Journal of Data Science,17(2). P. 331 - 348,2019
DOI:10.6339/JDS.201904_17(2).0005

PARAMETRIC FRACTIONAL IMPUTATION FOR LONGITUDINAL DATA WITH INTERMITTENT MISSING VALUES

Ahmed M. Gad*, Hanan E. G. Ahmed[2]

Department of Statistics, Faculty of Economics and Political Science, Cairo University

Abstract

Longitudinal data analysis had been widely developed in the past three decades. Longitudinal data are common in many fields such as public health, medicine, biological and social sciences. Longitudinal data have special nature as the individual may be observed during a long period of time. Hence, missing values are common in longitudinal data. The presence of missing values leads to biased results and complicates the analysis. The missing values have two patterns: intermittent and dropout. The missing data mechanisms are missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR). The appropriate analysis relies heavily on the assumed mechanism and pattern. The parametric fractional imputation is developed to handle longitudinal data with intermittent missing pattern. The maximum likelihood estimates are obtained and the Jackkife method is used to obtain the standard errors of the parameters estimates. Finally a simulation study is conducted to validate the proposed approach. Also, the proposed approach is applied to a real data.

Keywords: Longitudinal data; missing data; nonrandom missing; parametric fractional imputation; repeated measures.

*Corresponding author:
Email:ahmed.gad@feps.edu.eg

1 Introduction

Longitudinal studies are common in many fields, where data are collected from each subject repeatedly over time, or under different conditions. Missing values are very common in such studies due to many reasons. Many factors need to be considered to handle missing observations in longitudinal studies, such as: missing data mechanism, the missing data pattern, the covariance structure within individuals, and the used model for joint distribution of the observed and unobserved responses. The missing data mechanism is missing completely at random (MCAR) if the probability of missingness is not related to the unobserved and the observed values. It is missing at random (MAR) if the probability missingness is related to the observed values and if the probability of missingness is related to both observed and unobserved values, it is denoted as missing not at random (MNAR).

There are two patterns of missingness; the dropout and the intermittent. The dropout pattern is when an individual leaves prematurely the study, and the intermittent pattern is when an individual shows up again after leaving the study. Hence, in the dropout pattern no observed value after a missing one and the intermittent pattern allows to have observed value(s) after a missing value.

A way to analyze longitudinal data requires jointly modeling the distribution of the observed and unobserved responses. The unobserved response is expressed using indicator variable R, which takes the value 1 when the response is observed, and the value 0 when the response is missing. There are three models obtained by different factorization of the joint distribution of the response variable Y, and the missing indicator R. The first is the selection model, where the joint distribution of complete data (Y, R) is factorized as a product of the response variable model, and the conditional distribution of R given Y (Diggle and Kenward, 1994). The second is the pattern mixture model, where the joint distribution of the complete data is factorized as the conditional distribution of the complete data given the missing data, and the missingness model (Little, 1993). The third is the shared parameter model which depends on the idea that there are some common parameters that affect both repeated measures and missingness (Follmann and Wu, 1995). Gao and Thiebaut (2009) introduce a shared parameter model, for longitudinal data, in case of non-ignorable missing data.

Many techniques have been proposed to deal with incomplete longitudinal data such as; the complete case analysis (CC) which depends on analyzing the cases without missing and ignoring the others with missing (Donders., et al., 2006). The available case analysis (AC) which ignores the missing values not the whole case (Donders., et al., 2006). The weighting

method gives weight to the observed cases to compensate for the unobserved cases (Little and Rubin, 1987). The imputation methods which include different techniques that work on filling the missing values with other imputed values, and finally the likelihood based methods can be used for both modeling and estimation.

The likelihood based methods depend on maximizing the log-likelihood function of the joint model of the complete data. When the log-likelihood function does not have closed formula, iterative methods are used to obtain the ML estimates such as; the Newton-Raphson method, the scoring method, the Jennrich and Schluchter algorithm, and the expectation maximization (EM) algorithm. In the EM algorithm when the E-step is intractable, stochastic versions of the EM algorithm are possible. These include the stochastic expectation maximization algorithm (SEM) (Celeux and Diebolt, 1985), the stochastic approximation of EM (SAEM) algorithm (Delyon, et al., 1999), the Monte Carlo EM (MCEM) algorithm (Wei and Tanner, 1990), and the parametric fractional imputation (PFI) (Kim & Fuller, 2008).

The aim of this paper is to develop the PFI method to estimate the parameters in the presence of the intermittent missingness. The standard errors of the estimated parameters are obtained using the jackknife method. The rest of the paper is organized as follows. In Section 2 the basic notations are introduced. In Section 3 the selection model for the longitudinal data with intermittent missingness is introduced. In Section 4 the PFI method is developed in the case of intermittent missing pattern under the selection model. In Section 5 the Jackkinfe method is introduced as a method to find the standard error for the estimated parameters. In Section 5 the proposed method is applied to a real data. In Section 6 a simulation study is conducted to check the performance of the proposed method. Finally in Section 7 a brief conclusion is given.

2 Notation and Models

Let y_{ij} be the response variable and x_{ij} is a p-vector of fully observed covariates for the i^{th} individual at the j^{th} time point made at time t_{ij}, $j = 1,2,\dots,n_i$, $i = 1,2,\dots,m$. It is assumed that the time is common for all individuals. The mean and the variance of y_{ij} are respectively $E(y_{ij}) = \mu_{ij}$ and $V(y_{ij}) = \sigma_{ij}$, and the covariance between y_{ij} and y_{jk} is $cov(y_{ij}, y_{jk}) = \sigma_{jk}$. The vector $y_i = (y_{i1}, \dots, y_{in_i})$ is the response of i^{th} individual through all time points, and it assumed that $y_i \sim MVN(\mu_i, V_i)$, where $\mu_i = X_i\beta$, X_i is $n_i \times p$ matrix of the covariates, β is $p \times 1$ vector of unknown parameters and V_i is the covariance matrix of dimension $n_i \times n_i$. The matrix $y = (y_1, \dots, y_m)$ of size $m \times N$

represents the responses of all individuals, where $N = \sum_{i=1}^{m} n_i$.

The general linear regression model can be used to model the longitudinal data. The general linear model for y_{ij} can be written as

$$y_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \cdots + \beta_p x_{ijp} + \varepsilon_{ij} \quad for\ j = 1,2,\dots,n_i\ \ and\ \ i = 1,2,\dots,m \tag{1}$$

In vector notation $Y_i = X_i\beta + \varepsilon_i$, where $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ is p-vector of unknown regression coefficients, $\varepsilon_i = (\varepsilon_{i1}, \dots, \varepsilon_{in_i})$ is the error term which assumed to follow multivariate normal with mean zero and covariance matrix V_i. The matrix V is a block diagonal matrix with non-zero blocks V_i. The matrix V_i may be unstructured containing $n_i(n_i + 1)/2$ parameters, or structured; its elements are function of a number of parameters α_i and it will be written as $V_i(\alpha)$. The response variable Y follows the multivariate normal distribution, i.e. $Y \sim MVN(X_i\beta, V)$, where Y is $N \times 1$ vector of responses, X_i is a $N \times P$ matrix of covariates, and V is a block diagonal matrix with m non-zero blocks V_i (Diggle et al., 1994).

In the case of incomplete longitudinal data, the observed measurement for individual i is denoted as $y_{i,obs}$, while the missing one is denoted as $y_{i,mis}$. The complete data can be written as $Y = (Y_{obs}, Y_{mis})$. A binary variable R_{ij} is used to represent the missingness process, where R_{ij} takes the value 1 if y_{ij} is observed, and the value zero if y_{ij} is missing. Let L be the likelihood function of the underlying parameters, and ℓ be the corresponding log-likelihood function.

3 Selection Model for Non-Random Intermittent Missingness

Diggle and Kenward (1994) propose the selection model for contiuous longitudinal data with non-random missingness. The probability of missingness for the i^{th} individual at the time t_{d_i} depends on the history of the measurements including time t_{d_i}. The probability of missingness can be formulated as

$$P(D_i = d_i | history) = P_{d_i}(H_{id_i}, y_{id_i}, \phi),$$

where D is a random variable identifying the dropout time such that $2 \le D \le n$, and $H_{id_i} = (y_{i1}, \dots, y_{id_i-1})$ denotes the observed measurment up to time $t_{d_{i-1}}$. Diggle and Kenward (1994) formulate the probability of the dropout missingness using a logistic or probit linear model as

$$\text{logit}\{P_{d_i}(H_{id_i}, y_{id_i}, \phi)\} = \phi_0 + \phi_1 y_{id_i} + \sum_{j=2}^{d_i} \phi_j y_{id_i-j+1}..$$

Troxel et al. (1998) extend the selection model for continuous longitudinal data with non-random intermittent missingness. They assume a missing data model, which allows the probability of missingness to depend on the value of the current and/or previous measurement; $P(R_{ij}|y_i) = P(R_{ij}|y_{ij}, y_{i(j-1)})$. They formulate the probability of the intermittent missingness using a logistic linear model as

$$\text{logit}\{P_{ij}(y_{ij}, \emptyset)\} = \emptyset_0 + \emptyset_1 y_{i(j-1)} + \emptyset_2 y_{ij}.$$

In this case the missing data mechanisms can be expressed using the logistic model as follow:

- Nonrandom intermittent if $P_{it}(:)$ depends on Y_{it} .
- Random intermittent if $P_{it}(:)$ depends on $Y_{i(t-1)}$.
- Completely random intermittent if $P_{it}(:)$ depends on both Y_{it} and $Y_{i(t-1)}$.

Gad and Ahmed (2006) adopt the missingness model of Diggle and Kenward (1994) and obtain the parameter estimates using the stochastic EM (SEM) algorithm in intermittent pattern. Gad and Youssif (2006) apply the SEM algorithm to mixed linear models in longitudinal data. Gad (2011) propose a selection model in the case of missing values. Yaseen et al. (2016) use the model of Diggle and Kenward (1994) and apply the parametric fractional imputation (PFI) for the dropout missing pattern.

4 PFI for Intermittent Missingness Using Selection Model

The PFI method is introduced to handle longitudinal data under non-ignorable intermittent missingness. The selection model of Troxel et al. (1998) is adopted. Similar to Troxel et al. (1998), the probability of intermittent missingness is modeled as

$$\text{logit}\{P_{ij}(y_{it}, \emptyset)\} = \emptyset_0 + \emptyset_1 y_{i(j-1)} + \emptyset_2 y_{ij},$$

where $\emptyset = [\emptyset_0, \emptyset_1, \emptyset_2]$is a vector of the missingness parameters. The response variable is modeled using the general linear model in Eq. (1).

In the presence of intermittent missing values the proposed PFI is implemented using the following steps.

1- Generate M vector to impute for each missing value. We suggest using the Gibbs sampler to generate the imputed values for each missing value. The imputed values are generated from the conditional distribution of the missing data, $y_{i,mis} = (y_{i,mis1}, \dots, y_{i,misr})$, given the observed data. This conditional distribution has the mean and the covariance as

$$\mu_{i,m.o} = \mu_{i,m} + V_{i,mo}V_{\text{i,oo}}^{-1}(Y_{i,obs} - \mu_{i,o}),$$
$$V_{i,m.o} = V_{i,mm} - V_{i,mo}V_{i,oo}^{-1}V_{i,om},$$

where $\mu_{i,o}$, $\mu_{i,m}$, $V_{i,oo}$, $V_{i,mm}$, $V_{i,om}$ are suitable partitions of the mean vector $\mu_i$and the

covariance matrix V_i. The Gibbs sampling is applied as follow:

- At the $(t+1)$ iteration, $\mathrm{Y}_{i,mis}^{(t+1)} = \left(\mathrm{Y}_{i,mis1}^{(t+1)}, \dots, \mathrm{Y}_{i,misr}^{(t+1)}\right)$ is imputed from the conditional distribution of the missing given the observed, where $\mathrm{Y}_{i,mis1}^{(t+1)}$ is imputed from the conditional distribution $f\left(Y_{i,mis1} \middle| \mathrm{Y}_{i,mis2}^{(t)}, \dots, \mathrm{Y}_{i,misr}^{(t)}, Y_{i,obs}, R, \theta^{(t)}\right)$.
- $\mathrm{Y}_{\mathrm{i,mis2}}^{(\mathrm{t+1})}$ is imputed from the conditional distribution function $f\left(Y_{i,mis2} \middle| \mathrm{Y}_{i,mis1}^{(t+1)}, \mathrm{Y}_{i,mis3}^{(t)}, \dots, \mathrm{Y}_{i,misr}^{(t)}, Y_{i,obs}, R, \theta^{(t)}\right)$.
- $\mathrm{Y}_{i,mis3}^{(t+1)}$ is imputed from the conditional distribution function $f\left(Y_{i,mis3} \middle| \mathrm{Y}_{i,mis1}^{(t+1)}, \mathrm{Y}_{i,mis2}^{(t+1)}, \mathrm{Y}_{i,mis4}^{(t)}, \dots, \mathrm{Y}_{i,misr}^{(t)}, Y_{i,obs}, R, \theta^{(t)}\right)$.
- $\mathrm{Y}_{i,misr}^{(t+1)}$ is imputed from the conditional distribution function $f\left(Y_{i,misr} \middle| Y_{i,mis1}^{(t+1)}, Y_{i,mis2}^{(t+1)}, \dots, Y_{i,misr-1}^{(t)}, Y_{i,obs}, R, \theta^{(t)}\right)$.

2- Given the M imputed data sets and the current parameters estimates, calculate the fractional weight for the vector of the imputed values, $\mathrm{Y}_{i,mis}^{*(k)}$, for individual i in replicate k $,k = 1, \dots, M$, which takes the following form:

$$\mathrm{w}_{\mathrm{i(t)}}^{(\mathrm{k})} = \frac{\mathrm{f}\left(\mathrm{Y}_{\mathrm{i,mis}}^{*(\mathrm{k})} \middle| \theta^{(\mathrm{t})}\right) / \mathrm{f}\left(\mathrm{Y}_{\mathrm{i,mis}}^{*(\mathrm{k})} \middle| \mathrm{Y}_{\mathrm{i,obs}}; \theta_0\right) \prod_{\mathrm{j=d_i}}^{\mathrm{n}} \left(1 - \pi\left(\mathrm{Y}_{\mathrm{ij}}^{*(\mathrm{k})}; \phi^{(\mathrm{t})}\right)\right)}{\sum_{\mathrm{l=1}}^{\mathrm{M}} \mathrm{f}\left(\mathrm{Y}_{\mathrm{i,mis}}^{*(\mathrm{l})} \middle| \theta^{(\mathrm{t})}\right) / \mathrm{f}\left(\mathrm{Y}_{\mathrm{i,mis}}^{*(\mathrm{l})} \middle| \mathrm{Y}_{\mathrm{i,obs}}; \theta_0\right) \prod_{\mathrm{j=d_i}}^{\mathrm{n}} \left(1 - \pi\left(\mathrm{Y}_{\mathrm{ij}}^{*(\mathrm{l})}; \phi^{(\mathrm{t})}\right)\right)}$$

where $\pi\left(y_{ij}^{*(k)}; \phi^{(t)}\right)$ is the probability of being missing for the response $Y_{ij}^{*(k)}$ given the current parameter estimate of $\phi^{(t)}$. The denominator of the fractional weight guarantee that the sum of all fractional weight equals to 1, i.e.

$$\sum_{i=1}^{m} w_{i(t)}^{(k)} = 1$$

3- Under the selection model the joint density of the complete data and missing indicator is formulated as

$$f(Y, R|\theta, \phi) = f(Y/\theta) P(R|Y; \phi) = f(Y_{obs}, Y_{mis}|\theta)\ P(R|Y_{obs}, Y_{mis}; \phi),$$

where θ and ϕ are the parameters of Y and R respectively. The density function of the observed function can be obtained by integrating out the missing part,

$$f(Y_{obs}, R|\theta, \phi) = \int f(Y_{obs}, Y_{mis}/\theta)\, P(R|Y_{obs}, Y_{mis}; \phi) dY_{mis}\ .$$

Due to the missingness, to obtain the MLE for θ and ϕ we need to maximize the following

score function

$$Q^*\left(\theta^{(t)}, \phi^{(t)}\right) = \left[Q_1^*\left(\theta^{(t)}\right), Q_2^*\left(\phi^{(t)}\right)\right],$$

where

$$Q_1^*\left(\theta^{(t)}\right) = \sum_{i=1}^{m}\left[R_i \log f\left(Y_i|\theta^{(t)}\right) + (1-R_i)\sum_{k=1}^{M} w_{i(t)}^{(k)} \log f\left(Y_i^{*(k)}\middle|\theta^{(t)}\right)\right],$$

$$Q_2^*\left(\theta^{(t)}\right) = \sum_{i=1}^{m}\left[R_i \log P\left(R_i|Y_i;\phi^{(t)}\right) + (1-R_i)\sum_{k=1}^{M} w_{i(t)}^{(k)} \log P\left(R_i|Y_i^{*(k)};\phi^{(t)}\right)\right].$$

4- The log-likelihood function is maximized in two sub-steps; the normal step and the logistic step.

- The normal step:
  In this sub-step the score function from the previous step is maximized to update the parameter θ. The maximization is done using any appropriate optimization algorithm such as the Jennrich and Schluchter (1986) algorithm.
- The logistic step:
  The MLE of the logistic model parameters are obtained. These estimates are obtained using iterative reweighted least squares method (Collett, 2002).

5 The Standard Error Estimates

The PFI does not provide standard errors of the estimates. In this article the jackknife method is suggested to obtain the standard errors. It is a method of resampling from the original sample. It calculates the estimates from each resampled sample. Jiang et al. (2002) propose the jackknife method to estimate the standard errors in the general case and also apply it to the generalized linear mixed model as a special case. The jackknife method proceeds as follow:

- Assume that the original sample is $Y = [y_1, \dots, y_n]$. In this case n samples of size $n-1$; $S_1, S_2, \dots, S_n$ are generated from the original sample, where $S_1 = [y_2, \dots, y_n]$, $S_2 = [y_2, y_3, \dots, y_n]$, …, $S_n = [y_1, y_2, \dots, y_{n-1}]$.
- Calculate the parameter of interest $\hat{\theta}_{(i)}^*$ for $i = 1, \dots, n$, based on the same estimation method used to estimate the parameter $\hat{\theta}$ in the original sample.
- Estimate the standard error of $\hat{\theta}$ using the following formula

$$\text{S.}\widehat{\text{E}}\left(\hat{\theta}\right) = \sqrt{\frac{n-1}{n}\sum_{i=1}^{n}\left(\hat{\theta}_{(i)}^* - \hat{\theta}\right)^2}.$$

Applying the jackknife method on longitudinal data means that each time we omit individual $,Y_{i,obs}$, instead of deleting one observation. Hence, the generated sample at the

i^{th} iteration is $S_i = [Y_{1,obs}, Y_{1,obs}, \dots, Y_{i-1,obs}, Y_{i+1,obs}, \dots, Y_{1,nbs}]$.

5. Application (Breast Cancer Study)

This data set is collected by the International Breast Cancer Study Group (IBCSG). It is about the quality of life among four groups of women who diagnosed as breast cancer's patients. The target group is premenopausal women with breast cancer, after they were exposed randomly to four different chemotherapy regimens, namely; A, B, C, and D (Hurny et al., 1992).

The patients were asked to complete questionnaire about the quality of life before starting the chemotherapy, and every three months for fifteen months. So each patient should fill the questionnaire six times. In other words, the data set consists six time points for all patients including one time point before starting chemotherapy and other 5 during the four chemotherapy regimens.

In this data set the Personal Adjustment to Chronic Illness Scale (PACIS) is used as a measurement for the quality of life. The PACIS compare a global patient rating of the amount of effort costs to cope with the illness. It is scaled from 0 to 100, where a larger score indicates that a greater amount of effort is needed from the patient to cope with her illness.

The data set consists of 456 patients of breast cancer, 10 patients are excluded from the data because they died. Hence, the death is not a reason of missingness in the remaining data. Also 64 patients with missing values at the first time point are excluded. The remaining 382 patients are the data that have been analyzed. Participation in the questionnaire was not obligatory, and some patients did refuse to complete it. Even when a patient refuses to fill the questionnaire, she was asked to complete the questionnaire at her next scheduled follow-up visit.

Among the 382 patients, there are 90 (24%) patients who complete the PACIS for the 6 time points, while there are 292 (76%) patients have missing value in at least one of the five time points after the first one. The percentage of missingness in the second time point is 29% while in the sixth time point it reaches 62%. The percentages of patients who have 1, 2, 3, 4, 5 missing are 18%, 14%, 13%, 13%, and 19% respectively. The PACIS does not follow normal distribution, Hurny et al. (1992) use a square root transformation to normalize the data.

This data set had been analyzed by many authors. Hurny et al. (1992) use the complete case analysis. The analysis shows that the treatment differences are statistically insignificant.

It shows also that the quality of life increased over time while fixing the treatment effect. Troxel et al. (1998) use this data set and apply the Nerlder-mixed simplex algorithm. The analysis is based on the first 6 months of the study, and using AR(1) covariance structure. They use two models to apply the Nerlder approach. The first is the simple mean model which combines the response variable with three missingness models. They conclude that the coefficient of the response in the previous time point is statistically insignificant. The second examines the treatment effect on the missingness probabilities. The analysis shows that the treatment effect is statistically insignificant. Ibrahim et al. (2001) use a Monte Carlo EM algorithm, a random effect model and the AR(1) covariance structure. The results show that chemotherapy C and the response in the previous point are significant. Gad and Ahmed (2006) use the SEM algorithm in case of intermittent missingness, and the mean model to study the effect of the four treatments. They assume unstructured and structured, AR(1), covariance matrix. The results show that the missingness parameter for the current response is positive and statistically significant, which means that the missing mechanism is missing not at random. This implies that the woman who has higher values of PACIS tends to be missing. The results show that chemotherapy C has significant effect. This conclusion contradicts with Troxel et al. (1998), because of the fact that Troxel et al. (1998) depend on only 6 months in the analysis of breast cancer data.

The data set is modeled using the fixed effect linear model as

$$y_{ij} = \mu_j + \epsilon_{ij} \ \ ,i = 1, \dots ,382 \ \ ,j = 1, \dots ,6.$$

The y_{ij} is the response variable (PACIS score) for the i^{th} i^{th} patient at the j^{th}time point. The mean vector $\mu = (\mu_1, \dots , \mu_6)'$ is of dimension 6×1 where μ_j represents the mean of the PACIS score at the j^{th} time point. The unstructured and structured AR(1) covariance matrices are assumed. The unstructured covariance matrix is of dimension $6 \times 6; V_i = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{16} \\ \vdots & \ddots & \vdots \\ \sigma_{16} & \cdots & \sigma_6^2 \end{pmatrix}$, and the structured first order auto regressive AR(1) covariance of the form $\sigma_{ij} = \sigma^2 \rho^{|i-j|}, \ i = 1, \dots ,382; j = 1, \dots ,6.$

The missing data mechanism is modeled using the logistic model assuming that the probability of missingness depends on both current and previous observation as

$$\text{logit}\{r_{ij} = 1|\emptyset\} = \emptyset_0 + \emptyset_1 y_{i(j-1)} + \emptyset_2 y_{ij}.$$

The proposed PFI for intermittent missingness is applied to obtain estimates for the parameters. Also, the standard errors of the estimates are obtained. The parameter estimates

and their standard for unstructured covariance are shown in Table (1). Also, the parameter estimates and their standard for AR(1) covariance are shown in Table (2).

Table (1): The PFI estimates and their standard errors (SE) for unstructured covariance.

Parameter	Estimate	SE	Parameter	Estimate	SE	Parameter	Estimate	SE
μ_1	6.06	0.08	σ_{15}	2.96	4.12	σ_{36}	2.91	2.37
μ_2	6.09	4.09	σ_{16}	2.28	2.69	σ_{44}	4.83	3.9
μ_3	5.92	2.95	σ_{22}	5.13	1.68	σ_{45}	3.51	1.86
μ_4	5.64	5.8	σ_{23}	3.09	1.53	σ_{46}	3.01	3.34
μ_5	5.33	5.91	σ_{24}	2.85	1.85	σ_{55}	4.61	3.24
μ_6	5.66	9.3	σ_{25}	2.98	2.2	σ_{56}	2.8	2.61
σ_{11}	6.24	0.4	σ_{26}	2.01	2.04	σ_{66}	3.88	2.37
σ_{12}	2.65	1.63	σ_{33}	5.02	2.15	ϕ_0	-1.52	0.15
σ_{13}	2.99	2.13	σ_{34}	3.62	2.31	ϕ_1	0.02	0.01
σ_{14}	2.98	2.04	σ_{35}	2.86	1.9	ϕ_2	0.22	0.03

Table (2): The PFI estimates and their standard errors for the AR(1) covariance structure.

Parameter	Estimate	SE
μ_1	6.06	0.15
μ_2	5.98	0.55
μ_3	5.81	1.63
μ_4	5.5	0.92
μ_5	5.19	1.56
μ_6	5.48	1.15
ρ	3.92	0.82
σ	0.48	0.1
ϕ_0	-1.21	0.18
ϕ_1	0.05	0.02
ϕ_2	0.13	0.02

For both structures; the AR(1) and unstructured covariance matrix, the results show that parameter $\emptyset_2$ takes positive values, which implies that high values of the PACIS are more likely to be missing. This is logical because high value of PACIS means that more difficulty to cope with the illness. Hence, the woman who needs a great effort to cope with her illness, is more likely to refuse to complete the quality of life questionnaire. The parameter $\emptyset_2$ is significantly different from 0, which support the assumption that the missing mechanism is non-ignorable. The parameter $\emptyset_1$ is positive and significantly different from 0, which reflects the importance of the response in the previous time point.

6 Simulation

6.1 simulation setting

This simulation study is conducted to judge the performance of the PFI in the presence of intermittent missingness. Different sample sizes are used. The sample sizes are chosen to cover small, moderate and large sizes as 20, 50 and 100. The time points are fixed at five time points. Unstructured covariance model and first order auto regressive AR(1) covariance structure are applied. In the unstructured covariance the matrix is of dimension 5×5 and consists 15 parameters $\sigma = [\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2, \sigma_5^2, \sigma_{12}, \sigma_{13}, \sigma_{14}, \sigma_{15}, \sigma_{23}, \sigma_{24}, \sigma_{25}, \sigma_{34}, \sigma_{35}, \sigma_{45}]'$. The values of these parameters are fixed at $\sigma = [10,10,9,10,9.9,6,7,6,9,5,6,6,9,7,5.9]$. The structured AR(1) covariance is $\sigma_{ij} = \sigma^2 \rho^{|i-j|}$ that depends on two parameters. These parameters are fixed at $\sigma^2 = 4.5$ and $\rho = 0.3$.

The response variable y is simulated from the multivariate normal distribution with mean $X_i\beta$ and variance either unstructured or structured as defined above. The matrix X_i is a design matrix or matrix of covariates. Both vectors of β and ϕ are of length three: $\beta = [\beta_1, \beta_2, \beta_3]$ and $\phi = [\phi_1, \phi_2, \phi_3]$. The values of these parameters are fixed at the following values $\beta = (0.5,1,6)'$ and $\emptyset = (-5.5,0.07,0.04)'$.

The percentage rate of missingness ranges from 12% to 20%. The number of replications is 10000.

6.2 simulation results

Under the above settings the parameter estimates are obtained using the following methods:

1. The complete data analysis in which no missing where the analysis depends on the complete generated data.
2. The parametric fractional imputation (PFI) algorithm with M=10.
3. The stochastic expectation maximization (SEM) algorithm with M=10.

4. The Monto Carlo expectation maximization (MCEM) algorithm with M=10.
5. The fractional regression nearest neighbor imputation (FRNNI) with M=5.

The results are shown in Tables (3) – (8). From these results the relative bias decreases as the sample size increases for both unstructured and structured AR(1) covariance matrix. Under the unstructured covariance matrix, the estimates of covariance matrix tend to be underestimated as the sample size increases. While under structured covariance matrix, the parameter σ is the only parameter which tends to be underestimated.

Table (3): Relative bias (RB) percentage for unstructured covariance at $= 20$; CS, compete data; PFI, parametric fractional imputation; SEM, stochastic EM; MCEM, Monto Carlo expectation maximization, FRNNI; fractional regression nearest neighbor imputation.

Parameter	CS	PFI	SEM	MCEM	FRNNI
β_0	0.07	-0.07	0.19	-0.14	2.56
β_1	-0.08	0.10	-0.05	0.01	-0.87
β_2	0.00	0.06	0.01	-0.02	-3.16
σ_1^2	-0.10	-1.60	-0.96	-1.59	-3.11
σ_2^2	-0.47	-11.26	-1.44	-11.23	1.35
σ_3^2	-0.08	-7.55	-0.57	-7.63	3.99
σ_4^2	-0.21	-8.95	-2.23	-9.19	1.42
σ_5^2	-0.29	-8.65	-1.28	-8.41	1.81
σ_{12}	-0.12	-0.21	1.27	0.82	-3.33
σ_{13}	-0.03	-2.63	-1.43	-1.69	-6.33
σ_{14}	-0.13	3.64	4.87	4.64	-3.16
σ_{15}	-0.11	-6.15	-4.57	-5.22	-10.45
σ_{23}	-0.17	5.60	7.46	8.02	-0.64
σ_{24}	-0.31	-2.77	-1.27	-0.84	-4.13
σ_{25}	-0.17	-2.28	-0.21	-0.22	-4.30
σ_{34}	-0.04	-12.32	-11.34	-11.05	-11.61
σ_{35}	-0.11	-4.87	-2.94	-2.96	-7.48
σ_{45}	-0.20	2.34	4.10	4.36	-4.31
$\emptyset_0$	8.92	8.86	8.24	8.43	----
$\emptyset_1$	9.22	9.01	8.43	8.83	----
$\emptyset_2$	9.77	8.98	8.31	8.23	----

Table (4): Relative bias (RB) percentage for unstructured covariance at $n = 50$; CS, compete data; PFI, parametric fractional imputation; SEM, stochastic EM; MCEM, Monto Carlo expectation maximization, FRNNI; fractional regression nearest neighbor imputation.

Parameter	CS	PFI	SEM	MCEM	FRNNI
β_0	-0.03	0.07	0.00	0.02	1.01
β_1	0.00	0.00	0.00	-0.02	-1.11
β_2	0.00	0.06	0.00	-0.01	-2.52
σ_1^2	-0.30	-0.01	-0.79	-0.86	-17.19
σ_2^2	-0.10	-0.13	-1.51	-12.83	-13.87
σ_3^2	-0.36	-0.09	0.05	-8.34	-12.51
σ_4^2	-0.23	-0.10	-1.78	-10.05	-13.80
σ_5^2	-0.25	-0.10	-1.69	-9.33	-13.40
σ_{12}	-0.30	-0.01	0.44	0.53	-18.09
σ_{13}	-0.39	-0.03	-1.41	-1.37	-18.84
σ_{14}	-0.40	0.03	4.31	4.38	-17.80
σ_{15}	-0.20	-0.06	-4.56	-4.68	-20.25
σ_{23}	-0.60	0.04	6.51	6.95	-17.68
σ_{24}	-0.38	-0.04	-2.33	-1.86	-18.95
σ_{25}	-0.42	-0.03	-1.20	-1.07	-18.60
σ_{34}	-0.26	-0.13	-11.14	-11.11	-21.20
σ_{35}	-0.43	-0.06	-3.66	-3.51	-19.60
σ_{45}	-0.52	0.01	2.78	3.02	-18.43
$\emptyset_0$	3.27	0.03	0.02	2.77	---
$\emptyset_1$	3.32	0.03	0.03	2.97	---
$\emptyset_2$	3.59	0.03	0.02	2.93	---

Table (5): Relative bias (RB) percentage for unstructured covariance at $n = 100$; CS, compete data; PFI, parametric fractional imputation; SEM, stochastic EM; MCEM, Monto Carlo expectation maximization, FRNNI; fractional regression nearest neighbor imputation.

Parameter	CS	PFI	SEM	MCEM	FRNNI
β_0	-0.02	-0.07	-0.17	-0.14	0.61
β_1	-0.01	0.07	-0.02	-0.01	-0.93
β_2	0.00	0.07	-0.03	-0.01	-2.24
σ_1^2	-0.08	-0.33	-0.32	-0.33	-20.35
σ_2^2	-0.12	-13.68	-1.01	-13.19	-17.34
σ_3^2	-0.10	-8.64	0.68	-8.23	-16.76
σ_4^2	-0.10	-10.16	-1.29	-9.99	-17.65
σ_5^2	-0.21	-9.59	-1.17	-9.15	-17.16
σ_{12}	-0.14	-0.48	0.83	0.83	-21.01
σ_{13}	-0.05	-2.03	-1.03	-0.95	-21.57
σ_{14}	-0.06	3.65	4.46	4.76	-20.95
σ_{15}	-0.12	-5.15	-4.16	-4.23	-22.26
σ_{23}	-0.11	3.95	6.87	6.59	-20.81
σ_{24}	-0.07	-4.35	-2.19	-2.21	-21.58
σ_{25}	-0.26	-3.14	-0.90	-0.90	-21.44
σ_{34}	-0.07	-12.28	-10.84	-10.85	-23.16
σ_{35}	-0.14	-5.28	-3.30	-3.33	-22.01
σ_{45}	-0.17	1.12	2.93	3.17	-21.35
$\emptyset_0$	-0.02	0.01	0.78	1.18	---
$\emptyset_1$	-0.01	0.01	0.97	1.38	---
$\emptyset_2$	0.00	0.02	0.52	1.20	---

Table (6): Relative bias (RB) percentage for AR(1) covariance at $n = 20$; CS, compete data; PFI, parametric fractional imputation; SEM, stochastic EM; MCEM, Monto Carlo expectation maximization, FRNNI; fractional regression nearest neighbor imputation.

Parameter	CS	PFI	SEM	MCEM	FRNNI
β_0	-0.59	-0.31	-0.06	-0.03	7.75
β_1	-0.77	0.08	-0.18	-0.27	1.76
β_2	-0.67	0.02	-0.04	-0.04	-2.69
σ^2	-3.30	-20.71	-3.88	-18.95	3.63
ρ	-2.15	4.90	-14.11	9.68	53.61
$\emptyset_0$	9.09	7.98	7.44	8.12	---
$\emptyset_1$	9.45	8.23	7.83	8.43	---
$\emptyset_2$	10.03	8.67	7.10	7.94	---

Table (7): Relative bias (RB) percentage for AR(1) covariance at n=50; CS, compete data; PFI, parametric fractional imputation; SEM, stochastic EM; MCEM, Monto Carlo expectation maximization, FRNNI; fractional regression nearest neighbor imputation.

Parameter	CS	PFI	SEM	MCEM	FRNNI
β_0	-0.12	-0.14	-0.16	-0.20	6.37
β_1	-0.01	-0.11	-0.07	-0.04	1.56
β_2	0.01	0.03	-0.04	-0.04	-2.17
σ^2	-1.26	-19.53	-1.87	-1.73	-10.75
ρ	-0.29	3.03	-14.14	-14.40	-9.67
$\emptyset_0$	3.17	2.97	2.20	2.37	---
$\emptyset_1$	3.11	2.93	2.81	2.63	---
$\emptyset_2$	3.73	3.52	1.63	2.20	---

Table (8): Relative bias (RB) percentage for AR(1) covariance at n=100; CS, compete data; PFI, parametric fractional imputation; SEM, stochastic EM; MCEM, Monto Carlo expectation maximization, FRNNI; fractional regression nearest neighbor imputation.

Parameter	CS	PFI	SEM	MCEM	FRNNI
β_0	-0.04	-0.26	-0.18	-0.17	5.28
β_1	-0.05	-0.06	-0.17	-0.09	1.30
β_2	0.01	0.03	-0.03	-0.04	-1.83
σ^2	-0.57	-19.27	-0.97	-17.56	-14.98
ρ	-0.14	2.71	-14.80	8.43	-8.35
$\emptyset_0$	1.43	1.52	0.78	1.09	---
$\emptyset_1$	1.52	1.42	1.16	1.31	---
$\emptyset_2$	1.48	1.86	0.20	1.03	---

The fractional regression nearest neighbor imputation (FRNNI) approach appears to have the highest relative bias among the other methods under the unstructured covariance matrix. In case of structured AR(1) covariance matrix, it shows better performance compared to the other presented methods.

The stochastic expectation maximization (SEM) algorithm and the Monte Carlo expectation maximization (MCEM) algorithm show high performance under the unstructured covariance matrix. They produce relatively close results to each other. Regarding the structured AR(1) covariance matrix the MCEM show higher relative bias especially for the variance estimates comparable with the SEM algorithm.

The parametric fractional imputation (PFI) estimates have relatively low bias, especially with the unstructured covariance matrix at the moderate sample; n=50. The PFI method shows the best performance compared to the other methods. Regarding the structured AR(1) covariance matrix, the mean parameters estimates are the most efficient between the presented techniques. The relative bias of the variance parameters are positively related to the sample size. The parametric fractional imputation (PFI) method shows the same or relatively close results to the SEM algorithm and MCEM algorithm. As imputing the missing only one time and using the fractional weights accelerate the convergence process. So, the PFI gives the same performance and sometimes it gives better performance than the SEM algorithm and MCEM algorithm in less time using different starting points.

Subsequently, we can conclude based on the presented simulation results, that the parametric fractional imputation (PFI) method can guarantee relatively unbiased estimates, in the case of intermittent missingness, with different sample sizes and under structured and unstructured covariance matrix.

7 Conclusion

The parametric fractional imputation method is proposed as an innovative tool for parameters estimation in the presence of the missing values. The PFI method shows to be superior to the MCEM algorithm or the SEM algorithm. This is due to the fact that the imputed values are not regenerated at each iteration, This guarantees the convergence of the algorithm and accelerate its rate. The simulation results show that the proposed technique provides reasonable estimates.

References

[1] Celeux, G. and Diebolt, J. (1985) The SEM algorithm: A probabilistic Teacher algorithm derived from the EM Algorithm for the mixture problem, Computational Statistics Quarterly, 2, 73-82.

[2] Collett, D. (2002) Modelling Binary Data, 2nd Edition, Chapman and Hall, London.

[3] Delyon, B, Laird, N. M. and Rubin, D. B. (1999) Convergence of a stochastic approximation version of the EM algorithm, The Annals of Statistics, 27, 94-128.

[4] Diggle, P. J. and Kenward, M. G. (1994) Informative dropout in longitudinal data analysis, Journal of the Royal Statistical Society B, 43, 49-93.

[5] Diggle, P.J. Liang, K. Y. and Zeger, S. L. (1994) Analysis of Longitudinal Data, Oxford: Oxford Science, UK.

[6] Donders, A. R., Van der Heijden, G. J., Stijnen, T. & Moons, K. G., (2006) Review: a gentle introduction to imputation of missing values. Journal of Clinical Epistemology, 59, 1087-1091.

[7] Follmann, D. and Wu, M. (1995) An approximate generalized linear model with random effects for informative missing data, Biometrics, 51, 151-168.

[8] Gad, A. M. (2011) A selection model for longitudinal data with missing values, Journal of Data Science, 9, 171-180.

[9] Gad, A. M. and Ahmed, A. S. (2006) Analysis of longitudinal data with intermittent missing values using the stochastic EM algorithm, Computational Statistics & Data Analysis, 50, 2702 – 2714.

[10] Gad, A. M. and Youssif, N. A. (2006) Linear mixed models for longitudinal data with nonrandom dropouts, Journal of Data Science, 4, 447 - 460.

[11] Gao, S.-J. and Thiébaut, R. (2009) Mixed-effect Models for Truncated Longitudinal Outcomes with Nonignorable Missing Data, Journal of Data Science, 7, 13-25.

[12] Hurny, C., Bernhard, J., Gelber, R. D. & Coates, A. (1992) Quality of life measures for patients receiving adjuvant therapy for breast cancer and international trail, Eur. J. Cancer, 28, 118-124.

[13] Ibrahim, J. G., Chen, M. & Lipsitz, S. R. (2001) Missing responses in generalised linear mixed models when the missing data mechanism is nonignorable, Biometrika, 88, 551-564.

[14] Jennrich, R. I. and Schluchter, M. D. (1986) Unbalanced repeated-measures models with structured covariance matrices, Biometrics, 42, 805–820.

[15] Jiang, J., Lahiri, P. & Wan, S. (2002) A Unified Jackknife Theory for Empirical Best Prediction with M- Estimation, Annuals of Statistics, 30, 1782-1810.

[16] Kim, J. K. and Fuller, W. (2008) Parametric fractional imputation for missing data analysis. Proceeding of the section on survey research method, Joint Statistical Meeting, 158-169.

[17] Kim, J. Y. and Kim, J. K. (2012) Parametric fractional imputation for nonignorable missing data, Journal of the Korean Statistical Society, 41, 291-303.

[18] Little, R. J. and Rubin, D. B. (1987) Statistical Analysis with Missing Data, John Wiley & Sons, New York.

[19] Little, R. J. (1993) Pattern mixture models for multivariate incomplete data, Journal of American Statistical Association, 88, 125-134.

[20] Troxel, A. B., Lipsitz, S. R. & Harrington, D. P., (1998) Marginal models for the analysis of longitudinal measurements with nonignorable non-monotone missing data, Biometrika, 85, 661-672.

[21] Wei, G. C. G. and Tanner, M. A. (1990) A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithm, Journal of the American Statistical Association, 85, 699-704.

[22] Yaseen, A. S., Gad, A. M. and Ahmed, A. S. (2016) Fractional Imputation Methods for Longitudinal Data Analysis, American Journal of Applied Mathematics and Statistics, 4, 59 - 66.