4. Simulation for Lasso and OLS Regression

November 30, 2020

0.1 4. Simulation for Lasso and OLS Regression

We consider the multivariate regression model

$$Y = X\beta + \varepsilon = \sum_{j=1}^{p} \beta_j X_j + \varepsilon_i,$$

where $X = (X_1, \ldots, X_p) \in \mathbb{R}^p$ is a p-dimensional vector of regressors and β a p-dimensional vector of regression coefficients. The β-vector is *sparse*, i.e., only the first s elements of the vector are different from zero, and all others are exactly equal to zero.

1. In your own words, explain the idea behind Lasso, Ridge and Elastic Net and in what respect they differ from ordinary least squares.
2. Implement a data generating process according to the following setting:
   - $n = 100$,
   - $p = 50$,
   - $s = 5$,
   - $\beta = (5, 4, 3, 2, 1, 0, 0, \ldots, 0)$
   - $\varepsilon \sim N(0, 1)$
   - The covariates are drawn from a multivariate normal distribution with a $(p \times p)$-covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 0.25 & ... & 0.25 \\ 0.25 & 1 & \ldots & 0.25 \\ \ldots & \ldots & \ldots & \ldots \\ 0.25 & 0.25 & \ldots & 1 \end{pmatrix}$$

   (all entries in the covariance matrix are equal to 0.25 except for those on the diagonal. These values are equal to 1.)
   Hint: It will be useful if you implement the data generating process in a function.
3. Generate a sample according to the data generating process above. Estimate the regression coefficients by ordinary least squares and lasso regression. Calculate and compare the squared error, i.e., $(\hat{\beta}_j - \beta_j)^2$ for $j = 1, 5, 50$ for the lasso and ols estimators.

4. Perform a simulation study, i.e., repeat your calculation from part 3. $R = 1000$ times. Calculate the mean squared error for the lasso and ols coefficients for the first, fifth, and 50th regressor, i.e., average the squared error over the R repetitions. For the first, fifth, and 50th regressor, how many times is the corresponding coefficient (estimated with lasso) set exactly to zero?

5. How do your result in part 4. change, if you increase the number of regressors p to 70, 90, and, 110? Illustrate your findings with an appropriate figure and summarize your results.

0.1.1 4. Simulation for Lasso and OLS Regression

[151]:
```
from IPython.core.interactiveshell import InteractiveShell     #allows printing␣
↪multiple lines
InteractiveShell.ast_node_interactivity = "all"
```

[152]:
```
#def print(*args):
#    __builtins__.print(*("%.4f" % a if isinstance(a, float) else a
#                         for a in args))
```

1. Explain Lasso, Ridge and Elastic Net and how they differ from OLS:

2. Implement data generating process

[153]:
```
import numpy as np

#parameter
n= 100
p=70
s=5
mu = np.array([0]*p)            #oder np.zeros(p)

#covariance matrix
cov = np.linspace(0.25,0.25, p*p)      #np.identity(p) creates an identity␣
↪matrix with 0s and 1s on the diagonal
cov =np.array(cov)
cov = cov.reshape(p,p)
np.fill_diagonal(cov, 1.)

#coefficient
null = np.array([0]*(p-5))
coef=np.array([5,4,3,2,1])
coef=np.concatenate((coef,null))
```

[154]:
```
cov
cov.shape
coef
coef.shape
mu.shape
```

[154]:
```
array([[1.  , 0.25, 0.25, …, 0.25, 0.25, 0.25],
       [0.25, 1.  , 0.25, …, 0.25, 0.25, 0.25],
       [0.25, 0.25, 1.  , …, 0.25, 0.25, 0.25],
       …,
       [0.25, 0.25, 0.25, …, 1.  , 0.25, 0.25],
       [0.25, 0.25, 0.25, …, 0.25, 1.  , 0.25],
       [0.25, 0.25, 0.25, …, 0.25, 0.25, 1.  ]])
```

[154]:
```
(70, 70)
```

[154]:
```
array([5, 4, 3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
       0, 0, 0, 0])
```

[154]:
```
(70,)
```

[154]:
```
(70,)
```

[155]:
```
#create function for generating a sample

def gen_sample(mu, n, p, cov):
    X = np.random.multivariate_normal(mu, cov, n)
    error = np.random.normal(loc=0.0, scale=1.0, size=n)
    y = np.matmul(X, coef)          #oder np.dot(x, beta) (only yield different␣
 ↪results when mult. 3D matrices)
    y = np.add(y, error)
    return X,y
```

[156]:
```
#generate a sample

X, y = gen_sample(mu, n, p, cov)
#X[:5]
#y[:5]
X.shape
y.shape
```

[156]:
```
(100, 70)
```

[156]:
```
(100,)
```

0.1.2 Regularisierung

Estimate coefficients by linear regression and lasso

[157]:
```
#linear regression

from sklearn.linear_model import LinearRegression
```

```
model = LinearRegression()
model.fit(X, y)
#print("Coefficient:", model.coef_)
coef_lin_reg = model.coef_
#coef_lin_reg.shape
```

[157]:
```
LinearRegression()
```

[158]:
```
#lasso with CV (better MSE)

from sklearn.linear_model import LassoCV

model = LassoCV()
model.fit(X, y)
print("Coefficient:", model.coef_)
coef_lasso = model.coef_
#coef_lasso.shape
```

[158]:
```
LassoCV()
```

```
Coefficient: [ 4.78795693e+00  3.76362407e+00  3.00573214e+00  1.92530583e+00
  8.32301269e-01  0.00000000e+00 -0.00000000e+00  1.13872274e-04
  1.07327795e-01 -0.00000000e+00 -0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  4.20243221e-02  3.38446672e-05
  1.53043059e-01  0.00000000e+00  0.00000000e+00 -0.00000000e+00
  0.00000000e+00 -0.00000000e+00  0.00000000e+00  1.81460860e-01
  0.00000000e+00  0.00000000e+00 -0.00000000e+00  0.00000000e+00
  0.00000000e+00 -0.00000000e+00  0.00000000e+00  0.00000000e+00
  8.13228148e-02 -0.00000000e+00  0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00 -0.00000000e+00  0.00000000e+00
  0.00000000e+00 -0.00000000e+00  0.00000000e+00 -0.00000000e+00
  0.00000000e+00  0.00000000e+00 -0.00000000e+00  4.05596077e-02
  9.48969037e-02  0.00000000e+00 -0.00000000e+00 -0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00  6.77654197e-02
  0.00000000e+00  0.00000000e+00  2.21974021e-02 -0.00000000e+00
 -0.00000000e+00  0.00000000e+00 -0.00000000e+00  0.00000000e+00
  0.00000000e+00  0.00000000e+00  0.00000000e+00 -0.00000000e+00
 -0.00000000e+00  0.00000000e+00]
```

[159]:
```
from sklearn import linear_model

#lasso without CV

model = linear_model.Lasso(alpha=1, fit_intercept=True)
model.fit(X, y)
print("Coefficient:", model.coef_)
#print("Coefficient:", model.sparse_coef_)
```

```
#coef_lasso = model.coef_
#coef_lasso.shape
```

[159]:
```
Lasso(alpha=1)
```

```
Coefficient: [4.63216726 3.08657808 2.76028047 1.39478653 0.52435591 0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.         0.         0.
 0.         0.         0.         0.        ]
```

[160]:
```
#squared error for j=1,5,50

error_lin_reg = np.round((coef_lin_reg-coef)**2,2)
error_lasso = np.round((coef_lasso-coef)**2,2)

print("Error for j=1", np.c_[error_lin_reg[0], error_lasso[0]])
print("Error for j=5",np.c_[error_lin_reg[4], error_lasso[4]])
print("Error for j=50",np.c_[error_lin_reg[-1], error_lasso[-1]])
#print("\n")
#print("Total Error for j= 1,...,50", np.c_[np.sum(error_lin_reg), np.
 ↪sum(error_lasso)])
```

```
Error for j=1 [[0.1  0.04]]
Error for j=5 [[0.   0.03]]
Error for j=50 [[0.09 0.  ]]
```

Lasso has lower total error than linear regression

0.1.3 Create function for comparing the squared error and outputting lasso coefficients

[173]:
```
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LassoCV

def compare_squared_error(mu, n, p, cov):
    k=100
    error_lin_reg1_all =[]
    error_lin_reg5_all =[]
    error_lin_reg50_all =[]
```

```
    error_lasso1_all =[]
    error_lasso5_all =[]
    error_lasso50_all =[]

    lasso_coef1 =[]
    lasso_coef5 =[]
    lasso_coef50 =[]

    for i in range(k):
        X, y = gen_sample(mu, n, p, cov)
        model1 = LinearRegression()
        model1.fit(X, y)
        error_lin_reg1 = (model1.coef_[0]-coef[0])**2
        error_lin_reg5 = (model1.coef_[4]-coef[4])**2
        error_lin_reg50 = (model1.coef_[-1]-coef[-1])**2

        model2 = LassoCV()
        model2.fit(X, y)
        error_lasso1 = (model2.coef_[0]-coef[0])**2
        error_lasso5 = (model2.coef_[4]-coef[4])**2
        error_lasso50 = (model2.coef_[-1]-coef[-1])**2

        #save OLS and lasso errors
        error_lin_reg1_all.append(error_lin_reg1)
        error_lin_reg5_all.append(error_lin_reg5)
        error_lin_reg50_all.append(error_lin_reg50)

        error_lasso1_all.append(error_lasso1)
        error_lasso5_all.append(error_lasso5)
        error_lasso50_all.append(error_lasso50)

        #save lasso coefficients
        lasso_coef1.append(model2.coef_[0])
        lasso_coef5.append(model2.coef_[4])
        lasso_coef50.append(model2.coef_[-1])
    print(X.shape)
    return error_lin_reg1_all, error_lin_reg5_all, error_lin_reg50_all,␣
↪error_lasso1_all, error_lasso5_all, error_lasso50_all, lasso_coef1,␣
↪lasso_coef5, lasso_coef50
```

[162]:
```
import time

t0 = time.time()
error_lin_reg1_all, error_lin_reg5_all, error_lin_reg50_all, error_lasso1_all,␣
↪error_lasso5_all, error_lasso50_all, lasso_coef1, lasso_coef5, lasso_coef50␣
↪= compare_squared_error(mu, n, p, cov)
```

```
t1 = time.time()
total = t1-t0
print("Run Time:", total)
```

```
Run Time: 19.70381188392639
```

0.1.4 Return Coefficients

[163]:
```
type(error_lin_reg1_all)
error_lin_reg1_all[:5]
len(error_lin_reg1_all)

lasso_coef1[:5]
lasso_coef5[:5]
lasso_coef50[:5]

len(lasso_coef1)

#arrange coefficients next to each other
list =[lasso_coef1[:5], lasso_coef5[:5],lasso_coef50[:5]]

for a in zip(*list):
    print(*a)
```

[163]:
```
list
```

[163]:
```
[0.1513556548024474,
 0.10774451608632818,
 0.03644625337157365,
 0.07024689315073072,
 0.00021107470947513484]
```

[163]:
```
100
```

[163]:
```
[4.746985724916531,
 4.7357435701158055,
 4.950060531741323,
 4.9476958159140585,
 4.937513545658403]
```

[163]:
```
[0.7785109026286136,
 0.9396373902302702,
 0.8954315736133334,
 0.920798983985351,
 0.6942245408710509]
```

[163]:
```
[0.0, 0.0, 0.0, 0.13021288099401881, -0.0]
```

[163]: 100

```
4.746985724916531 0.7785109026286136 0.0
4.7357435701158055 0.9396373902302702 0.0
4.950060531741323 0.8954315736133334 0.0
4.9476958159140585 0.920798983985351 0.13021288099401881
4.937513545658403 0.6942245408710509 -0.0
```

0.1.5 Average Errors

[164]:
```
def average_err(reg1,reg5,reg50,las1,las5,las50):
    mse_ols_1 = np.mean(error_lin_reg1_all)
    mse_ols_5 = np.mean(error_lin_reg5_all)
    mse_ols_50 = np.mean(error_lin_reg50_all)

    mse_lasso_1 = np.mean(error_lasso1_all)
    mse_lasso_5 = np.mean(error_lasso5_all)
    mse_lasso_50 = np.mean(error_lasso50_all)
    return mse_ols_1, mse_ols_5, mse_ols_50, mse_lasso_1, mse_lasso_5,␣
↪mse_lasso_50
```

[165]:
```
mse_ols_1, mse_ols_5, mse_ols_50, mse_lasso_1, mse_lasso_5, mse_lasso_50 =␣
↪average_err(error_lin_reg1_all, error_lin_reg5_all, error_lin_reg50_all,␣
↪error_lasso1_all, error_lasso5_all, error_lasso50_all)
```

[166]:
```
print("Compare Errors:", "OLS - Lasso")

mse_ols_1 - mse_lasso_1

mse_ols_5 - mse_lasso_5

mse_ols_50 - mse_lasso_50

print("The OLS-error is higher for all coefficients, than the Lasso-error")
```

```
Compare Errors: OLS - Lasso
```

[166]: 0.016099302619472393

[166]: 0.014701067344755424

[166]: 0.05148583438800002

```
The OLS-error is higher for all coefficients, than the Lasso-error
```

0.1.6 Count how many times the 1st, 5th and 50th coefficient are set exactly to zero

[167]:
```
#count how many times the lasso coefficient is set exactly to zero

def count(first, fifth, fiftieth):
    count_1 = 0
    count_5 = 0
    count_50 = 0

    for i in lasso_coef1:
        if i == 0:
            count_1 += 1
    for i in lasso_coef5:
        if i == 0:
            count_5 += 1
    for i in lasso_coef50:
        if i == 0:
            count_50 += 1
    return count_1, count_5, count_50

#print(count_1)
#print(count_5)
#print(count_50)


#error_lasso1_all.shape
#print(np.round(error_lasso1_all, 4))
```

[168]:
```
count(lasso_coef1, lasso_coef5, lasso_coef50)
```

[168]:
```
(0, 0, 79)
```

The coefficient of the 50th regressor is 79 (number varies) out of the 100 times exactly set to zero.

[ ]:
```

```