Problem 1

a)

```
> sample_data <- read.csv("C://Users//ZIANG//Desktop//Data Mining//hw02_q1_p1_fall14.c
sv", header=T)
> cmean
      X1        X2        X3        X4       X5
15.866547  4.922694 13.143645 19.943426 24.595993
> rmean
  [1]  -1.4963055  24.7877832  -9.4443165  29.9734646 -10.4769774
  [6]  68.0682593  25.8629463  22.3463254  32.0205342 -12.6225711
 [11]  51.2438451   5.6902059  17.5638801  15.4494793  28.3819340
 [16]  10.2458163  -6.2263939  11.0289858  36.4028631  24.2813225
 [21]  41.8923840  34.4501909  29.1546149  -0.6599466   8.6705280
 [26]  29.6476549  -0.5557479  41.2031234  19.4950994   6.1090312
 [31]   8.9500287  22.3056921  -7.5220821   6.5751533  11.3321260
 [36]   9.9569694  21.9595265  14.6915955 -37.9384409  -2.4045728
 [41]  14.2572719  44.1719903  19.7666482  17.0388407  -5.8168371
 [46] -28.1142312  19.6730670  59.8437154  14.7865828   0.9392085
 [51]   9.5990458 -12.0350815   4.7637179  33.1262819   7.6702293
 [56]  -4.0669976  -3.7643030  10.8919496  34.3066107  23.4847180
 [61]  12.8944661   7.7919638  11.4116519  11.3914620  31.7027582
 [66] -14.7776667  40.4560945  28.0807579  -3.2489286  -3.5642117
 [71]  29.1712641  51.6379669  23.8415396  22.9853374   5.1969964
 [76]  18.8598974  19.6464816 -53.5346252   6.2789934  30.5016983
 [81]   0.7811300  29.3540016  -1.3947234  16.9782045  20.7448684
 [86]  25.9598794  24.0144457  70.0341197  12.6005819  32.9692630
 [91]  10.3509370  29.5625354 -11.5106471  26.4919323  21.2450021
 [96]   5.7807402  20.4478471  -2.3253249   6.6729239  73.0480414
```

There are 100 samples in data set, each with 5 dimension values.

The column mean tells us the sample mean in each dimension; The row mean tells us the mean of 5 dimension in each sample.

So the data needs to be centered.

b)

```
> emp_cov_matrix
         [,1]     [,2]    [,3]    [,4]    [,5]
[1,] 430.5585 127.60601 348.8711 593.7409 619.3303
[2,] 127.6060  39.11868 105.1028 164.8167 193.8888
[3,] 348.8711 105.10277 285.0393 466.9367 515.1161
```

```
[4,] 593.7409 164.81669 466.9367 916.5362 764.1611
[5,] 619.3303 193.88884 515.1161 764.1611 973.7700
```

The diagonal values of the covariance matrix is the variance of X_i, since

$$Var(X_i) = \frac{1}{n}\sum_i (X_i - \overline{X})^2, \overline{X} = 0$$

But it is biased because the sample variance should be

$$Var(X_i) = \frac{1}{n-1}\sum_i (X_i - \overline{X})^2$$

The off diagonal values of the matrix is the covariance of $(X_i, X_j), i \neq j$

c)

```
> egvalue_emp
[1] 2.461116e+03 1.837326e+02 1.725521e-01 1.074336e-03 3.430386e-05
> egvector_emp
           V1          V2          V3          V4          V5
[1,] -0.4182265  0.01354844  0.492631175  0.76143366 -0.04935904
[2,] -0.1241798 -0.07950568 -0.174935407 -0.01670828 -0.97333276
[3,] -0.3392392 -0.09576374 -0.837230088  0.36969822  0.19523085
[4,] -0.5750285  0.74781676  0.005118682 -0.33135407  0.01704676
[5,] -0.6032619 -0.65199286  0.160411062 -0.41649408  0.10854193
```

This matrix have the same left eigenvectors as right eigenvectors mainly because it is a symmetric matrix.

d)

With the definition, the loading is the matrix of eigenvectors versus first k large eigenvalues.(k < p)

$$W = \begin{bmatrix} w_{11} & \dots & w_{1p} \\ w_{21} & \dots & w_{2p} \\ w_{k1} & \dots & w_{kp} \end{bmatrix}$$

And the scores is defined as

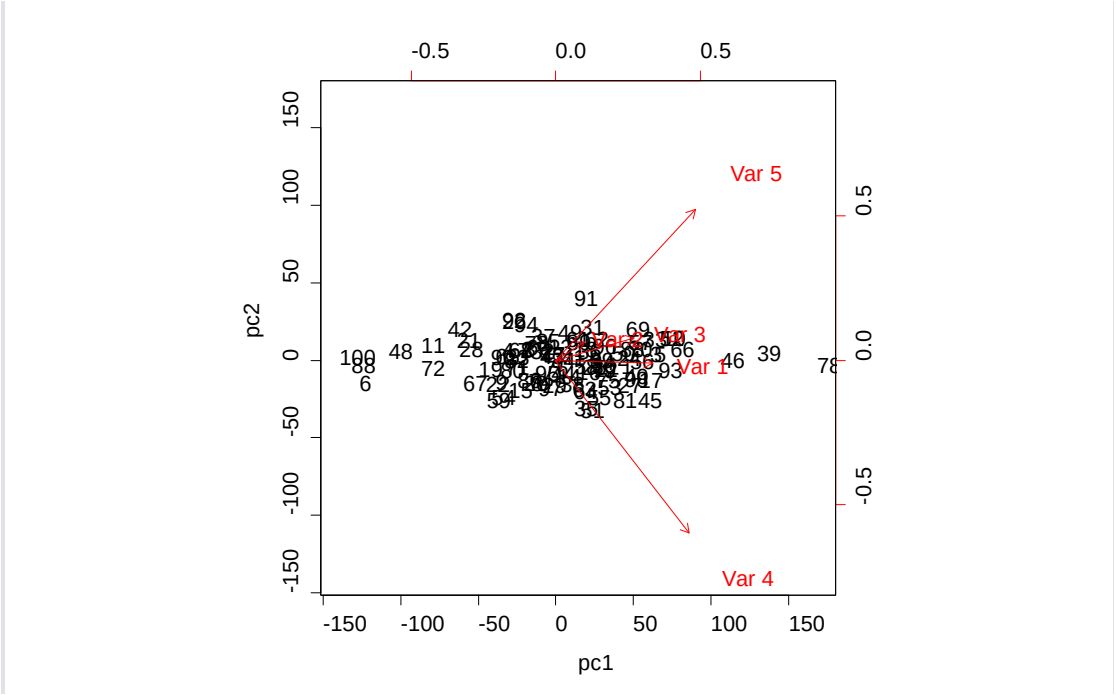
$$Y = \begin{bmatrix} Y_1 & \dots & Y_k \end{bmatrix}^T = W \begin{bmatrix} X_1 & \dots & X_p \end{bmatrix}^T$$

```
> loadings_pca
           V1          V2          V3          V4          V5
[1,] -0.4182265  0.01354844  0.492631175  0.76143366 -0.04935904
[2,] -0.1241798 -0.07950568 -0.174935407 -0.01670828 -0.97333276
[3,] -0.3392392 -0.09576374 -0.837230088  0.36969822  0.19523085
```

```
[4,] -0.5750285  0.74781676  0.005118682 -0.33135407  0.01704676
[5,] -0.6032619 -0.65199286  0.160411062 -0.41649408  0.10854193
> scores_pca
              PC1         PC2          PC3          PC4           PC5
 [1,]  41.8871281  -3.2434231 -0.3311582113 -0.0070941578 -3.712614e-03
 [2,] -22.2945096   8.3906804 -0.2913241580 -0.0074474608  8.732272e-03
 [3,]  60.8822403   4.8744255 -0.0694293388  0.0444565126  5.129981e-05
 [4,] -34.8352017   7.5567732 -0.3638895759  0.0043118418 -6.132391e-03
 [5,]  63.3914699   4.5707796 -0.0902014884 -0.0041798625  5.730350e-03
 [6,] -126.6908970 -14.1490476  0.1397208424  0.0254255478 -2.306360e-03
 [7,] -25.0054984   7.2750998  0.5331795888  0.0061345470  1.982241e-03
 [8,] -15.8229838 -14.0484161  0.7538612273  0.0230207526  3.317957e-03
 [9,] -39.0427493 -13.2168308 -0.9441325723 -0.0249205665 -1.248192e-02
[10,]  68.1816946  15.1879716  0.4355163627  0.0388199440 -7.909478e-03
[11,] -86.6775551  10.9569332  0.2556798247  0.0274499811  1.156879e-03
[12,]  24.3584971  -4.4373911  0.3735131920 -0.0227709589 -5.457288e-03
[13,]  -4.7575414   9.4078614 -0.4642985160  0.0096732188  1.558394e-03
[14,]   0.8791229  -9.1091508  0.0533418287 -0.0066418536 -1.008777e-02
[15,] -30.2987570 -18.2455319  0.4816475942 -0.0148012343  3.604208e-03
[16,]  13.4876909  -6.6327443 -0.3286546301 -0.0203106156  7.724907e-03
[17,]  53.6077806 -11.7710602 -0.0829796282  0.0554165928  1.492213e-03
[18,]  11.2147439   8.1079253 -0.8979743860 -0.0242808055  5.973202e-04
[19,] -50.0060477  -4.5264943 -0.5546377581  0.0775633196 -1.381381e-02
[20,] -20.4659872 -13.4186930  0.3325067695  0.0126200739 -6.432934e-03
[21,] -64.0223334  14.1172793 -0.1206209765 -0.0058083562  3.668049e-03
[22,] -45.0476540 -14.3732510 -0.1138176027 -0.0017892870  4.877393e-03
[23,] -32.7720347   2.8042604  0.0566522752 -0.0017642453  2.889241e-03
[24,]  39.6262463   1.9361872  0.0322066689 -0.0120633677  1.560887e-03
[25,]  17.1795142  -3.8031640 -0.0284388001  0.0290219630  3.261970e-03
[26,] -34.8224112  26.3644606  0.6598116230  0.0135481037  2.288368e-05
[27,]  39.9850725 -15.2607256 -0.2791579764  0.0172133620 -6.943953e-03
[28,] -62.2022571   7.7398328  0.2319444504  0.0170374430  4.925735e-03
[29,]  -8.8200698 -15.2538843  0.4739760532 -0.0246856056 -8.642565e-03
[30,]  23.2937510  -2.6897359  0.3112658451 -0.0393165269 -5.475175e-03
[31,]  15.7333547  22.5392799 -0.4661913043  0.0108594804 -5.413941e-03
[32,] -16.3610563   7.0463591  0.5385230136  0.0042540037 -9.749105e-03
[33,]  55.9073579  14.6917644 -0.1222708203  0.0055364865 -8.109748e-03
[34,]  22.0156850   0.1714335  0.6174090785 -0.0427222561  1.619161e-04
[35,]  11.5575119 -29.6527401 -0.1409907636 -0.0199908041  7.874120e-04
[36,]  13.8087805   6.7016265 -0.5382750571  0.0168262032 -3.766139e-03
[37,] -15.6361235  16.3369692 -0.5920658774 -0.0491696344 -1.287979e-03
[38,]   2.8616790 -14.7029400  0.2222226491 -0.0185065843  4.103990e-03
[39,] 129.8837941   5.8348082  0.6479276780  0.0118478665  3.831259e-03
[40,]  44.1661319  -9.5375970  0.4237163546  0.0100815610 -2.900960e-03
```

```
[41,]   3.4093930   2.6542802  0.0002099179  0.0365042727  1.726684e-03
[42,] -69.7910839  20.8274448  0.0412119308  0.0382049691  6.360098e-03
[43,] -10.0722431   3.8019421  0.3130442582 -0.0649817841  2.938511e-03
[44,]  -2.9741503  -7.4178208 -0.2888873090  0.0006448539 -5.731849e-03
[45,]  53.1312746 -24.6990864 -0.7362534952 -0.0001759880 -3.365916e-03
[46,] 106.2844093   1.1227960  0.0650746421 -0.0091013619  5.731701e-03
[47,]  -9.8063730   4.2301174  0.1044083628  0.0309670316  4.428173e-03
[48,] -107.5124841   6.8471840  0.8134734078  0.0386075975 -3.521322e-03
[49,]   1.5704933  19.0745596  0.0671921298 -0.0568767129  3.819966e-03
[50,]  35.5084035   5.2272471  0.8012842207 -0.0004554016  4.188183e-03
[51,]  15.7542286 -31.4357027  0.3477520219  0.0670328271 -3.758717e-03
[52,]  66.8468692  14.9389455 -0.1211937224 -0.0090288374  3.709506e-03
[53,]  27.1501322 -16.4536221 -0.5337846858 -0.0171906858 -9.848003e-03
[54,] -41.5220569 -23.0242594 -0.3374469171 -0.0126561276  8.153512e-03
[55,]  20.1806973 -22.9813055  0.1232693728 -0.0314559151  1.345906e-02
[56,]  47.8545419   0.6169291  0.5903275646  0.0456249966 -4.140606e-03
[57,]  46.9382259  12.5549426 -0.5887335478  0.0411635755  3.564889e-04
[58,]  11.8365020  -2.7013528 -0.5250755263  0.0109926639 -2.792016e-03
[59,] -44.4580048 -24.5380345  0.3358061726 -0.0462183343  1.508587e-03
[60,] -19.2223686   9.2110014  0.0836569402 -0.0274520719 -1.558312e-03
[61,]   6.3705704  14.5170901 -0.2226497642  0.0182324575  8.371734e-03
[62,]  18.7718165  14.5383600 -0.3073003351  0.0249158104  5.171174e-03
[63,]  10.9526065 -17.6326088  0.0348597011 -0.0076674309 -1.192052e-03
[64,]  11.1362514 -19.3916258 -0.3975438804  0.0096395608 -4.064129e-03
[65,] -38.8746308   3.6506842 -0.5753954115 -0.0087362627 -3.299853e-03
[66,]  73.7078507   8.3480756 -0.1227424818 -0.0412115695  7.717463e-03
[67,] -59.6154726 -13.4515892 -0.3316930232  0.0161870709  8.067226e-03
[68,] -30.3890089   7.1796940  0.5093535764 -0.0237424391  8.248469e-03
[69,]  45.2718131  20.8469401  0.3279062903  0.0729363448 -4.536266e-03
[70,]  46.4701677   7.9198421  0.1533233259  0.0299783444  3.721182e-03
[71,] -32.4802813  -3.4687648 -0.7123547619 -0.0169041509 -3.070191e-03
[72,] -87.0896817  -3.9597704 -0.1068043160  0.0419047861 -6.063423e-03
[73,] -20.2503621  12.2445451  0.5030089518  0.0314055703  8.008349e-03
[74,] -17.2963844 -12.2891908 -0.0734653601 -0.0068570019  1.193524e-02
[75,]  25.8273252 -11.8215278  0.1450686188 -0.0174068634  8.484362e-03
[76,]  -7.6045477  -3.9249706  0.2399460446 -0.0279681944  1.636474e-03
[77,]  -9.6154842   4.7662698 -0.8125967364 -0.0492570298  3.848540e-03
[78,] 168.0683235  -2.0055017  0.2383726993  0.0415255235 -2.218807e-03
[79,]  23.0410662  -5.2666990 -0.1278621663 -0.0065352546  5.906008e-03
[80,] -35.7248690  -5.6092142 -0.1878605960  0.0166642508 -4.277030e-03
[81,]  36.9758303 -25.1343878  0.2174172251  0.0617083470  1.067642e-03
[82,] -33.3121683   0.8071598  0.6777028967 -0.0801531941  1.065337e-03
[83,]  41.2773757   8.2278362 -0.3641002078 -0.0164690055 -3.081881e-03
[84,]  -3.0729187   1.1380014 -0.4429288388  0.0187425270 -3.373909e-05
```

```
[85,]  -12.5542496  13.2441430 -0.7107501741  0.0601462888  1.398987e-02
[86,]  -24.5928779 -12.0659305  0.4242169904  0.0643076001  1.665857e-03
[87,]  -20.6240264   8.5931873  0.8627571590 -0.0558255946 -1.125682e-02
[88,] -131.8307555  -2.0141126 -0.0307595173 -0.0180883967  2.281768e-03
[89,]    7.0073865  12.2907824  0.6167864057 -0.0359549327 -9.722926e-03
[90,]  -42.0805461   3.2359407  0.2285259954 -0.0386455738  1.302940e-03
[91,]   11.7122824  40.6346788 -0.2662129508 -0.0223396370 -8.367447e-04
[92,]  -34.4971900  26.7618569 -0.1040493225  0.0113994245 -1.262703e-03
[93,]   66.1680708  -5.6983250  0.2351676621 -0.0656105975 -1.174960e-02
[94,]  -26.9244225  23.9910917 -0.3296758467 -0.0517043294 -1.134201e-02
[95,]  -13.1936515  -7.3850314 -0.2701742944 -0.0167614598  2.168203e-03
[96,]   23.7629177   9.9097310 -0.1154699472 -0.0124921521  2.550852e-03
[97,]  -10.9475070 -17.2814732 -0.2435844585 -0.0235130232  9.144441e-04
[98,]   44.1353156 -10.7580628 -0.3516893411 -0.0339421612  5.681609e-03
[99,]   21.7997728   0.8748118  0.3833244532 -0.0233790919 -5.230643e-04
[100,] -139.3196924   3.0419686  0.1004765187  0.0344672508 -2.351879e-03
```



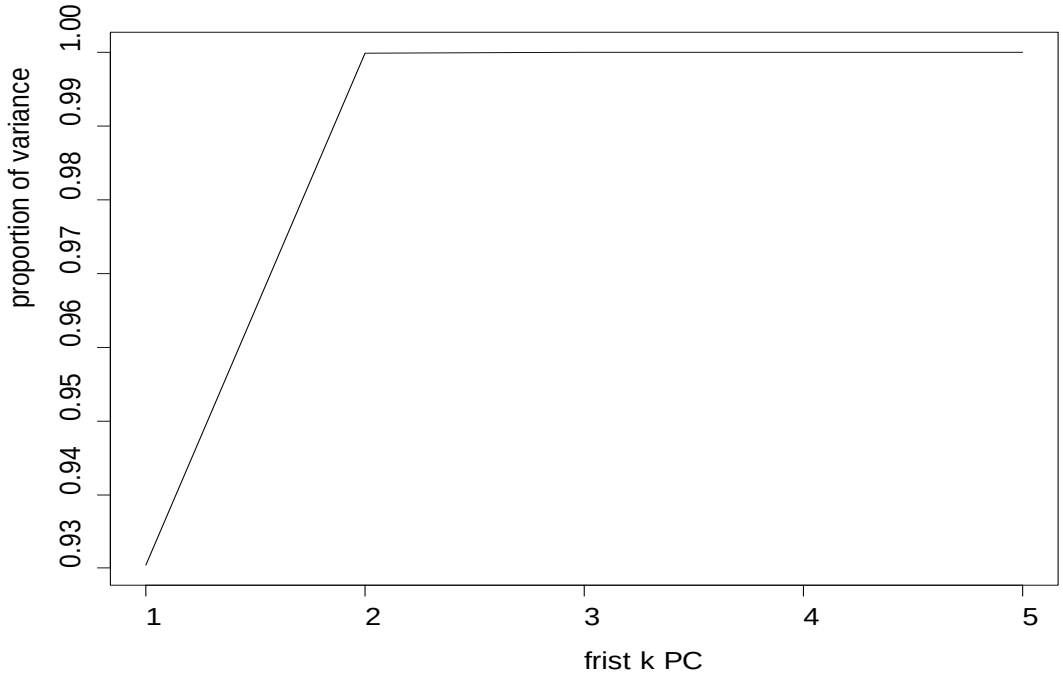
e)

Proportion of variance included first k dimensions is the contribution of first k principal component.

$$C_k = \frac{\sum_{i=k}^{k} \lambda_i}{tr\{\sum\}}$$

```
> ctrb_this
K        1         2         3         4         5
[1] 0.9304708 0.9999343 0.9999996 1.0000000 1.0000000
```

Then we plot the cumulative proportion of principal component included.



As the first two principal component have included 99.9% of the information in original data, we can use the first two component to describe the trait of original data, in which way we reduce data dimension from five to two.

f)

```
> new.scores
          PC1       PC2        PC3         PC4          PC5
[1,] 26.468450 15.161604 -0.2812722 -0.06447507 -0.0013566630
[2,] -7.764959 -8.047896 -0.2169575 -0.02003201  0.0002787115
[3,] -4.323813  2.285090 -0.1495802 -0.06268190 -0.0028327836
[4,] 49.879498 -5.293422 -0.3982574  0.05095983  0.0105162280
[5,] 18.695721 -5.593545 -0.6772342 -0.06048565 -0.0051370948
```

g)

```
> scores_2
          PC1       PC2
[1,] 26.468450 15.161604
[2,] -7.764959 -8.047896
[3,] -4.323813  2.285090
[4,] 49.879498 -5.293422
[5,] 18.695721 -5.593545
```

As the scores is the projection of original data in Euclidean space, we know that

$Y = (Y_1, Y_2)$ is the coordinates of projection in data space x'. For example

$y_1 = (26.468450, 15.161604, 0, 0, 0)$

$y_2 = (-7.764959, -8.047896, 0, 0, 0)$

So the distance of each sample x and scores y could be defined as

$$Dist(x_i, y_i) = \sqrt{(\sum_{j=1}^{5} x_{ij}^2) - (\sum_{j=1}^{2} y_{ij}^2)}$$

```
> dist
sample    xy1     xy2       xy3       xy4       xy5
[1] 0.2885705 0.2178805 0.1622075 0.4016422 0.6799493
```

h)

For each sample x in original space x, the x' is defined as

$$x' = W^T \cdot y$$

where $y = \begin{bmatrix} Y_1 = y_1 \\ Y_2 = y_2 \end{bmatrix}$

Then the error of 5 new observations would be $x - W^T \cdot y$

```
> error
         [X1]       [X2]       [X3]        [X4]        [X5]
[1,] -0.1875900 0.05160222 0.2113884  0.019901207 -0.018412946
[2,] -0.1221468 0.03801696 0.1742919  0.005531904 -0.026428909
[3,] -0.1212761 0.02997141 0.1015066  0.019955959  0.001804852
[4,] -0.1579105 0.05858208 0.3543259 -0.018745033 -0.083967907
[5,] -0.3794289 0.12448295 0.5436365  0.016488048 -0.084001534
```

The direction of the error should be vertical to the plane constructed by directions of first two basis vector of W (the first two principal component), because, for example the first principal component $[Y_1, Y_2]^T = \begin{bmatrix} w_{11}, w_{12}, ..., w_{1p} \\ w_{21}, w_{22}, ..., w_{2p} \end{bmatrix} \cdot \begin{bmatrix} X_1 & ... & X_p \end{bmatrix}^T$ included more than 99% information of the origin data so that the 7% information remain in the error, which should be explained by other three components.

To test our hypothesis, we can calculate the inner product of error and first two principal component. And we have

```
> inner_product <- error%*%loadings_pca[,1:2]
> inner_product
            [,1]         [,2]
[1,] -2.827599e-16 -6.919812e-15
[2,] -7.771561e-16  4.309053e-15
[3,] -1.216692e-15 -1.269384e-15
[4,]  8.201773e-15 -4.683753e-15
[5,]  2.914335e-15  1.117162e-15
```

So we can conclude our hypothesis that the direction of error should be vertical to the plane constructed by directions of first two basis vector of W.

Problem 2

a)

```
> print(dim(sample.matrix))
[1]   152 32256
```

So there are 38*4 rows and 192*168 columns in the matrix which has a size of 152*32256.

b)

The figure of the mean face:

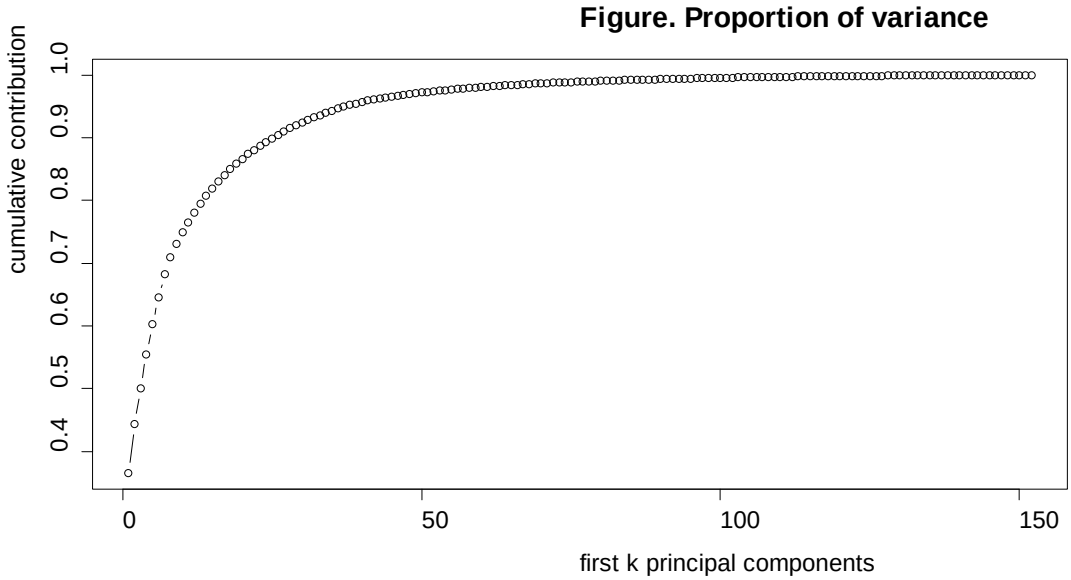
Figure. Mean face

c)

Let $C_{cum\,k}$ be the cumulative contribution (cumulative variance) of first k principal components. C_i be the contribution(variance or eigenvalue) of No.k principal component. Then we have,

$$C_{cum\,k} = \frac{C_k}{\sum_{k=1}^{p} C_i}$$

It is as same as proportion of variance of principal components.



d)

```
> print(dim(sample.loadings))
[1] 32256   152
```

Since loadings is a 32256*152 matrix, in which each column is a 192*168 size picture, the first 9 pictures shown as follow

Figure. Eigenfaces

The eigenfaces are like eigenvalues, which describe the most important traits among

all of the faces.

e)

First, we use first 24 eigenfaces to reconstruct the original face, as follow:

Figure. Use 24 Eigenfaces

Then we use the first 120 eigenfaces to reconstruct the original face, as follow:

Figure. Use 120 Eigenfaces

Generally, we always reduce the original p dimensions to first k principal components if these first k large components can descript more than $1-\alpha = 0.95*100\%$ of the whole data set. Since we deal with some pictures in this case, it would be better for reconstruction if we control $\alpha < 0.01$. The cumulative contribution of the principal component is shown as follow:

```
> sample_contribution
  [1] 0.3649001 0.4437039 0.5007963 0.5550613 0.6022067 0.6451596 0.6821229 0.7093335
  [9] 0.7309588 0.7489765 0.7651229 0.7804113 0.7946280 0.8076941 0.8194221 0.8303418
 [17] 0.8408504 0.8504347 0.8587935 0.8665934 0.8737712 0.8807132 0.8872907 0.8934396
 [25] 0.8992588 0.9047981 0.9100002 0.9150709 0.9198139 0.9243274 0.9284223 0.9324743
 [33] 0.9362720 0.9400380 0.9433515 0.9465605 0.9494824 0.9522651 0.9547972 0.9572239
 [41] 0.9593707 0.9614430 0.9631259 0.9647431 0.9661357 0.9674279 0.9686621 0.9698850
 [49] 0.9710353 0.9721517 0.9731907 0.9742106 0.9751680 0.9760768 0.9769448 0.9777660
 [57] 0.9785685 0.9793544 0.9801200 0.9808548 0.9815551 0.9822310 0.9828774 0.9835203
 [65] 0.9841391 0.9847207 0.9852743 0.9858072 0.9863031 0.9867858 0.9872248 0.9876480
 [73] 0.9880589 0.9884455 0.9888233 0.9891937 0.9895586 0.9899189 0.9902731 0.9906078
 [81] 0.9909356 0.9912532 0.9915619 0.9918670 0.9921650 0.9924535 0.9927285 0.9929872
 [89] 0.9932401 0.9934885 0.9937285 0.9939631 0.9941907 0.9944113 0.9946241 0.9948348
 [97] 0.9950378 0.9952367 0.9954292 0.9956149 0.9957979 0.9959741 0.9961434 0.9963104
[105] 0.9964760 0.9966342 0.9967876 0.9969351 0.9970800 0.9972179 0.9973519 0.9974842
[113] 0.9976119 0.9977357 0.9978561 0.9979706 0.9980821 0.9981841 0.9982847 0.9983806
[121] 0.9984666 0.9985507 0.9986331 0.9987133 0.9987891 0.9988621 0.9989295 0.9989948
```

```
[129] 0.9990554 0.9991152 0.9991739 0.9992292 0.9992818 0.9993335 0.9993837 0.9994328
[137] 0.9994807 0.9995274 0.9995736 0.9996169 0.9996595 0.9997002 0.9997390 0.9997775
[145] 0.9998142 0.9998498 0.9998841 0.9999150 0.9999447 0.9999732 1.0000000 1.0000000
```

So we can use the first 80 eigenfaces to reconstruct original face (the first 80 principal components contribute 99.06% of the information in original picture.

f)

We first remove rows 17:20 (which is the index of subject 05)from the sample matrix and do principal component analysis, obtaining the loadings W

Then we calculate the scores of yaleB05 P00A+010E+00.pgm. by

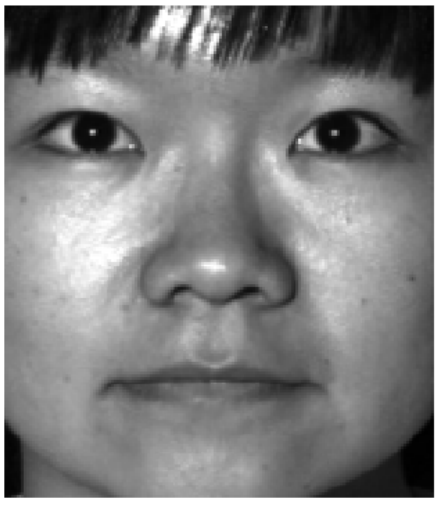
$$y = W \cdot x$$

$$y = \left[Y_1 = y_1, \ldots . Y_k = y_k\right]^T, x = \left[X_1 = x_1, \ldots . X_p = x_p\right]^T$$

Then we reconstruct the face by

$$x' = W^T \cdot y$$

The reconstruct picture is as follow, and we compare it with the original one.



It is not very like the original face, at least not as good as the reconstruction in Q.f. In my opinion, the main reason may be that we remove the 4 faces of her from original

data so that the original data does not contain her information as same as in those principal components. This is like we use original data to predict or simulate subject 05's face. Since there is no enough information about her, we cannot reconstruct her face very precisely.