Distributed Monotonic Overrelaxed Method for Random Effects Model with Missing Response

Yu Li, Guangbao Guo

Abstract—**In recent years, data has been developing rapidly. However, Monte Carlo expectation maximization (MCEM) method and parameter extended expectation maximization (PX-EM) method experience slow running speeds and cannot meet the requirements of data processing in the context of big data and distributed data, due to the lack of response in random effects model. This article introduces a distributed monotonic overrelaxation PX-EM (DMOPX-EM) method based on the PX-EM method by using a overrelaxation factor. The proposed method not only has less iteration time, but also has higher estimation accuracy when processing big or distributed data.**

Index Terms—**distributed processing, expectation maximization, random effects model, residual maximum likelihood.**

I. Introduction

THE missing data issue in random effects model is gaining increasing attention from scholars. Yucel [1] (2008) and Kline et al. [2] (2017) have already studied the imputation methods for level 1 variables. Many scholars have conducted in-depth research on response missing problems. Field and Welsh [3] (2007) shows various methods for handling clustered data, but these methods have dependency assumptions. The use of the expectation maximization (EM) method to calculate maximum likelihood (ML) estimates by Hall [4] (2000), but the EM method has advantages over the Newton-Raphson or Fisher scoring method when dealing with models with a large number of covariance parameters. The Monte Carlo EM (MCEM) method proposed by Ibrahim et al. [5] (2001) has been used for parameter estimation in choice models with non-negligible missing response data, with improvements made by Ibrahim and Molenberghs [6] (2009). Yu [7] (2012) shows the idea of overrelaxed methods to accelerate the EM algorithm, with a focus on preserving its simplicity and monotonic convergence. Diffey et al. [8] (2017) shows the parameter extension EM (PX-EM) method under the residual maximum likelihood (REML) of the random effects model and proved that this method converges to a local maximum of the residual log-likelihood function, resulting in better estimation performance. However, this method requires a large number of iterations. Guo et al. [9] (2020) shows statistical calculations under a distributed framework. Guo [10] (2012) shows the representation of parallel statistical computation, was provided to facilitate the statistical computation process. Guo et al. [11] (2015) shows the parallel maximum likelihood estimator for multiple linear regression models, which has experienced significant development in the field of statistics.

In this paper, we propose an monotonic overrelaxed PX-EM (MOPX-EM) method based on the PX-EM method. This method is similar to the Newton acceleration PX-EM method. The basic idea is to obtain the parameter estimation results of multiple processes in the M step based on the sufficient utilization of information in the E step, and introduce overrelaxation factors to accelerate the PX-EM method. Under distributed data, we perform distributed processing on the PX-EM method to obtain distributed PX-EM (DPX-EM) method. By combining the DPX-EM method with the MOPX-EM method, we obtain the distributed monotonic overrelaxed PX-EM (DMOPX-EM) method. This method divides the data into blocks and uses extremum statistics to improve the accuracy of estimation. In this way, not only the computational program is simplified and the computation time is reduced, but also the estimation accuracy is improved.

II. The Method

A. Random Effects Model Of Missing Response

First, we introduce the response missing data. Considering the $n \times 1$ dimensional truth data matrix Y^*, when the truth value exists, Y^* is a complete matrix, expressed as:

$$Y^* = (y^*_{ij})_{n\times 1} = (Y^*_1, ..., Y^*_n)^\top. \tag{1}$$

We construct the $n \times 1$ dimensional indicator matrix $\Re$, expressed as:

$$\Re = (\gamma_1, ..., \gamma_n)^\top,$$

where

$$\gamma_i = \begin{cases} 1, \text{if} \quad Y_{\text{mis}} \quad \text{is} \quad \text{unobserved.} \\ 0, \text{if} \quad Y_{\text{obs}} \quad \text{is} \quad \text{observed.} \end{cases}$$

The $n \times 1$ dimensional missing data matrix Y is expressed as:

$$Y = Y^* \bigotimes \Re = (Y_1, ..., Y_n)^\top, \tag{2}$$

where $\bigotimes$ is the Hadamard product, for matrix $\Re$, if $\gamma_i = 1$, position i corresponds to the observed value, if $\gamma_i = 0$, position i corresponds to the missing value, and the observed and missing values in matrix Y are expressed as Y_{obs} and Y_{mis}, respectively. In addition, we define n_{ob} as the number of observed data, and n_{na} as the number of missing data, with $n = n_{ob} + n_{na}$.

The linear mixed model at two levels is expressed as:

$$Y = X\beta + Zb + e, \tag{3}$$

where $X = (X_1, ..., X_n)^\top$ is a matrix of known fixed-effect covariates of $n \times p$, $\beta = (\beta_1, ..., \beta_p)^\top$ is an unknown parameter vector of $p \times 1$, the random effect matrix $Z =$

Manuscript received 3 July, 2023; revised 5 December, 2023.

This work was supported by a grant from National Social Science Foundation Project under project ID 23BTJ059, a grant from Natural Science Foundation of Shandong under project ID ZR2020MA022, and a grant from National Statistical Research Program under project ID 2022LY016.

Yu Li is a postgraduate student of Mathematics and Statistics, Shandong University of Technology, Zibo, China. (e-mail: Marlboro0608@163.com).

Guangbao Guo is a professor of Mathematics and Statistics, Shandong University of Technology, Zibo, China (corresponding author to provide phone:15269366362; e-mail: ggb11111111@163.com).

$(Z_1, ..., Z_n)^\top$ is a matrix of known random effect covariates of $n \times q$, $b = (b_1, ..., b_q)^\top$ is a $q \times 1$ parameter vector of random effects, distributed as $b_i \sim N(0, \sigma_b^2)$, random error $e = (e_1, ..., e_n)^\top$, distributed as $e_i \sim N(0, \sigma_e^2)$. D is the variance matrix of b, G is the variance matrix of e, $I_{n\times n}$ is the array of units.

For PX-EM method (Diffey et al, 2017), the random effects model is reformulated by introducing a secondary parameter λ, which extending the random effects model to

$$Y = X\beta + Z\Lambda f + e, \tag{4}$$

where $\Lambda = \Lambda(\lambda)$ is a $q \times q$ real reversible matrix, which is a function of the $v \times 1$ auxiliary parameter vector λ.

$$\begin{bmatrix} f \\ e \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} D & 0 \\ 0 & R \end{bmatrix})$$

with $b = \Lambda f$ and $G = \Lambda D \Lambda^\top$. For missing response variable Y, the method has fewer iterations and better outcomes compared to the EM.

B. The DMOPX-EM Method

The full data set $\{X, Z, Y\}$ in Equation (4) is divided into M data subsets, that is $\{X_{I_m}, Z_{I_m}, Y_{Im}\}$, here

$$X = (X_1, ..., X_n)^\top = (X_{I_1}, ..., X_{I_M}), \tag{5}$$

$$Z = (Z_1, ..., Z_n)^\top = (Z_{I_1}, ..., Z_{I_M}), \tag{6}$$

$$Y = (Y_1, ..., Y_n)^\top = (Y_{I_1}, ..., Y_{I_M}), \tag{7}$$

where $\bigcup I_m = I$, and $I \in (1, ..., n)$.

The amount of data on each machine can be equal or not equal, here we have an aliquot, that is, there are n_{I_m} subset number on mth machine, expressed as $Y_{I_m} = (Y_{I_m1}, ..., Y_{I_m n_{I_m}})$ and $\sum n_{I_m} = n$. The data are chunned, $\{Y_{obs,I_m}, X_{obs,I_m}, Z_{obs,I_m}\}$ are the observed data after projection and $\{Y_{mis,I_m}, X_{mis,I_m}, Z_{mis,I_m}\}$ are the missing data after projection. β_{I_m} and b_{I_m} are unknown parameter vectors of fixed effects and random effects in distributed manner, respectively. To estimate random effects model, the conditional derivation of REML starts by considering the transformation:

$$L_{I_m}^\top Y_{I_m} = \begin{bmatrix} L_{1,I_m}^\top Y_{obs,I_m} \\ L_{2,I_m}^\top Y_{obs,I_m} \end{bmatrix} = \begin{bmatrix} Y_{1,I_m} \\ Y_{2,I_m} \end{bmatrix},$$

where $L_{I_m} = (L_{1,I_m}, L_{2,I_m})$ is a non-singular matrix. L_{1,I_m} and L_{2,I_m} are $n_{ob,I_m} \times p$ and $n_{ob,I_m} \times (n_{ob,I_m} - p)$ matrices, respectively, are all column rank and satisfy $L_{1,I_m}^\top X_{obs,I_m} = I_p$ and $L_{2,I_m}^\top X_{obs,I_m} = 0$. The distribution of the transformed data as follows:

$$\begin{bmatrix} Y_{1,I_m} \\ Y_{2,I_m} \end{bmatrix} \sim N(\begin{bmatrix} \beta_{I_m} \\ 0 \end{bmatrix}, \begin{bmatrix} L_{1,I_m}^\top H_{I_m} L_{1,I_m} & L_{1,I_m}^\top H_{I_m} L_{2,I_m} \\ L_{2,I_m}^\top H_{I_m} L_{1,I_m} & L_{2,I_m}^\top H_{I_m} L_{2,I_m} \end{bmatrix}),$$

where

$$H_{I_m} = Z_{obs,I_m} G_{I_m} Z_{obs,I_m}^\top + R_{I_m}. \tag{8}$$

For the REML estimation of the variance parameters, a log-likelihood function on Y_{2,I_m} is expressed as

$$\ell(\theta; Y_{2,I_m}) = -\frac{1}{2}[\log(\det(L_{2,I_m}^\top H_{I_m} L_{2,I_m})) + Y_{2,I_m}^\top (L_{2,I_m}^\top H_{I_m} L_{2,I_m})^{-1} Y_{2,I_m}].$$

and $\theta = (\beta_{I_m}, b_{I_m})$ is the vector of variance parameters. From Verbyla [12] (1990), we can get

$$\begin{aligned} P_{I_m} &= L_{2,I_m}(L_{2,I_m}^\top H_{I_m} L_{2,I_m})^{-1} L_{2,I_m}^\top \\ &= H_{I_m}^{-1} - H_{I_m}^{-1} X_{obs,I_m} (X_{obs,I_m}^\top H_{I_m}^{-1} X_{obs,I_m})^{-1} \\ &\quad X_{obs,I_m}^\top H_{I_m}^{-1}. \end{aligned} \tag{9}$$

So we can get log-likelihood function

$$\begin{aligned} \ell(\theta; Y_{2,I_m}) = -\frac{1}{2}[&\log(\det(H_{I_m})) + \log(\det(X_{obs,I_m}^\top \\ &H_{I_m}^{-1} X_{obs,I_m})) + Y_{obs,I_m}^\top P_{I_m} Y_{obs,I_m}]. \end{aligned}$$

The least squares method, the initial estimator of β_{I_m} is expressed as:

$$\hat{\beta_{I_m}} = (X_{obs,I_m}^\top H_{I_m}^{-1} X_{obs,I_m})^{-1} X_{obs,I_m}^\top H_{I_m}^{-1} Y_{obs,I_m}. \tag{10}$$

When represented by matrix notation, the coefficient matrix of the Henderson mixture model equation can be written as

$$\begin{aligned} C &= \begin{bmatrix} X_{obs,I_m}^\top R_{I_m}^{-1} X_{obs,I_m} & X_{obs,I_m}^\top R_{I_m}^{-1} Z_{obs,I_m} \\ Z_{obs,I_m}^\top R_{I_m}^{-1} X_{obs,I_m} & Z_{obs,I_m}^\top R_{I_m}^{-1} Z_{obs,I_m} + G_{I_m}^{-1} \end{bmatrix} \\ &= \begin{bmatrix} C_{XX} & C_{XZ} \\ C_{ZX} & C_{ZZ} \end{bmatrix}. \end{aligned} \tag{11}$$

The inverse of C_{ZZ} in Equation (11) is:

$$C^{ZZ} = (Z_{obs,I_m}^\top S_{I_m} Z_{obs,I_m} + G_{I_m}^{-1})^{-1}, \tag{12}$$

where $S_{I_m} = R_{I_m}^{-1} - R_{I_m}^{-1} X_{obs,I_m} (X_{obs,I_m}^\top R_{I_m}^{-1} X_{obs,I_m})^{-1} X_{obs,I_m}^\top R_{I_m}^{-1}$.

The derivation of DPX-EM and DMOPX-EM methods is achieved by considering the fixed effect vector as a random effect where the variance tends to infinity, see Zhou and Tang [13] (2021). We call this implementation the random effects method. Using this method, we will assume that the $\beta_{I_m} \sim N(0, B)$. We define the joint distribution of Y_{obs,I_m}, b_{I_m} and e_{I_m} as

$$\begin{bmatrix} Y_{obs,I_m} \\ b_{I_m} \\ e_{I_m} \end{bmatrix} \sim N(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} F_{I_m} & Z_{obs,I_m} G_{I_m} & R_{I_m} \\ G_{I_m} Z_{obs,I_m}^\top & G_{I_m} & 0 \\ R_{I_m} & 0 & R_{I_m} \end{bmatrix}).$$

Among them, have

$$\begin{aligned} F &= X_{obs,I_m} B X_{obs,I_m}^\top + Z_{obs,I_m} G_{I_m} Z_{obs,I_m}^\top + R_{I_m} \\ &= X_{obs,I_m} B X_{obs,I_m}^\top + H_{I_m}. \end{aligned}$$

with

$$\begin{aligned} F^{-1} = H_{I_m}^{-1} - H_{I_m}^{-1} X_{obs,I_m} (B^{-1} + X_{obs,I_m}^\top H_{I_m}^{-1} X_{obs,I_m})^{-1} \\ X_{obs,I_m}^\top H_{I_m}^{-1}. \end{aligned}$$

The random effects method assumes that the variance of β_{I_m} tends to infinity, namely $B^{-1} \longrightarrow 0$. Therefore ,

$$\begin{aligned} F^{-1} \longrightarrow H_{I_m}^{-1} - H_{I_m}^{-1} X_{obs,I_m} (X_{obs,I_m}^\top H_{I_m}^{-1} X_{obs,I_m})^{-1} \\ X_{obs,I_m}^\top H_{I_m}^{-1} = P_{I_m}. \end{aligned}$$

The joint distribution is

$$e_{I_m} | Y_{obs,I_m} \sim N(R_{I_m} P_{I_m} Y_{obs,I_m}, W C^{-1} W^\top). \tag{13}$$

$$b_{I_m} | Y_{obs,I_m} \sim N(G_{I_m} Z_{obs,I_m}^\top P_{I_m} Y_{obs,I_m}, C^{ZZ}), \tag{14}$$

where $W = (X_{obs,I_m}, Z_{obs,I_m}) \in \mathbf{R}^{n_{ob,I_m} \times (p+q)}$.

E-step: For given $\theta^{(0)}_{I_m}$, at the iteration t,

$$Q_{I_m} = -\frac{1}{2}[\log(\det(L^\top_{2,I_m} R_{I_m} L_{2,I_m})) + \log(\det(G_{I_m})) + (Y_{obs,I_m} - Z_{obs,I_m}\widetilde{b}^{(t)}_{I_m})^\top S_{I_m}(Y_{obs,I_m} - Z_{obs,I_m}\widetilde{b}^{(t)}_{I_m}) + tr(Z^\top_{obs,I_m} S_{I_m} Z_{obs,I_m} C^{ZZ^{(t)}}) + tr(G^{-1}_{I_m} C^{ZZ^{(t)}})],$$

where

$$\widetilde{b}^{(t)}_{I_m} = G^{(t)}_{I_m} Z^\top_{obs,I_m} P^{(t)}_{I_m} Y_{obs,I_m},$$
$$\widetilde{e}^{(t)}_{I_m} = R^{(t)}_{I_m} P^{(t)}_{I_m} Y_{obs,I_m}.$$

M-step: Update the parameter estimates of each component under the DPX-EM method as

$$d^{(t+1)}_{I_m} = -\frac{1}{q}[(\widetilde{b}^{(t)}_{I_m})^\top \widetilde{b}^{(t)}_{I_m} + tr(C^{ZZ^{(t)}})], \tag{15}$$

$$\hat{\sigma}^{2^{(t+1)}}_{e,I_m,PXEM} = \frac{1}{n_{ob,I_m} - p}[(Y_{obs,I_m} - Z_{obs,I_m}\widetilde{b}^{(t)}_{I_m})^\top K (Y_{obs,I_m} - Z_{obs,I_m}\widetilde{b}^{(t)}_{I_m}) + T], \tag{16}$$

$$\lambda^{(t+1)}_{I_m} = \frac{Y^\top_{obs,I_m} K Z_{obs,I_m}\widetilde{b}^{(t)}_{I_m}}{(\widetilde{b}^{(t)}_{I_m})^\top Z^\top_{obs,I_m} K Z_{obs,I_m}\widetilde{b}^{(t)}_{I_m} + T}, \tag{17}$$

where

$$K = I_{n,I_m} - X_{obs,I_m}(X^\top_{obs,I_m} X_{obs,I_m})^{-1} X^\top_{obs,I_m},$$
$$T = tr(Z^\top_{obs,I_m} K Z_{obs,I_m} C^{ZZ^{(t)}}).$$

$$\hat{\sigma}^{2^{(t+1)}}_{b,I_m,PXEM} = (\lambda^{(t+1)}_{I_m})^2 d^{(t+1)}_{I_m}. \tag{18}$$

For DPX-EM method, given the initial value $\beta^{(0)}_{I_m}$, $b^{(0)}_{I_m}$, $\sigma^{2^{(0)}}_{b,I_m}$, $\sigma^{2^{(0)}}_{e,I_m}$, $m = 1, ..., M$, first update the parameter value $\sigma^{2^{(t+1)}}_{b,I_m,MOPXEM}$ of the $t+1$ iteration is

$$\sigma^{2^{(t+1)}}_{b,I_m,MOPXEM} = (1 + \omega^{(t)}_{I_m})\sigma^{2^{(t+1)}}_{b,I_m,PXEM} - \omega^{(t)}_{I_m}\sigma^{2^{(t)}}_{b,I_m,MOPXEM}. \tag{19}$$

Generally speaking, we call $\omega^{(t)}_{I_m}$ the actual overrelaxation factor and ω_{I_m} the nominal overrelaxation factor. The relationship of them is as follows:

$$\omega^{(t)}_{I_m} = \frac{\omega_{I_m} r_{I_m}}{1 + \omega_{I_m} - \omega_{I_m} r_{I_m}},$$

$$r_{I_m} = \frac{\sigma^{2^{(t+1)}}_{b,I_m,PXEM}}{\sigma^{2^{(t+1)}}_{b,I_m,MOPXEM}}, \omega_{I_m} \in [0, 1].$$

With the nominal overrelaxation factor we give the parameter value $\sigma^{2^{(t+1)}}_{e,I_m,MOPXEM}$ for the $t+1$ iteration as

$$\sigma^{2^{(t+1)}}_{e,I_m,MOPXEM} = (1 + \omega_{I_m})\sigma^{2^{(t+1)}}_{e,I_m,PXEM} - \omega_{I_m}\sigma^{2^{(t)}}_{e,I_m,MOPXEM}. \tag{20}$$

From this, we can obtain that covariance estimation matrices G_0 and R_0 of b and e.

After the iteration stops, from the estimated $\sigma^{2^{(t+1)}}_{b,I_m,MOPXEM}$ in Equation (19) and $\sigma^{2^{(t+1)}}_{e,I_m,MOPXEM}$ in Equation (20), we can obtain

$$\hat{\beta}_0 = (X^\top_{obs,I_m} H^{-1}_0 X_{obs,I_m})^{-1} X^\top_{obs,I_m} H^{-1}_0 Y_{obs,I_m}, \tag{21}$$

$$\hat{b_0} = G_0 Z^\top_{obs,I_m} P_0 Y_{obs,I_m}, \tag{22}$$

where

$$H^{-1}_0 = Z_{obs,I_m} G_0 Z^\top_{obs,I_m} + R_0,$$
$$P_0 = H^{-1}_0 - H^{-1}_0 X_{obs,I_m}(X^\top_{obs,I_m} H^{-1}_0 X_{obs,I_m})^{-1} X^\top_{obs,I_m} H^{-1}_0.$$

Missing values were interpolated, with

$$\hat{Y}_{mis,I_m} = X_{mis,I_m}\hat{\beta}_0 + Z_{mis,I_m}\hat{b_0}. \tag{23}$$

So, $\hat{Y_{I_m}} = \{\hat{Y}_{mis,I_m}, \hat{Y}_{obs,I_m}\}$.

III. Numerical Analysis

A. Prepare Knowledge

The experiments in this chapter were conducted on a computer with a Win 10 64-bit operating system and used simulation software with R software. Based on this foundation, five interpolation methods were employed to handle missing responses: MCEM, PX-EM, MOPX-EM, DPX-EM and DMOPX-EM. In the simulation analysis, the mean squared error (MSE) and mean absolute error (MAE) were used to evaluate the deviation between the true value and the estimated value. In real data analysis, the mean squared forecast error (MSFE) and mean relative error (MRE) are used to evaluate the deviation between the true value and the estimated value.

In non-distributed environment, performance metrics are expressed as

$$\text{MSE}(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \text{MAE}(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i|.$$

$$\text{MSFE}(\hat{Y}) = \frac{1}{3n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2, \text{MRE}(\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n}|\frac{Y_i - \hat{Y}_i}{Y_i}|.$$

where Y_i represents the true value after projection, and $\hat{Y}_i$ represents the estimated value after projection. The range of MSE, MAE, MSFE and MRE is $[0, 1]$. When the value is closer to 0, it indicates a better interpolation effect.

In distributed environment, performance metrics are expressed as

$$\text{MSE}(\hat{Y_{I_m}}) = \frac{1}{n_{I_m}}\sum_{m=1}^{M}(Y_{I_m} - \hat{Y_{I_m}})^2,$$

$$\text{MAE}(\hat{Y_{I_m}}) = \frac{1}{n_{I_m}}\sum_{m=1}^{M}|Y_{I_m} - \hat{Y_{I_m}}|,$$

$$\text{MSFE}(\hat{Y_{I_m}}) = \frac{1}{3n_{I_m}}\sum_{m=1}^{M}(Y_{I_m} - \hat{Y_{I_m}})^2,$$

$$\text{MRE}(\hat{Y_{I_m}}) = \frac{1}{n_{I_m}}\sum_{m=1}^{M}|\frac{Y_{I_m} - \hat{Y_{I_m}}}{Y_{I_m}}|.$$

where Y_{I_M} represents the true value after projection, and $\hat{Y_{I_m}}$ represents the estimated value after projection.

B. Simulation

In this section, we will conduct two simulation analyses to validate the stability and sensitivity of the five interpolation methods in handling missing responses.

1) Stability analysis: Varying (M, ω_{I_m}, MR) with fixed (n, p, q)

For $(n, p, q) = (600, 10, 5)$. We verified the impact of M on DPX-EM method, the impact of M and ω_{I_m} on DMOPX-EM method, and the impact of missing ratio (MR) on EM, MCEM, PX-EM, MOPX-EM, DPX-EM, and DMOPX-EM methods. The range of values for the aforementioned control factors is shown in Table 1.

TABLE I
THE NUMERICAL SELECTION OF EACH CONTROL FACTOR IN SIMULATION 1

M	5	10	15	20
ω_{I_m}	0.15	0.3	0.45	0.6
MR	0.1	0.2	0.3	0.4

Case 1. Varying M on DPX-EM

For $MR = 0.1$, change $M = \{5, 10, 15, 20\}$, Figure 1 presents the comparison of MSE values and MAE values of DPX-EM with respect to the changes in M. Subfigure (a) illustrates the variations in MSE values of the estimated response variable, while subfigure (b) represents the changes in MAE values. From the observation of Figure 1, it can be seen that the trends in the change of MSE values and MAE values are generally similar. As M increases, the values of MSE and MAE first decrease and then increase. When M increases from 5 to 10, the MSE value decreases from 0.0917928 to 0.08130032, and the MAE value decreases from 0.08367735 to 0.08272665, with a small magnitude of change. When M increases from 10 to 15, the MSE value decreases from 0.08130032 to 0.004041909, and the MAE value decreases from 0.08272665 to 0.01005225, with a larger magnitude of decrease. When M increases from 15 to 20, the MSE value increases from 0.004041909 to 0.01391535, and the MAE value increases from 0.01005225 to 0.02847215. When $M = 15$, DPX-EM achieves the minimum values for MSE and MAE.

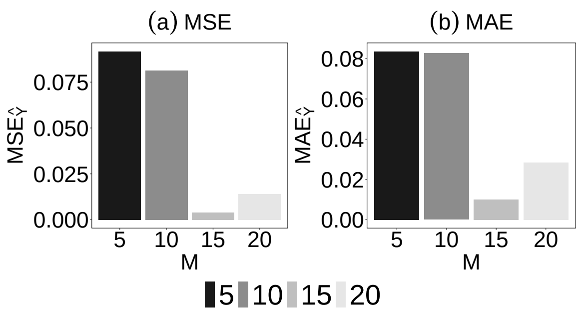

Fig. 1. Comparison Results Of DPX-EM In The Simulation

Case 2. Varying M and ω_{I_m} on DMOPX-EM

For $MR = 0.1$, changes $M = \{5, 10, 15, 20\}$ and $\omega_{I_m} = \{0.15, 0.3, 0.45, 0.6\}$. Figure 2 presents the comparison of MSE values and MAE values for DMOPX-EM method with respect to the changes in M and ω_{I_m}. Subfigure (a) illustrates the variations in MSE values of the estimated response variables, while subfigure (b) represents the changes in MAE values. Based on figure 2, it can be observed that the value of ω_{I_m} has little impact on the DMOPX-EM method. Specifically, when the parameter M is fixed and different values of ω_{I_m} are considered, the MSE values and MAE values remain quite consistent and stable. On the other hand, with a fixed parameter ω_{I_m} and varying values of M, the MSE values and MAE values of DMOPX-EM method decrease as M increases. For instance, when $M = 5$, the maximum MSE value is 0.1001999 and the maximum MAE value is 0.08783184; as M further increases, the MSE values and MAE values fluctuate significantly. When $M = 20$, the minimum MSE value is 0.00139224, while the minimum MAE value is 0.006812342.

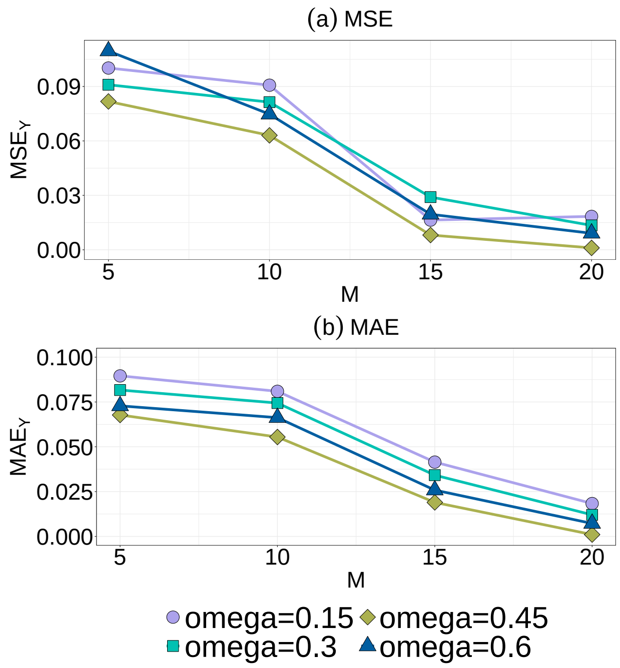

Fig. 2. Comparison Results Of DMOPX-EM In The Simulation

Case 3. Varying MR on EM, MCEM, PX-EM, MOPX-EM, DPX-EM and DMOPX-EM

For $(M, \omega_{I_m}) = (5, 0.15)$, change $MR = \{0.1, 0.2, 0.3, 0.4\}$. Figure 3 presents the comparison results of MSE values and MAE values for five interpolation methods with respect to the changes in MR. Subfigure (a) illustrates the variations in MSE values of the estimated response variables, while subfigure (b) represents the changes in MAE values. Based on figure 3, it can be seen that when MR is fixed, the MSE values and MAE values of the five methods are not significantly different, but it is still noticeable that the DMOPX-EM method has lower MSE values and MAE values compared to the other methods. As MR increases, the MSE values and MAE values of the five interpolation methods also increase, and the DMOPX-EM method consistently has lower MSE values and MAE values than the other methods. When $MR = 0.1$, the DMOPX-EM method achieves the minimum MSE value of 0.1001999 and the minimum MAE value of 0.08756552. When $MR = 0.4$, the DMOPX-EM method obtains the maximum MSE value of 0.4299362 and the maximum MAE value of 0.3087811.

2) Sensibility analysis: Varying (n, p, q) with fixed (M, ω_{I_m}, MR)

For $(M, \omega_{I_m}, MR) = (5, 0.15, 0.1)$. We examine the MSE values and MAE values of the five interpolation methods under different sample size n, dimensionality p, and q for sensitivity analysis.

Case 4. Varying n with fixed $(M, \omega_{I_m}, MR, p, q)$

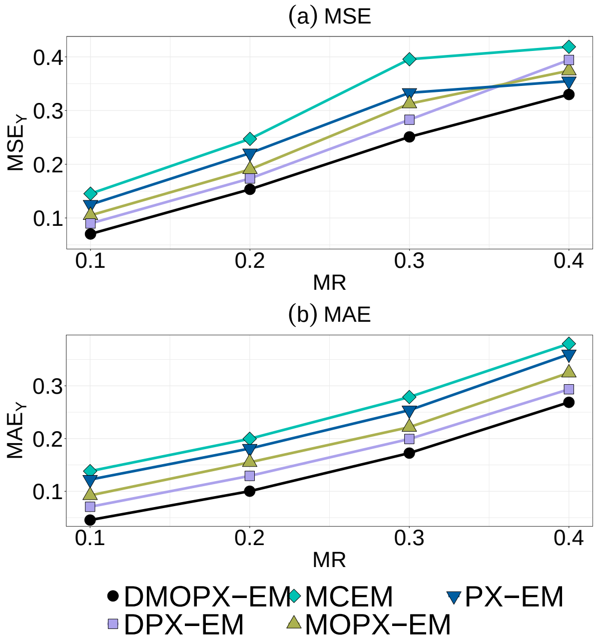

Fig. 3. Comparison Results Of The Five Methods In The Simulation

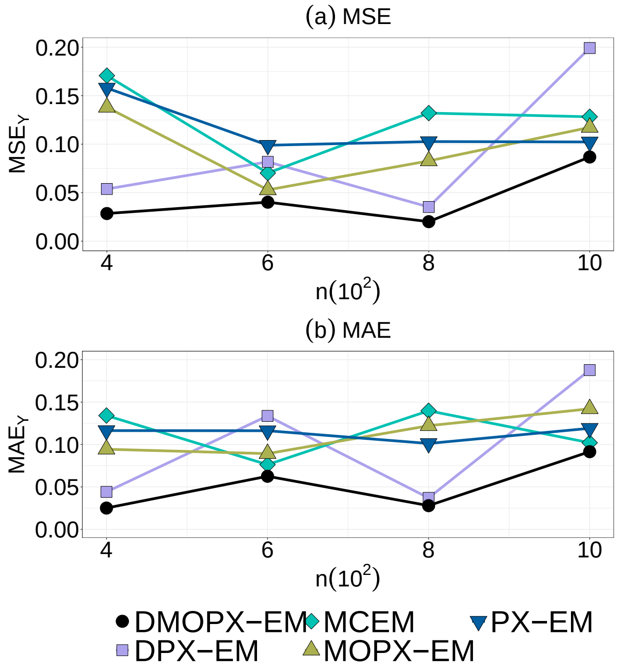

Fig. 4. Comparison Results Of The Five Methods In The Simulation

For $(M, \omega_{I_m}, MR, p, q) = (5, 0.15, 0.1, 10, 5)$, change $n = \{400, 600, 800, 1000\}$. Figure 4 presents the comparison results of MSE values and MAE values for five interpolation methods with respect to the changes in n. Subfigure (a) illustrates the variations in MSE values of the estimated response variables, while subfigure (b) represents the changes in MAE values. From figure 4, it can be observed that as n increases, the MSE values and MAE values of MCEM,PX-EM and MOPX-EM method show a significant variation, initially decreasing, then increasing, and then decreasing. The MSE values and MAE values of the DPX-EM and DMOPX-EM methods, on the other hand, show minimal differences. As n increases, their MSE values and MAE values first increase, then decrease, and then increase again. It is evident that the DMOPX-EM method consistently yields lower MSE values and MAE values compared to the other methods, indicating that the DMOPX-EM method has the best interpolation effect. When $n = 400$, the DMOPX-EM method achieves the minimum MSE value of 0.02850028 and the minimum MAE value of 0.02509272. When $n = 1000$, the MSE value is 0.09680235 and the MAE value is 0.07953884.

Case 5. Varying p with fixed $(M, \omega_{I_m}, MR, n, q)$

For $(M, \omega_{I_m}, MR, n, q) = (5, 0.15, 0.1, 400, 5)$, change $p = \{10, 15, 20, 25\}$. Figure 5 presents the comparison results of MSE values and MAE values for five interpolation methods with respect to the changes in p. Subfigure (a) illustrates the variations in MSE values of the estimated response variables, while subfigure (b) represents the changes in MAE values. From figure 5, it can be seen that as p increases, the MSE values and MAE values of the MCEM, PX-EM, and MOPX-EM methods initially decrease, then increase, and then decrease again. The minimum values are reached when $p = 25$. On the other hand, the MSE values and MAE values of the DPX-EM and DMOPX-EM methods initially increase, then decrease, and then increase again as p increases. The minimum values are achieved when $p = 10$. It can be observed that the DMOPX-EM method consistently has lower MSE values and MAE values compared to other methods, indicating better interpolation performance. Specifically, when $p = 10$, the DMOPX-EM method achieves a minimum MSE value of 0.02850028 and a minimum MAE value of 0.02509272.

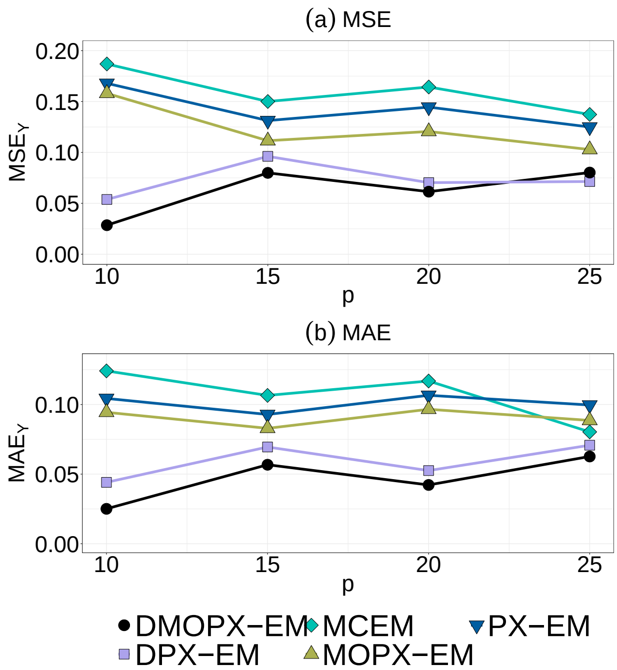

Fig. 5. Comparison Results Of The Five Methods In The Simulation

Case 6. Varying q with fixed $(M, \omega_{I_m}, MR, n, p)$

For $(M, \omega_{I_m}, MR, n, p) = (5, 0.15, 0.1, 400, 10)$, change $q = \{5, 6, 7, 8\}$. Figure 6 presents the comparison results of MSE values and MAE values for six interpolation methods with respect to the changes in q. Subfigure (a) illustrates the variations in MSE values of the estimated response variables, while subfigure (b) represents the changes in MAE values. From figure 6, it can be observed that there is a difference in the MSE values and MAE values among the five interpolation methods. It is evident that the DMOPX-EM method has lower MSE values and MAE values compared to the other methods, indicating that the DMOPX-EM method has the best interpolation performance. As q increases, the MSE values and MAE values of the DMOPX-EM method first increase, then decrease, and then increase again. When $q = 7$, the DMOPX-EM method achieves the minimum MSE value of 0.01894107, and when $q = 5$, it achieves the minimum MAE value of 0.02509272. When $q = 6$, the DMOPX-EM method reaches the maximum MSE value of 0.06750091, and the maximum MAE value of 0.07335642.

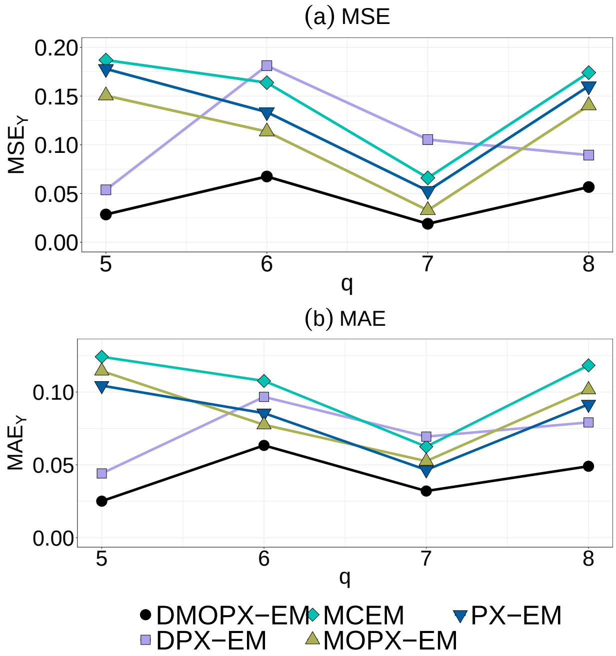

Fig. 6. Comparison Results Of The Five Methods In The Simulation

C. Real Data Analysis

In this section, we analyze popmis dataset, comparing the performance of the DMOPX-EM method and its comparative methods by testing two performance indicators in estimating response variables in random effects model.

The dataset was generated by Hox and records the characteristics of students in different classes, including missing data in terms of student popularity. It consists of a 2000×7 matrix. The popmis dataset was fitted with a random effects model, with the fixed effects covariate matrix $X = \{pupil, texp, teachpop\}$, the random effects covariate matrix $Z = \{school, sex, const\}$, and popular as the response variable. There are 848 missing values. Therefore, the dimensions are $(n, p, q) = (2000, 3, 3)$. When analyzing this dataset, the control factors for each algorithm were fixed as $(M, \omega_{I_m}) = (20, 0.15)$. The comparison results of the imputation methods as follows:

Figure 7 shows the comparison results of five interpolation methods, MCEM, PX-EM, MOPX-EM, DPX-EM and DMOPX-EM in the popmis dataset. Subfigure (a) illustrates the variations in MSFE values of the estimated response variables, while subfigure (b) represents the changes in MRE

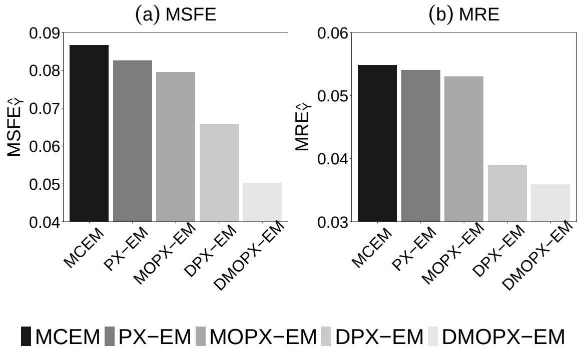

Fig. 7. MSFE and MRE comparison results of the popsis data set

values. From the test results in figure 7, it is evident that the MSFE values and MRE values for MCEM, PX-EM, and MOPX-EM methods are quite similar. In comparison to other imputation methods, the DMOPX-EM method has the lowest MSFE values and MRE values, and MSFE value is 0.0502788 and MRE value is 0.0359211. This indicates better performance and higher estimation accuracy, demonstrating superior imputation effectiveness.

IV. Discussion

The response missing of random effects model is an essential research topic in various academic and application fields. To address the missing response variable issue, we propose a novel distributed monotonic overrelaxation PX-EM method based on distributed theory and overrelaxation idea to accurately and quickly estimate the missing values. For random effects model, we simulate and analyze the control factors and dimensions using block numbers and overrelaxation factors, which show that the DMOPX-EM method has stronger rationality and requires less iteration time compared to other methods, verifying the effectiveness of the DMOPX-EM method.

The DMOPX-EM method in this paper improves the performance of the PX-EM algorithm under certain circumstances and can be applied to handle various types of data. However, there are still some shortcomings:

(1) Affected by the initial value, in the simulation analysis of section III, especially when analyzing the influence of n, p, and q on DMOPX-EM method in dealing with response missing, the mean absolute error (MAE) and mean squared error (MSE) values of the MCEM, PX-EM, DPX-EM, and DMOPX-EM methods are not significantly different, and the differences cannot be clearly seen. It is worth further studying to improve the performance of these methods from the perspective of initial values.

(2) In the future, it may be worth while to extend the methods mentioned in this article to more mixed models, such as logistic mixed regression models, which also deserves further research.

References

[1] R. M. Yucel, "Multiple imputation inference for multivariate multilevel contnuous data with ignorable non-response," *Philosophical Transactions of the Royal Society A:Mathematical, Physical and Engineering Sciences*, vol. 366, no. 1874, pp. 2389–2403, Apr. 2008.

[2] D. Kline, R. Andridge, and E. Kaizar, "Comparing multiple imputation methods for systematically missing subject-level data," *Research Synthesis Methods*, vol. 8, no. 2, pp. 136–148, 2017.

[3] C. A. Field and A. H. Welsh, "Bootstrapping clustered data," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 69, no. 3, pp. 369–390, 2007.
[4] D. B. Hall, "Zero-inflated poisson and binomial regression with random effects: a case study," *Biometrics*, vol. 56, no. 4, pp. 1030–1039, 2000.
[5] J. G. Ibrahim, M.-H. Chen, and S. R. Lipsitz, "Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable," *Biometrika*, vol. 88, no. 2, pp. 551–564, 2001.
[6] J. G. Ibrahim and G. Molenberghs, "Missing data methods in longitudinal studies: a review," *Test*, vol. 18, no. 1, pp. 1–43, 2009.
[7] Y. Yu, "Monotonically overrelaxed em algorithms," *Journal of Computational and Graphical Statistics*, vol. 21, no. 2, pp. 518–537, 2012.
[8] S. M. Diffey, A. B. Smith, A. Welsh, and B. R. Cullis, "A new reml (parameter expanded) em algorithm for linear mixed models," *Australian & New Zealand Journal of Statistics*, vol. 59, no. 4, pp. 433–448, 2017.
[9] G. Guo, Y. Sun, and X. Jiang, "A partitioned quasi-likelihood for distributed statistical inference," *Computational Statistics*, vol. 35, pp. 1577–1596, 2020.
[10] G. Guo, "Parallel statistical computing for statistical inference," *Journal of Statistical Theory and Practice*, vol. 6, pp. 536–565, 2012.
[11] G. Guo, W. You, G. Qian, and W. Shao, "Parallel maximum likelihood estimator for multiple linear regression models," *Journal of computational and applied mathematics*, vol. 273, pp. 251–263, 2015.
[12] A. P. Verbyla, "A conditional derivation of residual maximum likelihood," *Australian Journal of Statistics*, vol. 32, no. 2, pp. 227–230, 1990.
[13] L. Zhou and Y. Tang, "Linearly preconditioned nonlinear conjugate gradient acceleration of the px-em algorithm," *Computational Statistics & Data Analysis*, vol. 155, p. 107056, 2021.