STA 414/2104, Spring 2012, Answers to Practice Problem Set #3

Question 1: Suppose that we are fitting a Gaussian mixture model for data items consisting of a single real value, x, using $K = 2$ components. We have $N = 5$ training cases, in which the values of x are as follows:

$$5, \quad 15, \quad 25, \quad 30, \quad 40$$

We are using the EM algorithm to find the maximum likeihood estimates for the model parameters, which are the mixing proportions for the two components, π_1 and π_2, and the means for the two components, μ_1 and μ_2. The standard deviations for the two components are fixed at 10.

Suppose that at some point in the EM algorithm, the E step found that the responsibilities of the two components for the five data items were as follows:

r_{i1}	r_{i2}
0.2	0.8
0.2	0.8
0.8	0.2
0.9	0.1
0.9	0.1

What values for the parameters π_1, π_2, μ_1, and μ_2 will be found in the next M step of the algorithm?

The new estimates will be

$$\pi_1 = (0.2 + 0.2 + 0.8 + 0.9 + 0.9)/5 = 0.6$$

$$\pi_2 = (0.8 + 0.8 + 0.2 + 0.1 + 0.1)/5 = 0.4$$

$$\mu_1 = (0.2 \times 5 + 0.2 \times 15 + 0.8 \times 25 + 0.9 \times 30 + 0.9 \times 40) \,/\, (0.2 + 0.2 + 0.8 + 0.9 + 0.9) = 29$$

$$\mu_2 = (0.8 \times 5 + 0.8 \times 15 + 0.2 \times 25 + 0.1 \times 30 + 0.1 \times 40) \,/\, (0.8 + 0.8 + 0.2 + 0.1 + 0.1) = 14$$

Question 2: Consider the factor analysis model, $x = \mu + Wz + \epsilon$, where x is an observed vector of p variables, μ is the mean vector for x, z is an unobserved vector of m common factors, W is the matrix of "factor loadings", and ϵ is a random residual. We assume that $z \sim N(0, I)$ and independently $\epsilon \sim N(0, \Sigma)$, where Σ is diagonal with diagonal entries $\sigma_1^2, \ldots, \sigma_p^2$.

Let the number of observed variables be $p = 4$ and the number of common factors be $m = 1$.

a) Give an explicit example (specifying μ, W, and Σ) showing that it is possible for the correlation of x_1 and x_2 to be negative, the correlation of x_1 and x_3 to be positive, and the correlation of x_1 and x_4 to be zero. Compute the covariance and correlation matrices of x for your example.

One possible example is $\mu = [0\ 0\ 0\ 0]^T$, $\Sigma = I$, *and* $W = [1\ \ -1\ 1\ 0]^T$. *The covariance matrix of* x *will then be*

$$E[(Wz + \epsilon)(Wz + \epsilon)^T] = E[Wzz^TW^T + \epsilon\epsilon^T] = WW^T + \Sigma = \begin{bmatrix} 2 & -1 & 1 & 0 \\ -1 & 2 & -1 & 0 \\ 1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The correlation matrix will be

$$\begin{bmatrix} 1 & -1/2 & 1/2 & 0 \\ -1/2 & 1 & -1/2 & 0 \\ 1/2 & -1/2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

b) Suppose that $\mu_j = 0$ and $\sigma_j^2 = 4$ for $j = 1, 2, 3, 4$, and $W = [\,3\ 2\ 1\ 0\,]^T$. Find the covariance matrix for x, the direction of the first principal component of that covariance matrix, and the variance in that direction.

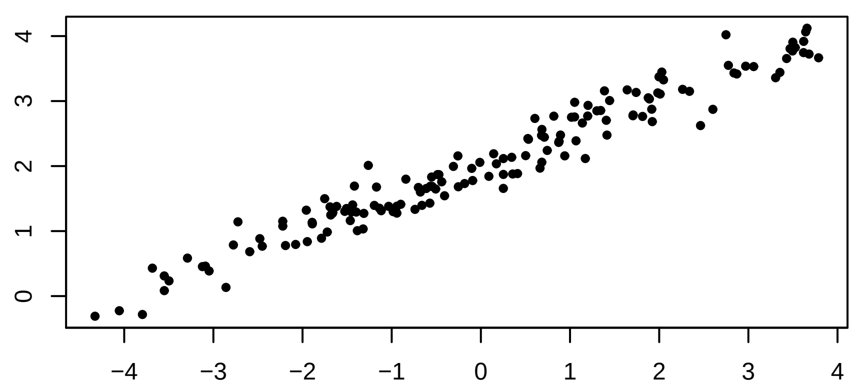
The covariance matrix is $WW^T + 4I$, *which is*

$$\begin{bmatrix} 13 & 6 & 3 & 0 \\ 6 & 8 & 2 & 0 \\ 3 & 2 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

One eigenvector of this matrix is W, *with eigenvalue* $3^2+2^2+1^2+0^2+4 = 19$, *since* $(WW^T + 4I)W = (W^TW + 4)W$. *The other eigenvectors will be orthogonal to this eigenvector, and hence will have eigenvalue 4, since for such an eigenvector,* V, $(WW^T + 4I)V = W(W^TV) + 4V = 4V$.

So the first principal component direction is $[3\ \ 2\ \ 1\ \ 0]$, *and the variance in this direction is 4.*

Question 3: Below is a scatterplot of 150 observations of two variables:



a) Write down a vector pointing in the direction of the first principal component for this data. An approximate answer found by eye is sufficient. The vector need not have length one.

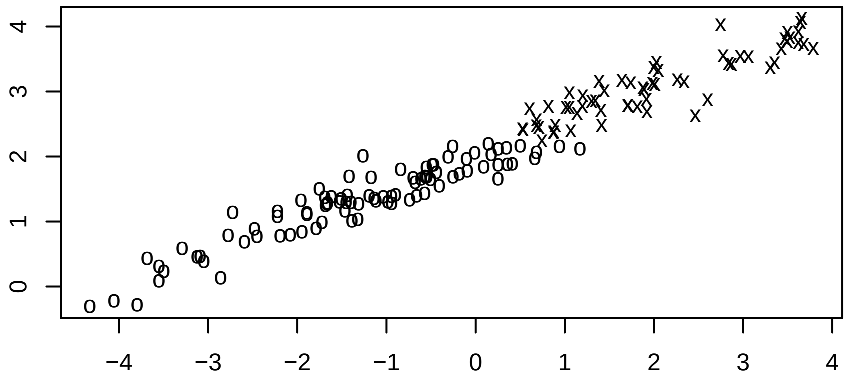
Also, draw the direction of the first principal component on the scatterplot above.

One answer is $[2\ 1]$. *I won't try to draw this on the plot.*

b) What is the approximate standard deviation in the first principal component's direction?

Somewhere around 2 or 3.

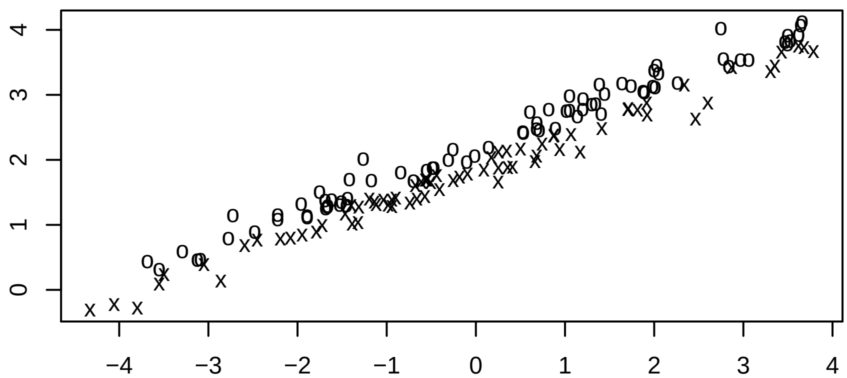
c) Suppose that each of these data points are associated with one of two classes, as shown below (with one class marked by "o" and the other by "x"):



If we reduce the data to just the projection on the first principal component, how well will we be able to classify the data points using this one number, compared to how well we would have been able to classify using the two original numbers?

We will be able to classify almost as well as with the original data.

d) Suppose instead that the two classes are as shown below:



In this case, how well will we be able to classify using just the projection on the first principal component, compared to using the two original numbers?

The projection on the first principal component will give almost no information about the class. One could do much better using the original data, since one can see in the plot that points in the circle class are usually above those in the x class. So there is a a diagonal line that separates the classes fairly well.