CSCI 5521: Pattern Recognition

Fall 2007, Prof. Schrater

Problem 4:

12/1/09 Due: 12/15/09

The problem is worth 40% of a normal homework (i.e. the preceeding 3 homeworks).

Problem Implement dimensionality reduction methods on the digit images.

a) Implement PCA. Apply to the digit images. To reduce computational complexity, reduce the amount of data. First load the digit images data. Restrict yourself to 200 images from each class:

```
indices = [1:200 1001:1200 2001:2200];
digitims = digitims(:,:,indices);
```

For pca, stretch the images into a matrix.

```
D = reshape(digitims(:,:,indices),28^2,600);
```

REMOVE THE MEAN IMAGE, then perform pca.

How many components required to capture 95% of the data?

b) Implement Kernel PCA using a Gaussian kernel. Apply to the reduced digit images you constructed above.

a. Compute the Gram matrix: $K_{ij} = \exp\left(-\frac{1}{2\sigma^2}\left(x_i - x_j\right)^T \left(x_i - x_j\right)\right)$

Let $R_{ij} = \left(x_i - x_j\right)^T \left(x_i - x_j\right)$

The matrix can be computed efficiently using

```
Rp = D'*D;
b=diag(Rp)*ones(1,600);
R = b+b'-2*Rp;
```

A good range for the variance σ^2 is between 10^6.3 and 10^6.7.

b. You should subtract off the mean, using the kernel result.

$$K_{ij}^{meanremoved} = K_{ij} - \frac{1}{M}\sum_{k=1}^{M} K_{ik} 1_{kj} - \frac{1}{M}\sum_{k=1}^{M} 1_{ik} K_{kj} + \frac{1}{M^2}\sum_{l=1}^{M}\sum_{k=1}^{M} 1_{il} K_{lk} 1_{kj}$$

where K is the Gram matrix and 1_{ik} is the ik^{th} entry in a matrix of ones.

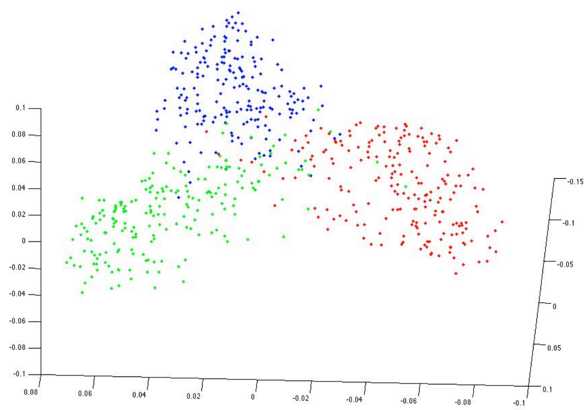
The derivation is in Appendix A "Centering in High-Dimensional Spaces" in *Nonlinear Component Analysis as a Kernel Eigenvalue Problem* by Bernhard Schölkopf, Alexander Smola, and Klaus Müller

c. Perform an eigenanalysis on the Gram Matrix
d. Save the top 3 eigenvectors = {v1,v2,v3}
e. To compute the coordinates $\{y_1,\ldots,y_m\}$ of any images in the new space, project the kernel evaluated at the new point and all the training points onto the m^{th} eigenvector.

$$y_m = \sum_{j=1}^{M} K(x_{new}, x_j) v_m^j$$

c) Take the leading 3 eigenvectors of both kpca and pca analyses. Project the data into this 3-dimensional space.

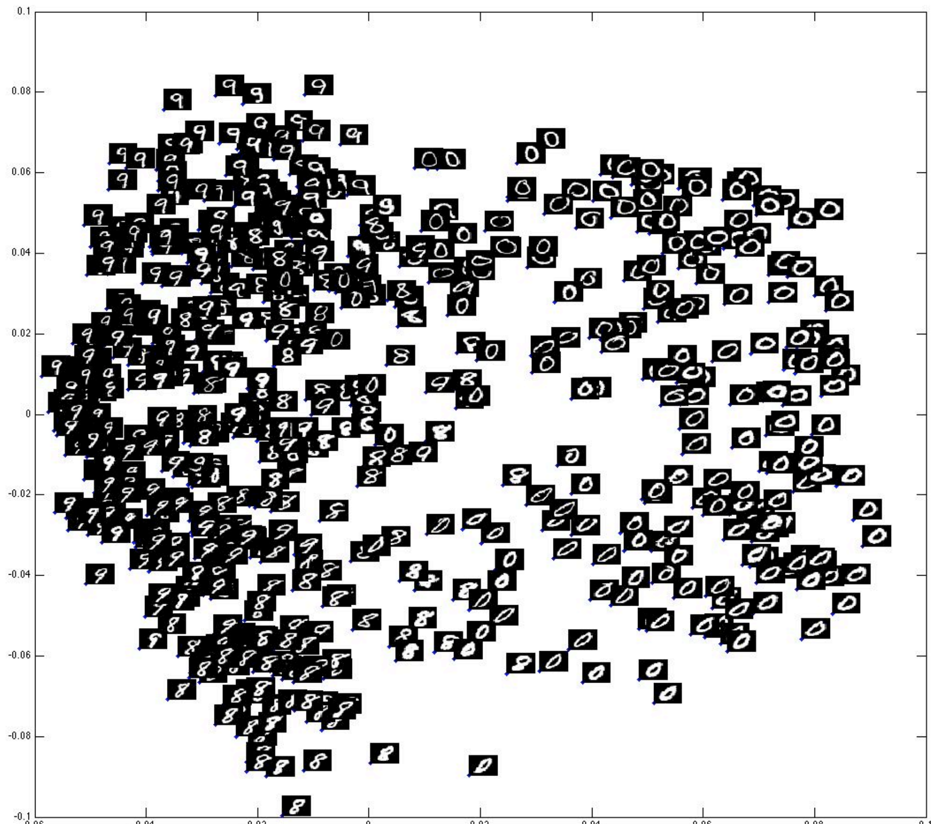
For variance of 10^6.5, the projections should look about like this for kpca:



Plot images at each coordinate location two components at a time (generating 3 images) for both kpca and pca. Here is some code to help visualize in this format:

```
ranges = max(y)-min(y);
% renormalize
y(:,1) = y(:,1)/ranges(1);
y(:,2) = y(:,2)/ranges(2);
y(:,3) = y(:,3)/ranges(3);
imcoorrange = 0.0:0.025/27:0.025;

hold on;
for j = 1:size(y,1),
    imagesc(y(j,1)+ imcoorrange/2, y(j,2)+ imcoorrange(end:-1:1)/2, digitims(:,:,indices(j)))
end
```



Which dimensionality reduction method, PCA vs. kPCA works better for recognition, in the sense that it preserves more of the information about the digit classes? Prove this using the classifier of your choice. Which method is better for reconstruction?

d) Use the Gaussian mixture model code in the Netlab toolbox to learn a soft cluster assignment using a 3 component model for the image data mapped into a 3D kpca space as above. The soft cluster assignments are the posterior probabilities of each data point (the probability of point k belonging to cluster j). Use the posterior probabilities to classify the points, and compute the unsupervised error rate by comparing with the true digit labels.

EXTRA CREDIT 20% of a normal homework.

Use Gaussian process regression on the galaxy data. You can adapt the code by Rasmussen and Williams found:
http://www.gaussianprocess.org/gpml/code/matlab/doc/

For full credit, experiment with using different kernel functions, and optimize the kernel weights and parameters.