Statistical and Learning Techniques in Computer Vision
Homework 4: Due Monday, October 2, 2006

1. (**20 points**) Implement the EM algorithm for estimating the parameters of a Gaussian mixture model, as described in Handout 4 and discussed in class. This will include using the K-means algorithm to initialize the estimates. You will have to investigate ways of initializing the K-means algorithm on your own, however. You will also have to investigate different choices of K, although you do not need to come up with a way of automatically choosing K.

   Your program will read in a set of N points. To keep it simple and to allow visualization of the results, the points will be taken from the 2-d plane. The first line of input will have 2 integer values: N and K. N is the number of points and K is correct the number of components. The remaining lines of input will be the x and y coordinates of the points, with one point per line.

   For each input, you should test your algorithm on several different initializations. (This assumes that the input method uses some randomness in the initialization process.) You must run your implementation of the EM algorithm using both the given value of K and $K+1$ and $K-1$ as the number of components. Your output should be both graphical and numeric. You should give numeric output for each of your tests. For each test there should be K lines of output, one for each component. The output line should give the estimated location, $\boldsymbol{\mu}_k$, the eigenvalues of the estimated covariance matrix (larger first), and the estimated mixture fractions π.

   The output should also be graphical, with a plot showing the results of all of the initializations. For each initialization, plot the resulting K means and covariance matrices as ellipses. Each ellipse should give the values of $\mathbf{x}$ such that

   $$(\mathbf{x} - \boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1} (\mathbf{x} - \boldsymbol{\mu}_k) = 4.$$

   This plot will show how much (if at all) the estimated ellipses changes with different initialization. Generate a separate plot for each number of components — i.e. $K-1$, K and $K+1$ — tested.

   Submit the code, the outputs and the plots. Include a description of how you initialized the k-means algorithm and a discussion of your overall results.

2. (**20 points**) In this problem you will implement an importance sampling method to estimate the mean of a distribution. Let p is the target distribution and g is the proposal density. Suppose a set of N samples, $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$, is drawn from distribution g. Compute the weights

   $$w_i = \frac{p(\mathbf{x}_i)/g(\mathbf{x}_i)}{W},$$

   where

   $$W = \sum_{i=1}^{N} p(\mathbf{x}_i)/g(\mathbf{x}_i).$$

(Note this differs slightly from the definition given in the lecture notes.)

Then the estimate of the mean of p is

$$\mu_p = \sum_{i=1}^{N} w_i \mathbf{x}_i.$$

The function is

$$p(x) = x^4 e^{-(x-2)^2/20} + 300 e^{-(x-30)^2/10}$$

For the proposal density, use a normal distribution. As part of your experiments, vary the mean, the variance and the number of samples. Discuss your results.