Question 5:

Download a collection of handwritten digit images from `http://www.cs.nyu.edu/~roweis/data/mnist_all.mat`. The dimensionality of each image is 28×28 pixels. In this exercise, you will apply principal component analysis to reduce the rank of the images. You will use only the training images for digits 0, 1, 2, and 3 (ignore the images for other digits). You should also ignore the test images in the data set.

(a) Create a data matrix $\mathbf{X}$ of size 200×784 by choosing the first 50 rows of train0, train1, train2, and train3 and append them together. Note that the matrix train0 contains images for digit 0, train1 contains images for digit 1, and so on. Each column in the matrix represents one of the 784 pixels (28×28) of the corresponding images.

(b) Plot the resulting images using the Matlab script given below:

```
N = 50;         % number of images associated with each digit
numCols = 10;
numRows = ceil(4*N/numCols);
d = sqrt(size(X,2));

figure;
set(gcf,'color','white');
set(gcf,'Position',[520 85 1020 720]);  % This command will resize the plot
for i=1:size(X,1);
    subplot(numRows,numCols,i);
    img = reshape(X(i,:),d,d)';     % convert each row into 28 x 28 matrix
    imagesc(img);                   % plot the image
    set(gca,'xtick',[]);
    set(gca,'ytick',[]);
end;
colormap(gray);         % convert the images into gray scale
```

Save the images into a jpg file as follows:

```
matlab> saveas(gcf,'digit_image.jpg','jpeg');
```

(c) Use the `pca` command to generate the principal components of the matrix $\mathbf{X}$.

```
[U,Z,S] = pca(X);
```

Note that the matrix $\mathbf{U}$ contains the eigenvectors (i.e., principal components) of the covariance matrix for $\mathbf{X}$, $\mathbf{Z}$ represents the projection of each row in $\mathbf{X}$ onto the subspace spanned by the principal components, and $\mathbf{S}$ is a vector containing the variance explained by each principal component. Plot the images associated with the first two

principal components. Hint: use the `reshape` and `imagesc` commands shown in part (b) above. Each image should be plotted in a separate figure. Save the figures as jpeg images and attach them with your homework solution.

(d) Reduce the dimensionality of each data point from 784 to 2 by projecting the data to its first two principal components. This is given by the first two columns of the matrix **Z**. Draw a scatter plot of the data points, using different markers to represent each digit.

```
figure;
set(gcf,'color','white');
plot(Z(1:50,1),Z(1:50,2),'r*');         % images for digit 0 is shown as *
hold on
plot(Z(51:100,1),Z(51:100,2),'b+');     % images for digit 1 is shown as +
plot(Z(101:150,1),Z(101:150,2),'ko');   % images for digit 2 is shown as o
plot(Z(151:200,1),Z(151:200,2),'gv');   % images for digit 3 is shown as triangles
hold off
```

Save the resulting plot into a jpeg image and append it to your solution file. Based on the plot, answer the following question: which classes are easier to be discerned by the first two components and which are harder to be discerned?

(e) Using the script in part (b), plot the resulting digit images when the data is reduced to a matrix of rank 2. To create the reduced-rank matrix **W**, you need to do the following:

```
rank = 2;
W = Z(:,1:rank)*diag(S(1:rank))*U(:,1:rank)';
```

Save the resulting images as a jpeg file and attach it to your solution. Which digits can be more easily discerned and which are harder? Is it consistent with your answer in part (d)?

(f) Repeat part (e) to re-create the digit images using a matrix of rank 50. Can you visually discern more digit images correctly with the increasing rank of the matrix **W**?

Make sure you insert all the figures to your pdf solution file (and label them appropriately so we know which figure is for which question) instead of submitting each figure as a separate jpeg file to handin. You should put all your Matlab code in a single file named `q3.m`. Make sure you add comments to the different parts of your code (using 'represent a comment). Submit the Matlab code to handin as a separate file from the rest of the homework.