STA 414/2104, Spring 2014, Practice Problem Set #3

Note: these problems are not for credit, and not to be handed in

Question 1: Recall that a multilayer perceptron network with m hidden units using the tanh activation function computes a function defined as follows:

$$f(x,w) \;=\; w_0^{(2)} \;+\; \sum_{j=1}^{m} w_j^{(2)} \phi_j(x,w), \qquad \phi_j(x,w) \;=\; \tanh\Big(w_{0j}^{(1)} \;+\; \sum_{k=1}^{p} w_{kj}^{(1)} x_k\Big)$$

where w is the set of parameters (weights) for the network, and x is the vector of p inputs to the network.

Suppose we train such a network with $m = 1$ hidden units on the following set of $n = 4$ training cases, with a single input, x_1 (so $p = 1$), and one real-valued response, y:

x_1	y
-1	1
0	1
1	5
2	5

We use a Gaussian model for the response, in which y given x has a Gaussian distribution with mean $y(x,w)$ and variance one.

a) Suppose that we initialize the weights to $w_{01}^{(1)} = 0$, $w_{11}^{(1)} = 0$, $w_0^{(2)} = 0$, and $w_1^{(2)} = 0.1$. Define $E(w)$ to be the minus the log likelihood, dropping terms that don't depend on w, so that $E(w)$ is 1/2 times the sum of the squares of the residuals in the four training cases.

Find the gradient of $E(w)$, as would be needed to do gradient descent learning, evaluated at the initial value of w specified above. In other words, find the partial derivatives of E with respect to all the components of w, at the initial value of w.

b) If gradient descent learning to minimize minus the log likelihood is done from the initial weights specified in part (a) above, what weights will the learning converge to (assuming that the learning rate used is small enough to ensure stability)? You may not be able to say exactly what the values of all the weights will be, but say as much as you can.

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.

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.

Question 3: We have two i.i.d. observations of seven variables, as follows:

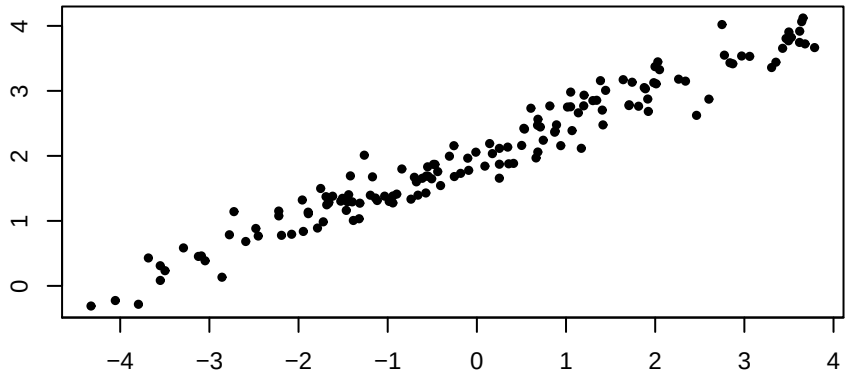
5 7 8 2 3 5 2
3 3 6 6 1 1 0

a) Find a 7-dimensional vector of length one that points in the direction of the first principal component of this data. Explain how you obtained it.

b) Find the projection on this principal component of the new observation shown below:

4 1 9 3 2 2 1

Question 4: Below is a scatterplot of 150 observations of two variables:

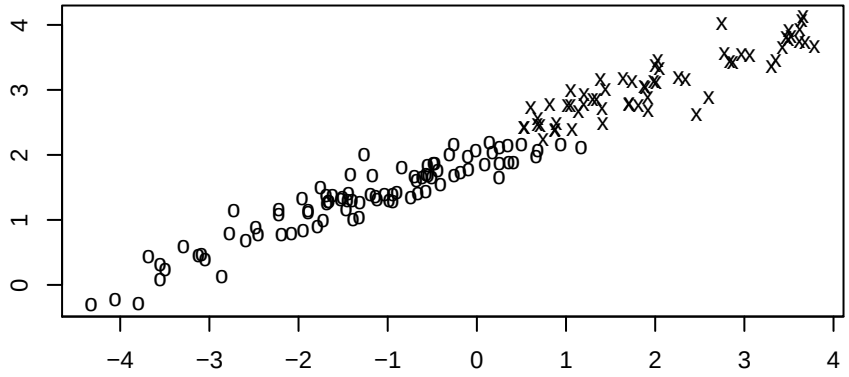


a) Write down a vector pointing in the direction of the first principal component for this data. An approximate answer found by eye is sufficient. The vector need not have length one.

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

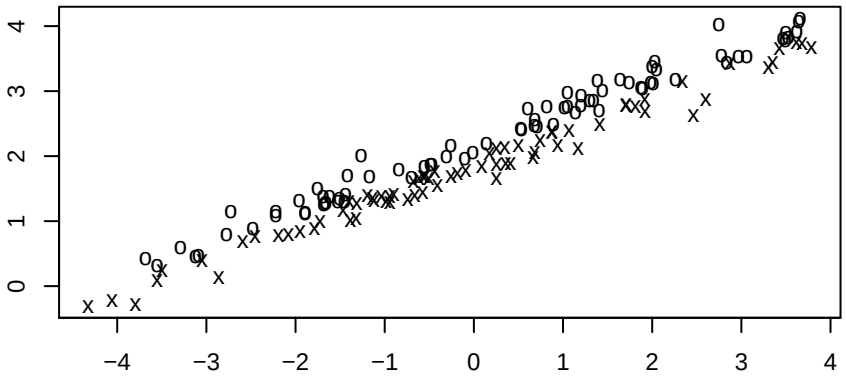
b) What is the approximate standard deviation in the first principal component's direction?

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?

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?

Question 5: Recall that in a factor analysis model an observed data point, x, is modeled using M latent factors as

$$x \;=\; \mu \;+\; Wz \;+\; \epsilon$$

where μ is a vector of means for the p components of x, W is a $p \times M$ matrix, z is a vector of M latent factors, assumed to have independent $N(0,1)$ distributions, and ϵ is a vector of p residuals, assumed to be independent, and to come from normal distributions with mean zero. The variance of ϵ_j is σ_j^2.

Suppose that $p = 5$ and $M = 2$, and that the parameters of the model are mean $\mu \;=\; [0\;0\;0\;0\;0]^T$, residual standard deviations $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 2$, $\sigma_4 = 2$, $\sigma_5 = 2$, and

$$W \;\;=\;\; \begin{bmatrix} 1 & 2 \\ -1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

a) Find the covariance matrix for x.

b) Suppose that we don't observe vectors x of dimension five, but rather we observe vectors y of dimension four, where $y_1 = x_1$, $y_2 = 3x_2$, $y_3 = -x_3$, and $y_4 = 2x_4 + x_5$. Assuming that the distribution of x is given by the factor analysis model with parameters above, write down a factor analysis model (including values of its parameters) for the distribution of y.