Computational Methods (16:540:540)
Homework 4: Solution

1. Generate random 150 MVN data, $p = 20$. The covariance matrix should be positive semi-definite symmetric matrix.

1. Calculate sample covariance matrix.
2. Find out the first three principal components.
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.

Solution: Use Matlab to generate a random 20×20 covariance matrix Σ as below.

$$\Sigma = \begin{bmatrix}
261 & 63 & 74 & -60 & -49 & -27 & -28 & 44 & 22 & 24 & -30 & 4 & -24 & 3 & -98 & 16 & -26 & -25 & 20 & 4 \\
63 & 130 & -56 & -33 & -3 & 55 & -55 & 49 & -13 & 8 & -52 & -17 & 68 & 55 & -52 & 41 & 24 & 44 & -9 & -51 \\
74 & -56 & 196 & -3 & -121 & -32 & 15 & -39 & 10 & 66 & 58 & 47 & -85 & -22 & -36 & -70 & -14 & -29 & 93 & 56 \\
-60 & -33 & -3 & 201 & 30 & -5 & 34 & -7 & -63 & 52 & 37 & -25 & 39 & 89 & 75 & 13 & 0 & 42 & 7 & -3 \\
-49 & -3 & -121 & 30 & 215 & 23 & 44 & -29 & -78 & -43 & 12 & -42 & 43 & 3 & 47 & 63 & -19 & 20 & -88 & -14 \\
-27 & 55 & -32 & -5 & 23 & 254 & -45 & -32 & -48 & -52 & 68 & 1 & 35 & 65 & -32 & 9 & 25 & 34 & 30 & -40 \\
-28 & -55 & 15 & 34 & 44 & -45 & 194 & 32 & -26 & -49 & 26 & -96 & 0 & -50 & 64 & -10 & 11 & -70 & -50 & 11 \\
44 & 49 & -39 & -7 & -29 & -32 & 32 & 241 & 31 & 4 & -48 & -58 & -47 & 73 & 68 & -20 & 28 & -18 & -7 & -101 \\
22 & -13 & 10 & -63 & -78 & -48 & -26 & 31 & 130 & -39 & -18 & 59 & -54 & -10 & 9 & -18 & -63 & -15 & 44 & -23 \\
24 & 8 & 66 & 52 & -43 & -52 & -49 & 4 & -39 & 144 & 7 & -10 & -26 & 15 & -25 & 18 & 25 & 29 & 58 & 3 \\
-30 & -52 & 58 & 37 & 12 & 68 & 26 & -48 & -18 & 7 & 156 & 38 & -43 & 21 & 8 & 4 & -38 & -3 & 67 & 17 \\
4 & -17 & 47 & -25 & -42 & 1 & -96 & -58 & 59 & -10 & 38 & 179 & -32 & 35 & -4 & -54 & -115 & 42 & 81 & 52 \\
-24 & 68 & -85 & 39 & 43 & 35 & 0 & -47 & -54 & -26 & -43 & -32 & 210 & 5 & -21 & 65 & 54 & 80 & -81 & 15 \\
3 & 55 & -22 & 89 & 3 & 65 & -50 & 73 & -10 & 15 & 21 & 35 & 5 & 182 & 55 & -32 & -28 & 32 & 38 & -86 \\
-98 & -52 & -36 & 75 & 47 & -32 & 64 & 68 & 9 & -25 & 8 & -4 & -21 & 55 & 234 & -46 & -16 & -29 & 76 & -15 \\
16 & 41 & -70 & 13 & 63 & 9 & -10 & -20 & -18 & 18 & 4 & -54 & 65 & -32 & -46 & 175 & -22 & 77 & -27 & -8 \\
-26 & 24 & -14 & 0 & -19 & 25 & 11 & 28 & -63 & 25 & -38 & -115 & 54 & -28 & -16 & -22 & 239 & 4 & -78 & -66 \\
-25 & 44 & -29 & 42 & 20 & 34 & -70 & -18 & -15 & 29 & -3 & 42 & 80 & 32 & -29 & 77 & 4 & 156 & -13 & -28 \\
20 & -9 & 93 & 7 & -88 & 30 & -50 & -7 & 44 & 58 & 67 & 81 & -81 & 38 & 76 & -27 & -78 & -13 & 207 & 35 \\
4 & -51 & 56 & -3 & -14 & -40 & 11 & -101 & -23 & 3 & 17 & 52 & 15 & -86 & -15 & -8 & -66 & -28 & 35 & 155
\end{bmatrix}$$

Use $x = mvnrnd(mu, sigma, n)$; to generate 150 samples with mean μ and covariance Σ in Matlab.

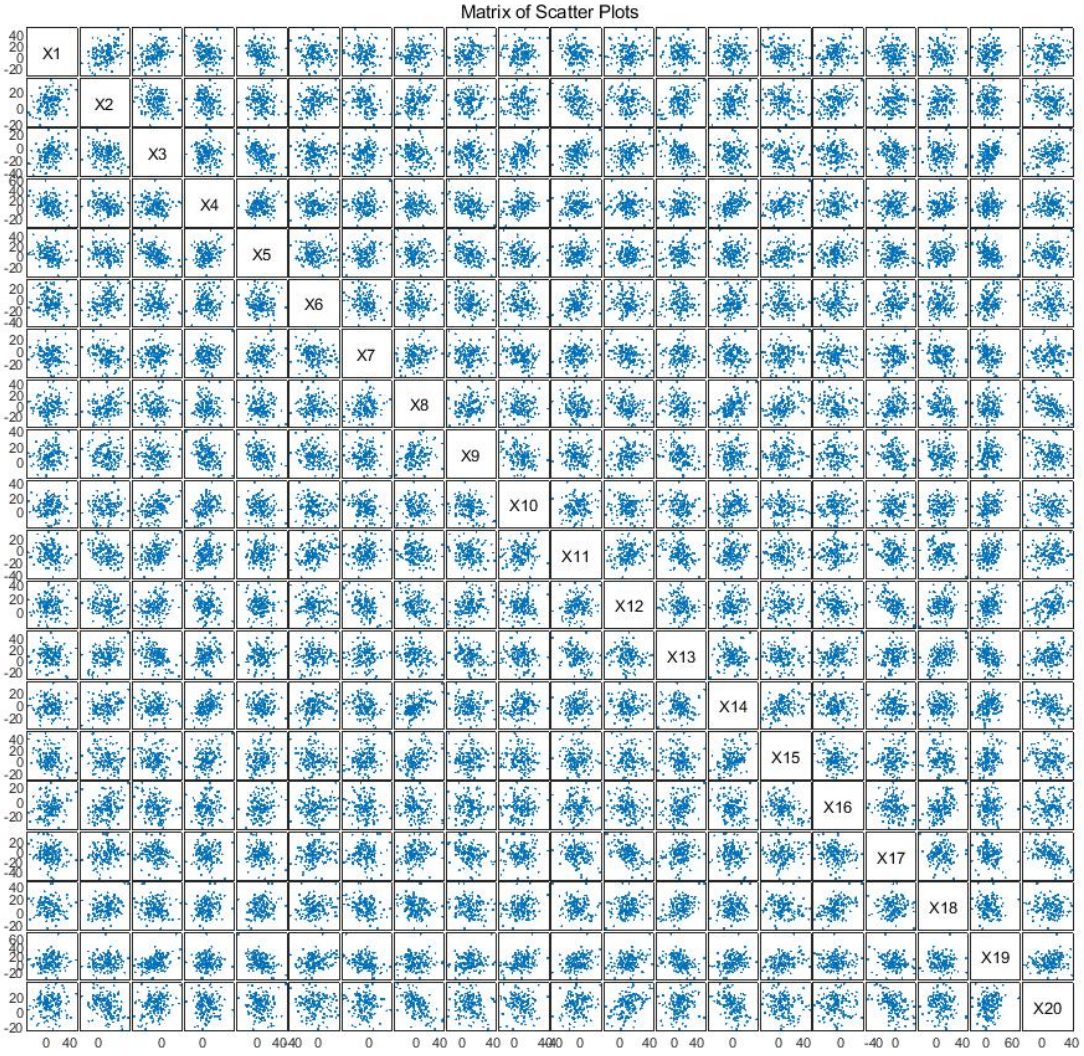
1. The sample variance is

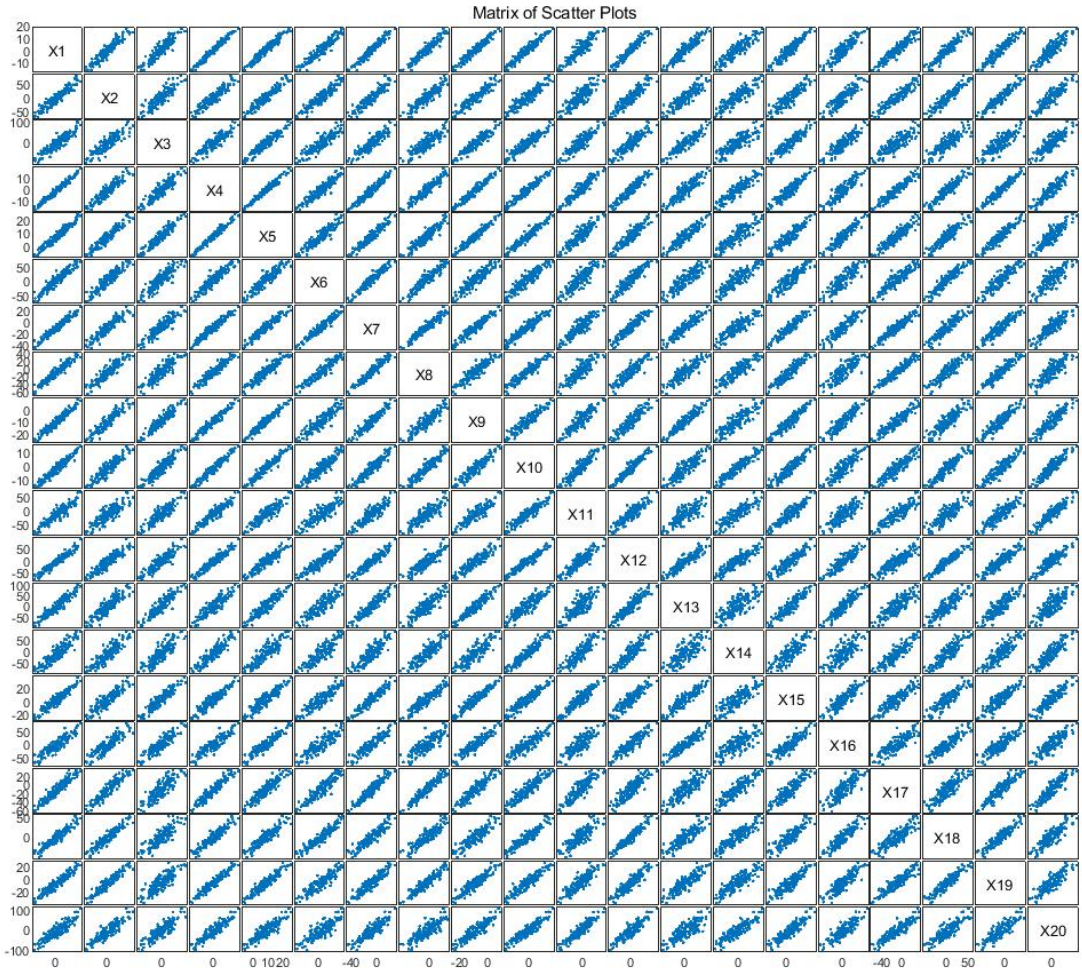
$$s_j^2 = s_{jj} = \frac{1}{n-1}\sum_{i=1}^{5}(x_{ij} - \bar{\mathbf{x}}_j)^2, j = 1, 2, ..., 20$$

$$s_{jk} = \frac{1}{n-1}\sum_{i=1}^{5}(x_{ij} - \bar{\mathbf{x}}_j)(x_{ik} - \bar{\mathbf{x}}_k), \quad \bar{\mathbf{x}}_k \text{ is the sample mean.}$$

$$\mathbf{S_n} = \begin{bmatrix}
221 & 56 & 74 & -45 & -51 & -18 & -13 & 18 & 27 & 10 & -31 & 8 & -11 & 0 & -73 & 31 & -42 & -13 & 35 & 14 \\
56 & 120 & -36 & -34 & -24 & 50 & -44 & 41 & -9 & 11 & -60 & -29 & 61 & 41 & -38 & 38 & 37 & 35 & 5 & -46 \\
74 & -36 & 174 & -6 & -97 & -14 & 11 & -28 & 1 & 61 & 59 & 41 & -78 & -7 & -36 & -44 & -31 & -19 & 91 & 48 \\
-45 & -34 & -6 & 214 & 35 & 15 & 18 & -14 & -65 & 55 & 29 & -32 & 40 & 89 & 74 & 15 & -4 & 44 & 17 & 1 \\
-51 & -24 & -97 & 35 & 201 & -4 & 40 & -32 & -66 & -36 & 9 & -17 & 29 & -3 & 39 & 39 & -21 & 35 & -98 & -12 \\
-18 & 50 & -14 & 15 & -4 & 250 & -41 & -26 & -61 & -20 & 84 & -26 & 22 & 53 & -26 & 22 & 42 & 14 & 64 & -21 \\
-13 & -44 & 11 & 18 & 40 & -41 & 152 & 39 & -15 & -48 & 26 & -69 & -3 & -29 & 37 & -13 & 9 & -49 & -65 & -9 \\
18 & 41 & -28 & -14 & -32 & -26 & 39 & 228 & 33 & 1 & -33 & -64 & -55 & 80 & 107 & -33 & 57 & -38 & 15 & -112 \\
27 & -9 & 1 & -65 & -66 & -61 & -15 & 33 & 129 & -37 & -22 & 48 & -52 & -16 & 4 & -4 & -65 & -20 & 34 & -25 \\
10 & 11 & 61 & 55 & -36 & -20 & -48 & 1 & -37 & 128 & 13 & -5 & -26 & 27 & -13 & 10 & 14 & 30 & 69 & 0 \\
-31 & -60 & 59 & 29 & 9 & 84 & 26 & -33 & -22 & 13 & 176 & 27 & -60 & 9 & -9 & 16 & -50 & -18 & 72 & 29 \\
8 & -29 & 41 & -32 & -17 & -26 & -69 & -64 & 48 & -5 & 27 & 152 & -28 & 0 & -13 & -34 & -122 & 28 & 60 & 70 \\
-11 & 61 & -78 & 40 & 29 & 22 & -3 & -55 & -52 & -26 & -60 & -28 & 211 & -9 & -28 & 57 & 54 & 78 & -83 & 32 \\
0 & 41 & -7 & 89 & -3 & 53 & -29 & 80 & -16 & 27 & 9 & 0 & -9 & 153 & 79 & -29 & -6 & 13 & 53 & -75 \\
-73 & -38 & -36 & 74 & 39 & -26 & 37 & 107 & 4 & -13 & -9 & -13 & -28 & 79 & 255 & -58 & 26 & -29 & 88 & -38 \\
31 & 38 & -44 & 15 & 39 & 22 & -13 & -33 & -4 & 10 & 16 & -34 & 57 & -29 & -58 & 159 & -42 & 73 & -6 & 8 \\
-42 & 37 & -31 & -4 & -21 & 42 & 9 & 57 & -65 & 14 & -50 & -122 & 54 & -6 & 26 & -42 & 279 & -1 & -69 & -96 \\
-13 & 35 & -19 & 44 & 35 & 14 & -49 & -38 & -20 & 30 & -18 & 28 & 78 & 13 & -29 & 73 & -1 & 140 & -18 & -13 \\
35 & 5 & 91 & 17 & -98 & 64 & -65 & 15 & 34 & 69 & 72 & 60 & -83 & 53 & 88 & -6 & -69 & -18 & 240 & 35 \\
14 & -46 & 48 & 1 & -12 & -21 & -9 & -112 & -25 & 0 & 29 & 70 & 32 & -75 & -38 & 8 & -96 & -13 & 35 & 167
\end{bmatrix}$$

The figure below shows the matrix of scatter plots for all variables $X_i, i = 1, 2, ..., 20$


Matrix of Scatter Plots
X1
X2
X3
X4
X5
X6
X7
X8
X9
X10
X11
X12
X13
X14
X15
X16
X17
X18
X19
X20



The figure above is the matrix of scatter plots for all variables $X_i, i = 1, 2, ..., 20$ from a highly correlation data set. And the figure below is the matrix of scatter plots of transformed variables $Z_1, Z_2, ..., Z_{20}$. After the transformed, the variables are independent, and the variance is decreasing.

2. The first three principal components are the eigenvector with the three largest eigenvalue.

e1	0.20	-0.12	0.36	-0.11	-0.25	-0.04	-0.13	-0.17	0.17	0.10	0.20	0.33	-0.28	-0.06	-0.12	-0.04	-0.42	-0.08	0.39
e2	0.05	0.02	-0.09	-0.11	0.16	-0.03	-0.05	-0.46	-0.07	-0.09	-0.05	0.09	0.31	-0.31	-0.47	0.24	-0.14	0.18	-0.32
e3	-0.24	-0.04	-0.03	0.47	0.18	0.42	-0.12	-0.23	-0.28	0.15	0.31	0.05	0.15	0.24	0.15	0.14	-0.09	0.22	0.25

$$Z_1 = \mathbf{e_1}^T \mathbf{X}$$

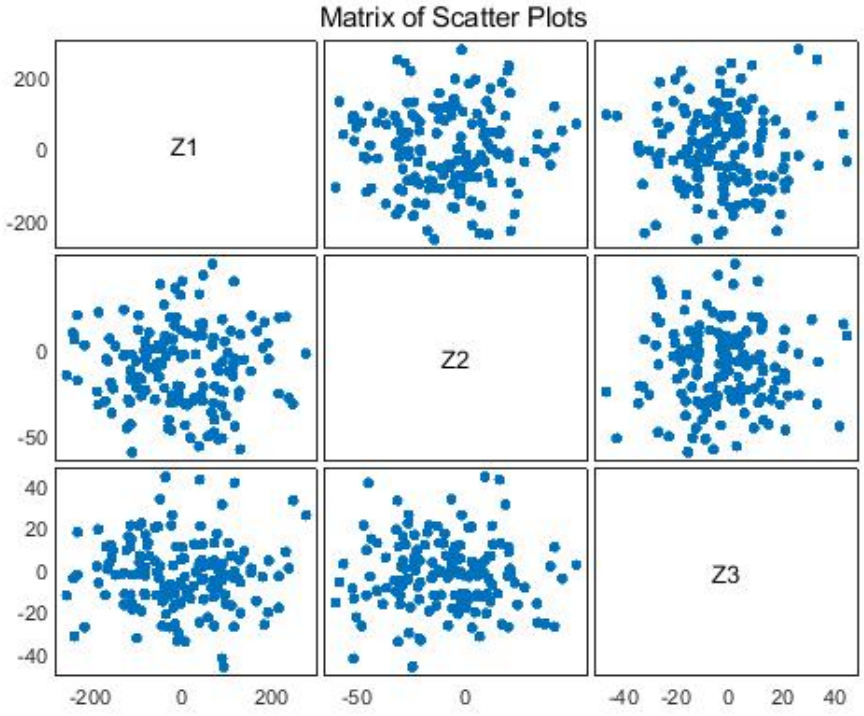
$$Z_2 = \mathbf{e_2}^T \mathbf{X}$$

$$Z_3 = \mathbf{e_3}^T \mathbf{X}$$

The matrix below is the covariance matrix of first three principal components, which the diagonal elements decrease.

$$\mathbf{S_{1,2,3}} = \begin{bmatrix} 12595.48 & -0.01 & 0.18 \\ -0.01 & 535.93 & -0.03 \\ 0.18 & -0.03 & 275.74 \end{bmatrix}$$

The figure also shows the variables Z_1, Z_2 and Z_3 are independent. The scale of the axis also implies the covariance is decreasing. (The result is from a more correlated covariance matrix shows in (3).)



3. When the covariance matrix is randomly generated, the correlation between variables is low, the table below shows the contribution of each component from high to low. When total 15 components are considered, it reaches 99% of the variability.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
653.4	568.6	460.9	432.4	325.7	272.9	243.7	183.5	155.4	131.1	97.6	83.7	54.2	43.8	20.0	14.7	4.8	2.6	0.9	0.1

In the contrast, when the correlation value between variables is set higher only a few of components can reaches to 99% of the variability. As shown by the table below, it only takes 6 components. And the first component explains 92.8%.

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
12595.5	535.9	275.7	218.1	127.1	100.6	97.3	42.1	35.6	24.5	6.5	5.7	1.6	1.3	0.8	0.6	0.2	0.1	0.1	0.0

2. Generate Y values using the following regression functions: $Y = 5 + 2X_1 + 5X_3 + 3X_{19} + \epsilon$ for the generated X matrix in #1, where $\epsilon \sim N(0, 2^2)$.

1. Estimate the regression line using the least square method with the first 100 observations as training data and find out the predicted values, residuals for each observation(training and testing) and the mean square errors.
2. Estimate the regression line using the least square method based on the first 5 principal components with the first 100 observations as training data and find out the predicted values, residuals for each observation(training and testing) and the mean square errors.
3. Compare the results in (1) and (2).

Solution:

1.

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_{100} \end{bmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,20} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,20} \\ 1 & x_{3,1} & x_{3,2} & \cdots & x_{3,20} \\ 1 & x_{4,1} & x_{4,2} & \cdots & x_{4,20} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{100,1} & x_{100,3} & \cdots & x_{100,20} \end{bmatrix}}_{\boldsymbol{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_{20} \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_{100} \end{bmatrix}}_{\boldsymbol{\epsilon}}.$$

The OLS solution has the form

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T \boldsymbol{Y} = \begin{bmatrix} 2.247 \\ 2.051 \\ -0.291 \\ 4.732 \\ -0.255 \\ 0.020 \\ 0.007 \\ 0.310 \\ -0.089 \\ 0.178 \\ 0.180 \\ -0.084 \\ -0.095 \\ -0.062 \\ 0.461 \\ -0.177 \\ -0.008 \\ 0.192 \\ 0.183 \\ 3.148 \\ 0.321 \end{bmatrix}.$$

The fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}.$$

The residuals

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \hat{\boldsymbol{y}}.$$

The residual sum of squares $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/n$.
For the training data the $MSE_{training} = 2.754.$, while for the testing data the $MSE_{testing} = 3.711$.

2.

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_{100} \end{bmatrix}}_{\boldsymbol{Y}} = \underbrace{\begin{bmatrix} 1 & z_{1,1} & z_{1,2} & \cdots & z_{1,k} \\ 1 & z_{2,1} & z_{2,2} & \cdots & z_{2,k} \\ 1 & z_{3,1} & z_{3,2} & \cdots & z_{3,k} \\ 1 & z_{4,1} & z_{4,2} & \cdots & z_{4,k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{100,1} & z_{100,3} & \cdots & z_{100,k} \end{bmatrix}}_{\boldsymbol{Z}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}}_{\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \epsilon_4 \\ \vdots \\ \epsilon_{100} \end{bmatrix}}_{\boldsymbol{\epsilon}}.$$

The OLS solution has the form

$$\hat{\boldsymbol{\beta}} = (Z^T Z)^{-1} Z^T \boldsymbol{Y}.$$

The fitted values

$$\hat{\boldsymbol{y}} = \boldsymbol{Z}\hat{\boldsymbol{\beta}}.$$

k	MSE_training	MSE_testing	MSE_total
5	8.877	11.555	9.756
6	7.169	8.003	7.461
7	5.285	6.309	5.656
8	5.268	6.388	5.667
9	5.264	6.317	5.639
10	5.262	6.286	5.628
11	5.060	6.237	5.491
12	4.043	4.403	4.192
13	3.965	4.635	4.216
14	3.171	3.850	3.423
15	3.166	3.860	3.423
16	3.100	4.102	3.462
17	3.100	4.119	3.467
18	2.821	3.711	3.140
19	2.789	3.594	3.081
20	2.754	3.711	3.096

The residual sum of squares $\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}/n$.
For the training data the $MSE_{training} = 8.877.$, while for the testing data the $MSE_{testing} = 11.555$.

3. The table above provides the different MSEs of different k (number of PCs). As the number of k increases the values of MSEs decrease. The MSE from (1) is always lower than (2). When all the components are considered, the MSEs of two methods are equal. However, when $k = 19$ the MSE of the testing data from (2) is little better than the method(1). The component 20 may be considered as some noise and could be discarded.

   When we adjust the value of error which is σ to a higher value. The result of MSE of training, testing and total data is shown as below. For this set of data, when $k = 9$, the MSE of testing data is the smallest which is 544.2. This value is smaller than 624.2 when all the variables are used in the regression model. Thus, for some data, the PCR can avoid over-fitting.

k	MSE_training	MSE_testing	MSE_total
1	4609.0	8038.3	5700.3
2	1360.8	1716.7	1490.2
3	1280.1	1684.6	1425.9
4	964.6	1313.3	1089.5
5	681.8	749.6	707.1
6	661.9	755.3	696.6
7	361.3	576.5	436.7
8	361.3	576.6	436.7
9	348.1	544.2	416.8
10	343.6	554.1	417.3
11	340.2	566.5	419.4
12	317.2	605.2	417.2
13	314.1	605.7	415.3
14	313.8	607.4	415.7
15	312.7	599.2	412.1
16	305.9	621.2	415.1
17	288.8	630.5	406.5
18	280.6	627.2	399.7
19	280.2	624.2	398.4
20	280.2	624.2	398.4