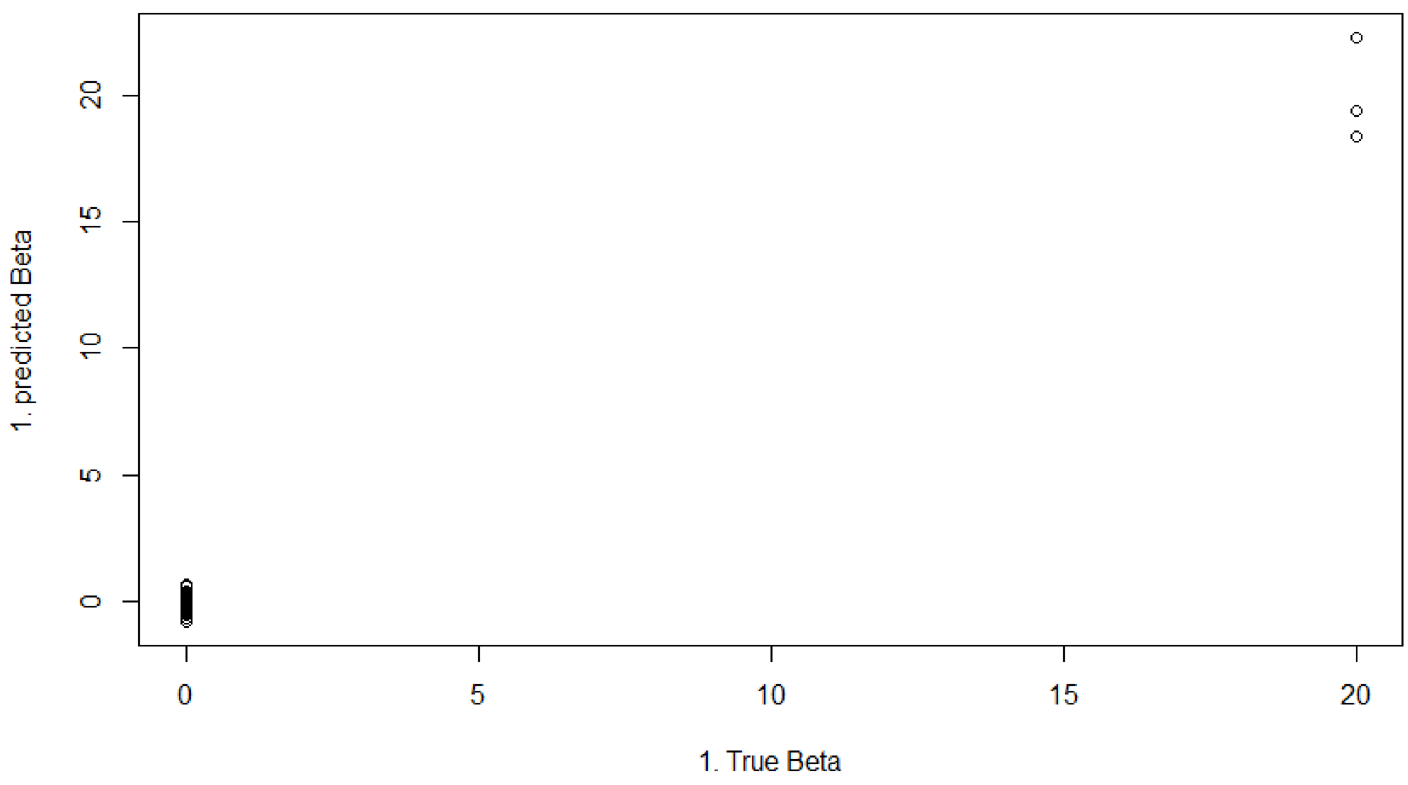
$$Y = 100 * X\beta^{k_{ID}+1} + \epsilon \ ,$$

  where $\epsilon \sim N(0, I)$ and empirically verify that there does not exist λ which allows for separating zero and nonzero elements of β.

First we generate the response variable according to the formula above.

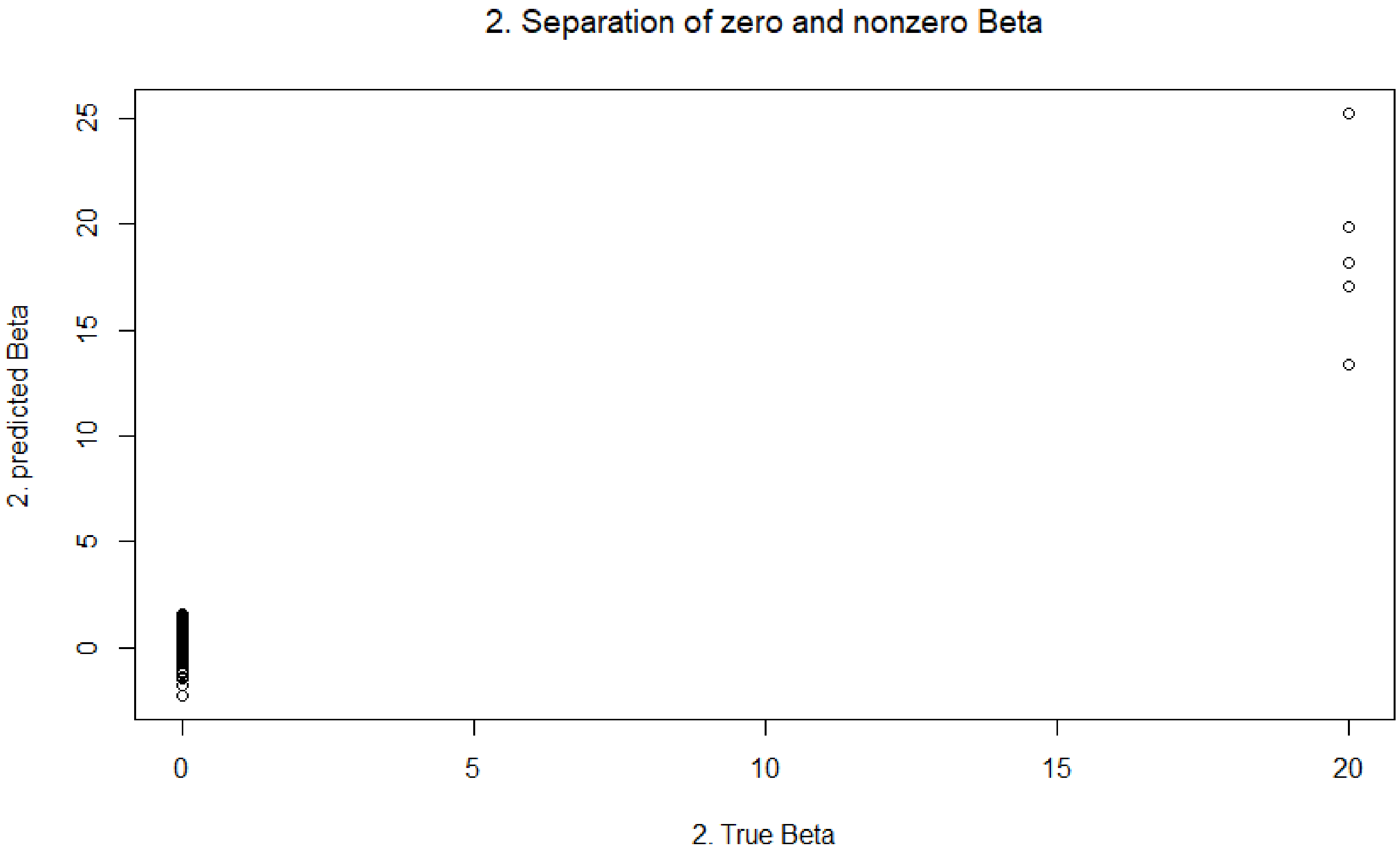
Maximal k for which the LASSO irrepresentability condition was satisfied was around \(k_{IR} = 5\). After generating data this way, when trying to empirically find the minimal \(\lambda\) such that LASSO can recover the sign of \(\beta\) we found \(\lambda = 5 * 10^{-10}\)

1. Separation of zero and nonzero Beta



In the second case maximal k for which the LASSO irrepresentability condition was satisfied was around $k_{ID} = 30$.
After generating data this way, when trying to empirically find the minimal λ such that

LASSO can recover the sign of \(\beta\) we found \(\lambda = 0.0004\)



Now we want to generate response variable according to the third formula.
We test \(\lambda\) in range from \(10^{-5}\) to \(50\).
For each lambda we fit LASSO model and check for non-zero betas.
We count how many of them are in range from -1 to 1 and how many of them are greater than 10.
This experiment resulted with founding that for around 80% of lambdas model estimates betas greater than 10.

For \(\lambda = 5 * 10^{-5}\) all the non zero betas are recovered. As one can see below on the plot for this \(\lambda\) it doesn’t allow for separating non zero betas from zero betas.

Task 6

6. For this problem use the set realdata.Rdata from List 2 and the same split of the data into the training and the test set as the one you used for the previous assignment.

   a) Use the training set (180 individuals) to construct the regression model explaining the expression level of gene 1 (first column in the data set) as the function of expression levels of other genes. Use Ridge regression, LASSO and elastic net with $\alpha = 0$ and apply crossvalidation to select the tuning parameter (verify that cv.glmnet indeed identifies the minimum of the prediction error). Use the test set to verify the predictive accuracy of considered models. Compare to the predictive performance of model selection criteria from the previous assignment. Compare also the number of variables selected by different methods.

   b) Preselect interesting explanatory variables. Select 300 variables with the largest marginal correlation with the response variable and add variables selected by mBIC2. Then apply regularization methods (ridge, LASSO and elastic net) to build a predictive model on such reduced set of variables. Use the test set to verify the predictive performance of the obtained models and compare to the predictive properties of models obtained in earlier experiments.

a)

We randomly select 30 individuals for the test. The remaining 180 will be used as training samples. We will compare three models, Ridge, LASSO and elastic net with \(\alpha=0\). We will use cross-validation to select the tuning parameter. Then we will test models on the test set.

	Ridge	**LASSO**	**ElasticNet**
RMSE	0.779	1.543	1.553
selected variables	3123	2	3

b)

Here we select 300 variables with the largest marginal correlation with the response variable and add variables selected by mBIC2.

	Ridge	**LASSO**	**ElasticNet**
RMSE	0.221	1.782	1.872
selected variables	243	1	1

Only Ridge regression worked better after applying these conditions.