Name: Jingyi Yuan
UNI: jy2736
Class: STAT W4240

Homework 02

Problem 1

(a).

Column mean contains 5 numbers, with each number means the mean of each column in the original data sets.

```
> colmean_original
       x1        x2        x3        x4        x5
6.049104 -8.277221  4.665532  7.914270 62.138753
```

Column mean contains 100 numbers, with each number means the mean of each row in the original data sets.

```
> rowmean_original
  [1]  -0.1277116  20.8162864  -8.8984358  25.5999204  -9.7472153  64.0626702  22.0392371
  [8]  23.3914888  31.7598224 -13.8680290  43.8318898   6.5478369  14.1665143  16.1945993
 [15]  29.6357898  11.0316832  -2.5453007   8.6124471  33.8364419  24.9647839  34.8385372
 [22]  34.1951748  25.8869897  -0.4545730   9.0418836  21.4051827   3.2291136  35.5748021
 [29]  21.1031545   6.5535668   3.7478608  18.9230712  -9.2447158   6.3811655  16.8358750
 [36]   7.9628124  16.6264489  16.7027735 -34.4147885   0.4138282  12.6572899  35.4589880
 [43]  17.3456417  17.2383651   0.5124620 -24.7073649  17.1498949  52.3665782   9.6993053
 [50]   0.3079195  15.6758568 -13.3093667   8.2062088  34.8247664  12.1909900  -3.1939531
 [57]  -5.4779341  10.7689107  36.2253846  19.5034554   8.9492321   4.4008921  14.3901288
 [64]  14.7207124  27.9510161 -14.3617846  39.3331820  24.0356530  -6.7256757  -4.2948679
 [71]  27.1881673  47.2951022  19.1932996  23.5607379   7.6480638  18.1517706  16.9872267
 [78] -46.6660940   7.2223867  28.8378401   6.5043155  26.5206768  -2.4442159  15.3802055
 [85]  16.1739005  26.1705488  20.1409435  63.2646829   9.1977728  29.2026018   1.2105932
 [92]  21.2145724  -8.4896595  19.0639963  20.9767512   3.5962333  22.3461063   0.7145014
 [99]   6.3080005  64.8829556
```

(b).

The empirical covariance matrix is shown below:

```
> cov_matrix
          x1         x2        x3         x4         x5
x1  72.96417  -83.90858  53.23708  120.1162  568.4105
x2 -83.90858  110.89101 -63.89570 -115.9430 -817.3388
x3  53.23708  -63.89570  39.60282   83.7386  445.2511
x4 120.11620 -115.94304  83.73860  232.1333  683.5587
x5 568.41046 -817.33884 445.25112  683.5587 6288.8569
```

The diagonal values of the covariance matrix are the variances of 5 variables. The off

diagonal elements are the covariance between the 2 corresponding variables, with A_{ij} means the covariance between the i^{th} variable and the j^{th} variable.

(c).

The eigenvalues and associated eigenvectors (5 columns of the loading) are shown below:

```
> eigenvalues
[1] 6.557348e+03 1.868951e+02 2.038354e-01 9.775594e-04 9.373658e-05
> loading
            [,1]       [,2]         [,3]        [,4]        [,5]
[1,]  0.09009603 -0.3247102 -0.383470773  0.82286709  0.24957150
[2,] -0.12797842  0.1364755  0.227047683 -0.11412319  0.94890526
[3,]  0.07028767 -0.1941349  0.894987159  0.37278501 -0.13191135
[4,]  0.11077853 -0.9008231 -0.019718518 -0.40719485  0.10024632
[5,]  0.97892389  0.1636064  0.002946326 -0.07133967  0.09921159
```

This matrix has the same left eigenvectors as right eigenvectors because matrixes are symmetrical and can be used to describe transformations and some of the original directions may be preserved. Eigenvector is the direction of a matrix and each matrix has only one direction.

(d).

The loadings is a 5x5 matrix and the scores is a 100x5 matrix:

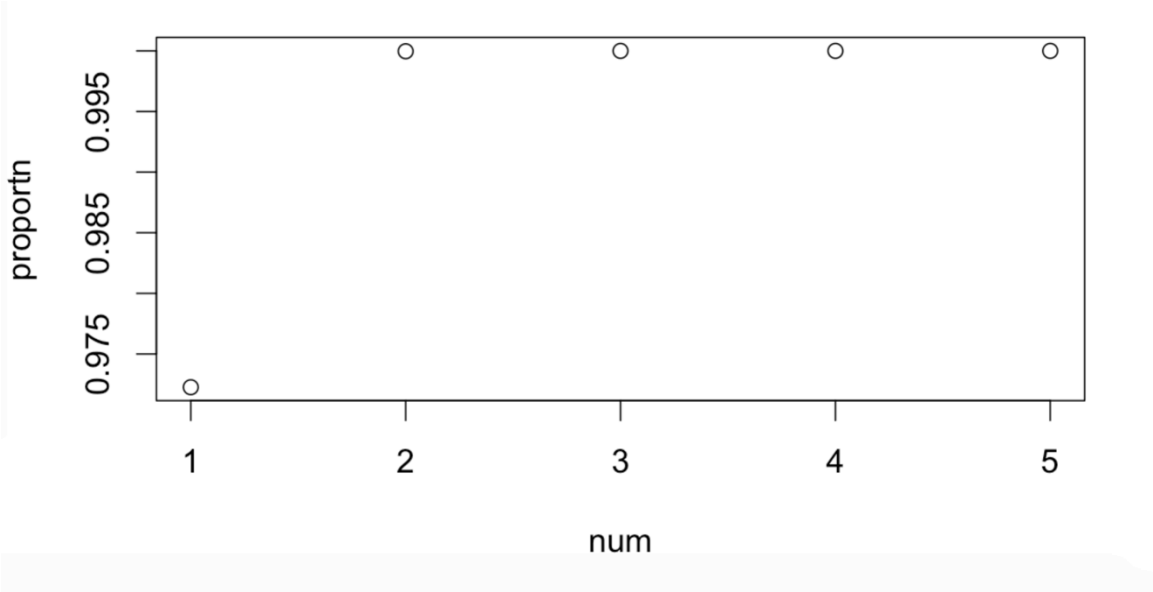
loading	num [1:5, 1:5] 0.0901 -0.128 0.0703 0.1108 0.9789 …
scores	num [1:100, 1:5] -58.6 18 -103.6 38.9 -106.8 ...

The loadings and part of the scores are shown below:

```
> loading
           [,1]       [,2]        [,3]       [,4]        [,5]
[1,]  0.09009603 -0.3247102 -0.383470773  0.82286709  0.24957150
[2,] -0.12797842  0.1364755  0.227047683 -0.11412319  0.94890526
[3,]  0.07028767 -0.1941349  0.894987159  0.37278501 -0.13191135
[4,]  0.11077853 -0.9008231 -0.019718518 -0.40719485  0.10024632
[5,]  0.97892389  0.1636064  0.002946326 -0.07133967  0.09921159
> scores
             [,1]       [,2]        [,3]          [,4]          [,5]
  [1,]  -58.606720   6.8128841  0.358690823 -0.0080251123  6.008437e-03
  [2,]   17.967890 -10.0253314  0.313590316 -0.0051005161 -1.454011e-02
  [3,] -103.557582   0.7721199  0.080912663  0.0422848507  5.511172e-04
  [4,]   38.865124 -10.3589218  0.396822082  0.0021659166  1.016818e-02
  [5,] -106.785255   1.3009950  0.094545132 -0.0025889590 -9.527577e-03
  [6,]  223.094855   2.1706728 -0.142606288  0.0237964971  4.184366e-03
  [7,]   24.866274  -9.1292222 -0.579070066  0.0068324074 -3.154517e-03
  [8,]   52.165664  12.0480945 -0.813698583  0.0235157234 -5.105202e-03
  [9,]   85.268224   9.0398772  1.026154372 -0.0278183294  2.019815e-02
 [10,] -134.472367  -8.3620478 -0.469219982  0.0352751309  1.364136e-02
 [11,]  112.720830 -18.1914430 -0.272898296  0.0266712638 -1.498326e-03
 [12,]  -28.717690   6.4568495 -0.408980503 -0.0227692145  8.701904e-03
 [13,]  -11.231216  -9.4480600  0.504892877  0.0091787452 -2.458606e-03
 [14,]   16.340741   8.7785057 -0.056563844 -0.0088876150  1.656248e-02
 [15,]   82.515039  14.7456820 -0.523378213 -0.0126785133 -6.137646e-03
 [16,]   -8.157462   7.5083979  0.353281349 -0.0176219768 -1.306550e-02
 [17,]  -60.075356  15.9949539  0.099597926  0.0531282578 -1.670995e-03
 [18,]  -33.627755  -6.8073458  0.970449080 -0.0238127849 -1.385280e-03
 [19,]   85.475222  -0.2322110  0.619350495  0.0698057329  2.389607e-02
 [20,]   57.823737  10.9987319 -0.356183214  0.0106818326  1.082133e-02
 [21,]   71.400819 -19.2036888  0.129225070 -0.0047113439 -6.144754e-03
 [22,]   97.346321   9.6712000  0.125602067 -0.0004988105 -8.084343e-03
 [23,]   45.205046  -5.6010380 -0.061884581 -0.0008785511 -4.791884e-03
 [24,]  -64.893384   1.7042777 -0.038372645 -0.0110658820 -2.750028e-03
 [25,]   19.206101   5.1502677  0.035441792  0.0284796665  4.969443e-03
```

(e).

The proportion of variance captured against the number of components included is shown below:



We should include 3 components so that more than 99.5% of values can be explained thus the data set is sufficiently accurate. At the same time, the proportion of variance is stable.

(f).

The scores of these new 5 observations are shown below:

```
> scores2
           [,1]       [,2]      [,3]        [,4]          [,5]
[1,] -70.371540 -12.098356 0.2924970 -0.06207175  0.0012951121
[2,]  27.468925   6.969873 0.2336020 -0.01918974 -0.0007608147
[3,]   2.223577  -2.556569 0.1522306 -0.06059532  0.0037656303
[4,] -67.083557   9.451715 0.4395150  0.05093323 -0.0166514432
[5,] -18.414179   6.968268 0.7268590 -0.05956837  0.0075723055
```

(g).

The coordinates of the projections in the original space are shown below:

```
> proj_x
           [,1]       [,2]       [,3]       [,4]       [,5]
[1,] -2.4117373  7.3549090 -2.5975381   3.102824 -70.867750
[2,]  0.2116526 -2.5642125  0.5776310  -3.235656  28.030303
[3,]  1.0304795 -0.6334791  0.6526094   2.549341   1.758442
[4,] -9.1130304  9.8751758 -6.5500546 -15.945741 -64.123336
[5,] -3.9217119  3.3076158 -2.6470738  -8.317073 -16.886027
```

after adding the mean of the column:

```
> projct_x
          [,1]        [,2]      [,3]       [,4]      [,5]
[1,]  3.637367  -0.9223123  2.067993 11.0170934 -8.728997
[2,]  6.260757 -10.8414338  5.243163  4.6786141 90.169056
[3,]  7.079584  -8.9107004  5.318141 10.4636109 63.897195
[4,] -3.063926   1.5979545 -1.884523 -8.0314715 -1.984582
[5,]  2.127392  -4.9696055  2.018458 -0.4028028 45.252726
```

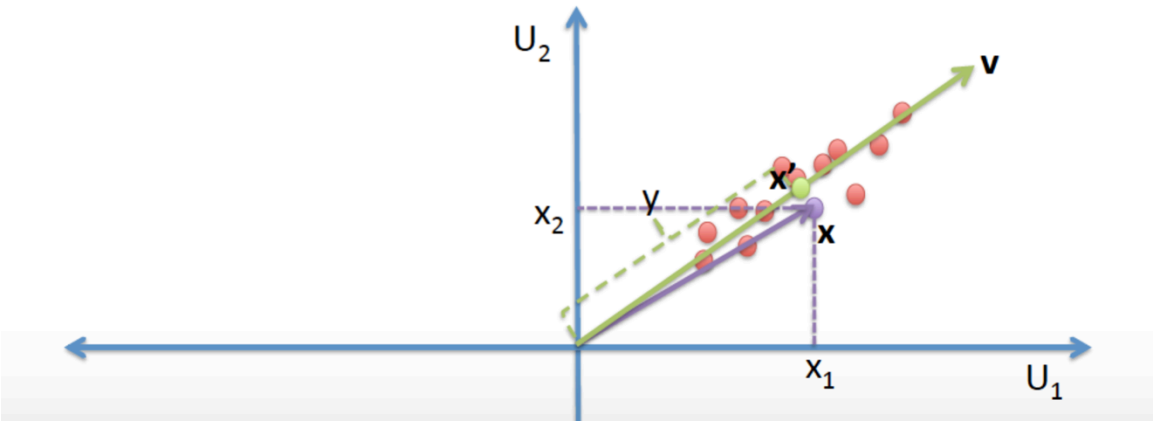
their Euclidean distance from the original data points:

```
> euclidean
[1] 0.9482524
```

(h).

```
> error
            x1          x2         x3           x4           x5
[1,] 0.1629176 -0.07472354 -0.2384708 -0.019637518 -0.005418460
[2,] 0.1055600 -0.05450684 -0.2020175 -0.003131408 -0.001981775
[3,] 0.1072981 -0.04505217 -0.1131587 -0.022049832 -0.005144966
[4,] 0.1307856 -0.07817755 -0.4145439  0.031075577  0.003990621
[5,] 0.3258562 -0.17901519 -0.6273244 -0.010682446 -0.007142412
```

The error d(x', x) for the 5 new points is orthogonal to the principal component. Because x' is the projection of x on the principal component as shown in the picture below.



(Picture is from Lecture5_Sec01 page 14)

Problem 2

(a).

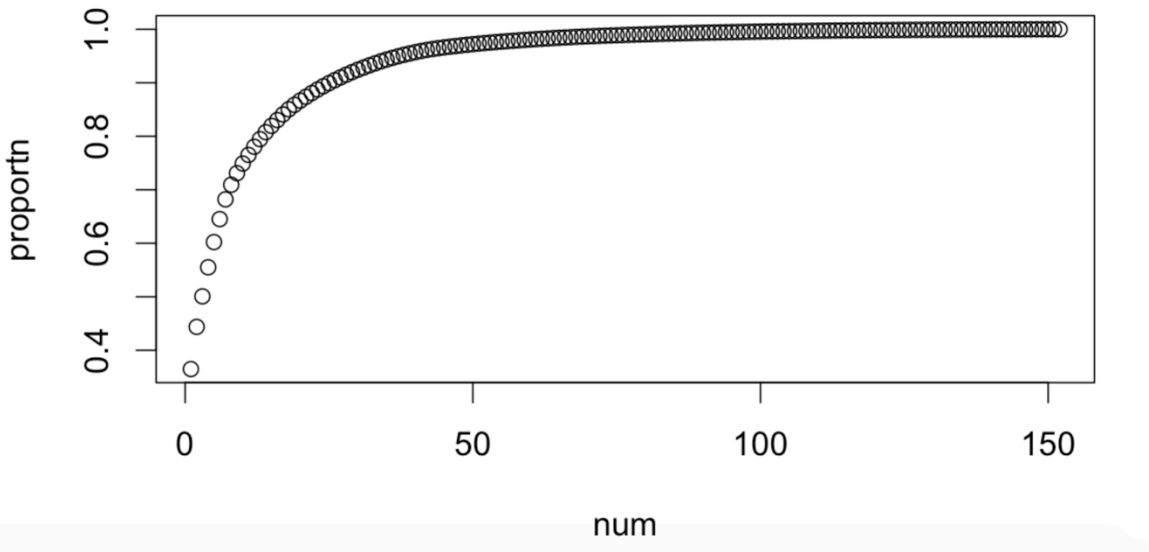
The matrix is a 152x32256 matrix. There are 38 people with each one having 4 pictures. Each picture has 192x168 pixels.

```
> dim(faces_matrix)
[1]   152 32256
```

(b).

(c).

The number of components on the x-axis against the proportion of the variance explained on the y-axis is plotted below:



(d).

Each eigenface describes some features that can represent a human face. The features that contribute little about the data are disposed thus the dimension is deducted. The eigenfaces themselves form a basis set of all images used to construct the covariance matrix. This produces dimension reduction by allowing the smaller set of basis images to represent the original training images.

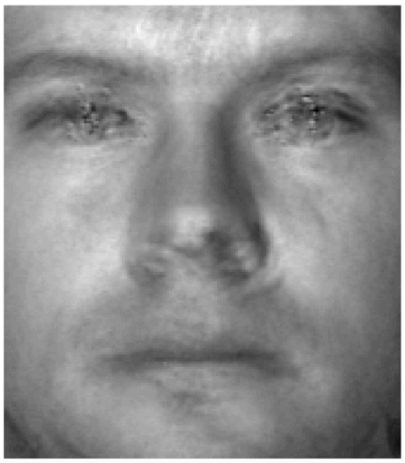
(e).

Add in one eigenface at a time:

Add in five eigenfaces at a time:

I feel that I need 10 to 205faces to recognize the person.

(f).



It does not look like the original image. Since the picture s of subject 01 are removed from the image matrix, the eigenface of the first person is not in the loadings. Thus when reconstructing, we can just use other eigenfaces and form a very ambiguous face, which cannot be recognized as the first person.