ST 790, Homework 2
Spring 2017

1. Let $Z = (Y_1, Y_2)$, where Y_1 and Y_2 are categorical variables such that Y_1 takes on values in $\{1, \dots, G\}$ and Y_2 takes on values in $\{1, \dots, H\}$. Let

$$\theta_{gh} = \text{pr}(Y_1 = g, Y_2 = h), \quad g = 1, \dots, G, \; h = 1, \dots, H, \tag{1}$$

subject to the constraint

$$\sum_{g=1}^{G} \sum_{h=1}^{H} \theta_{gh} = 1. \tag{2}$$

(a) Suppose we have a sample of full data, (Y_{i1}, Y_{i2}), $i = 1, \dots, N$. Show that the MLE for θ_{gh} in (1) is given by

$$\widehat{\theta}_{gh} = N^{-1} \sum_{i=1}^{N} I(Y_{i1} = g, Y_{i2} = h);$$

that is, the sample proportion for $g = 1, \dots, G$, $h = 1, \dots, H$. Be sure to take into account the constraint (2).

(b) Now suppose that it is possible for either Y_1 or Y_2 to be missing. Define as usual $R = (R_1, R_2)^T$, taking possible values r, and suppose that there are three possible situations: (i) Y_1 and Y_2 are both observed, $r = (1, 1)^T$; (ii) Y_1 is observed, Y_2 is missing, $r = (1, 0)^T$; and (iii) Y_1 is missing, Y_2 is observed, $r = (0, 1)^T$. The observed data are then $(R_i, Z_{(R_i)i})$, $i = 1, \dots, N$, which may be written as $(R_{i1}, R_{i1}Y_{i1}, R_{i2}, R_{i2}Y_{i2})$, $i = 1, \dots, N$.

Assume that the mechanism governing this missingness is MAR. Show how to use the EM algorithm to obtain the MLE for θ_{gh}, $g = 1, \dots, G$, $h = 1, \dots, H$, based on the observed data. Specifically, indexing iterations by t, given the tth iterate, provide explicit expressions for the E-step and M-step to obtain the $(t + 1)$th iterate.

(c) In your favorite programming language, write a program to implement the EM algorithm you derived in (b) and to obtain standard errors for the resulting estimates of θ_{gh} using a nonparametric bootstrap. Try it out on the data set `mulitnom.dat` available (with missing values indicated by "`NA`") on the course website, for which $G = 3$ and $H = 2$. Turn in your program and the results.

2. In the argument justifying the EM algorithm on page 63 of the notes, it is necessary to show the equality in (3.42), namely,

$$E_{\theta'}\Big[\log\{p_Z(Z_i;\theta)\}|R_i, Z_{(R_i)i}\Big] = \sum_{r} I(R_i = r)\, E_{\theta'}\Big[\log\{p_Z(Z_i;\theta)\}|Z_{(r)i}\Big].$$

Provide a full argument demonstrating this equality.

3. On the course webpage, you will find data from a multicenter clinical trial carried out to compare an experimental (active) treatment, interferon-α, with a placebo for the treatment of patients with age-related macular degeneration (AMD). AMD is a common eye condition and leading cause of vision loss among people age 50 and older. It causes damage to the macula, a spot near the center of the retina and the part of the eye needed for sharp, central vision. Patients with AMD progressively lose vision, at varying rates.

In the trial, visual acuity was assessed at baseline (week 0) and then at clinic visits at 4, 12, 24, and 52 weeks through patients' ability to read lines of letters on standardized vision charts. The charts display lines of five letters of decreasing size, which the patient must read from top (largest letters) to bottom (smallest letters). The raw visual acuity measure is the total number of letters correctly read. In this problem, we will focus on the visual acuity measure as the outcome of interest.

Another measure is the number of "lines of vision," where a line of vision is one with at least four letters correctly read. In the trial, this was also recorded at baseline and weeks 4, 12, 24, and 52 weeks. We will not consider this measure in this problem.

The trial involved $N = 240$ participants, each of whom was to provide these measures at baseline and 4, 12, 24 , and 52 weeks after randomization to placebo or active treatment. However, as usual, not all participants have full data. Most of those who do not have full data do not because they dropped out of the trial prior to completing all four post-baseline study visits. In addition, some subjects have intermittently missing visits.

On the course website, you will find data sets `armd.dat`, with missing values indicated using the SAS "." convention, and `armd.R.dat`, with missing values indicated by "`NA`." The columns are (1) patient ID number; (2) baseline lines of vision; (3)-(6) change from baseline lines of vision at 4, 12 24, and 52 weeks; (7)-(11) visual acuity at baseline, 4, 12, 24, and 52 weeks; (12) lesion grade; and (13) treatment, coded as 1 (placebo) and 4 (active treatment). These data sets are in the "wide" format of one record per individual. As noted above, we are interested in an analysis of the visual acuity outcomes in columns (7)-(11).

Here, the full data are $Z = (A, Y_1, Y_2, Y_3, Y_4, Y_5)$, where Y_1 is visual acuity at baseline, and $Y_2, \dots, Y_5$ are visual acuity at weeks 4, 12, 24, and 52, and A is the treatment indicator such that $A = 0$ if a patient was assigned to placebo an $A = 1$ if assigned to active treatment. Letting $Y = (Y_1, \dots, Y_5)^T$ and treating the visual acuity measures as continuous, it is not unreasonable to assume that Y has an approximate multivariate normal distribution with possibly different mean vectors for each treatment. (Given that this is a clinical trial, by randomization, the means at baseline should be the same for each treatment, but we ignore this aspect here.)

We thus assume the following full data model: Although the mean vectors may differ between the treatments, the covariance matrix is the same for each treatment. The model can be expressed as

$$Y \sim \mathcal{N}(\mu_0, \Sigma) \text{ for placebo,} \qquad Y \sim \mathcal{N}(\mu_1, \Sigma) \text{ for active treatment,} \tag{3}$$

where $\mu_0 = (\mu_{01}, \dots, \mu_{05})^T$ and $\mu_1(\mu_{11}, \dots, \mu_{15})^T$ are (5×1) vectors of means whose components may differ, and Σ is the assumed common (5×5) covariance matrix.

Our objective is to fit the model (3) based on the observed data from the trial. As noted above, Z is not fully observed for some trial participants; while A and Y_1 are observed on everyone, $Y_2, \dots, Y_5$ can be missing, either because of dropout or due to intermittent missed visits. Note that, if we had a sample of full data, (Y_i, A_i), $i = 1, \dots, N$, we could fit (3) to these data by fitting

$$Y_{ij} = \mu_{0j} + \beta_j A_i + \epsilon_{ij}, \quad \epsilon_i = (\epsilon_{i1}, \dots, \epsilon_{i5})^T \sim \mathcal{N}(0, \Sigma), \quad j = 1, \dots, 5, \; i = 1, \dots, N, \tag{4}$$

where $\beta_j = \mu_{1j} - \mu_{0j}$. Here (4) is simply a reparameterization of (3). We do not assume any particular structure for Σ (e.g., compound symmetric), so Σ is a symmetric matrix with

15 distinct variance and covariance parameters (i.e., an "unstructured" covariance specification).

As demonstrated in ***EXAMPLE 2*** in Chapter 3 of the notes in the simpler case of a bivariate normal, the density of any subset of Y is also normal, so that, under MAR, the observed data likelihood based on (3) boils down to the likelihood for the available data. Accordingly, the observed data likelihood can be maximized directly by fitting model (4) to the available data. This can be implemented using `proc mixed` in SAS with the `method=ml` option in the `proc mixed` statement and `type=un` in the `repeated` statement, or using the `gls` function in the `nlme` package in R with the `method="ML"` option and `corSymm` covariance structure. The only difference between the code for the example and model (4) is the specification of the slightly more complex model.

(a) Summarize the distinct patterns of missingness in the observed data. E.g., how many trial participants have full data? How many participants exhibit each of the observed patterns? *Hint:* If you are using R, the `mice` package has a handy function, `md.pattern`, that takes as input the relevant columns of a data frame or matrix and outputs a summary of the patterns. Alternatively, there are many more sophisticated packages in R to visualize missing data, for example, `vim`. If you are using SAS, `proc mi` creates a similar summary automatically.

(b) Using your favorite software, fit (4) to the observed data.

(c) Using your favorite software, fit (4) to the data from participants who have full data, thus implementing a *complete case* analysis.

(d) Using your favorite software, create an imputed data set using the Last Observation Carried Forward (LOCF) technique. That is, for participants who drop out, replace their missing visual acuity measures by the last observed value. For intermittent missing values, fill these in using the most recent observed values. Fit (4) to these data.

(e) Of primary interest in the trial was inference on β_5, the difference in mean visual acuity at one year (52 weeks). Compare the inferences on this quantity based on each of the analyses in (b) - (d).

(f) We will now fit (3) separately by treatment group so that we can compare direct maximization of the observed data likelihood to using the EM algorithm under MAR. (The software we have discussed for the latter only fits a single multivariate normal.) When considered separately by treatment, the model for the full data for participants i who were assigned to treatment a (i.e., for whom $A_i = a$), is

$$Y_{ij} = \mu_{aj} + \epsilon_{ij}, \tag{5}$$

where $a = 0$ or 1, and all other model components are as in (4). Fit (5) to the available data on each treatment separately using SAS `proc mixed` or the R `gls` function, thus obtaining estimates of μ_0 and μ_1 and estimates of Σ separately by treatment.

(g) Using either SAS `proc mi` or the `norm` package in R, fit each model in (3) based on the observed data using the EM algorithm, thus obtaining estimates of μ_0 and μ_1 and estimates of Σ separately by treatment.

(h) Compare your estimates from (f) and (g). Do they agree? Should they?

4. Consider the situation of ***EXAMPLE 4***, in which $Z = (Y_1, Y_2)$,

$$Y = (Y_1, Y_2)^T \sim \mathcal{N}(\mu, \Sigma), \quad \mu = (\mu_1, \mu_2)^T, \quad \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix},$$

$\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_{12}, \sigma_2^2)^T$, where Σ is positive definite. Suppose that we are interested in inference on $\mu = (\mu_1, \mu_2)^T$, and, as in the example, that Y_1 is always observed and only Y_2 is possibly missing.

Let $S_\mu^F(Z)$ be the component of the full data score vector (3.2) corresponding to partial derivatives of the full data log likelihood with respect to μ. Similarly, let $S_\mu(R, Z_{(R)})$ be the analogous component of the observed data score vector (3.31).

(a) Find $S_\mu(Z)$ by direct differentiation of the observed data loglikelihood.

(b) Find an expression for $S_\mu^F(R, Z_{(R)})$, and show that, under MAR, it is indeed true that

$$S_\mu(R, Z_{(R)}) = E_\theta\{S_\mu^F(Z) | R, Z_{(R)}\} \tag{6}$$

as in (3.33). That is, show that the expression you found in (a) is identical to the right hand side of (6) under a MAR missingness mechanism.

(c) Now consider the components of the observed information matrix in (3.76)-(3.78) on page 80. As noted in the discussion there, under MAR, it can be shown that the expectations of these terms are not necessarily equal to zero, so that, as discussed on page 81, the usual approximate standard errors based on the expected information matrix reported by standard software are not appropriate.

To partially verify this claim, consider (3.77) and (i) derive the expression for

$$-\frac{\partial^2 \ell}{\partial \mu \partial \sigma_{12}}$$

given in (3.77), and then (ii) derive an expression for the (unconditional) expectation of (3.77) in terms of $\pi = \text{pr}(R = 1) > 0$ and Σ and argue that a necessary and sufficient condition for the expectation of (3.77) to be equal to zero is that the missingness mechanism is MCAR, so that the expectation need not be equal to zero under MAR.