Mark Klobukov

Professor Matthew Burlick

CS 383 Homework 1

January 17, 2018

Part I: Theory Questions

1) Consider the following data:

$$\begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \\ -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

a) Find the principle components of the data.

First step in the PCA is to standardize the data. Standardization means subtracting mean from each data point and then dividing it by the standard deviation.

$$\mu_{column\ 1} = \frac{-2-5-3+0-8-2+1+5-1+6}{10} = -0.9$$

$$\mu_{column\ 2} = \frac{1-4+1+3+11+5+0-1-3+1}{10} = 1.4$$

$$\sigma_{column\ 1} = 4.2282$$

$$\sigma_{column2} = 4.2740$$

The data looks as follows after standardization:

$$
\begin{bmatrix}
-0.2602 & -0.0936 \\
-0.9697 & -1.2635 \\
-0.4967 & -0.0936 \\
0.2129 & 0.3744 \\
-1.6792 & 2.2462 \\
-0.2602 & 0.8423 \\
0.4494 & -0.3276 \\
1.3954 & -0.5615 \\
-0.0237 & -1.0295 \\
1.6319 & -0.0936
\end{bmatrix}
$$

Now that the data is standardized, it is necessary to compute the covariance matrix defined as:

$$
\Sigma = \begin{bmatrix}
\mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\
\mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)]
\end{bmatrix}.
$$

We only have two features in the provided dataset, so the covariance matrix will be 2x2.

Because the data is already centered, a special formula can be used to calculate covariance. Let the data matrix be called A.

$$
\Sigma(A) = \frac{A^T A}{N-1} = \frac{1}{9} \times
\begin{bmatrix}
-0.2602 & -0.0936 \\
-0.9697 & -1.2635 \\
-0.4967 & -0.0936 \\
0.2129 & 0.3744 \\
-1.6792 & 2.2462 \\
-0.2602 & 0.8423 \\
0.4494 & -0.3276 \\
1.3954 & -0.5615 \\
-0.0237 & -1.0295 \\
1.6319 & -0.0936
\end{bmatrix}^T
\begin{bmatrix}
-0.2602 & -0.0936 \\
-0.9697 & -1.2635 \\
-0.4967 & -0.0936 \\
0.2129 & 0.3744 \\
-1.6792 & 2.2462 \\
-0.2602 & 0.8423 \\
0.4494 & -0.3276 \\
1.3954 & -0.5615 \\
-0.0237 & -1.0295 \\
1.6319 & -0.0936
\end{bmatrix} =
$$

$$
= \begin{bmatrix} 1 & -0.4083 \\ -0.4083 & 1 \end{bmatrix}
$$

Performing the PCA involves solving the equation of the form $\Sigma w = \lambda w$, where Σ is the covariance matrix. Finding λ, the vector of eigenvalues, requires solving the following equation:

$$
\det(\Sigma - \lambda I) = 0
$$

$$\det\left(\begin{bmatrix} 1-\lambda & -0.4083 \\ -0.4083 & 1-\lambda \end{bmatrix}\right) = (1-\lambda)^2 - (-0.4083)^2 = (1-\lambda)^2 - 0.4083^2$$

Using quadratic formula, we get the values for λ:

$$\lambda = \begin{bmatrix} 0.5917 \\ 1.4083 \end{bmatrix}$$

These eigenvalues can now be pasted into the equation $(\Sigma - \lambda I)w = 0$ to find w, the eigenvectors.

$$(\Sigma - \lambda I)w = \left(\begin{bmatrix} 1 & -0.4083 \\ -0.4083 & 1 \end{bmatrix} - \begin{bmatrix} 0.5917 & 0 \\ 0 & 1.4083 \end{bmatrix} \right) w = 0$$

This gives us two equations:

$$0.4083w_1 - 0.4083w_2 = 0$$

$$-0.4083w_1 - 0.4083w_2 = 0$$

First equation has a solution [1, 1] and the second equation can be solved as follows:

$$w_1 = \frac{-0.4083w_2}{0.4083}$$

Let $w_2 = 1$. Then $w_1 = -1$

Thus, the two eigenvectors are $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$ and $\begin{bmatrix} -1 \\ 1 \end{bmatrix}$. They need to be normalized:

$$|e_1| = \sqrt{2}, PC_1 = [\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}]$$

$$|e_2| = \sqrt{2}, PC_2 = \left[-\tfrac{1}{\sqrt{2}}, \tfrac{1}{\sqrt{2}}\right]$$

b) Project the data onto the principal component corresponding to the largest eigenvalue found in the previous part.

The largest of the two eigenvalues is $\lambda = 1.4083$. Its corresponding eigenvector is $\left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T$

To project data onto the principal component, it is necessary to calculate the dot product of every standardized data point with this eigenvector. The resulting 1x1 matrices will give us the projection of data onto 1-dimensional space.

All of the dot products can be done in one matrix multiplication:

$$\begin{bmatrix} -0.2602 & -0.0936 \\ -0.9697 & -1.2635 \\ -0.4967 & -0.0936 \\ 0.2129 & 0.3744 \\ -1.6792 & 2.2462 \\ -0.2602 & 0.8423 \\ 0.4494 & -0.3276 \\ 1.3954 & -0.5615 \\ -0.0237 & -1.0295 \\ 1.6319 & -0.0936 \end{bmatrix} \left[-\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right]^T = \begin{bmatrix} 0.1178 \\ -0.2077 \\ 0.2850 \\ 0.1142 \\ 2.7756 \\ 0.7796 \\ -0.5494 \\ -1.3838 \\ -0.7112 \\ -1.2201 \end{bmatrix}$$

2) Consider the following data:

$$\text{Class } 1 = \begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \end{bmatrix}, \text{Class } 2 = \begin{bmatrix} -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

a) Compute information gain for each feature

IG is computed using the following formula:

$$IG(A) = H\left(\frac{p}{p+n}, \frac{n}{p+n}\right) - remainder(A)$$

where remainder(A) is:

$$remainder(A) = \sum_{i=1}^{k} \frac{p_i + n_i}{p+n} H\left(\frac{p_i}{p_i + n_i}, \frac{n_i}{p_i + n_i}\right)$$

There are two features in this dataset. Feature one has the following set of possible values:

$$\{-2, -5, -3, 0, -8, 1, 5, -1, 6\}$$

Let p and n represent the number of samples with label 1 and label 2, respectively.

$$p_{-2} = 1, n_{-2} = 1$$

$$p_{-5} = 1, n_{-5} = 0$$

$$p_{-3} = 1, n_{-3} = 0$$

$$p_0 = 1, n_0 = 0$$

$$p_{-8} = 1, n_{-8} = 0$$

$$p_1 = 0, n_1 = 1$$

$$p_5 = 0, n_5 = 1$$

$$p_{-1} = 0, n_{-1} = 1$$

$$p_6 = 0, n_6 = 1$$

Total number of samples = 10

samples with class 1 = 5

samples with class 2 = 5

$$remainder(F_1) = \frac{2}{10} \times \left[-\frac{1}{2}\log\left(-\frac{1}{2}\right) - \frac{1}{2}\log\left(-\frac{1}{2}\right)\right] +$$
$$+ 8 \times \left(\frac{1}{10}[-1\log(1) - 0\log(0)]\right) = \frac{2}{10}$$

$$Information\ Gain\ (F_1) = H\left(\frac{1}{2}, \frac{1}{2}\right) - \frac{2}{10} = 1 - \frac{1}{5} = \frac{4}{5}$$

Now repeat the procedure for feature 2.

Feature two has the following set of possible values:

$$\{1, -4, 3, 11, 5, 0, -1, -3\}$$

Let p and n represent the number of samples with label 1 and label 2, respectively.

$$p_1 = 2, n_1 = 1$$

$$p_{-4} = 1, n_{-4} = 0$$

$$p_3 = 1, n_3 = 0$$

$$p_{11} = 1, n_{11} = 0$$

$$p_5 = 0, n_5 = 1$$

$$p_0 = 0, n_0 = 1$$

$$p_{-1} = 0, n_{-1} = 1$$

$$p_{-3} = 0, n_{-3} = 1$$

Total number of samples = 10

$$remainder(F_2) = \frac{3}{10} \times \left[-\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right)\right] +$$
$$+7 \times \left(\frac{1}{10}[\,-1\log(1) - 0\log(0)\,]\right) = 0.2755$$

$$Information\ Gain\ (F_2) = H\left(\frac{1}{2},\frac{1}{2}\right) - 0.1909 = 1 - 0.2755 = 0.7245$$

b) $IG(F_1) = 0.8, \quad IG(F_2) = 0.7245$

First feature has a greater information gain value; therefore it is more discriminating than feature 2.

c) Using LDA, find the direction of projection. Normalize this vector to be unit length.

$$\text{Class 1} = \begin{bmatrix} -2 & 1 \\ -5 & -4 \\ -3 & 1 \\ 0 & 3 \\ -8 & 11 \end{bmatrix}, \text{Class 2} = \begin{bmatrix} -2 & 5 \\ 1 & 0 \\ 5 & -1 \\ -1 & -3 \\ 6 & 1 \end{bmatrix}$$

The first step is to standardize the data by subtracting mean from each observation and dividing by the standard deviation (just like in question 1 with the PCA). It is important to standardize the entire data matrix (instead of within classes). Otherwise, the LDA procedure will break down.

$$\mu_{feature1} = -0.9$$

$$\mu_{feature2} = 1.4$$

$$\sigma_{feature1} = 4.2282$$

$$\sigma_{feature2} = 4.2740$$

After standardization, the matrices look as follows:

$$\begin{bmatrix} -0.2602 & -0.0936 \\ -0.9697 & -1.2635 \\ -0.4967 & -0.0936 \\ 0.2129 & 0.3744 \\ -1.6792 & 2.2462 \end{bmatrix} \qquad \begin{bmatrix} -0.2602 & 0.8423 \\ 0.4494 & -0.3276 \\ 1.3954 & -0.5615 \\ -0.0237 & -1.0295 \\ 1.6319 & -0.0936 \end{bmatrix}$$

Compute the means for each class. Class 1:

$$\mu_{feature1} = \frac{-0.2602 - 0.9697 - 0.4967 + 0.2129 - 1.6792}{5} = -0.6386$$

$$\mu_{feature2} = \frac{-0.0936 - 1.2635 - 0.0936 + 0.3744 + 2.2462}{5} = 0.2340$$

Similarly, for class 2:

$$\mu_{feature1} = 0.6386$$

$$\mu_{feature2} = -0.2340$$

Put the means for each class into their own vector:

$$\mu_1 = [-0.6386, 0.2340]\ , \qquad \mu_2 = [0.6383, -0.2340]$$

Now the scatter matrices can be calculated for each class.

$$\sigma_i^2 = (|C_i| - 1)\ cov(C_i)$$

$$\sigma_1^2 = (5-1)\begin{bmatrix} 0.5202 & -0.4123 \\ -0.4123 & 1.6314 \end{bmatrix} = \begin{bmatrix} 2.0808 & -1.6490 \\ -1.6490 & 6.5255 \end{bmatrix}$$

$$\sigma_2^2 = (5-1)\begin{bmatrix} 0.7104 & -0.1328 \\ -0.1328 & 0.4818 \end{bmatrix} = \begin{bmatrix} 2.8415 & -0.5312 \\ -0.5312 & 1.9270 \end{bmatrix}$$

From these scatter matrices, compute the "within class" and "between class" scatter matrices.

$$S_W = \sigma_1^2 + \sigma_2^2$$

$$S_w = \begin{bmatrix} 4.9223 & -2.1803 \\ -2.1803 & 8.4526 \end{bmatrix}$$

$$S_B = (\mu_1 - \mu_2)^T(\mu_1 - \mu_2)$$

$$S_B = \begin{bmatrix} 1.6311 & -0.5976 \\ -0.5976 & 0.2190 \end{bmatrix}$$

Now do the eigen-decomposition on $S_w^{-1}S_B$:

$$S_w^{-1} = \begin{bmatrix} 0.2294 & 0.0592 \\ 0.0592 & 0.1336 \end{bmatrix}$$

$$S_w^{-1}S_B = \begin{bmatrix} 0.2294 & 0.0592 \\ 0.0592 & 0.1336 \end{bmatrix}\begin{bmatrix} 1.6311 & -0.5976 \\ -0.5976 & 0.2190 \end{bmatrix} = \begin{bmatrix} 0.3387 & -0.1241 \\ 0.0167 & -0.0061 \end{bmatrix}$$

Results of eigen-decomposition:

$$Eigenvalues = [0.3326, 0]$$

$$Eigenvectors = \begin{bmatrix} 0.9988 & 0.3440 \\ 0.0492 & 0.9390 \end{bmatrix}$$

The eigenvectors are already normalized. The first column of the matrix corresponds to the eigenvector with an associated non-zero eigenvalue. Therefore, that is the principle component on which the data will be projected.

$$PC = [0.9988, 0.0492]^T$$

d) Project the data X onto the principal component by multiplying it by the projection matrix W (a vector in this case):

$$Z = XW$$

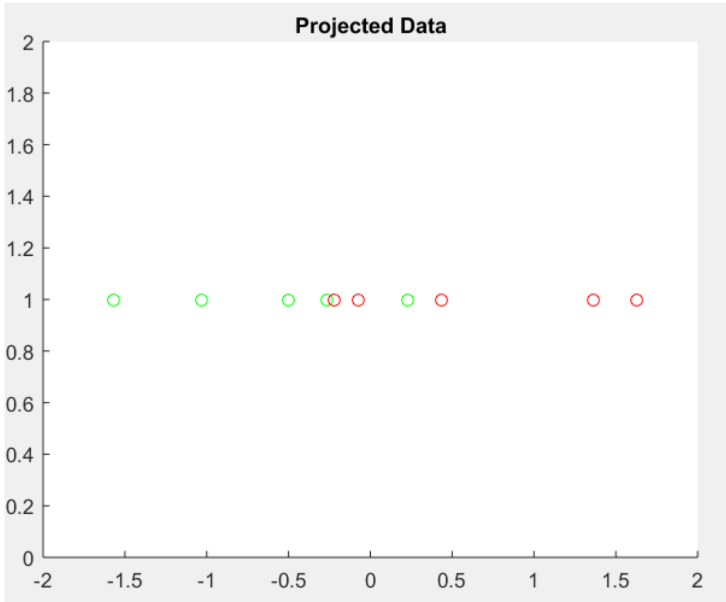
$$Class\ 1\ Projection = \begin{bmatrix} -0.2602 & -0.0936 \\ -0.9697 & -1.2635 \\ -0.4967 & -0.0936 \\ 0.2129 & 0.3744 \\ -1.6792 & 2.2462 \end{bmatrix} [0.9988, 0.0492]^T = \begin{bmatrix} -0.2644 \\ -1.0306 \\ -0.5007 \\ 0.2310 \\ -1.5667 \end{bmatrix}$$

$$Class\ 2\ Projection = \begin{bmatrix} -0.2602 & 0.8423 \\ 0.4494 & -0.3276 \\ 1.3954 & -0.5615 \\ -0.0237 & -1.0295 \\ 1.6319 & -0.0936 \end{bmatrix} [0.9988, 0.0492]^T = \begin{bmatrix} -0.2184 \\ 0.4327 \\ 1.3661 \\ -0.0742 \\ 1.6253 \end{bmatrix}$$

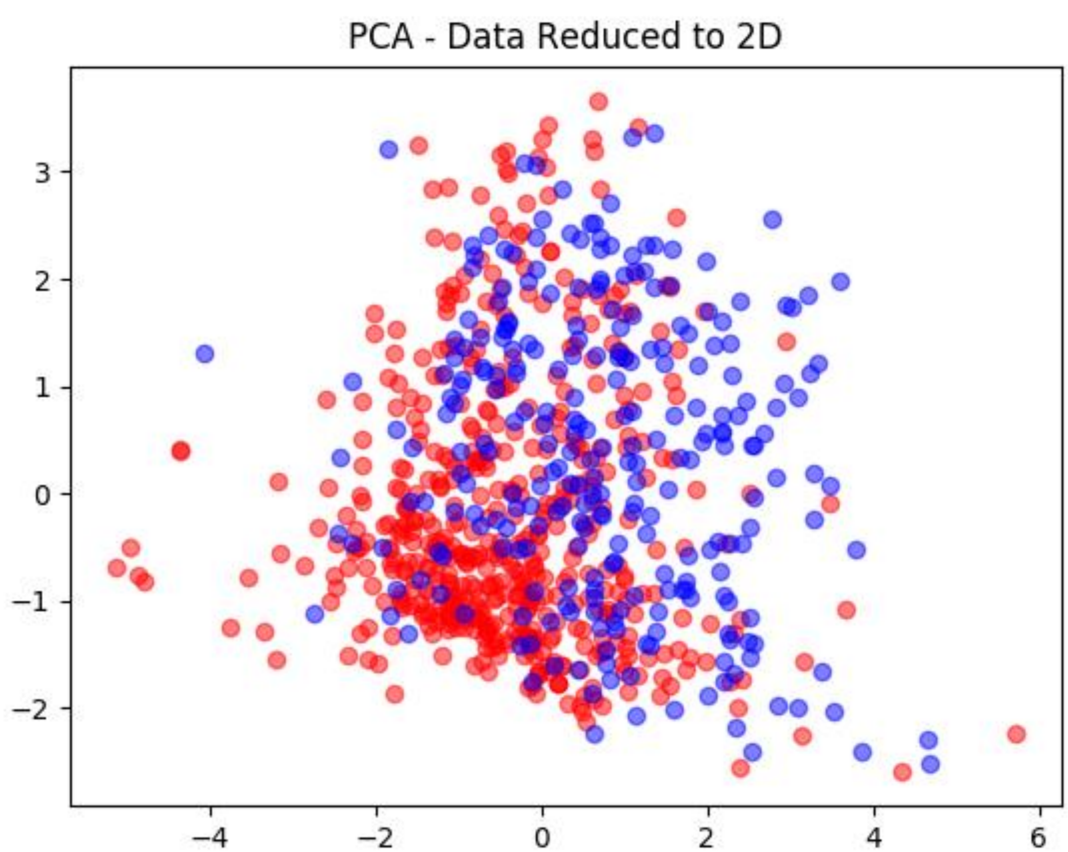
e) Does the projection provide good class separation?

The figure below shows a plot of the projected class data onto the principal component. The data points are plotted along the line y = 1 (in order to create a one-dimensional plot). Green circles represent Class 1, and red circles represent Class 2. Although there is some overlap between the classes, the data from Class 1 is mostly on the left side, and the data from Class 2 is mostly on the right side. The projection did not provide a perfect class separation but it is good enough. Assuming the data provides an accurate sample of the distribution of these classes, most new cases will be identified correctly.

More data would be necessary to interpret this projection correctly. It is currently not possible to answer the question whether the rightmost green dot is in the red range, or if the two leftmost red dots are in the green range.



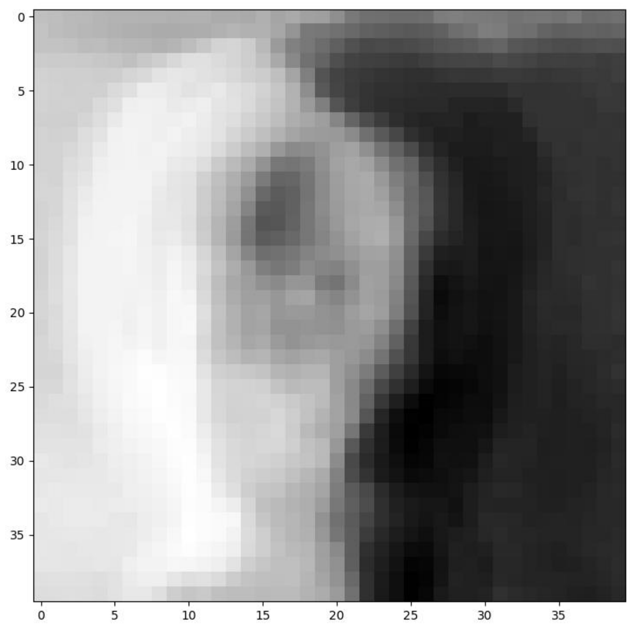
Part 2: Dimensionality Reduction via PCA



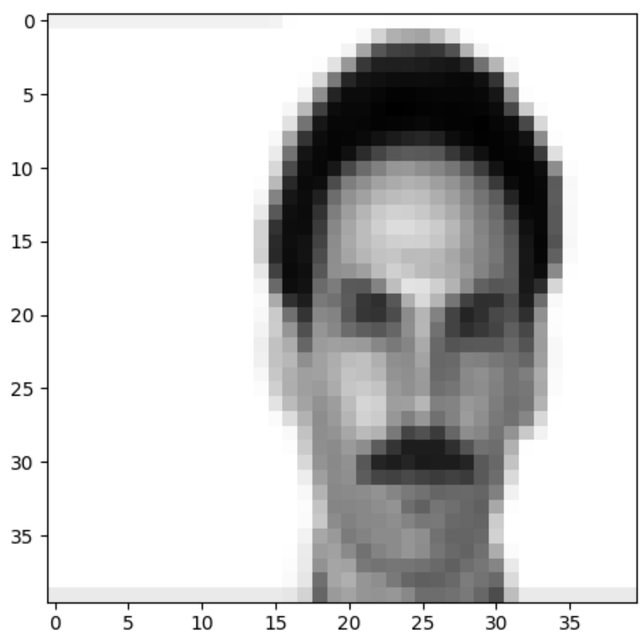
Part 3: Eigenfaces

The number of principal components *k* that was necessary to encode 95% of information was 33.

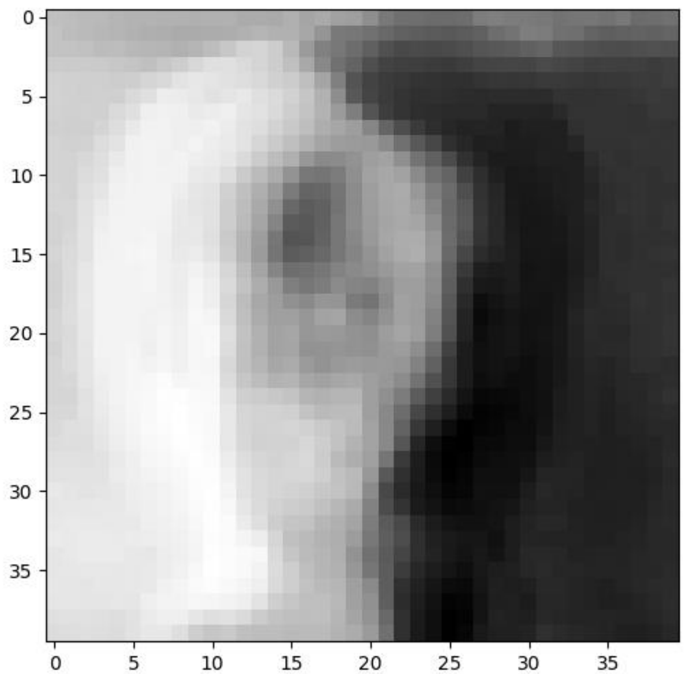
Visualization of Primary Principal Component:



Original Image



Reconstruction Using One Principal Component



Reconstruction Using k=33 Principal Components