Homework 3

Names:

Student IDs:

1 Problem Description

Read article: *Maximum Likelihood Algorithms for Generalized Linear Mixed Models* (McCulloch 1997), and try to understand the basic concept of generalized linear mixed model (GLMM). Section 3.1 in this paper described a Monte Carlo EM (MCEM) method to derive Maximum Likelihood Estimates (MLE). Please use your own software (Matlab, R, C++, etc.) to perform following simulation study and answer related questions.

1.1 Model and Notations

In this project, we consider a clustering problem. Suppose we have observed n observations, each observation is a binary process, i.e. the response $Y_{ij} = 0$ or 1, $i = 1, ..., n$, $j = 1, ..., T$. Here n is the number of subjects and T is the length of observation. In general, T might vary across subjects, time points may also be different. In this project, however, we simply assume that all subjects have common time length and time points. We also assume that these subjects belong to two clusters. For each cluster, the conditional expectation of response variable is:

$$
\begin{aligned}
P_{ij} &\equiv E(Y_{ij}|U_i = 1, X_{1,ij}, Z_{1,i}) = g^{-1}(\beta_1 X_{1,ij} + Z_{1,i}) \\
P_{ij} &\equiv E(Y_{ij}|U_i = 2, X_{2,ij}, Z_{2,i}) = g^{-1}(\beta_2 X_{2,ij} + Z_{2,i}) \qquad (1)
\end{aligned}
$$

where U is cluster membership, $X_{c,ij}$ and $Z_{c,i}$ ($c = 1, 2$) are fixed and random effects, respectively. The link function $g^{-1}(x) = \frac{\exp(x)}{1+\exp(x)}$ is given. In a typical clustering problem, U is usually unknown, and hence we treat U as another random effect.

For random effects, we assume that $Z_{c,i} \sim N(0, \sigma_c^2)$ and $P(U = 1) = \pi_1$ (then $\pi_2 = 1-\pi_1$). Then the parameter to be estimated is $\Omega \doteq \{\beta_1, \beta_2, \sigma_1, \sigma_2, \pi_1\}$. Treating random effects as missing data, one can write the complete data likelihood function as

$$
L(\mathbf{\Omega}|Y_{ij}, U_i, Z_{1,i}, Z_{2,i}) = \prod_{i=1}^{n} \prod_{c=1}^{2} \left\{ \pi_c f_c(Z_{c,i}) \left[\prod_{j=1}^{T} f_c(Y_{ij}|Z_{c,i}) \right] \right\}^{\omega_{ic}}, \qquad (2)
$$

where $f_c(Z_{c,i})$ is the density function of Normal distribution, $f_c(Y_{ij}|Z_{c,i}) = P_{ij}^{Y_{ij}}(1-P_{ij})^{1-Y_{ij}}$. ω_{ic} is the dummy variable of U_i, i.e.

$$
\omega_{ic} = \begin{cases} 1 & \text{if subject i belongs to cluster c} \\ 0 & \text{otherwise,} \end{cases}
$$

1.2 Simulation Setup and Requirement

Generate 100 simulations. In each simulation, set $n = 100$ and $T = 10$. The true values of parameter are: $\beta_1 = \beta_2 = 1$, $\pi_1 = 0.6$,$\sigma_1 = 2$, and $\sigma_2 = 10$.

Before you start, use $N(0, 1)$ to generate the fixed effect X, and use them for all 100 simulations. Please follow the paper mentioned earlier and use MCEM to evaluate the log-likelihood function. In the E-step, perform $K = 500$ Gibbs sampling incorporated with a Metropolis-Hastings step, and drop the first 100 as a burn-in procedure.

1.3 Your Report

(1) Please write down the form of Monte Carlo average of log-likelihood (which you are going to evaluate)

(2) Please write down details of the EM algorithm you use for this simulation, especially the Metropolis-Hastings steps.

(3) What are your initial values? What is your convergence rule?

(4) How to accelerate your EM algorithm? Any improvement you can observe?

(4) Try different numbers of simulations: 200,300,...,1000. And plot the corresponding MSE.

(5) Write a report in either Chinese or English. Please attach your code to your report.