Computational Method, Hosung Lee
Homework_4

1. Generate random 100 MVN data, p=20. The covariance matrix should be positive semi-definite symmetric matrix.

1. (1) Calculate sample covariance matrix.

   (2) Find out the first three principal components.

   (3) Calcuate the proportions of the variability of data that can be explained by the first K principal components and find the value of K that it reaches to 99% of the variability.

```
R = randn(20);

sigma = R*R'; %symmetric sigma
mu = [ 5,1,2,3,5 , 4,1,8,3,5, 5,4,6,7,8, 1,11,5,2,1];%set the mu
n=150;

X = mvnrnd(mu,sigma,n); % observations are generated


%1-a)

var = cov(X)
sig  =corrcov(var)

e = eig(sig); %eigenvalue of sig
[V,D] = eig(sig); % V is eigen vetors of sig(correlation matrix)
```

(1)
Cov(X)

var
20x20 double

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
1	15.4497	0.4564	1.3497	-2.1537	-4.9102	0.8224	-3.5111	-6.6709	-0.9782	-3.4445	1.9665	-0.9059	0.3552	-2.9340	-1.9577	-1.8037	2.5537	1.1206	-7.0101	-5.8229
2	0.4564	11.1701	2.6244	3.7152	-10.4055	1.6966	1.0522	1.1147	0.8878	-3.3279	1.1361	-2.6064	-6.0301	5.7606	-2.5030	3.1432	2.4510	-2.2830	-9.1339	-3.4205
3	1.3497	2.6244	11.6888	4.3205	-0.1686	-0.2521	-0.2788	1.9578	-2.6956	-0.1804	4.6513	7.9374	-6.8456	6.0087	3.0858	2.2927	-4.6581	0.2021	-6.2113	1.7661
4	-2.1537	3.7152	4.3205	30.5402	-1.2587	-4.3651	5.9921	4.3175	3.3201	1.4799	3.4362	-8.0529	-0.6823	5.5645	-0.3765	5.7472	-4.4645	8.8598	-2.7013	2.6278
5	-4.9102	-10.4055	-0.1686	-1.2587	29.5863	-6.2397	11.9993	1.4810	-6.3495	-0.2171	-4.2691	2.8222	7.8963	0.0460	12.3564	0.0605	-0.7393	4.5035	14.3602	4.1072
6	0.8224	1.6966	-0.2521	-4.3651	-6.2397	17.8579	0.3154	-0.9830	-3.4776	0.2309	2.8128	-3.7818	4.4105	3.2583	-1.9189	-11.1423	-1.3084	-0.9707	0.4092	4.5339
7	-3.5111	1.0522	-0.2788	5.9921	11.9993	0.3154	24.1682	5.8798	-1.0549	1.7140	-5.1218	-3.2643	8.8680	9.3018	3.1483	5.4860	-0.4789	5.7413	7.3503	-0.2810
8	-6.6709	1.1147	1.9578	4.3175	1.4810	-0.9830	5.8798	16.3374	3.9008	3.8131	-1.6285	0.7392	-1.1920	-0.2096	-5.3964	8.9982	-2.6149	3.2528	0.0798	2.8631
9	-0.9782	0.8878	-2.6956	3.3201	-6.3495	-3.4776	-1.0549	3.9008	20.7545	-3.7402	-1.6999	-6.8080	-5.1420	-2.0327	-4.5208	4.3311	-4.9368	-4.0927	-2.7733	-1.7580
10	-3.4445	-3.3279	-0.1804	1.4799	-0.2171	0.2309	1.7140	3.8131	-3.7402	16.7950	-0.2326	2.5071	6.1804	-4.3214	-0.6449	-0.9106	-0.7378	1.1024	0.8421	6.9177
11	1.9665	1.1361	4.6513	3.4362	-4.2691	2.8128	-5.1218	-1.6285	-1.6999	-0.2326	10.0590	4.0652	-0.5316	1.5746	-0.1449	-3.7392	-4.2304	-1.7660	-0.8436	3.8571
12	-0.9059	-2.6064	7.9374	-8.0529	2.8222	-3.7818	-3.2643	0.7392	-6.8080	2.5071	4.0652	30.9733	-9.2443	3.9886	3.6013	-0.5359	-2.2971	2.5910	-5.0437	-2.9408
13	0.3552	-6.0301	-6.8456	-0.6823	7.8963	4.4105	8.8680	-1.1920	-5.1420	6.1804	-0.5316	-9.2443	30.4518	-5.6796	2.8388	-6.5881	-3.3971	-2.6261	14.5441	3.4030
14	-2.9340	5.7606	6.0087	5.5645	0.0460	3.2583	9.3018	-0.2096	-2.0327	-4.3214	1.5746	3.9886	-5.6796	19.5284	-0.4552	2.0481	-2.3094	1.2376	-3.2054	-1.0668
15	-1.9577	-2.5030	3.0858	-0.3765	12.3564	-1.9189	3.1483	-5.3964	-4.5208	-0.6449	-0.1449	3.6013	2.8388	-0.4552	18.7634	-5.1106	2.2791	-2.4710	1.9603	0.3560
16	-1.8037	3.1432	2.2927	5.7472	0.0605	-11.1423	5.4860	8.9982	4.3311	-0.9106	-3.7392	-0.5359	-6.5881	2.0481	-5.1106	26.0029	0.8028	1.5162	-1.8570	-5.5812
17	2.5537	2.4510	-4.6581	-4.4645	-0.7393	-1.3084	-0.4789	-2.6149	-4.9368	-0.7378	-4.2304	-2.2971	-3.3971	-2.3094	2.2791	0.8028	13.8659	2.9977	-5.0731	-7.7031
18	1.1206	-2.2830	0.2021	8.8598	4.5035	-0.9707	5.7413	3.2528	-4.0927	1.1024	-1.7660	2.5910	-2.6261	1.2376	-2.4710	1.5162	2.9977	13.6057	0.0685	-0.6033
19	-7.0101	-9.1339	-6.2113	-2.7013	14.3602	0.4092	7.3503	0.0798	-2.7733	0.8421	-0.8436	-5.0437	14.5441	-3.2054	1.9603	-1.8570	-5.0731	0.0685	21.0125	6.9498
20	-5.8229	-3.4205	1.7661	2.6278	4.1072	4.5339	-0.2810	2.8631	-1.7580	6.9177	3.8571	-2.9408	3.4030	-1.0668	0.3560	-5.5812	-7.7031	-0.6033	6.9498	15.0496

(2)

```
Z=zeros([20,20]);

for i = 1:20
Z(i,:)= V(i,1)*X(i,:);
end

Cov_Z  = cov(Z);

%%
%B)

lamda = zeros([1,20]);

for  i=1:20
    lamda(1,i) = Cov_Z(i,i);

end

lamda_sort = sort(lamda, 'descend');
lamda_sort_tr = lamda_sort'
```

lamda_sort_t
20x1 double

	1
1	7.3224
2	3.7299
3	3.1905
4	2.9252
5	2.2440
6	2.0045
7	1.8989
8	1.8273
9	1.7404
10	1.6271
11	1.3005
12	1.1365
13	1.0988
14	1.0662
15	0.9197
16	0.8745
17	0.7569
18	0.4580
19	0.4387
20	0.3501

%first three largest PCA component means three largest variability
%[7.3224, 3.7299, 3.1905]   = [Z17, Z14, Z15]

(3) Calculate the proportions of the variability of data that can be explained by the first K principal components and find the value of K that it reaches to 99% of the variability.

```
total_var = sum(lamda_sort); %36.9097

ninty_nine = total_var*0.99
difference = total_var-ninty_nine % minmum variability(0.35) is smaller than the difference(0.36)
% Therefore we need all K except for mimimum one(0.3501)
```

```
%%

%total variability is sum of lamda

total_var = sum(lamda_sort); %36.9097

ninty_nine = total_var*0.99
difference = total_var-ninty_nine % minmum variablity is smaller than the difference
% Therefore we need all K except for mimimum one(0.3501)
```

2. Generate Y values using the following regression functions:

```
%%
%Q2)


%x = randn(150,21);
mu2 = mean(X);
Cov2 = cov(X);
coeff = pca(X);
er = normrnd(0,3,[150,1]);
X2 = [ones(150,1) , X];
X2_train = X2(1:100 , 1:21);
X2_test = X2(101:150 , 1:21);

%%

Y = 5 + 2*X2(:,2) + 5*X2(:,4) + 3*X2(:,20) + er;
```

1) Estimate the regression line using the least square method and find out the predicted values, residuals for each observation and the mean square errors.

```
%%
%Fit regression

beta = inv(X2'*X2)*X2'*Y ;

Y_pred = X2_test * beta;
plot(Y_pred,'*')

%%

Y_MSE  = Y_test - Y_pred;%residual
Y_MSE1 = Y_MSE.*Y_MSE;
plot(Y_MSE1, '*')

MSE = sum(Y_MSE1)/49;
%11.55
```

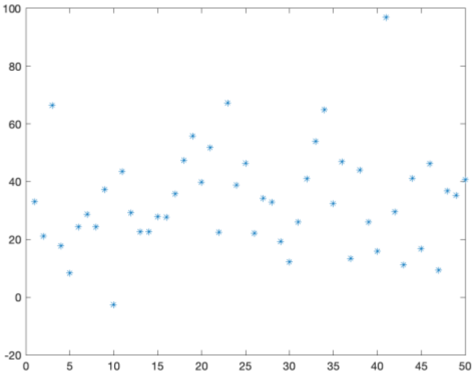
a.Fit for regression line = Beta.

Computational Method, Hosung Lee
Homework_4

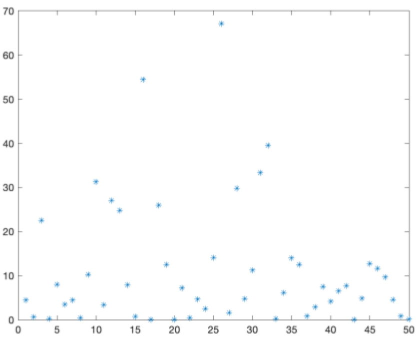
beta
21x1 double

	1
1	-3.4341
2	2.0882
3	0.0769
4	5.4483
5	0.1224
6	0.1585
7	0.3536
8	-0.0315
9	0.0825
10	-0.0853
11	-0.1855
12	0.0999
13	0.2783
14	-0.2648
15	0.4704
16	-0.0024
17	-0.0102
18	0.1877
19	0.1383
20	2.4270
21	0.1326

b.Y-pred by plot from the code(1 by 50)



c. residual plot from the code(1 by 50)



d. MSE

```
MSE = sum(Y_MSE1)/49;   %n-1=49
%11.55
```

2) Estimate the regression line using the least square method based on the first 5 principal components and find out the predicted values, residuals for each observation and the mean square errors.

```
%%
%2-b
%Find PCA Beta

for t = 1:20

    PCA5 = [];
    PCA5_tr = [];
    PCA5_test = [];
    beta_pca = [];
    Y_pred_pca = [];

    PCA_5 = X * coeff(:,1:t);
    PCA_5 = [ones(150,1) , PCA_5];

    PCA5_tr = PCA_5(1:100 ,:);
    PCA5_test = PCA_5(101:150 , :);

    beta_pca = inv(PCA5_tr'*PCA5_tr)*PCA5_tr'*Y_tr
    Y_pred_pca = PCA5_test * beta_pca;

end
```

a.

Fit for regression line, LSE = Beta.

b. predicted value

beta_pca
21x1 double

	1
1	3.6363
2	0.2475
3	-0.8869
4	0.4303
5	0.3582
6	-0.2768
7	0.8971
8	3.6471
9	1.1557
10	0.8268
11	0.8079
12	-0.4826
13	-3.0899
14	0.1433
15	-1.5290
16	-1.5354
17	1.3287
18	1.0054
19	-0.4648
20	3.0937
21	1.1818

Computational Method, Hosung Lee
Homework_4

Y_pred_pca
50x1 double

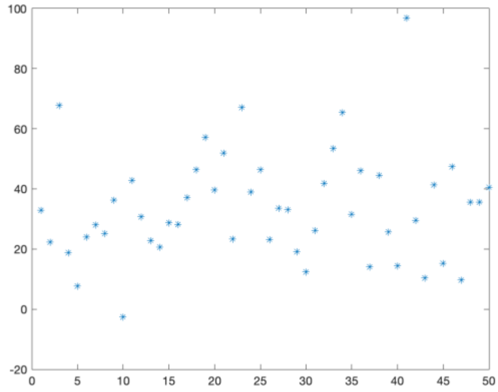
	1
1	32.8902
2	22.3129
3	67.6907
4	18.8198
5	7.6983
6	24.0680
7	28.1318
8	25.2328
9	36.3521
10	-2.5989
11	42.8137
12	30.7353
13	22.8755
14	20.6438
15	28.6523
16	28.1545
17	37.0842
18	46.4016
19	57.1977
20	39.6214
21	51.8950

50x1 double

	1
23	67.0668
24	39.0358
25	46.3573
26	23.0892
27	33.6309
28	33.1214
29	19.1545
30	12.4204
31	26.1698
32	41.9105
33	53.5379
34	65.4663
35	31.5359
36	46.0288
37	14.0483
38	44.4703
39	25.6945
40	14.3405
41	96.8072
42	29.5457
43	10.4278
44	41.3893
45	15.2169

50x1 double

	1
33	53.5379
34	65.4663
35	31.5359
36	46.0288
37	14.0483
38	44.4703
39	25.6945
40	14.3405
41	96.8072
42	29.5457
43	10.4278
44	41.3893
45	15.2169
46	47.3251
47	9.7459
48	35.5810
49	35.6708
50	40.5248
51	
52	
53	
54	
55	



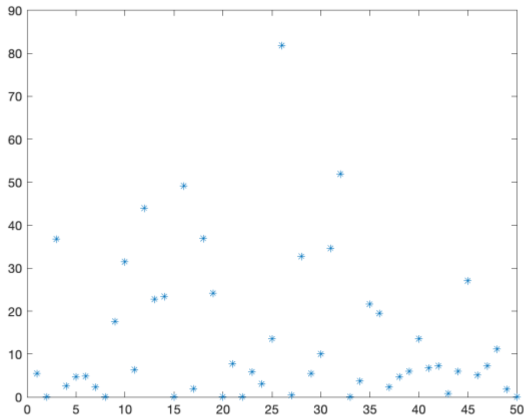
Plot for Y-predicted value

c. Residual and MSE

```
%%
plot(Y_pred_pca,'*')
```

```
%%


Z_MSE  = Y_test - Y_pred_pca;
Z_MSE1 = Z_MSE.*Z_MSE;
plot(Z_MSE1, '*')
MSE_Z = sum(Z_MSE1)/50;
```



Plot for residuals from the code(Z_MSE)

MSE = 14.1523

3) Compare the results

MSE from the data, 11.15
MSE from the PCA-5 14.15
PCA-5 could explain most of the data.
PCA-% could fit the regression line with smaller data.
In that sense, using all original data might be overfitted.