20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

Integrative biostatistical approach shows better performance in missing data analysis using longitudinal study of pediatric head trauma

C. Yoo [a], A. Cheeti [b], D. Brooten [c], and J. Youngblut [c]

[a] Department of Biostatistics, [b] Department of Computer Science, [c] Nursing and Health Science, Florida International University, Miami, FL, USA
Email: cyoo@fiu.edu

Abstract: Missing data are a prevailing problem especially in longitudinal studies. A case of a variable is considered missing if the value of the case of the variable for the case is not observed, usually in random. There are many different imputation methods to use for missing data analysis. However, noticing that longitudinal studies measures variables over time, time is a key aspect that we can utilize in missing data analysis. Thus, we propose two Markov Bayesian imputation methods that incorporate time into the process. We propose imputation method based on first degree Markov – the model that impose causal relationship only between immediately adjacent time points – model and second degree Markov – the model that impose causal relationship between immediately adjacent time points and two time points before or after -- model. We also use Bayesian networks, a widely used probabilistic graphical method to express relationships (including that caused by time) among the variables, to implement those Markov imputation models (we call them Integrative Markov Bayesian Method (IMBM)).

To compare the IMBM imputation performance, we compare them with widely used missing data analysis methods – simple averaging and Expected Maximization (EM) – using data collected from a longitudinal study of child and family functioning after pediatric head trauma. We selected group of three variables that were measured in different time points from the original dataset from the longitudinal study. We then remove all the missing cases of the three variables, resulting in a complete dataset with no missing data.

We then randomly selected different portions (20%, 30%, and 40%) of the complete dataset to be randomly missing and report how each missing data analysis method performs in predicting the selected missing cases. We show that even with high (≥ 30%) missing rate for each variable, selecting proper subset of variables and using the relationships among them – that they were measured in different time points – integrative Markov Bayesian statistical methods better estimate the distribution of the missing data that are missing at random. Especially, IMBM with first degree Markov model performed surprising well, resulting in a method that could be easily scaled up to be used in imputation methods for a very large longitudinal studies.

Keywords: *Missing Data Imputation, Expected Maximization, Bayesian Markov Models*

1. INTRODUCTION

Many longitudinal studies end up with missing data due to various of reasons, e.g., participations drop out of the study or data collector mistakes. A case of a variable is considered missing if the value of the case of the variable for the case is not observed, usually in random. Many missing data imputation methods have been suggested from simple averaging to complicated methods such as Expected Maximization (EM) (Dempster, et al., 1977) and Gibbs Sampling (Geman and Geman, 1984).

In this article, using data collected from a longitudinal study of child and family functioning after pediatric head trauma (Youngblut and Brooten, 2008), we compare different missing data analysis methods such as filling missing values with the sample mean of the variable (simple averaging), EM, and a method denoted as Integrative Markov Bayesian Method (IMBM) that uses experts knowledge (here we only model the expert knowledge of which variable precedes in time). We have used causal Bayesian networks (Charniak, 1991; Pearl and Verma, 1987) to implement EM and IMBM and compared their performance using the dataset from a longitudinal study of child and family functioning after pediatric head trauma. Simple averaging method performance served as a baseline performance. Both EM and IMBM showed better performance than the baseline method, moreover, causal Bayesian networks with simpler IMBM performed better than EM. This suggests merely incorporating simple objective knowledge shows promising ways in performing missing data imputations in longitudinal studies.

2. METHODS

A causal Bayesian network (or *causal network* for short) is a Bayesian network in which each arc is interpreted as a direct causal influence between a parent node (variable) and a child node, relative to the other nodes in the network (Pearl, 1988). Figure 1 illustrates the structure of a hypothetical causal Bayesian network structure that contains five nodes. The probabilities associated with this causal network structure are not shown.

The causal network structure in Figure 1 indicates, for example, that *Parent Characteristics* can influence *Parent Grief*, which in turn can influence Family *Functioning*.

A node is independent of its non-descendants (i.e., non-effects) given its parents (i.e., its direct causes).

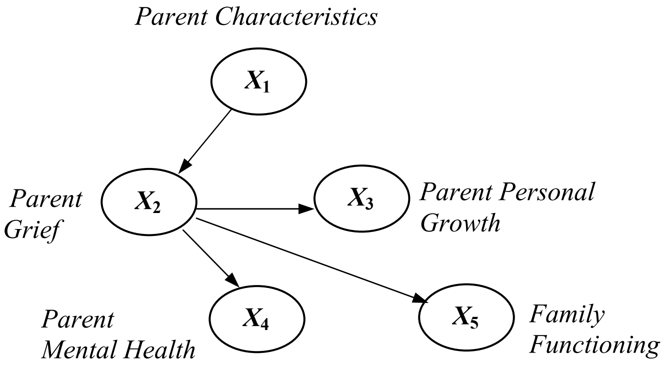


Figure 1. A causal Bayesian network that represents an expert's knowledge.

The causal Markov condition permits the joint distribution of the *n* variables in a causal Bayesian network to be factored as follows (Pearl, 1988):

$$P(x_1, x_2, ..., x_n \mid K) = \prod_{i=1}^{n} P(x_i \mid \pi_i, K) \qquad (1)$$

where x_i denotes a state of variable X_i, π_i denotes a joint state of the parents of X_i, and *K* denotes background knowledge that represents expert's knowledge.

Among many variables to consider from the longitudinal study of child and family functioning after pediatric head trauma dataset, in this study, we have concentrate on Mother's Distress (dstrm) throughout different time points. Mother's Distress had five different time points, namely, dstrm1, dstrm3, dstrm4, dstrm5, and dstrm6. However we have concentrated on variables dstrm3, dstrm4, and dstrm5 due to the fact they had the least missing cases when the variables were combined together for the missing data analysis. Initially, the missing case dataset has 187 cases out of which 90 missing cases under dstrm3, 112 missing cases under dstrm4, 107 missing cases under dstrm5. To achieve a complete dataset all the three variables (dstrm3, dstrm4, and dstrm5), we have removed cases that contain at least one missing values in any of the three different time points to achieve a complete dataset with three variables. At the end, we have a complete master dataset (we will refer this as MASTER DATA) with three variables and 41 cases. All the datasets were discretized in a hierarchical manner. That is, using the range of values under each variable, discretization is performed in the following way:

```
Input: N=number of records (41), K=number of desired bins (3).
Step 1: Let k denote the running number of bins, initialized to k=N (each record starts in its own cluster).
Step 2: If k=K quit, else set k=k-1 by combining the two bins whose mean value has the smallest separation.
Step 3: Repeat step 2.
```

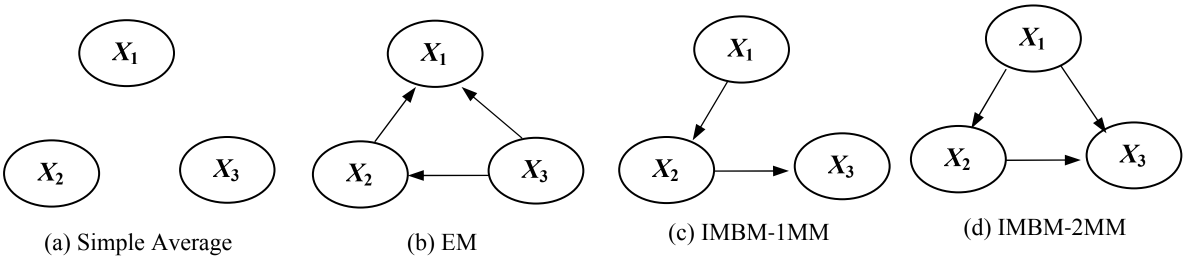
All the values under each variable are discretized to three states, i.e., state0, state1, state2. From MASTER DATA, we have randomly removed 20%, 30% 40% cases of each variables. Values under each variable were randomly selected and removed independently of other variables. Thus, from MASTER DATA, we have created three simulated missing datasets (we denote them as MISSING-20-1, MISSING-30-1, and MISSING-40-1 respectively). We repeated this process twice more to produce MISSING-20-2, MISSING-30-2, MISSING-40-2, MISSING-20-3, MISSING-30-3, and MISSING-40-3. For simplicity, for example, we denote MISSING-20-1, MISSING-20-2, and MISSING-20-3 as MISSING-20-*. We have removed the missing cases from MISSING-20-*, MISSING-30-*, and MISSING-40-* to produced nine complete datasets COMPLETE-20-*, COMPLETE-30-*, and COMPLETE-40-* respectively. It results in the nine complete simulated datasets. The cases for COMPLETE-20-*, COMPLETE-30-*, and COMPLETE-40-* datasets were 33, 29, and 25 respectively.

Each complete simulated dataset was used for developing models such as EM, IMBM with First degree Markov Model (IMBM-1MM), and IMBM with Second degree Markov Model (IMBM-2MM). For each model, we estimate the missing cases of MISSING-20-*, MISSING-30-*, and MISSING-40-* with the following expected value:

$$\widehat{m} = \sum_{i=0}^{2} i \cdot P(x_i \mid E_i)$$

where x_i represents the i-th state of the variable and E_i represent the other two variables' known states; if states are missing, we simply not include in the equation.

The Expected Maximization method (EM) is an iterative method for finding maximum likelihood estimates of parameters in statistical models, where the model depends on unobserved latent variables (Dempster, et al., 1977). The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to determine the distribution of the missing cases in the next E step. Thus the model with highest log-likelihood is used for parameter learning in EM-model and shown as in Figure 2(b).



(a) Simple Average (b) EM (c) IMBM-1MM (d) IMBM-2MM

Figure 2. Causal Bayesian network model that is used in each model. X1, X2, and X3 represents dstrm3, dstrm4, and dstrm5 respectively.

In IMBM-1MM, we use simple chain structure shown in Figure 2(c) that states mother's distress at time point 5 (dstrm5) is only related to mother's distress at time point 3 (dstrm3) only through mother's distress at time point 4 (dstrm4). In other words, if we know what mother's distress at time point 4 (dstrm4) is, knowing mother's distress at time point 3 (dstrm3) will not help predicting mother's distress at time point 5 (dstrm5). This is a simplified model than the IMBM-2MM that will be discussed next.

In IMBM-2MM process the probabilities for the next choice depend on the last two events. In a second order process each transition would refer to a pair of past outcomes. Note that to implement a second order table each past value would be a list of past values. In a second order analysis the window is three elements wide, the first two elements are the past outcomes and the third element is that past's successor. The analysis proceeds iteratively, by moving the window sequentially over the elements and counting the number of times each unique outcome happens for each unique past. The model is shown in Figure 2(d).

Since we know the actual value from MASTER DATA that is missing in the MISSING-20-*, MISSING-30-*, and MISSING-40-*, we can compare how the missing analysis methods are performing by evaluating sum of the square difference of prediction $\widehat{m}$ and actual value m:

$$\sum(\widehat{m} - m)^2$$

We calculate the sum of the square difference for Simple Average (using Figure 2(a), we simply calculate the sample mean of each variable in COMPLETE-20-*, COMPLETE-30-*, and COMPLETE-40-*, and use them as $\widehat{m}$), EM, IMBM-1MM, and IMBM-2MM.

3. RESULTS

Table 1 shows the results of each imputation method in predicting the missing data from the longitudinal data. It is clear that IMBM-1MM predicts the missing value better than other methods. These are especially encouraging results, because in the longitudinal studies we can always use the knowledge of which variable precedes in time. Also, IMBM-1MM uses the simplest Markov Model that requires small number of parameters, thus resulting in fast imputation that will easily scale up with larger number of variables.

Table 1. Sum of square difference in predicting the missing cases for each imputation method.

Dataset Methods	MISSING-20-*	MISSING-30-*	MISSING-40-*
Simple Average	12.06228	16.91125	22.98765
EM	11.31001	13.21965	21.11565
IMBM-1MM	9.276745	12.79382	18.71769
IMBM-2MM	11.39269	13.34584	19.12636

4. DISCUSSION

Since we have only used three variables, EM and IMBM-2MM had a similar performance. However, if we have more variables, we believe IMBM-2MM will perform better than EM because of the simpler model it will use than EM. Because of the amount of the missing data in this longitudinal study, using beyond three variables did not yield a complete MASTER DATA that has meaningful number of cases. We are also planning to apply these method using Gibbs Sampling (Geman and Geman, 1984) and compare their results. Since promising results from this study, we are planning to impute the missing data of the longitudinal study and compare the results with other imputation methods.

ACKNOWLEDGMENTS

This study was funded by the NIH grant NIH R01-NR04430.

REFERENCES

Charniak, E. (1991) Bayesian networks without tears, **12**, 50–63.
Dempster, A., Laird, N. and Rubin, D. (1977) Maximum likelihood from incomplete data via EM algorithm, *Journal of the Royal Statistical Society*, **B39**, 1-38.
Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721-742.
Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.
Pearl, J. and Verma, T.S. (1987) The logic of representing dependencies by directed graphs. *Proceedings of AAAI*. Morgan Kaufmann, Seattle, WA, 374–379.
Youngblut, J.M. and Brooten, D. (2008) Mother's mental health, mother-child relationship, and family functioning 3 months after a preschooler's head injury, *Journal of Head Trauma Rehabilitation*, **23**, 92-102.