**Homework Assignment 2**

**Statistical Methods for Analysis with Missing Data, Winter 2019**

Instructor: Mauricio Sadinle, Department of Biostatistics, U. of Washington – Seattle

Submit your solutions via Canvas. Due by 12:00pm (noon) on Feb 13, 2019.

From this assignment you can get a maximum of 20 points. The assignment contains a list of problems, each worth a different number of points. You may choose any combination of problems that you like. I recommend that you solve a combination of problems that is worth more than 20 points as a way of gaining insurance against errors in some of your problem solutions. If you are submitting solutions to theoretical problems, feel free to hand-write them and submit a scanned copy.

---

1. (1 point) Let $Z_1 \in \{1, 2\}$, $Z_2 \in \{A, B\}$, $R \in \{0, 1\}^2$. Say the full-data probability density is given by $p(r, z) \equiv p(r_1, r_2, z_1, z_2) \equiv \pi_{r_1 r_2 z_1 z_2}$. Derive the observed-data probability density $p(r, z_{(r)})$ for all elements $(r, z_{(r)})$ in the sample space of $(R, Z_{(R)})$.

2. (2 points) Say $Z^T = (Z_1, Z_2)^T \sim \mathcal{N}(\mu, \Sigma)$, $R \in \{0, 1\}^2$. Say $p(r \mid z) = p(r)$. Derive $p(r, z_{(r)})$ for all $r \in \{0, 1\}^2$.

3. (2 points) Say $Z = (Z_1, Z_2)$ follows a generic bivariate distribution with density $p(z)$. $R \in \{0, 1\}^2$. Say $R_1 \perp\!\!\!\perp R_2 \mid Z$, with

$$\text{logit } p(R_j = 1 \mid z) = \beta_{j0} + \beta_{j1} z_1 + \beta_{j2} z_2, \quad j = 1, 2.$$

   Derive $p(r, z_{(r)})$ for all $r \in \{0, 1\}^2$.

4. (3 points) Refer to the notation in slide 7 of lecture 6. Show that $h(\vartheta^{(t)} \mid \vartheta^{(t)}) = \log \ell_{obs}(\vartheta^{(t)})$.

5. (5 points) Say $p(z \mid \theta)$ belongs to an exponential family with $\theta = (\theta_1, \ldots, \theta_d)$, that is,

$$p(z \mid \theta) = b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] / c(\theta)$$

with $c(\theta) = \int b(z) \exp\left[\sum_{s=1}^{d} \eta_s(\theta) T_s(z)\right] dz$. Show that the EM algorithm can be written as:

(a) E step:

$$Q_\theta(\theta \mid \theta^{(t)}) = \sum_{s=1}^{d} \eta_s(\theta) E\ [\ T_s(z)\ \mid\ Z_{(r)} = z_{(r)}, \theta^{(t)}] - \log c(\theta)$$

(b) M step: find $\theta^{(t+1)}$ as the solution to

$$E\ [\ T_s(Z)\ \mid\ Z_{(r)} = z_{(r)}, \theta^{(t)}] = E\ [\ T_s(Z)\ \mid\ \theta],\ s = 1, \dots, d$$

---

For problems 6–8, let $Z_i = (Z_{i1}, Z_{i2})$, $Z_{i1}, Z_{i2} \in \{1, 2\}$, $Z_i$'s are i.i.d. Denote

$$p(Z_{i1} = z_{i1}, Z_{i2} = z_{i2} \mid \theta) = \pi_{z_{i1} z_{i2}},$$

and the likelihood of the study variables as $L(\theta) = \prod_i \pi_{z_{i1} z_{i2}}$. Let $R_i = (R_{i1}, R_{i2})$, $R_{i1}, R_{i2} \in \{0, 1\}$, $R_i$'s are i.i.d.

6. (1 point) Show that the observed-data likelihood for the study variables can be written as $L_{obs}(\theta) = \prod_i \pi_{z_{i1} z_{i2}}^{I(r_i=11)} \pi_{z_{i1}+}^{I(r_i=10)} \pi_{+z_{i2}}^{I(r_i=01)}$.

7. (3 points) Parameterize $L(\theta)$ in terms of the odds ratio

$$\phi = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$$

and the marginal probabilities $\pi_{1+}$ and $\pi_{+1}$.

8. (1 points) Show that $\phi$ only appears in the observed-data likelihood $L_{obs}(\theta)$ for those observations with $r_i = 11$. What is the meaning of this result?

---

Problems 9 and 10 are computational.

9. (5 points) Under the setup used in Part 1 of `Lecture07code.R`, run a simulation study to compare estimators of the probabilities $\{\pi_{kl}\}_{k,l}$ and odds ratio, based on EM vs complete cases. Compare the performance in terms of bias and variance. Submit a report with your results and code.

10. (5 points) Read Example 2 in Chapter 3 of the lecture notes of Davidian and Tsiatis (pages 59 and 69). Implement the EM algorithm described in that example, and illustrate its use with a simulated dataset. Submit a report with your results and code.

---

For problems 11–14 we have the setup of a two-class mixture model. Think about the following generative process for the data:

- Each individual $i$ is randomly assigned to one of two classes. Let $C_i \sim \text{Bernoulli}(\pi)$ represent the class assigned to individual $i$.
- Given the value of $C_i$, the individual gets assigned a bivariate measurement $Z_i^T = (Z_{i1}, Z_{i2})^T \mid C_i \sim \mathcal{N}(\mu_{C_i}, \Sigma_{C_i})$. Here $\mu_j$ and $\Sigma_j$ represent the parameters for class $j$, where $j = 0, 1$.
- Individuals are generated independently from each other.

11. (3 points) Write down the likelihood function for this generative process.
12. (3 points) Assume that none of the $C_i$'s are observed. Write down the observed-data likelihood.
13. (5 points) Derive an EM algorithm to estimate the parameters in this model.
14. (5 points) Code the EM algorithm in R and test it with some data generated using this generative process. Submit a report with your results and code.