Statistics 601, Spring 2018

Assignment 4

In the course notes the EM algorithm is illustrated for a problem involving finite mixtures. In lecture I indicated that "missing information" could mean a number of different quantities such as latent variables or actually missing observations. In this assignment you will use the EM algorithm to find maximum likelihood estimates of data assumed to be from a bivariate normal distribution, but with some of the values not recorded.

To set up the problem, let $(X_i, Y_i); \quad i = 1, \ldots, n$ denote independent and identically distributed bivariate random variables assumed to have the probability density function

$$\begin{aligned} f(x, y|\boldsymbol{\theta}) &= \frac{1}{2\pi(1-\rho^2)\sigma_x\sigma_y} \\ &\times \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\left(\frac{x-\mu_x}{\sigma_x}\right)^2 - 2\rho\left(\frac{x-\mu_x}{\sigma_x}\right)\left(\frac{y-\mu_y}{\sigma_y}\right)\right.\right. \\ &+ \left.\left.\left(\frac{y-\mu_y}{\sigma_y}\right)^2\right\}\right]. \end{aligned} \tag{1}$$

In (1) $\boldsymbol{\theta} = (\mu_x, \mu_y, \sigma_x, \sigma_y, \rho)$, where

$$\begin{array}{ccccc} -\infty & < & \mu_x & < & \infty \\ -\infty & < & \mu_y & < & \infty \\ 0 & < & \sigma_x & < & \infty \\ 0 & < & \sigma_y & < & \infty \\ -1 & < \rho & < & 1 & \end{array}$$

and

$$\rho = \frac{cov(X, Y)}{[var(X)\, var(Y)]^{1/2}} = \frac{\sigma_{xy}}{[\sigma_x^2\, \sigma_y^2]^{1/2}}.$$

On the course web page in the Data folder is a set of observed data for this model under the heading "Bivariate Normal Observed Data". You will note that for this

data set there are $n = 30$ observations, but $m_x = 10$ have missing values for the variable X and $m_y = 7$ have missing values for the variable Y. Thus, there are only 13 observations with both X and Y recorded.

Your assignment is:

1. Find marginal maximum likelihood estimates of μ_x, μ_y, σ_x^2, σ_y^2 and σ_{xy} (or ρ) using the EM algorithm. You do not need to obtain inferential quantities (i.e., standard errors), just point estimates. Show derivation of the quantites you will need to compute and write your own function rather than searching R packages. The point of this is to help you understand the EM algorithm, not see if you are profficient at searching the web. Although I usually do not wish to see such things, for this assignment, turn in a copy of your R functions with your answer. Be sure to record parameter estimates at each iteration of the algorithm, and how many iterations your algorithm required.

2. Find maximum likelihood estimates using only the data that have observed values for both X and Y variables and compare to the estimates from the EM algorithm.

3. I class, I made the comment that for exponential families, the E step in the algorithm consisted of replacing missing values with their conditional expected values. The M step then would consist of maximizing the complete data log liklelihood with the "imputed" values of the missing data at each iteration. This was basically true, but not general enough. We were discussing an example in which the complete data corresponded to a *natural* exponential family (in which there is only one parameter and the sufficient statistic is the variable, Y, say). For exponential families in general replacing the missing data with their conditional expected values is not quite right (although it is close). In this problem, produce estimates from an algorithm in which the missing

data are replaced with their conditional expectations (re-computed at each iteration) and then the complete data log likelihood including those imputed values is maximized for the M step. Compare to the other estimates you have obtained.

4. If we wanted to compare the behavior of the first two estimators considered above (those from parts 1 and 2) we could conduct a Monte Carlo study. That would actually be possible here because this EM agorithm runs very rapidly in real time, but don't do it. To simulate data sets for this Monte Carlo procedure we would simulate a full set of $n = m + m_x + m_y$ observations from a bivariate normal, and then delete m_x of the X variable values and m_y of the Y variable values, taking care that no observation has missing values for both variables. We would probablly select the values to be made "missing" at random, so that there is no relation between the data distribution (bivariate normal) and the missingness mechanism – this is called "missing completely at random". Given this, what would you **anticipate** the results would be in terms of (1) bias, and (2) precision, which are the two components of mean squared error.

A little help on all of this:

1. In the EM algorithm,

   (a) Write the logarithm of the complete data density by expanding squared terms, pulling out constants whenever possible (in the density, parameters are constants) and trying to isolate variables (x and y) as additive terms. This will make taking the expected value of the complete data log likelihood easier than it would otherwise be. Also, use the following notation for missing data:

- Let $(x_i,\, y_i)$; $\quad i = 1, \ldots, m$ denote the observations that have recorded values for both X and Y variables.
- Let $(x_j^m,\, y_j)$; $\quad j = 1, \ldots, m_x$ denote the observations that have missing values for X and recorded values for Y.
- Let $(x_k,\, y_k^m)$; $\quad k = 1, \ldots, m_y$ denote the observations that have recorded values for X and missing values for Y.

The joint pdf of the complete data is then

$$f(\boldsymbol{x}, \boldsymbol{y}) = \left[\prod_{i=1}^{m} f(x_i,\, y_i)\right] \left[\prod_{j=1}^{m_x} f(x_j^m,\, y_j)\right] \left[\prod_{k=1}^{m_y} f(x_k,\, y_k^m)\right].$$

(b) Note that in the problem we start with the complete data density. Because of results for bivariate normal distributions (which you can use without derivation) the conditional density of unobserved given observed data is readily available.

(c) As with many applications of the EM algorithm, derivation of the necessary quantities is more involved than is programming the results. The computational algorithm should acutally not be horrendous for this problem, although there is a bit of length due to the fact that there are 5 parameters.

(d) There are a number of forms of the convergence criteria that could be used in this problem. If $\boldsymbol{\theta}_k$ denotes the estimated parameter at iteration k, use

$$||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}|| \leq 10^{-8},$$

where

$$||\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}|| = \left[\sum_{j=1}^{p} (\theta_{j,k} - \theta_{j,k-1})^2\right]^{1/2}.$$

(e) To get starting values for the EM algorithm, use the estimates you obtain from using only the data with no missing values. Then you already have

part 2 above done as well. This also depends on well-known results for bivariate normal distributions, so you do not need to derive maximum likelihood estimators here. Just use the standard forms, which you can find in many books.

2. There are a number of reasonable approaches one could take toward constructing computational functions for this problem. My approach was to write one function for the E step applied to observations missing the X variable, another (parallel) function for the E step applied to observations missing the Y variable (note from the joint pdf presented above that these will constitute different portions of the overall complete data log likelihood), one function for the M step, and a bigger function to control computations, collect results, and take care of other "bookkepping" tasks.

3. For question 3 you should be able to make a few simple modificaions to the functions you used to compute estimates under the EM algorithm.

4. For question 4 (at least part of it) it is crucial that we are are assuming missing values are missing completely at random.