Electronic Journal of Applied Statistical Analysis
EJASA, Electron. J. App. Stat. Anal.
http://siba-ese.unisalento.it/index.php/ejasa/index
e-ISSN: 2070-5948
DOI: 10.1285/i20705948v16n2p410

A comparative study on repeated measurements data in the presence of missing data
By Al-Rawwash and Alquran

14 October 2023

Electronic Journal of Applied Statistical Analysis
Vol. 16, Issue 2, 2023, 410-422
DOI: 10.1285/i20705948v16n2p410

A comparative study on repeated measurements data in the presence of missing data

Mohammad Y. Al-Rawwash*[a] and Haneen Alquran[a]

[a]*Department of Statistics, Yarmouk University, Jordan*

14 October 2023

The occurrence of missing observations is nearly unavoidable in longitudinal studies where repeated measurements are taken over time on the same subject who may miss appointments or drop out during the study period. In this article, we use the Gaussian estimating objective function to estimate the regression and correlation parameters and handle missing data using multiple imputation. The estimation of these parameters is carried out simultaneously using the iterative Newton-Raphson algorithm and the expectation-maximization algorithm. These ideas are implemented using two real data sets and both algorithms showed comparable results with respect to the standard errors of the parameters of interest.

keywords: Correlation structure; Gaussian estimation; Imputation; Longitudinal data; Missing values.

1 Introduction

Longitudinal studies are prevalent in many areas including biological, medical and social studies where subjects are measured repeatedly over time. In addition, modeling the covariance matrix and handling missing data are among the most challenging problems that may encounter such analysis. The basic goal is to model and describe the mean response as a function of time as well as other related explanatory variables in the presence of missing values. One of the common techniques adopted for estimating the regression and covariance parameters is the Gaussian estimation (GE) that requires no

*Corresponding author:rawwash@yu.edu.jo.

ISSN: 2070-5948
http://siba-ese.unisalento.it/index.php/ejasa/index

distributional assumptions. It uses the normal log-likelihood as the estimation objective function without assuming the data to be normally distributed. The GE procedure has been shown to have good properties in a number of applications. For example, Crowder (1985) showed by simulation that a Gaussian estimate of the correlation parameter of equi-correlated clustered binary data has high efficiency. In another direction, Chaganty and Shults (1999) proposed their own approach to estimate the mean and correlation parameters based on a specific objective function that ensure both estimators to be unbiased. GE was introduced and discussed in several articles to estimate the mean response and correlation parameters ((Crowder, 2001); (Al-Rawwash, 2005a) and (Al-Rawwash and Pourahmadi, 2006)). In this work, we use the Gaussian function as an objective function for the estimation of the regression, correlation and dispersion parameters and attempt to obtain the minimizers of the Gaussian function with respect to the parameters of interest. The idea will handle all types of data sets even when the observations are not normally distributed. Ignoring the correlation structure and parameters may cause a severe loss of efficiency. It is vital to model the correlation among the measurements in time series, spatial data and longitudinal data which is considered as significant as modeling the mean ((Diggle et al., 1994); (Seif and Rawwash, 2006) and (Elashoff et al., 2016)).

The problem of missing observations while collecting data sets is common in clinical trials, biological studies in addition to many research areas including panel data and longitudinal data studies. The loss of information due to the missing values is crucial and it is unacceptable in certain cases to ignore such problems. One of the methods used to treat the missing values in longitudinal data analysis is the case deletion which basically allows the analyst to drop the individual records completely from analysis once we notice the missing values in that record. This method may work well if the sample size is large and dropping one individual's record will be insignificant. The framework for missing data, introduced by Rubin (2004), consists of three missing data mechanisms that describe the relationships between measured variables and the probability of missing data: missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). MCAR is very strict and uncommon in research while MAR assumes that the missing values depend on the observed values which makes it a realistic and flexible assumption. It was reported that, under certain circumstances, missing data may cause some bias and thereby increasing the chance of misleading inferences about changes in the mean response ((Rubin, 2004)). Many imputation techniques have been proposed to compensate for the nonresponse in surveys and to handle the missing values problem in clinical trials. Bootstrap technique was also proposed to treat the nonresponse and missing values cases by using the bootstrap sample as if they are the original sample values ((Shao and Sitter, 1996)). Many sample-based imputation techniques commonly used in the literature including mean imputation, last observation carried forward, multiple imputation (see (Rubin, 2004) and (Raghunathan, 2015)). Eventually dealing with missing or incomplete data is a common problem in many areas of applied statistics and it is common to handle the missing values in the study rather than excluding any case that shows some missing values. The imputation of the missing values has been studied extensively in the literature and researchers concluded that the

multiple imputation method handles the missing values better than other imputation techniques. The multiple imputation is one of the most common methods used to overcome this problem ((Rubin, 2004) and (Li et al., 2015)). In another direction, Garcia et al. (2012) proposed a data driven and graphical method to handle the missing values problem and they incorporated the EM algorithm in the analysis to obtain the estimates of the mean and covariance parameters.

In this article, data analysis will be conducted to investigate the efficiency of the GE approach compared to other methods of estimation in the presence of missing data. The EM method is proved to be efficient in the presence of missing values and it will iteratively obtain the maximum likelihood estimates of the parameter of interest ((Dempster et al., 1977) and (Lindstrom and Bates, 1988)). The iterative Newton-Raphson method as well as the EM method will be used in this article to obtain the parameters estimates in the presence of missing values.

The paper is structured as follows. In section 2, we setup the notation and review some estimation approaches using the framework of GE. In section 3, we discuss the model, the estimation equations as well as the imputation techniques in the longitudinal data setup. The methodology will be illustrated using two real data sets in section 4 and a comparison will be conducted using different strategies. Finally, section 5 provides a discussion on the merits of imputation and modeling of correlation matrix in longitudinal data analysis.

2 The objective function

To setup the notation, we assume having m different subjects each of which has n_i repeated measurements. Also, we consider $Y_i = (y_{i1}, y_{i2}, \ldots, y_{in_i})'$ to be the measurements for the ith subject taken at times $t_i = (t_{i1}, t_{i2}, \ldots, t_{in_i})'$ where $i = 1, 2, \ldots, m$ and y_{ij} is the repeated measurement of subject i on time j. Moreover, we assume that there is a covariate $X_{ij} = (x_{ij1}, x_{ij2}, \ldots, x_{ijp})'$ associated with each repeated measurement. In this article, no distributional assumptions are available except for those about the first two moments where $\mu_i(\beta) = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{in_i})'$ is the expected value of Y_i and Σ_i is the $(n_i \times n_i)$ covariance matrix of Y_i. In certain cases where $n_i = n$, the data set will be balanced and there will be no missing values.

The mean vector $\mu = \mu(\beta) = (\mu_1, \cdots, \mu_n)'$ will be modeled assuming a relation with the associated covariate via a link function such that $g(\mu_i) = X_i\beta$ where g is an invertible link function, $\beta = (\beta_1, \cdots, \beta_p)'$ is a $(p \times 1)$ unknown regression parameters of primary interest. The variance of the repeated measurements is modeled as $\text{var}(y_j) = \phi_j v(\mu_j), j = 1, 2, \cdots, n$, where $v(\cdot)$ is a known variance function and $\phi_1, \cdots, \phi_n$ are dispersion parameters that does not depend on β. Note that the variance of y_j is the product of two components; $v(\mu_j)$ expresses the part of the variance functionally dependent on the mean μ_j, while ϕ_j expresses the variability beyond the mean. The correlation matrix $R = R(\alpha)$ is introduced to model the correlations between the measurements on the same subject assuming that the correlation and regression parameters (α, β) will not depend on each other, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_q)$ is a $(q \times 1)$ vector of

unknown correlation parameters. The correlation structure might be selected from a wide range of possible structures and models available in the literature including some well-known parsimonious structures such as the autoregressive, moving average and compound symmetry. In addition, data driven techniques that involves nonparametric or Bayesian approaches are considered as strong competitors to be used while finding the best way to model the variances and correlation in repeated measures scenarios ((Al-Rawwash, 2005b) and (Al-Rawwash and Pourahmadi, 2013)). Graphical methods such as regressograms introduced by Pourahmadi (1999) may be considered as another way to model the components of the covariance matrix via meaningful and unconstraint set of parameters. This is, in fact, a generalization of the Yule's correlogram for stationary data setup. The covariance matrix of Y_i will be decomposed in order to view the correlation matrix, the variance components and the dispersion parameters as separate components which will be modeled based on the associated covariates. We write the covariance matrix as follows:

$$\Sigma_i = V_i^{\frac{1}{2}} R V_i^{\frac{1}{2}}. \tag{1}$$

where $V_i =$ diag $(\phi_1 v(\mu_1), \cdots, \phi_n v(\mu_n))$, $i = 1, 2, \ldots, m$.

To obtain the estimates of the parameters (β, α, ϕ), we minimize the Gaussian objective function proposed by Al-Rawwash and Pourahmadi (2006) with respect to the parameters of interest. The objective function is

$$G(\beta, \alpha, \phi|Y) = \sum_{i=1}^{m} [log|V_i^{\frac{1}{2}} R V_i^{\frac{1}{2}}| + Z_i'(\beta) R^{-1} Z_i(\beta)], \tag{2}$$

where $Z_i(\beta) = V_i^{-\frac{1}{2}}(Y_i - \mu_i(\beta))$, $i = 1, 2, \ldots, m$.

3 Estimation

Many researchers proposed estimating the parameters of interest via different approach including but not limited to parametric, nonparametric methods and entropy approach (see (Diggle et al., 1994); (Al-Rawwash, 2005a); (Al-Rawwash, 2005b); (Al-Rawwash and Al-Nasser, 2011) and (Al-Rawwash and Al-Nasser, 2013)). The problem of obtaining the minimizers of the Gaussian objective function in a closed form solution is not attainable for general choice of correlation structure and unfortunately there is no analytical solution of this objective function. Consequently, numerical computation is usually used to carry out the estimation process. This section paves the way to obtain (numerically) the minimizers of this objective function and obtain the parameters estimates without any distributional assumptions except for the first two moments. Two well known methods will be adopted to iteratively reach to the minimizers of the objective function, namely the Newton-Raphson method and the EM algorithm. These methods will be compared via two data sets analysis and the results will be presented in the sequel.

3.1 The Newton-Raphson Algorithm

The iterative Newton-Raphson algorithm is one of the most common methods used to find the optimizers of the Gaussian objective function. Although the differentiation of this objective function with respect to β is complicated since β appears in $V_i(\beta)$, Al-Rawwash and Pourahmadi (2006) proposed avoiding this problem by assuming that $V_i^{-1/2}$ does not depend on β. Accordingly, differentiating the objective function with respect to the parameters of interest give the following:

$$D'(\beta)V^{-1/2}(\beta)\tilde{R}^{-1}(\alpha)Z = 0, \tag{3}$$

$$\frac{\partial}{\partial\alpha}\log|\tilde{R}(\alpha)| + Z'\frac{\partial\tilde{R}^{-1}(\alpha)}{\partial\alpha}Z = 0. \tag{4}$$

where $D(\beta) = \text{diag}(D_1, D_2, \cdots, D_m)$, $D_i = D_i(\beta) = \frac{\partial\mu_i}{\partial\beta'}$ for $1 \leq i \leq m$, $\tilde{R} = I_m \otimes R(\alpha)$, $V(\beta) = \text{diag}(V_1(\beta), V_2(\beta), \cdots, V_m(\beta))$, $Z = (Z_1', \ldots, Z_m')'$ and $\otimes$ is the kronecker product. Following Al-Rawwash (2001) the Newton-Raphson iterative algorithm is carried out as follows:

Step 1: Select and initial value of β say $\tilde{\beta}$.

Step 2: Evaluate the following quantities at $\tilde{\beta}$: $\tilde{V}_i = V_i(\tilde{\beta})$, $\tilde{\mu}_i = \mu_i(\tilde{\beta})$, $\tilde{Z}_i = \tilde{V}_i^{-1/2}(Y_i - \tilde{\mu}_i)$ and $\tilde{D}_i = D_i(\tilde{\beta})$.

Step 3: Solve equation (4) for $\tilde{\alpha}$,

Step 4: Evaluate $\tilde{R} = R(\tilde{\alpha})$ and $\tilde{\Sigma}_i = \tilde{V}_i^{1/2}\tilde{R}\tilde{V}_i^{1/2}$ for $i = 1, 2, \ldots, m$ and construct the covariance matrix $\tilde{\Sigma} = \text{diag}(\tilde{\Sigma}_1, \tilde{\Sigma}_2, \cdots, \tilde{\Sigma}_m)$ as well as $D(\tilde{\mu})$ and $\mu(\tilde{\beta})$.

Step 5 Update the value $\hat{\beta}$ using $\hat{\beta} = \tilde{\beta} + (\tilde{D}'\tilde{\Sigma}^{-1}\tilde{D})^{-1}\tilde{D}'\tilde{\Sigma}^{-1}(Y - \tilde{\mu})$ and the iterative process will stop once $\hat{\beta} \simeq \tilde{\beta}$.

The least squares estimates $\tilde{\beta} = (X'X)^{-1}X'Y$ of β is a good candidate for the initial value of β, while estimates of α must be chosen so that the positive definiteness of the correlation matrix is guaranteed. The estimate of ϕ depends on the GE results of β and α obtained using Newton-Raphson iterative method which is reduced to $\hat{\phi} = \frac{1}{mn}\sum_{i=1}^{m}\hat{Z}_i'\hat{Z}_i$

3.2 The EM algorithm

Following the notation setup in Section 2, we assume that $Y_{i,ob}$ is the vector of observed repeated measurements of the i^{th} subject and the missing values for this subject will be denoted by $Y_{i,ms}$. Accordingly, we define $\tilde{Y_i}' = (Y_{i,ob}', Y_{i,ms}')$ where the missing values on certain occasions occur when the i^{th} subject fails to report a response or the researcher miss the record on that occasion. Hence, the mean vector will be given as $\tilde{\mu}' = (\mu_{ob}', \mu_{ms}')$ and the covariance matrix will be partitioned as follows

$$\tilde{\Sigma} = \begin{pmatrix} \Sigma_{ob,ob} & \Sigma_{ob,ms} \\ \Sigma_{ob,ms}' & \Sigma_{ms,ms} \end{pmatrix},$$

and we use equation (1) to explain the decomposition of the covariance matrix into the correlation matrix and the variance components. The EM algorithm ((Dempster et al., 1977)) eventually leads to a competitive estimates of the parameters in spite

of the slow rate of convergence. The score equations might be used (iteratively) to obtain the estimates of the parameters noting that this procedure will eventually lead to global convergence. Accordingly, the algorithm aims at finding the minimizers of the GE which is equivalent to finding the zeros of (3) and (4). This works initially by setting the correlation and dispersion parameters fixed and update the regression parameter by minimizing the objective function with respect to β. The iterative EM algorithm proceeds by updating the correlation and dispersion parameters depending on the negative value of conditional expectation of objective function(G) assuming the observed data to be given and depending on the updated parameter β which is denoted by Q. The E-step allows us to write:

$$-Q = E(G(\beta, \alpha, \phi)|Y_{1,ob}, \cdots, Y_{m,ob}) = \sum_{i=1}^{m} E[log|V_i^{\frac{1}{2}} R V_i^{\frac{1}{2}}| + Z_i'(\beta) R^{-1} Z_i(\beta)], \qquad (5)$$

where $Z_i(\beta) = V_i^{-\frac{1}{2}}(Y_i - \mu_i(\beta))$ and V, R and μ are obtained based on the observed values.
The M-step involves updating the parameters α and ϕ so that the conditional expectation of objective function given the observed values is maximized at every iteration. The M-step allows us to choose the correlation and dispersion parameters so that the quantity Q increases, consequently we switch back to updating the parameter β and so on until the parameters of interest converge.

Generally, we start with an initial value that is usually considered to be the least square estimate $\hat{\beta}_{(1)} = (X'X)^{-1}X'\tilde{Y}$ while the starting point of Σ is usually selected to be the sample covariance matrix of the observed response. Finally we keep updating the values of parameters of interest and the algorithm will stop when $|\hat{\beta}_{(k+1)} - \hat{\beta}_{(k)}| \approx 0$. It is noteworthy that the speed of convergence of the EM algorithm depends on the quantity of the missing information in the data set. ((Dempster et al., 1977) and (Taisir and Islam, 2014)). Moreover, Herzet et al. (2006) pointed out that the Newton-Raphson method converges faster in some cases compared to the EM algorithm.

3.3 Imputation methods

As mention earlier, missing values may be encountered in real life data sets and researchers introduced many ideas and solutions to overcome this problem, however imputation may act better if we use a proper method to replace the missing value. Some imputation methods such as mean imputation and last observation carried forward may produce biased results even though it may result in a small standard error of the parameter estimator. Among a variety of methods that have been developed in the literature to overcome the missing values problem, we use in this article the multiple imputation approach along with the Newton-Raphson method and the EM algorithm as iterative methods to help finding the parameters estimates and overcome such kind of situations ((Rubin, 2004); (Li et al., 2015) and (Huque et al., 2018)). Multiple imputation basically handles the missing values (Y_{ms}) by allowing different plausible values of these missing

values to create imputed data sets and appropriately combining results obtained from these setups. Once we create multiple copies of the data set with a proper imputed values inserted to treat the missing values, we obtain the estimates of the parameters by finding the minimizers of the objective function (2) using each of the imputed data sets via Newton-Raphson method. The parameters estimates and their standard errors are combined and formulated using Rubin's rules ((Little and Rubin, 2019)). Multiple imputation method assume that observations are missed at random (MAR) and researchers have reported some drawbacks of the imputation method, however it is still considered as an important and popular method to treat the missing values problem.

4 Data analysis

In this section, two data sets will be considered to investigate the effect of missing data and the imputation strategies explained earlier. Potthoff and Roy (1964) considered a data set that consists of 27 children in a dental study where each individual was subject to four measurements. The second data set is reported by Koziol et al. (1981) to study the effects of three different immunotherapy regimens on colon carcinoma in mice. The experiment was applied on 30 mice that have been randomly divided into three groups of 10 mice with eleven repeated measurement knowing that missing values occurred during the data collection.

Table 1: Estimation results for the dental data set using different methods

	GEE		C-QLS		GE	
	Est.	St.Error	Est.	St.Error	Est.	St.Error
β_1	16.5946	1.2788	16.5931	1.2781	16.5925	1.2778
β_2	17.3213	0.7780	17.3215	0.7776	17.3216	0.7774
γ_1	0.7965	0.1050	0.7695	0.1049	0.7696	0.1049
γ_2	0.4838	0.0629	0.4837	0.0629	0.4837	0.0629
ϕ	4.9107		4.9106		4.9105	
α	0.5997		0.6105		0.6082	

4.1 Dental data

The data set consists of 27 subjects in a dental study (11 girls and 16 boys). The response of the i^{th} subject measured on the j^{th} occasion (denoted by y_{ij}) is the distance from center of pituitary to pterygomaxillary fissure. Each child has been examined and the results are recorded at the ages of 8, 10, 12, and 14.

We consider the model presented by Al-Rawwash and Pourahmadi (2013) which can be written as

$$\tilde{\mu}_{ij} = \beta_1 x_{i1} + \beta_2 x_{i2} + \gamma_1 x_{i1} * x_{i3} + \gamma_2 x_{i2} * x_{i3}, j = 1, 2, 3, 4 \text{ and } i = 1, 2, \cdots, 27. \quad (6)$$

where x_{i1}, x_{i2} are indicator variables for boys and girls, respectively. The covariate x_{i3} is the subject's age at the j^{th} measurement time. The estimate of (β, α, ϕ) are computed using Newton-Raphson method assuming working correlation structure to be an autoregressive structure of order 1 (AR(1)). Al-Rawwash and Pourahmadi (2013) reported the estimation results including the parameters estimates and their standard error using three estimation strategies namely generalized estimation equations (GEE), corrected quasi least squares method (C-QLS), and the GE method. Table 1 shows the estimates of the regression parameters and their standard errors and we notice the results using the three methods to be in a reasonable agreement.

Table 2: Estimated parameters and standard errors in the presence of missing data

Parameter	25% missing data	Multiple imputation	EM algorithm
$\hat{\beta}_1$	17.1006	16.3030	16.3952
(SE)	1.5004	1.1056	1.1095
$\hat{\beta}_2$	17.3799	15.8890	16.9125
(SE)	0.9339	0.7578	0.7557
$\hat{\gamma}_1$	0.7109	0.8054	0.7928
(SE)	0.1456	0.0962	0.0962
$\hat{\gamma}_2$	0.4772	0.6485	0.4388
(SE)	0.0862	0.0576	0.0579
$\hat{\phi}$	4.9905	5.3395	5.0147
$\hat{\alpha}$	0.5630	0.6051	0.5932

In order to investigate the effect of the previously mentioned optimization algorithms and missing values treatments on the parameters estimates as well as their standard errors, we will reanalyze the dental data set after randomly deleting 25% from the observation and compare the imputation-based results using only the GE method. Table 2 shows the parameters estimates and the standard errors based on model (6) assuming the AR(1) correlation structure and depending on the previously mentioned imputation techniques. We notice that the standard error gets larger when we drop some of the existing data elements while the multiple imputation technique and the Newton-Raphson method reduce the standard errors of the parameters estimates. Also, the standard errors of the parameter estimates using the EM method are comparable to those obtained via multiple imputation method. The number of missing observation may affect the

speed of the algorithm and the rate of convergence during the optimization procedure ((Dempster et al., 1977) and (Herzet et al., 2006)). The results in Tables 1 and 2 show the effectiveness of the algorithms used to treat the missing values problem.

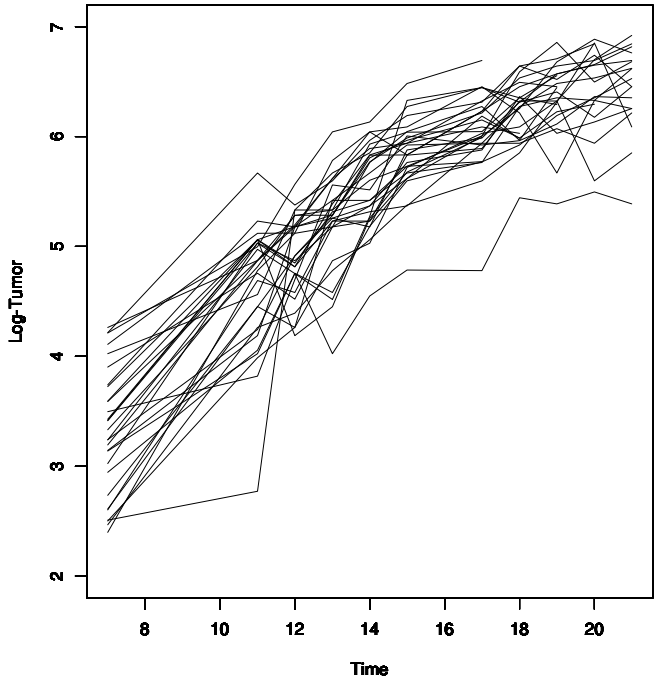

Figure 1: Log-tumor size of the colon carcinoma mice over time

4.2 Mice tumor data

In a study to test the efficiency of three therapies against colon carcinoma in mice, Koziol et al. (1981) reported the study of immunotherapy regimens on colon carcinoma where thirty mice have been injected with mouse colon carcinoma cells. They divided the 30 mice randomly into three equal groups such that repeated measurements were taken for the size of the tumor on unequally spaced time span as follows, days 7, 11, 12, 13, 14, 15, 17, 18, 19, 20 and 21. Unfortunately, some mice did not survive until the end of the experiment causing around 9% missing values in the final data set and therefore contributed only partial information of growth study. The tumor size of the 26^{th} mouse was decreasing with time which is considered illogical and therefore removed from the experiment. Figure 1 shows the tumor size measured over time and a general pattern of the tumor size over time is clearly noticeable. It is noteworthy to mention that we used the log-transformation of the tumor size to make patterns more visible. In fact, the transformation is given as follows:

$$y_{ij}^* = ln(y_{ij} + 1), i = 1, 2, \cdots, 29, j = 1, 2, \cdots, n_i.$$

Table 3: Estimated parameters and standard errors for the mice data

Parameter	Incomplete data	Multiple imputation	EM algorithm
$\hat{\beta}_1$	1.6947	2.1386	2.1877
(SE)	0.1899	0.1699	0.1691
$\hat{\beta}_2$	0.2441	0.2127	0.2093
(SE)	0.0106	0.0105	0.0103
$\hat{\phi}$	0.3936	0.2947	0.3154
$\hat{\alpha}$	0.5106	0.5109	0.5110

Al-Rawwash (2005b) and the references therein discussed modeling longitudinal data that exhibit some missing values. He analyzed the mice data set using parametric and nonparametric methods. The components of the covariance matrix were estimated via the variogram and variances using kernel smoothing. In this article, we treat the missing value problem using the aforementioned methods and we suggest fitting the data set using the model

$$y_{ij}^* = \beta_0 + \beta_1 t + \epsilon_{ij},$$

where $i = 1, 2, \cdots, 29$ and $j = 1, 2, \cdots, n_i$ and $t = 1, 2, \cdots, 11$. In addition, we assume the correlation structure to be the first order autoregressive (AR) with parameter α and we proceed in the estimation process using GE after repairing the missing values through the imputation techniques. In addition, we carry out the estimation techniques assuming that the data is missing completely at random (MCAR). In Table 3, we present the estimates of the parameters as well as the standard errors for the mice data set using the incomplete data in addition to the imputation techniques discussed earlier to adjust the missing values assuming AR(1) correlation structures. The results in Table 3 indicate that the imputation methods enhanced the estimation process and the standard errors are smaller compared to the incomplete data set situation. Treating the missing values and reaching to the parameters estimates is shown in Table 3 and the results therein support the significance of the imputation strategies and both approaches perform well.

5 Discussion

We have indicated that the GE discussed by Al-Rawwash and Pourahmadi (2006) is a practical method since it reduces to the maximum likelihood estimation in the presence of the normal distribution assumption. It also subsumes the recent improvements of

GEE designed to estimate the mean and covariance parameters efficiently. We considered treating the missing values using different imputation methods assuming the GE approach. A first order autoregressive correlation structure is considered to handle the correlation among the repeated measurements and to formulate the models for the mean and correlations using two real data sets. The parsimonious correlation structure reduces the number of parameters used to detect the correlation among the measurements in the estimation process in order to model the specification of covariance matrices that are closer to the unknown covariance and correlation matrices of the data. It is clear that modeling correlation structure carefully and simultaneously with the mean usually leads to improved model formulation and parameter estimation both for the mean and the correlation in the presence of missing data. The Newton-Raphson method and the EM algorithm exhibit comparable results in the presence of missing data and the standard errors are comparable using both methods. The results confirm that the adopted methods of treating the missing values are efficient choices for the incomplete data setups.

Acknowledgement

The authors would like to thank the editor and two anonymous referees for their valuable and constructive comments that helped improving this paper.

References

Al-Rawwash, M. (2001). *Gaussian estimation and modeling covariance in longitudinal data analysis*. Ph.D. Thesis, Northern Illinois University.

Al-Rawwash, M. (2005a). Modeling covariance parameters for purely autoregressive correlated longitudinal data. *Communications in Statistics - Simulation and Computation®*, 34(1):73–83.

Al-Rawwash, M. and Al-Nasser, A. D. (2013). Repeated measures and longitudinal data analysis using higher-order entropies. *Statistica Neerlandica*, 67(1):100–111.

Al-Rawwash, M. and Pourahmadi, M. (2006). Gaussian estimation and joint modeling of dispersions and correlations in longitudinal data. *Computer methods and Programs in Biomedicine*, 82(2):106–113.

Al-Rawwash, M. and Pourahmadi, M. (2013). Gaussian estimation of regression and correlation parameters in longitudinal data. *Journal of the Association of Arab Universities for Basic and Applied Sciences*, 13(1):28–34.

Al-Rawwash, M. Y. (2005b). Covariance matrix estimation using repeated measurements when data are incomplete. *Applied mathematics and computation*, 167(1):299–315.

Al-Rawwash, M. Y. and Al-Nasser, A. D. (2011). Longitudinal data analysis using generalized maximum entropy approach. *Jordan Journal of Mathematics and Statistics (JJMS)*, 4(1):47–60.

Chaganty, N. R. and Shults, J. (1999). On eliminating the asymptotic bias in the quasi-

least squares estimate of the correlation parameter. *Journal of Statistical Planning and Inference*, 76(1-2):145–161.

Crowder, M. (1985). Gaussian estimation for correlated binomial data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 47(2):229–237.

Crowder, M. (2001). On repeated measures analysis with misspecified covariance structure. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 63(1):55–62.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Diggle, P., Liang, K.-Y., and Zeger, S. L. (1994). *Longitudinal data analysis*. New York: Oxford University Press.

Elashoff, R., Li, G., and Li, N. (2016). *Joint modeling of longitudinal and time-to-event data*. CRC press.

Garcia, T. P., Kohli, P., and Pourahmadi, M. (2012). Regressograms and mean-covariance models for incomplete longitudinal data. *The American Statistician*, 66(2):85–91.

Herzet, C., Wautelet, X., Ramon, V., and Vandendorpe, L. (2006). Iterative synchronization: Em algorithm versus newton-raphson method. In *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*, volume 4, pages IV–IV. IEEE.

Huque, M. H., Carlin, J. B., Simpson, J. A., and Lee, K. J. (2018). A comparison of multiple imputation methods for missing data in longitudinal studies. *BMC medical research methodology*, 18(1):1–16.

Koziol, J. A., Maxwell, D. A., Fukushima, M., Colmerauer, M., and Pilch, Y. H. (1981). A distribution-free test for tumor-growth curve analyses with application to an animal tumor immunotherapy experiment. *Biometrics*, pages 383–390.

Li, P., Stuart, E. A., and Allison, D. B. (2015). Multiple imputation: a flexible tool for handling missing data. *Jama*, 314(18):1966–1967.

Lindstrom, M. J. and Bates, D. M. (1988). Newton-raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022.

Little, R. J. and Rubin, D. B. (2019). *Statistical analysis with missing data*, volume 793. John Wiley & Sons.

Potthoff, R. F. and Roy, S. (1964). A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 51(3-4):313–326.

Pourahmadi, M. (1999). Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690.

Raghunathan, T. (2015). *Missing data analysis in practice*. CRC press.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys*, volume 81. John Wiley & Sons.

Seif, M. A. and Rawwash, M. (2006). Surface roughness measurement using laser speckle statistical analysis. *Journal of testing and evaluation*, 34(5):410–415.

Shao, J. and Sitter, R. R. (1996). Bootstrap for imputed survey data. *Journal of the American Statistical Association*, 91(435):1278–1288.

Taisir, R. and Islam, M. A. (2014). Em algorithm for longitudinal data with non-ignorable missing values: an application to health data. *Bangladesh Journal of Scientific Research*, 27(2):133–142.