**Likelihood based missing data analysis in multivariate crossover trials**

Savita Pareek, Siuli Mukhopadhyay, and Kalyan Das

Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, 400076

We have run all the codes on

R version 3.6.1 (2019-07-05)

Platform: x86\_64-w64-mingw32/x64 (64-bit)

Running under: Windows >= 8 x64 (build 9200)

*Supplementary information for Simulation* i.e., Table 1 and Figure 3, 4 and 5:

For Table 1 (see ‘simulation code.R’): The below steps are for missing probability 15% (for different value of missing change the number in line #74). The .csv file to be read are ‘sample in seq1.csv’ at line # 126 and ‘sample in seq2.csv’ at line # 127.

1. Line # 1 to 9 is used for cleaning the environment and calling the required libraries (doParallel, MASS, psych).
2. Line # 10 to 58: we define a function which updates the parameter estimates at particular iteration using our proposed algorithm.
3. Line # 60 to 127: we define the true value of parameters and create the X and Z matrices.
4. Line # 131 to 267: a for loop is created which runs all 500 simulations one by one. For a particular simulation:
5. Line # 136 to 138 assigns initial values to the parameters
6. Line # 145 to 166 generates the missing data indicator using Bernoulli random variable and corresponding missing and observed matrices X, Z and vector y are obtained.
7. Line # 168 to 185: a sample of size 2000 is generated from the distribution .
8. Line # 186 to 197: parameter estimates are obtained using the function defined in (b).
9. Line # 201 to 265: at the convergence of MCEM algorithm we use multiple imputation technique (discussed in section 4) to obtain the standard errors.
10. Line # 270 to 286: once all the 500 simulations are performed, we make the desired table by computing the coverage probability, average mean square error, average bias.

For Figure 3 (see ‘asymptotic normality.R’): The below steps explains the working of the R code. The .csv file to be read is ‘15% missing all estimates and se.csv’ at line # 9 (read ‘25% missing all estimates and se.csv’ to asses 25% missing case).

1. Line #1 to 3 is used for cleaning the environment and calling the required libraries (ggplot2, EnvStats).
2. Line # 9 to 20: we store the data for 100 simulation and 500 simulations in different matrices and standardize the estimates.
3. Line # 23 to 38: we find the empirical density corresponding to each estimate.
4. Line # 41 to 46: a ggplot graph is made using geom\_line().

For Figure 4, 5 (see ‘test\_gene\_effects.R’, ‘test\_treatments.R’ respectively): The below steps explains the working of the R code.

1. Line # 1 to 11 is used for cleaning the environment and calling the required libraries (doParallel, MASS, psych).
2. Line # 13 to 60: we define a function which updates the parameter estimates at particular iteration using our proposed algorithm.
3. Line # 62 to 115: we define the true value of parameters and create the X and Z matrices.
4. Line # 118: X matrix is created under reduced model (or under ).
5. Line # 120 to 133: A sample of size one thousand is generated from the full model.
6. Line # 136 to 304: a for loop is created which runs all 1000 simulations one by one. For a particular simulation:
7. Line # 140 to 240 computes estimates from reduced model and full model.
8. Line # 244 to 300 calculates likelihood of reduced and full model.
9. Line # 301 to 304: critical value is compared with the test statistic value .
10. Line # 305 to 308: empirical power is evaluated.

*Supplementary information for the gene expression data* i.e., Table 2, 3 and Figure 1, 2:

For Figure 1, 2 (see ‘real data\_figure\_1\_2.R’): The below steps explains the working of the R code.

1. Line # 1 to 7 is used for cleaning the environment and calling the required libraries (ggpubr, ggplot2).
2. Line # 11: read the file ‘anova 3 way allseq.csv’.
3. Line # 12 to 29: prepare the data frame for ggplot2 function.
4. Line # 30 to 82: make the interaction plot and export them in png file.

For Table 2, 3 (see ‘gene\_data code.R’): The below steps are for Table 2, complete cases and MAR responses. For log responses, change response in line # 128, 155, 156, 157 to log(response). Similarly, for Table 3, read the files for 21% or 24% missing for sequence 1, 2 and 3 in line # 136, 137 and 138 respectively. The .csv file to be read are:

1. ‘anova 3 way all seq.csv’ at line # 114.
2. ‘Seq1\_9%’ or ‘seq1\_21%’ or ‘seq1\_24%’ at line # 136.
3. ‘seq2\_9%’ or ‘seq2\_21%’ or ‘seq2\_24%’ at line # 137.
4. ‘seq3\_9%’ or ‘seq3\_21%’ or ‘seq3\_24%’ at line # 138.

Working of the R code:

1. Line # 1 to 12 is used for cleaning the environment and calling the required libraries (doParallel, MASS, psych, nlme, lmeInfo, insight).
2. Line # 14 to 61: we define a function which updates the parameter estimates at particular iteration using our proposed algorithm.
3. Line # 63 to 111: we create X and Z matrices; at line # 102, we define a variable ‘lu’ and assign it 5, meaning that we want to find the p-value for . Change the value in line # 102 to obtain p-value for other parameter estimates.
4. Line # 114 to 133: complete case analysis using ‘lme’ function.
5. Line # 136 to 174: read the files seq1\_9%.csv, seq2\_9%.csv and seq3\_9%.csv at line # 136, 137 and 138 respectively, create the new X and Z matrices and assign the initial estimate of parameters as the complete case estimates.
6. Line # 175 to 289: find the parameter estimate and standard error using our proposed algorithm.
7. Line # 292 to 333: calculate likelihood under full model and compute AIC and BIC and RMSE values.
8. Line # 335 to 410: find the estimates and likelihood value under reduced model in order to get the p-value for .