

Modeling Correlation in Incomplete Longitudinal Data: The Case of Fruit Fly Mortality Data

Tanya Garcia, Priya Kohli, and Mohsen Pourahmadi ¹

Abstract

Longitudinal studies are prevalent in clinical trials, biological and social sciences where subjects are measured repeatedly over time. Modeling the correlations of repeated measurements on the same subject and handling missing data are challenging problems in the statistical analysis of such data. The situation is exacerbated knowing that the presence of missing data can hamper modeling of dependence, and improper accounting of dependence can negatively affect imputation of the missing values. Methods for handling missing data have been thoroughly studied, but data-based and graphical methods for modeling the covariance matrix of longitudinal data are relatively new. Our work illustrates the insensitivity of formulating models for the covariance matrix to different methods of handling missing values in longitudinal studies for the fruit fly mortality data which has about 22% missing values. We emphasize the role of graphical tools like the *regressograms* in formulating models for covariance matrices under different methods of handling missing data. Surprisingly, for five of commonly used methods, the regressograms remain robust and consistent in suggesting the same class of cubic polynomial models for the components of the modified Cholesky decomposition of the sample covariance matrix. We hope this aspect of the success of the regressograms will encourage statisticians to use them in conjunction with other graphical tools for displaying dependence in longitudinal data and formulating parametric models for the covariance matrix in the presence of missing data.

Key words: Covariance Structure; Graphical tools; Imputation; Missing Data; Regressograms.

Short title: Interplay Between Correlation and Missing Data

¹ Tanya Garcia (email: tpgarcia@stat.tamu.edu) and Priya Kohli (email: pkohli@stat.tamu.edu) are doctoral students and Mohsen Pourahmadi (email: pourahm@stat.tamu.edu) is Professor in the Department of Statistics, Texas A&M University, College Station, TX 77843-3143, USA. This work grew out of a class project in a course on Missing Data in Longitudinal Studies. The research is supported in part by the National Science Foundation Bridge to the Doctorate Fellowship (T. Garcia) and the National Science Foundation grant DMS-0906252 (M. Pourahmadi and P. Kohli).

1. INTRODUCTION

Missing or incomplete data is a common problem in many research areas (Rubin 1976; Little and Rubin 2002). The occurrence of missing observations is especially omnipresent in longitudinal studies where repeated measurements are taken on the same subject, and hence there is more opportunity for a subject to miss appointments or drop out during the study period; see Daniels and Hogan (2008) and Fitzmaurice, Laird and Ware (2004, Chap. 14). Another expected feature of longitudinal data is the correlations among successive measurements on the same subject. When there is substantial missingness and strong correlation in the data, serious statistical issues such as bias and inefficient estimates often lead to incorrect inferences (Fitzmaurice et al. 2004, Chap. 14). Whereas methods for handling complete data are easily accessible and can be implemented using commercial software packages, resources and methods for handling missing data are not readily available nor can be easily implemented by a novice data analyst. In addition, data-based modeling techniques for the dependence structure or the covariance matrix of longitudinal data is scarce even for complete data.

In this paper we investigate the sensitivity of model formulation for the covariance matrix of longitudinal data to five commonly used methods of handling missing data. The first three naive methods are: the complete-case analysis which uses data from those subjects with no missing values, available-data analysis where *all* the observed values are used, and finally, the last value carried forward (Fitzmaurice et al. 2004, Chap. 14) which uses the last observed value to fill-in (impute) the missing values so that the responses remain constant at the last observed value for a given missing segment. Of course, these are fairly simple to apply and require no models for the data. **The next two methods require explicit modeling of the data and are ideal starting points for understanding the conceptual underpinnings of single- and multiple-imputation methods based on the predictive distribution of the missing values.** First, the regression imputation replaces the missing values for a subject by predicted values from a regression of the missing values on the observed data. Finally, the stochastic regression imputation replaces the missing values by a value predicted from a regression imputation plus a random quantity, drawn to reflect the uncertainty in the predicted value. We shall work throughout with normal linear regression models, hence the random quantities added to the predicted values will be drawn independently from a normal distribution with mean zero and variance equal to the residual variance of the regression (Little and Rubin 2002).

We present methods for formulating models for the covariance matrix of the incomplete data using data-driven and graphical techniques. Two graphical examples are

the ordinary scatterplot matrices (OSM) commonly used in multivariate statistics and the partial regression on intervenors scatterplot matrix (PRISM) which consists of pairwise partial scatterplots of the standardized responses (Zimmerman 2000). We highlight the roles of the regressograms proposed by Pourahmadi (2000) and their interplays with the five methods of dealing with missing data. **Regressograms for longitudinal data involve its covariance matrix and are suitable plots of predictor coefficients and prediction error variances when a measurement is regressed on its predecessors.** The quantities involved in the construction of regressograms can **also** be obtained from the modified Cholesky decomposition of the related sample covariance matrix. Similar to the use of scatterplots in regression analysis and correlograms in time series analysis, these graphical tools allow formulating parsimonious parametric models for the dependence structure of the longitudinal data. This graphical approach makes it possible to avoid completely or can be of additional help in selecting a covariance matrix from a long menu of structured covariance matrices available in standard software packages (Fitzmaurice et al. 2004, Chap. 7). Once models are formulated for the components of the modified Cholesky decomposition, their parameters are estimated using ordinary least squares method which can be programmed in any software with linear regression capabilities.

To illustrate the relevance, sensitivity and versatilities of the regressograms in formulating models for the covariance matrix under various methods of handling missing data, we analyze the Fruit Fly Mortality (FFM) data (Zimmerman and Núñez Antón 2009, pp. 21) which has 112 cohorts (subjects) with about 22% missing values and a general pattern of missingness. Our analysis of the FFM data begins with preliminary methods of using all available-data (AD), the complete-cases (CC) and the last value carried forward (LVCF). As there are only 23 cohorts for which measurements are available at all time points, the CC or *completers* analysis is generally viewed to be wasteful as it ignores cohorts with even a single missing value. At the other extreme, the available-data analysis which uses all available observations is not necessarily superior to the **completers** analysis, and could produce estimated covariance matrices that are not positive semi-definite (Little and Rubin 2002, pp. 55). Thus, the feasibility of using these methods are very much dependent on the data at hand. For example, if there are no completers in the data then the CC analysis cannot be used. In general, it is of interest to know whether any of these or other intuitively appealing preliminary methods should be used in practice. In fact, according to Little and Rubin (2002, pp. 41), “Although these methods appear in statistical software and are widely used, we do not generally recommend any of them except in special cases where the amount of missing information is limited.”

Next, we discuss the option of handling missing data through completing the data via imputation. The first and simplest single-imputation method is the LVCF. Of course, the idea of LVCF is based on a very strong and unrealistic assumption that the **missing** responses in a segment are the same as the last observed response. Here, too, it is of interest to know whether **or not** one should use LVCF in practice. According to Fitzmaurice et al. (2004, pp. 393-394), “Despite frequent and well-founded criticism by statisticians, LVCF is still widely used to handle dropouts in clinical trials. Regulatory agencies such as the U.S. Food and Drug Administration (FDA) seem to encourage the continuing use of LVCF as a method for handling dropouts, despite all of its obvious shortcomings. Except in very rare cases, . . . , we do not recommend the use of LVCF as a method for handling dropout.”

Given the shortcomings of AD, CC and LVCF, it is natural to consider more sophisticated imputation methods like the regression imputation and the stochastic regression imputation which rely on the prediction of missing values based on a predictive distribution. Throughout this paper we take this predictive distribution to be a multivariate normal distribution for the FFM data.

An interesting find, perhaps peculiar to the FFM data, is that the regressograms from all five methods for handling missing data remain fairly robust. In each case, the regressograms suggest the same class of cubic polynomials for the components of the modified Cholesky decomposition regardless of the method used for handling missing data. **Throughout, the parameters of these cubic polynomial models are estimated using ordinary least squares (OLS). The resulting OLS estimates can subsequently serve as the initial values for the iterative maximum likelihood estimation (Pourahmadi 2000) and the EM (expectation-maximization) algorithm (Little and Rubin 2002) for handling missing values.** These two methods are not pursued here, **however**, so that the focus of the paper remains on the model formulation stage of the statistical model fitting process.

The remaining sections of this paper are as follows. Section 2 introduces the FFM data along with an explanation of aspects of its missingness and a brief summary of the taxonomy of missing data mechanisms. A detailed discussion of the graphical techniques and the ingredients for the construction of the regressograms are presented in Section 3. Our analysis of the FFM data in Section 4 is based on the simple idea of using all available-data, the 23 complete-cases for which measurements are available at all time points, and the LVCF. The more sophisticated imputation methods based on regression models are discussed in Section 5. Finally, Section 6 concludes the paper **along with** suggestions for further research.

2. FRUIT FLY MORTALITY DATA

The FFM data (Zimmerman and Núñez Antón 2009, pp. 21) consist of age specific mortality measurements from 112 cohorts of *Drosophila melanogaster* (the common fruit fly) obtained by replicating 56 recombinant inbred lines with 500 to 1000 flies within each cohort. Each day during the study, the number of dead flies within each cohort were counted and removed. The FFM data, however, consist of these counts pooled into eleven 5-day intervals.

For each cohort, we let $N(t)$ denote the number of flies still alive at the beginning of the 5-day interval, where $t = 1, \dots, 11$. Then, an appropriate measure for the raw mortality rate is $-\log \{N(t+1)/N(t)\}$. In the aim of making the data approximately normal, a log transform of the raw mortality rate was used as the response variable. That is, for any cohort at time t , the log mortality rate response is given by $y_t = \log [-\log \{N(t+1)/N(t)\}]$, so that $Y_i = (y_{i,1}, \dots, y_{i,11})'$ denotes the vector of repeated measures for cohort i , $i = 1 \dots, 112$. We assume that the observations Y_i are independent 11-dimensional vectors following a multivariate normal distribution with a common mean μ and covariance matrix Σ .

For unknown reasons, approximately 22% of the data values are missing. The missingness is intermittent so that some missing values are followed by observed values. The diverse patterns of missingness in the FFM data is obvious from the reported values for cohorts 7-13 in Table 1. Here, the location/time of missing values is denoted by ‘.’.

Table 1: Measurements for cohorts 7-13 of the FFM Data.

Time	1	2	3	4	5	6	7	8	9	10	11
	-3.39	-3.65	-3.62	-3.60	-3.57	-3.25	-2.98	-2.21	-1.22	-0.81	-0.70
	-3.57	-2.60	-3.47	-2.49	-0.08	-0.18	-0.51	-0.50	0.09	.	.
	-3.27	-2.89	-4.80	-4.10	-1.87	0.4	0.71	0.09	.	.	.
	-4.82	.	-4.11	-1.11	0.32
	.	-4.25	-3.83	-1.35	0.15	0.88	0.09
	-4.63	-3.92	-1.97	-0.78	-0.67	-0.79	-0.30
	-4.96	.	.	-4.95	.	.	-3.84	-3.82	-2.37	-2.04	-1.64

According to Zimmerman and Núñez Antón (2009, p. 21), the main objectives of the study are two-fold: first, to find a parsimonious model that adequately describes the change in mortality averaged over recombinant inbred lines with age (time), and secondly, to find the relation between

mortality at any given age with the mortality at a previous age. Of course, these objectives are closely related to the parsimonious modeling of the mean vector μ and the covariance matrix Σ , which are also the ultimate aim of this paper.

2.1 A Preliminary Available-Data Analysis

How does one start to analyze a data set with even one missing value? The first uncomfortable realization is that the familiar formulas/methodologies for complete/rectangular data **are** not applicable to the missing data situation. A novice data-analyst, however, may pretend that there are no missing values and push the standard formulas for the means, variances and covariances, skipping values or their products in the sums when not available and then divide the resulting sums by the number of available terms. This is a simple instance of *available-data* analysis where the analyst tries not to waste any observations.

For more complicated statistical analysis, it is prudent to start by examining the locations and patterns of missing values, and providing a general summary of the data as in Table 2. These steps are helpful in understanding the causes and types of missingness, and eventually in the choice of methods to handle the missing data. In the case of FFM data, the number of observed values at each occasion reveals that the last three time points, $t = 9, 10, 11$, have **the most amount of missingness**. The sample means and variances in Table 2 were computed using the available-data for each measurement time. Further information relevant to modeling the dependence in longitudinal data can be obtained from the sample correlation matrix and other summary statistics as given in Table 3 for the FFM data. Recall that the sample correlation for a pair of time points is computed using *pairwise* available-data. For example, it can be seen from Table 1 that in computing the sample correlation for the first two time points, cohorts 10-13 do not contribute any cross-product terms due to their patterns of missing values. Mostly due to the imbalance in the number of contributed terms, it is known that a covariance matrix computed using the available-data are not necessarily positive semi-definite (Little and Rubin 2002, pp. 55). Fortunately, for the FFM data such a covariance matrix, denoted by S from here on, is positive-definite since its smallest eigenvalue is 0.131.

As expected in most longitudinal studies and as can be seen easily in Table 3, the available-data correlations are mostly positive and measurements closer in time are more correlated; they become smaller as the time lags between measurements increases. Interestingly, for the FFM data, most of the correlations (appearing in bold face) up to lag 6 are fairly large. Given that the the correlations along the subdiagonals are also decreasing, the assumption of a stationary covariance structure is invalid for this data set. Thus, a method of modeling the covariance that **properly** accounts for

Table 2: Summary statistics for the FFM data; n_{obs} is the number of observed values at each time, \bar{Y} is the sample mean and s^2 is the sample variance of the available data.

Time	1	2	3	4	5	6	7	8	9	10	11
n_{obs}	85	89	94	98	103	104	100	85	73	67	56
\bar{Y}	-3.789	-3.409	-3.418	-2.641	-1.996	-1.283	-0.775	-0.481	-0.567	-0.511	-0.408
s^2	0.701	1.084	1.657	2.609	2.167	1.726	1.054	0.630	0.428	0.334	0.417

the potential nonstationarity is bound to lead to better inferences. For a discussion on the disparity of the available-data and complete-case correlations **evident in Table 3**, see Section 3.1.

2.2 The Taxonomy of Missing Data

In developing the vocabulary and the conceptual underpinnings of the missing data, it is useful to partition response vectors with missing observations as $Y = (Y'_{obs}, Y'_{miss})'$ with the obvious notation for observed and missing values. The missingness indicator in this situation is introduced as the random vector R with entries taking values 1 or 0 depending on whether the corresponding response is observed or not. For example, cohort 10 in the FFM data has $Y_{obs} = (y_1, y_3, y_4, y_5)' = (-4.82, -4.11, -1.11, 0.32)'$, and the remaining components are in Y_{miss} . The random vector R in this case takes the value $r = (1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0)'$. The pair $(y', r)'$ constitutes the *full data* for this cohort and the vector r carries considerable information. Understanding the nature of dependence between R and Y is at the heart of the theory of missing data analysis (Rubin 1976; Little and Rubin 2002). This entails, for example, understanding in a clinical trial if the chance that a patient misses a scheduled visit depends on his medical conditions, the weather, state of the economy, etc.

Following the taxonomy proposed by Rubin (1976) in his landmark paper, we typically categorize the missingness process (mechanism, reason) into one of three well-known classifications: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). The first, and often simplest classification is MCAR which requires that the probability that a particular entry is missing is independent of the values of the response vector y ; namely,

$$P(R = r|y) = P(R = r), \text{ for any } r \text{ and } y.$$

Since missingness is not caused by the response under MCAR, the observed and complete data share the same distributions and moments. More specifically, they share the same means, vari-

Table 3: FFM available-data sample variances (along the main diagonal) and correlations (below the main diagonal), and complete-case sample correlations (above the main diagonal).

Time	1	2	3	4	5	6	7	8	9	10	11
1	0.701	0.488	0.496	0.433	0.230	0.235	-0.124	-0.008	-0.235	0.046	0.023
2	0.585	1.084	0.807	0.458	0.317	0.333	0.093	0.129	-0.139	-0.031	-0.124
3	0.531	0.711	1.657	0.573	0.425	0.355	-0.093	-0.008	-0.450	-0.164	-0.129
4	0.476	0.621	0.776	2.609	0.651	0.413	-0.001	-0.002	-0.167	-0.173	0.384
5	0.340	0.525	0.591	0.787	2.167	0.688	0.287	0.154	-0.051	0.125	0.284
6	0.323	0.383	0.432	0.533	0.821	1.726	0.653	0.561	0.094	0.367	0.196
7	0.206	0.253	0.212	0.277	0.573	0.781	1.054	0.741	0.450	0.468	0.012
8	0.004	0.111	0.080	0.035	0.308	0.512	0.736	0.630	0.466	0.482	0.175
9	-0.025	0.070	-0.024	0.039	0.053	0.050	0.399	0.592	0.428	0.491	0.101
10	-0.006	0.161	-0.159	0.170	0.011	0.029	0.311	0.336	0.462	0.334	0.076
11	0.088	-0.052	-0.153	0.281	0.177	0.192	0.253	0.289	0.292	0.373	0.417

ances, covariances, correlations, etc. Pushing this idea a bit further, it gives a justification for the “complete-case analysis” because the completers (cohorts with no missing values) can be viewed as a *random sample* from the target population, though with a smaller sample size. While this simplifies the analysis by using solely the completers, it is wasteful in that it does not make the best use of the available data. For example, in the FFM data it relies only on the 23 completers, and throws away the data from the other 89 cohorts with some missing values. To circumvent this difficulty, one may rely on the available-data analysis which still can be justified under MCAR.

The second classification, MAR, is more stringent and requires the probability of missingness to depend on the responses only through the observed values:

$$P(R = r|y) = P(r|y_{\text{obs}}).$$

Finally, the missing data mechanism is MNAR when it is neither MCAR nor MAR. More precisely, the missingness depends on both the observed and missing data, thus rendering the likelihood analysis under this assumption difficult. **For more detailed discussions on the missing data mechanisms and their relevance and application to longitudinal data**

analysis, see the expository article by Kenward (1998), and Fitzmaurice et al. (2004, Chap. 14).

For our analysis of the FFM data, we shall assume and rely on the first two missingness mechanisms, MCAR and MAR. A more serious analysis of the FFM data demands answers to the following questions. How does one determine which missingness mechanism to assume for the FFM data? Are there graphical and data-based methods for deciding whether the missingness is MCAR or MAR?

Our preliminary analysis of the FFM data based on the profile plots and mean trends of complete-case, available-data and LVCF-imputed reveals some differences in their distributions under different methods of handling the missing values as depicted in Figure 1 (d). Note, in particular, the difference in the mean trends between the available-data and completers. Of course, deciding whether this observed difference is statistically significant requires estimating their covariance matrices, a topic discussed next.

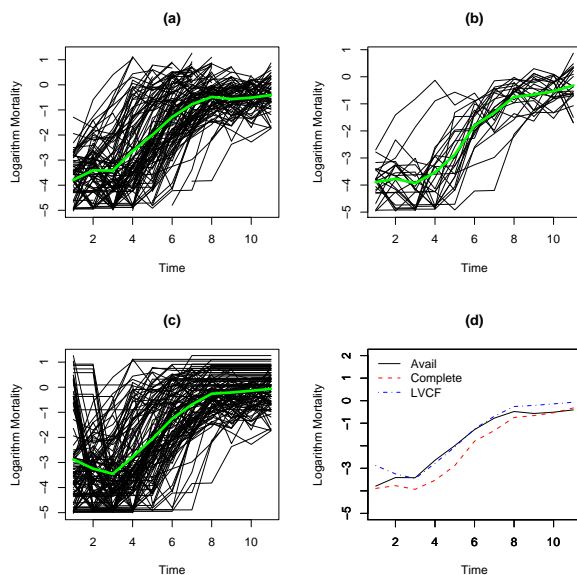


Figure 1: Profile plots: (a) Available-data, (b) Completers, (c) LVCF-imputed, (d) Mean profiles. Darker line in each profile plot exhibits the mean trend.

3. JOINT MODELING OF THE MEAN AND COVARIANCE MATRIX

Modeling the mean using covariates or regression analysis is a powerful area of statistics that took over two centuries to mature to its current status. Graphical tools like scatterplots and residual plots

play central roles in formulating models, and when the basic assumptions of inference are violated, Box-Cox transformations and the like are used to stabilize the variances or to bring the responses closer to normality. Unfortunately, the use of graphical tools or data-driven methods for formulating models for covariance matrices is lagging behind even for complete data. This deficiency is felt, for example, in longitudinal data analysis when using the linear mixed models or generalized estimating equations (GEE) (Fitzmaurice et al. 2004), where one must specify a covariance structure for the dependence of the random effects or the repeated measures on the subjects (Little and Rubin 2002, pp. 242). In most popular software packages in this area, one is directed to pick a structured covariance matrix like the compound symmetry (CS) or autoregression (AR) from a long menu of structured covariance matrices. This is so because data-driven modeling of the covariance matrices is difficult and more demanding due to the positive-definiteness constraint and the large number of parameters which grows quadratically with the number of repeated measures.

In this section we review various graphical techniques that can provide better insight into the dependence structure of the data over time. However, we focus more on the regressograms which are composed of two plots representing the two factors of the modified Cholesky decomposition of a covariance matrix. **For generality**, we assume throughout that the observations Y_i , $i = 1, \dots, m$, are independent n -vectors following a multivariate normal distribution with common mean μ and a common covariance matrix Σ . **Note that in the case of FFM data, we have $m = 112$ and $n = 11$.**

3.1 Some Graphical Tools for Model Formulation

In longitudinal data analysis, it is common to use the profile plot for visually assessing the changes in responses over time and formulating models for the mean. While profile plots are also capable of revealing patterns/models for variances of measurements over time, they are not as successful in revealing dependence patterns or temporal correlations. **Profile plots** for the FFM data, displayed in Figure 1 (a)-(c), show that the variance function takes larger values in the middle, but for the LVCF-imputed data it takes larger values in the early segment.

For modeling correlations over time, it is usually advisable to work either with the profile plot of standardized responses or their ordinary scatterplot matrix (OSM) which consists of pairwise scatterplots of the standardized responses (Dawson, Gennings and Carter 1997). Viewing the OSM as a graphical counterpart of the sample correlation matrix can provide valuable insight into the dependence structure of the data (Pourahmadi 2002) and other features like outliers, nonlinearities, etc. A complementary tool to the OSM is the partial regression on intervenors scatterplot matrix (PRISM) which consists of pairwise partial scatterplots of the standardized

responses (Zimmerman 2000) and is related to the partial correlations between variables. More precisely, the plot in row i and column j ($i \geq j$) is the partial regression plot (otherwise known as added variable plots) of standardized response variables y_{i+1} and y_j adjusted for standardized responses at the intervening times $j+1, j+2, \dots, i$. Therefore, PRISM is the graphical equivalent of the matrix of sample partial correlation between y_{i+1} and y_j adjusted for the intervening variables (not the remaining variables). A random scatter in the (i, j) th plot is evidence that y_{i+1} and y_j are partially uncorrelated adjusted for the intervening variables. For more information on the use of graphical tools in informal model formulation for longitudinal data see Zimmerman and Núñez Antón (2009, Chap. 4).

Figure 2 shows the OSM and PRISM of the FFM completers data. Analogous to our observations for the available-data, the correlations for the completers data also have higher values for smaller lags as can be seen in the OSM and the values in Table 3. The correlations for the completers are for the most part, smaller in magnitude than the corresponding correlations for the available-data and tend to have more negative correlations for larger lags. Notice that the OSM depicts larger, more negative scatter for sections farther from the diagonal indicating the smaller, negative correlations for larger time lags. This discrepancy between the correlations for the completers and available-data indicates a possible drawback of only considering the completers for analysis as most novice analysts may do. Doing so may lead to radically different conclusions for the structure of the correlation matrix.

Contrary to the OSM, the PRISM **of the FFM completers data** indicates that aside from those plots on the main diagonal, the rest appears as random scatter. Interestingly, similar observations were made for the available-data. In **the AD-case**, the partial correlations given in Zimmerman and Núñez Antón (2009, pp. 98, Table 4.4), exhibit approximately quadratic behavior with a point of maximum near the midpoint. The values beyond lag one show no pattern and are very small; **this indicates that a structured first-order antedependent model might be suitable for modeling the dynamics of the mortality rate**, see Section 4.5 for more discussion.

3.2 The Regressograms of a Covariance Matrix

The regressograms of a covariance matrix are graphical tools for formulating models for and revealing the dependence structure of a longitudinal data set. The ingredients used in the construction of the two plots in the regressograms are obtained using the familiar idea of regression or equivalently the modified Cholesky decomposition of $\Sigma = (\sigma_{ij})$, the common covariance matrix of the repeated measures.

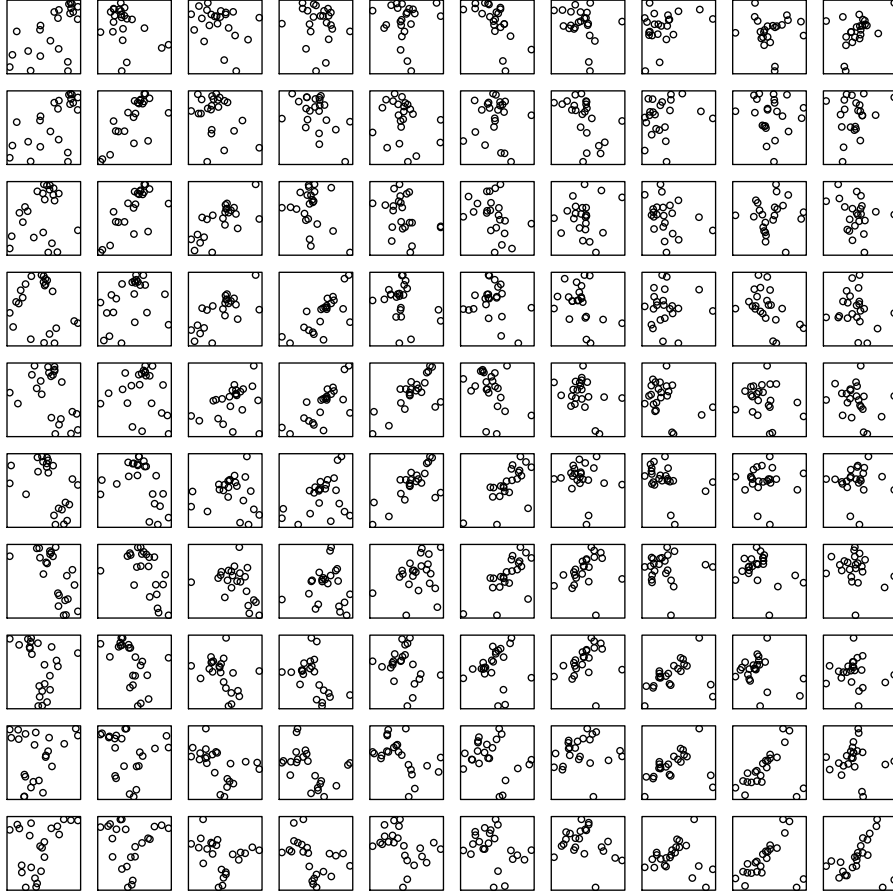


Figure 2: OSM (lower triangle including main diagonal) and PRISM (upper triangle including main diagonal) of completers data.

Recall $Y = (y_1, \dots, y_n)'$ is a random vector of repeated measures on a subject with the mean vector μ and the covariance matrix Σ . For $t = 1, \dots, n$, consider the linear predictor \hat{y}_t , the regression coefficients, denoted below by ϕ_{tj} for $j = 1, \dots, t - 1$, and residual (prediction, innovation) variances, denoted below by σ_t^2 , when the measurement y_t of a subject is regressed on its predecessors y_{t-1}, \dots, y_1 :

$$\begin{aligned} \hat{y}_t &= \mu_t + \sum_{j=1}^{t-1} \phi_{tj}(y_j - \mu_j), \\ \text{var}(y_t - \hat{y}_t) &= \sigma_t^2. \end{aligned} \tag{1}$$

When $t = 1$ the above sum is taken to be zero. Since the successive prediction errors

$$\epsilon_t = y_t - \hat{y}_t,$$

are uncorrelated random variables, **the covariance matrix for** $\epsilon = (\epsilon_1, \dots, \epsilon_n)'$ turns out to be a diagonal matrix $D = \text{diag}(\sigma_1^2, \dots, \sigma_n^2)$. **In matrix form, the above identities simplify to**

$$\epsilon = TY,$$

with T **being the following** unit lower triangular matrix,

$$T = \begin{pmatrix} 1 & & & & & \\ -\phi_{21} & 1 & & & & \\ -\phi_{31} & -\phi_{32} & 1 & & & \\ \vdots & & & \ddots & & \\ -\phi_{n1} & -\phi_{n2} & \cdots & -\phi_{n,n-1} & 1 & \end{pmatrix}.$$

Now, using the above linear transformation to compute the covariance matrix of Y leads to a useful and important factorization/diagonalization of the covariance matrix:

$$T\Sigma T' = D,$$

known as the modified Cholesky decomposition of Σ . Note that **the regression coefficients** ϕ_{tj} 's and the prediction variances σ_t^2 's are computed using the appropriate submatrices of the population covariance matrix Σ .

This decomposition makes it possible to swap the $n(n+1)/2$ constrained parameters of the covariance matrix of Y with the same number of nonredundant entries of the unique (T, D) -pair. For a general Σ , the nonredundant entries of T computed as regression coefficients are unconstrained. Similarly, $\log \sigma_t^2$'s are unconstrained. **It** is important to note that σ_t^2 is different from $\sigma_{t,t}$ the t -th diagonal entry of Σ ; they are the same only when Σ is diagonal in which case T is the identity matrix. Thus, in a sense the matrix T or its nonredundant entries capture the dependence in Σ and D captures the degree of predictability of a measurement using its past.

Motivated by the similarity of the roles of ϕ_{tj} 's and σ_t^2 's to those appearing in the autoregressive models in time series analysis, we refer to these new parameters as the generalized autoregressive parameters (GARP) and the innovation variances (IV), respectively. These terminologies further suggest that ϕ_{tj} 's are expected to be small when the lag $t-j$ is large. Thus, sensible plots of the entries of T and D should reveal useful information about the extent of dependence and predictability of the successive entries of the underlying vector of repeated measurements Y . We refer to the two plots introduced below as the (population) regressograms (plural) of Σ . The first graph called the *regressogram* (singular) is a plot of the same-lag regression coefficients against their

lags, i.e. we plot GARPs $\{\phi_{t+k,t}\}$ or the subdiagonals of T versus the lags $k = 1, \dots, t - 1$. The second graph called *innovariogram* following Zimmerman and Núñez Antón (2009, p.106) plots the log IV versus the times $t = 1, \dots, n$.

For a given data set, the sample regressograms can be constructed using the components (\hat{T}, \hat{D}) of the modified Cholesky decomposition of the sample covariance matrix. Then, similar to the uses of scatterplots in regression analysis, correlograms in time series analysis and variograms in spatial data analysis, the sample regressograms can be used in search of familiar and recognizable patterns for the pair (\hat{T}, \hat{D}) which can be captured with as few parameters as possible by writing linear models for these new parameters using covariates as we show in the next section.

It is worth noting that an important consequence of using the modified Cholesky decomposition for modeling the covariance matrix of a longitudinal data is that the positive-definiteness of the estimated covariance matrix given by

$$\hat{\Sigma} = \hat{T}^{-1} \hat{D} \hat{T}'^{-1}$$

is *guaranteed*. In fact, many recent alternative methods of estimating covariance matrices cannot guarantee the positive-definiteness of the estimated covariance matrix.

4. THE REGRESSOGRAMS AND MISSINGNESS IN THE FFM DATA

In this section we construct and display the regressograms of the FFM data corresponding to the three methods of handling missing data, namely available-data, complete-case and LVCF. By scanning the sample regressograms we look for familiar and parsimonious patterns for the pair (\hat{T}, \hat{D}) obtained from the sample covariance matrix of the respective data. It turns out that in spite of the apparent vast statistical differences that exist among the three ways of handling missing data, their sample regressograms suggest a common class of cubic functions of lag and time for components of the pair (\hat{T}, \hat{D}) in all three cases. This approach leads to certain structured covariance matrices for the data whose entries are polynomial functions of lag and time.

As far as parsimony is concerned, we note that for the FFM data with eleven repeated measurements on each cohort, fitting the saturated mean and covariance matrix requires $11 + 66 = 77$ possibly distinct parameters. However, our data-driven procedure **for** formulating a quadratic model for the mean trend and cubic models for the (\hat{T}, \hat{D}) manages to model them using only $3 + 4 + 4 = 11$ unconstrained parameters which is a considerable gain in terms of parsimony and ease of estimation. Next, we show the usefulness of the profile plot and regressograms in formulating models for the mean-covariance of longitudinal data.

4.1 Available-Data Analysis

The profile plot of the available-data in Figure 1 (a) indicates that the number of dead flies generally increases with time. To capture the overall increasing patterns of the the sample mean (indicated by the solid line), we propose a quadratic function of time as a model for the mean. The sample covariance matrix of the available-data and the corresponding GARPs and IVs were computed and reported in Table 4. The corresponding regressograms in Figure 3 (a)-(b) suggest cubic models for the sample GARPs and log IVs.

In summary, the profile plot and the regressograms suggest the following simple models for the mean trend and components of the modified Cholesky decomposition of the sample covariance matrix of the available-data in the FFM data:

$$\begin{aligned}\hat{\mu}_t &= \beta_0 + \beta_1 t + \beta_2 t^2 + \epsilon_t, \\ \log \hat{\sigma}_t^2 &= \lambda_1 + \lambda_2 t + \lambda_3 t^2 + \lambda_4 t^3 + \epsilon_{t,v}, \\ \hat{\phi}_{tj} &= \gamma_1 + \gamma_2(t-j) + \gamma_3(t-j)^2 + \gamma_4(t-j)^3 + \epsilon_{t,d},\end{aligned}\tag{2}$$

for $t = 1, \dots, 11$ and $j = 1, \dots, t-1$. The errors terms in these models are assumed to have means zero and unknown finite variances. Once models like (2) are formulated, their parameters are estimated using the ordinary least squares (OLS) or other methods like the maximum likelihood estimation.

Although our analysis of the FFM data has led to the specific model in (2), in general, model formulation based on profile plots and regressograms may lead to the following generalized linear models. For $t = 1, \dots, n$ and $j = 1, \dots, t-1$,

$$\mu_t = x_t' \beta, \quad \log \sigma_t^2 = z_t' \lambda, \quad \phi_{tj} = z_{tj}' \gamma,\tag{3}$$

where x_t , z_t , and z_{tj} are $p \times 1$, $q \times 1$, $d \times 1$ vectors of known covariates, with associated parameters $\beta = (\beta_1, \dots, \beta_p)'$, $\lambda = (\lambda_1, \dots, \lambda_q)'$, and $\gamma = (\gamma_1, \dots, \gamma_d)'$. In the estimation stage, it is helpful to make explicit the relation between the vectors x_t , z_t , and z_{tj} in (3) and the proposed models in (2) for the FFM data by writing the design matrices. The design matrix X for the mean model μ is an 11×3 matrix whose t^{th} row is $(1, t, t^2)$, for $t = 1, \dots, 11$. Likewise, the design matrix Z_λ for the log IV model is an 11×4 matrix whose t^{th} row is $(1, t, t^2, t^3)$. Construction of Z_γ , the design matrix for the GARP model, is a bit more complicated as it depends on how the “response” $\hat{\phi}_{tj}$ are arranged in a long vector. Here, we stack them according to the subdiagonals of T with increasing lags $t-j$. Thus, for the FFM data or model (2), Z_γ is a 55×4 design matrix with its first 10 rows corresponding to the lag $t-j = 1$; **that is the first first 10 rows have**

$z'_{tj} = \{1, t - j, (t - j)^2, (t - j)^3\} = (1, 1, 1, 1)$. The subsequent 9 rows correspond to lag 2 GARPs **and have** $z'_{tj} = (1, 2, 4, 8)$; the 8 rows correspond to lag 3 GARPs **and have** $z'_{tj} = (1, 3, 9, 27)$, and so on. Preliminary estimates of these parameters obtained using the OLS method are reported in Table 5.

Model selection techniques for model (3) within such class of polynomials based on BIC in Pan and MacKenzie (2003) can be used to identify the optimum degrees of the polynomials. The OLS estimates of the components of the covariance matrix reported here can be used as initial values for the more sophisticated maximum likelihood estimation (MLE). More precisely, representing all parameters succinctly as $\theta = (\beta', \lambda', \gamma)'$, we can resort to the MLE procedure developed in Pourahmadi (2000). Having assumed that the observations are independent multivariate normals, the algorithm relies on the results of standard likelihood methods for multivariate normal data and replacing Σ with its (T, D) -pair in the normal likelihood.

Table 4: Sample GARPs (below the main diagonals), fitted GARPs through OLS (above main diagonal), sample IVs (last row) and fitted IVs (along the main diagonal) for the FFM available-data.

Time	1	2	3	4	5	6	7	8	9	10	11
1	0.589	0.510	0.191	0.006	-0.081	-0.106	-0.105	-0.114	-0.169	-0.306	-0.561
2	0.722	0.845	0.510	0.191	0.006	-0.081	-0.106	-0.105	-0.114	-0.169	-0.306
3	0.331	0.728	0.978	0.510	0.191	0.006	-0.081	-0.106	-0.105	-0.114	-0.169
4	0.059	0.258	0.804	0.951	0.510	0.191	0.006	-0.081	-0.106	-0.105	-0.114
5	-0.014	0.178	-0.160	0.693	0.811	0.510	0.191	0.006	-0.081	-0.106	-0.105
6	0.165	0.016	0.039	-0.253	0.808	0.631	0.510	0.191	0.006	-0.081	-0.106
7	0.047	0.069	-0.088	-0.040	-0.004	0.577	0.469	0.510	0.191	0.006	-0.081
8	-0.141	0.028	0.093	-0.089	-0.020	-0.007	0.564	0.346	0.510	0.191	0.006
9	-0.007	0.049	-0.080	0.085	0.007	-0.213	0.153	0.447	0.264	0.510	0.191
10	-0.021	0.171	-0.301	0.248	-0.140	-0.039	0.120	0.074	0.261	0.219	0.510
11	0.149	-0.111	-0.298	0.299	-0.095	0.085	-0.063	0.183	0.089	0.185	0.204
	0.701	0.719	0.763	0.995	0.998	0.713	0.518	0.329	0.271	0.197	0.270

4.2 Complete-Case Analysis

The complete-case analysis or analysis of completers data consists of using only the 23 cohorts in the FFM data where all 11 measurements were observed. Small sample sizes like 23 with 11 repeated measures are not uncommon in longitudinal studies in biological sciences (Daniels and Hogan 2008).

The profile plot of the completers in Figure 1 (b) shares a similar pattern with that of the available-data in that the number of dead flies generally increases with time. Interestingly, the regressograms of the completers in Figure 4 (a)-(b) also suggests cubic models for the sample GARPs and log IVs as did the regressograms of the available-data. In summary, the profile plot and the regressograms for the completers suggests the same class of polynomial models (2) for the mean and components of the modified Cholesky decomposition of the covariance matrix.

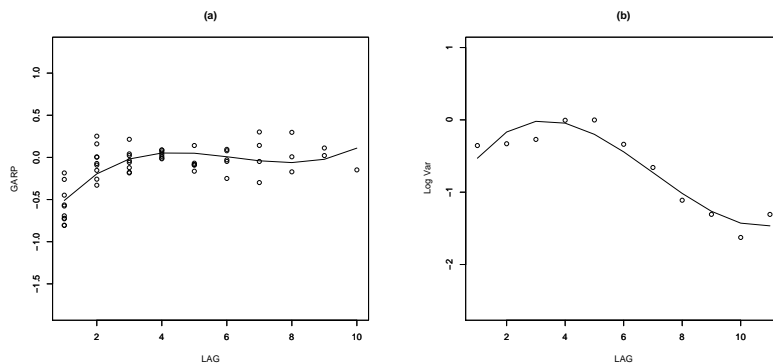


Figure 3: FFM Available-Data: (a) Sample Regressogram, (b) Sample Innovariogram.

The OLS estimates and their standard errors are given in Table 5 where all the coefficients are significant. The solid lines in the regressograms of Figure 4 (a)-(b) depict the fitted curves obtained from the OLS estimates. The closeness of the curve to the data points shows that the models **fit** reasonably well.

The PRISM (not shown) of the completers exhibits positive correlations along the main diagonal which increases as we move downwards along the diagonals, implying that assuming stationarity for the covariance structure might not be valid. For lag one and other higher lags the PRISM depicts clustering but with a more random behavior as we move down each diagonal.

4.3 Last Value Carried Forward

Next, we apply LVCF to the FFM data where there is the unusual feature that for some cohorts their baseline or the first few observations are also missing. We handle this issue by applying the

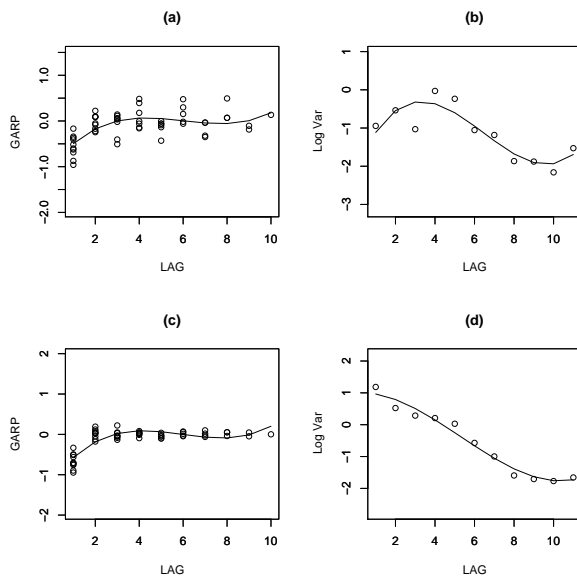


Figure 4: Sample Regressogram and Innovariogram for completers in (a) and (b). Sample Regressogram and Innovariogram for LVCF-imputed in (c) and (d).

idea of LVCF in the opposite time direction to the affected segment of the data. The profile plot given in Figure 1 (c) further exemplifies the concern for bias incurred due to LVCF. It is rather interesting to note that in spite of the bias, the profile plot suggests that a quadratic model for the mean might be adequate. An inspection of the regressograms of the completed data in Figure 4 (c)-(d) suggests cubic models for the GARPs and log IVs as in (2). The OLS estimates of these parameters are given in Table 5.

Added section below. –TG

4.4 Summary of Models fitted for AD, CC, and LVCF

Although the mean-covariance models for AD, CC, and LVCF and parameter estimates were similar, the precisions of the parameter estimates are quite different. The parameter estimates for the completers have the most variability, which could be due to its small sample size (only 23 cohorts were involved). It is rather surprising that the estimation results for the completers and LVCF are similar to each other. A particularly disturbing feature of the LVCF results is that the sign of λ_1 is positive while it is negative for the other two methods of handling missing data.

Table 5: OLS estimates along with their standard errors in parentheses for the parameters of the mean β , innovation variances λ , and dependence γ for AD, CC and LVCF-imputed.

Method	j	1	2	3	4
AD	β	-4.873 (0.120)	0.745 (0.046)	-0.029 (0.004)	
	λ	-1.150 (0.335)	0.763 (0.231)	-0.150 (0.044)	0.007 (0.002)
	γ	0.999 (0.130)	-0.586 (0.110)	0.103 (0.025)	-0.006 (0.002)
CC	β	-4.819 (0.210)	0.475 (0.081)	-0.0034 (0.007)	
	λ	-2.111 (0.645)	1.232 (0.445)	-0.250 (0.084)	0.013 (0.005)
	γ	1.003 (0.173)	-0.613 (0.146)	0.111 (0.034)	-0.006 (0.002)
LVCF	β	-4.037 (0.139)	0.463 (0.053)	-0.006 (0.004)	
	λ	0.997 (0.378)	0.050 (0.261)	-0.087 (0.049)	0.006 (0.003)
	γ	1.196 (0.110)	-0.752 (0.093)	0.139 (0.021)	-0.008 (0.001)

4.5 An AR(1) Model for the Mortality Rate

Recall that the second objective of the study leading to the FFM data was to find a relation between the mortality at any given age with the mortality at a previous age. From model (1) and the fact that the sample GARPs in the first subdiagonal of Table 4 are reasonably large, the following nonstationary AR(1) type model

$$y_t - \mu_t = \phi_t(y_{t-1} - \mu_{t-1}) + \epsilon_{t,m}, \quad (4)$$

with time-varying coefficients, $\phi_t = \hat{\phi}_{t,t+1}$, for $t = 1, \dots, 10$, seems adequate. Figure 5 (d) gives a plot of these coefficients, and shows an obvious decreasing trend in the coefficients of model (4) for $t = 1, \dots, 10$ except at $t = 3$ and $t = 5$. Furthermore, these coefficients vary significantly over time where the largest value is 0.8 versus the smallest one close to 0.1.

5. IMPUTATION METHODS

We now consider more general and flexible imputation methods than the LVCF for handling missing data. The goal is to complete the data by filling-in or imputing the missing responses with “reasonable” values. Then, the data-analyst can employ the standard analysis methods available for the complete data. A missing value can be imputed once or more than once giving rise to

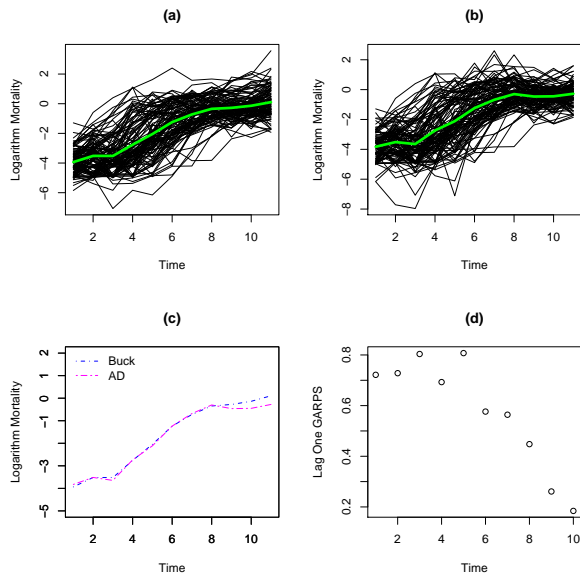


Figure 5: Profile plots: (a) Imputed using CC; (b) Imputed using AD sample mean-covariance; (c) Mean profiles; (d) Lag one sample GARPs.

single- and multiple-imputation methods. A key advantage of a multiple-imputation method is its ability to assess the imputation *uncertainty*. For the sake of brevity, we confine our attention to regression imputation and stochastic regression imputation methods using aspects of the predictive (conditional) distribution of the missing values given the observed values.

5.1 Regression Imputation

A plausible method for handling missing data that goes far beyond the LVCF is based on the idea of linear regression where the missing values are imputed by their linear least-squares predictions using the observed components as the predictors (Little and Rubin 2002).

More precisely, following our earlier convention for Y partitioned into its observed and missing components, $Y = (Y_{\text{obs}}, Y_{\text{miss}})'$, we assume Y has a multivariate normal distribution with (initial) mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$. Then, it follows from the standard results that the conditional distribution of $Y_{\text{miss}|\text{obs}}$ also has a multivariate normal distribution with estimated (initial) mean vector $\hat{\mu}_{\text{miss}|\text{obs}}$ and covariance matrix $\hat{\Sigma}_{\text{miss}|\text{obs}}$ where

$$\begin{aligned}\hat{\mu}_{\text{miss}|\text{obs}} &= \hat{\mu}_{\text{miss}} + \hat{\Sigma}_{\text{miss,obs}} \hat{\Sigma}_{\text{obs,obs}}^{-1} (Y_{\text{obs}} - \hat{\mu}_{\text{obs}}), \\ \hat{\Sigma}_{\text{miss}|\text{obs}} &= \hat{\Sigma}_{\text{miss,miss}} - \hat{\Sigma}_{\text{miss,obs}} \hat{\Sigma}_{\text{obs,obs}}^{-1} \hat{\Sigma}_{\text{obs,miss}}.\end{aligned}$$

There are several natural choices for $(\hat{\mu}, \hat{\Sigma})$ (Little and Rubin 2002, pp. 225). First, the use of

the completers data to estimate the population mean and covariance matrix has a longer history and leads to the so-called Buck (1960) method of imputing the missing data (Little and Rubin 2002, pp. 63). Second, for the purpose of comparison here, we also use the available-data (AD) sample mean and covariance matrix, provided that the latter is positive semi-definite. Under both choices for $(\hat{\mu}, \hat{\Sigma})$, the regression imputations were performed for each individual with at least one missing value by replacing missing values by their predicted values using $\hat{\mu}_{\text{miss}|\text{obs}}$ and covariance $\hat{\Sigma}_{\text{miss}|\text{obs}}$ from the above formulas. The profile plots in Figure 5 (a)-(b) and the regressograms in Figure 6 of the singly imputed data suggest quadratic models for the mean and cubic models for the GARPs and log IVs, analogous to the models in (2). The parameters estimated using the OLS procedure and their standard errors given in Table 6 are generally in agreement with those in Table 5.

A disadvantage of regression imputation is that it puts too much emphasis on the center of the conditional distribution of the missing values given the observed data. A way to mitigate this problem is to instead simulate values from the conditional distribution of the missing values given the observed. This idea relates neatly to the framework of stochastic regression or **multiple-imputation (MI)** discussed below, and allows imputed values coming from a wide range and hence assessing the uncertainty of MI becomes more realistic.

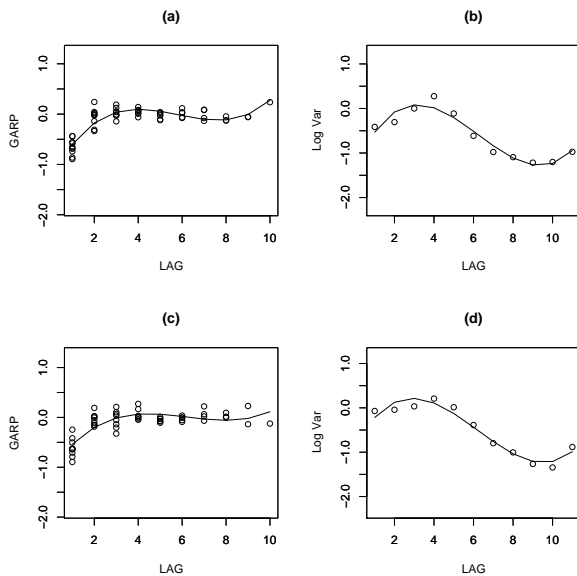


Figure 6: Sample Regressogram and Innovariogram for imputed data using CC in (a) and (b). Sample Regressogram and Innovariogram and imputed data using AD sample mean-covariance (c) and (d).

5.2 Stochastic Regression: Multiple Imputation (MI)

A possible improvement over single imputation is multiple imputation where essentially an imputation procedure is repeated M times, say $5 \leq M \leq 10$ (Fitzmaurice et al. 2004, Sec. 14.5). At each repetition, the parameter estimates and their standard errors are computed for the completed data, and an appropriate average is ultimately reported. By repeating the imputation procedure multiple times, we are able to account for the uncertainty of replacing an unobserved response with an imputed one.

More precisely, if interest lies in estimating the parameter α , at each iteration k , for $k = 1, \dots, M$, a parameter estimate $\hat{\alpha}^{(k)}$ and estimated covariance $\widehat{\text{cov}}(\hat{\alpha}^{(k)})$ are obtained. The end result of the multiple imputation is the averaged parameter estimate

$$\bar{\alpha} = \frac{1}{M} \sum_{k=1}^M \hat{\alpha}^{(k)},$$

along with its estimated covariance

$$\widehat{\text{cov}}(\bar{\alpha}) = \frac{1}{M} \sum_{k=1}^M \widehat{\text{cov}}(\hat{\alpha}^{(k)}) + (1 + 1/M) \frac{1}{M-1} \sum_{k=1}^M \left(\hat{\alpha}^{(k)} - \bar{\alpha} \right) \left(\hat{\alpha}^{(k)} - \bar{\alpha} \right)'.$$

The first term in $\widehat{\text{cov}}(\bar{\alpha})$ accounts for the within-imputation variance and the second accounts for the between-imputation variance (Fitzmaurice et al. 2004, Sec. 14.5). Together, they account for the total variability induced by imputation.

We applied the stochastic regression imputation to the FFM data using the complete-case and AD-based sample mean and covariance matrix, with $M = 5$ repetitions. Results in Table 6 show that the MI procedure assigns higher variability to the estimated parameters than the single-imputation regression method.

6. DISCUSSION

We have shown that the regressograms are powerful graphical tools for suggesting parametric models for the covariance matrix of (in)complete longitudinal data. They reveal nicely the interplay between different methods of handling the missing data and formulation of models for the dependence structure. For the FFM data, the regressograms of the three naive methods of handling missing data, viz. completers, available-data and LVCF, led us to the same class of cubic models for the components of the modified Cholesky decomposition of the corresponding covariance matrices. The same phenomenon was observed for completing the data using single- and multiple-imputation based on explicit multivariate normal distributions for the data and relying on the completers and

AD sample mean and covariance as initial values. At least for this data, the regressograms are not sensitive to the choice of methods for handling the missing data. Also, these methods lead to similar parameter estimates, but with different precisions. The parameter estimates from the completers had the greatest variability which is a possible consequence of its small sample size.

Throughout, our focus has been on model-formulation for the covariance matrix, and not so much on model estimation and diagnostics. We posited a quadratic model for the mean trend, so as to simplify the computations and focus on the impact of methods of handling missing data. However, a careful reexamination of the profile plots in Figures 1 and 5, reveals an “S”-shape which suggests fitting a logistic function for the mean as a possible alternative. This paper provides the necessary initial values and ingredients for a likelihood-based approach to dealing with missing data and the use of the EM (expectation-maximization) algorithm (Little and Rubin 2002, Chap. 8).

Table 6: OLS estimates along with their standard errors for the parameters of the mean β , innovation variances λ , and dependence γ for the single- and multiple-imputation.

Method	j	1	2	3	4
CC	β	-5.058 (0.121)	0.750 (0.046)	-0.024 (0.004)	
	λ	-1.066 (0.316)	0.766 (0.218)	-0.163 (0.041)	0.009 (0.002)
	γ	1.260 (0.088)	-0.795 (0.075)	0.149 (0.017)	-0.008 (0.001)
AD	β	-4.992 (0.127)	0.774 (0.049)	-0.030 (0.004)	
	λ	-1.224 (0.280)	1.216 (0.193)	-0.263 (0.037)	0.014 (0.002)
	γ	0.963 (0.107)	-0.544 (0.091)	0.093 (0.021)	-0.005 (0.001)
MI CC	β	-4.992 (0.132)	0.715 (0.053)	-0.022 (0.004)	
	λ	-1.178 (0.294)	0.892 (0.209)	-0.192 (0.040)	0.010 (0.002)
	γ	1.261 (0.108)	-0.783 (0.090)	0.144 (0.020)	-0.008 (0.001)
MI AD	β	-5.019 (0.144)	0.776 (0.052)	-0.030 (0.004)	
	λ	-1.183 (0.357)	1.014 (0.251)	-0.212 (0.046)	0.011 (0.003)
	γ	1.052 (0.129)	-0.633 (0.107)	0.113 (0.024)	-0.006 (0.002)

REFERENCES

- Buck, S. (1960), “A method of estimation of missing values in multivariate data suitable for use with an electronic computer,” *Journal of the Royal Statistical Society B*, 22, 302–306.
- Daniels, M., and Hogan, J. (2008), *Missing data in longitudinal studies: strategies for Bayesian modeling and sensitivity analysis*, Boca Raton: Chapman and Hall/CRC.
- Dawson, K., Gennings, C., and Carter, W. (1997), “Two graphical techniques useful in detecting correlation structure in repeated measures data,” *The American Statistician*, 51, 275–283.
- Fitzmaurice, G., Laird, N., and Ware, J. (2004), *Applied Longitudinal Analysis*, New York: Wiley.
- Kenward, M. (1998), “Selection models for repeated measurements with non-random dropout: an illustration of sensitivity,” *Statistics in Medicine*, pp. 2723–2732.
- Little, R., and Rubin, D. (2002), *Statistical Analysis with Missing Data*, 2nd edn, New York: Wiley.
- Pan, J., and MacKenzie, G. (2003), “On modeling mean-covariance structures in longitudinal studies,” *Biometrika*, 90, 239–244.
- Pourahmadi, M. (2000), “Maximum likelihood estimation of generalized linear models for multivariate normal covariance matrix,” *Biometrika*, 87, 677–690.
- Pourahmadi, M. (2002), “Graphical diagnostics for modeling unstructured covariance matrices,” *International Statistical Review*, 70(3), 395–417.
- Rubin, D. (1976), “Inference and missing data,” *Biometrika*, pp. 581–592.
- Zimmerman, D. (2000), “Viewing the Correlation Structure of Longitudinal Data Through a PRISM,” *The American Statistician*, 54(4).
- Zimmerman, D., and Núñez Antón, V. (2009), *Antedependence models for longitudinal data*, Florida: CRC Press.