Sādhanā Vol. 41, No. 8, August 2016, pp. 825–836
DOI 10.1007/s12046-016-0526-3

BAGEL: A non-ignorable missing value estimation method for mixed attribute datasets

R DEVI PRIYA[1,*], S KUPPUSWAMI[2] and R SIVARAJ[3]

[1]Department of Information Technology, Kongu Engineering College, Erode, Tamil Nadu 638 052, India
[2]Department of Computer Science and Engineering, Kongu Engineering College, Erode, Tamil Nadu 638 052, India
[3]Department of Computer Science and Engineering, Velalar College of Engineering and Technology, Erode, Tamil Nadu 638 012, India
e-mail: scrpriya@gmail.com; skuppu@gmail.com; rsivarajcse@gmail.com

MS received 19 October 2015; revised 23 February 2016; accepted 12 March 2016

Abstract. Surveys are mainly conducted to obtain valuable information on some criteria from a specified population. But, the survey results often become biased due to non-response of the subjects under study for highly significant attributes. Such non-ignorable missingness need to be treated and the actual values should be retrieved. Many methods have already been proposed for handling missing values in either discrete or continuous attributes. But, there exists a large gap in handling non-ignorable missing values in datasets with mixed attributes. With the intent of addressing this gap, this paper proposes a methodology called as Bayesian Genetic Algorithm (BAGEL) with hybridized Bayesian and Genetic Algorithm principles. In BAGEL, the initial population is generated using Bayesian model and fitness values of the chromosomes are evaluated using Bayesian principles. BAGEL is implemented in real datasets for imputing both discrete and continuous missing values and the imputation accuracy is observed. The experimental results show the superior performance of BAGEL than other standard imputation techniques. Statistical tests conducted to validate the experimental results also prove that BAGEL outperforms at all missing rates from 5% to 50%.

Keywords. Non-ignorable missing data; Bayesian techniques; genetic algorithm; Bayesian genetic algorithm; continuous attributes; discrete attributes.

1. Introduction

Surveys often suffer from the problem of dropouts and results in incomplete data for analysis. These kinds of dropouts are due to various reasons like movement of subjects to different locations, difficulty in assessing the study parameters, lack of support from the participants, etc. Whatever may be the reason, the data analysts need complete data for accurate inferences since it is one of the major challenges for statisticians [1]. In general, missing patterns are either ignorable or non-ignorable. Ignorable missingness includes Missing At Random (MAR) and Missing Completely At Random (MCAR) cases. Under MAR condition, the probability of missing values depends on other observed variables and under MCAR condition, the probability of missing values is totally independent of both observed and unobserved variables. Since in both MAR and MCAR, missingness does not depend on the missing variable, the mechanism of missingness can be ignored and are called as ignorable ones. But some incomplete data which are of vital importance cannot be ignored and should be included for analysis and are called as Non-Ignorable (NI) missing data. Under NI circumstances, missingness is often Not Missing At Random (NMAR) where the probability in which values are missing depends on the unobserved (missing) values themselves.

Non-Ignorable missingness, if untreated causes serious distorts in the estimates [2]. The negative impacts of such NMAR pattern are described in detail in [3]. Imputation of NMAR values needs strong assumptions and prior knowledge about the missing variable and its related variables under study. The original values are required since the observed data alone cannot be used for making decisions. The assumptions about non-ignorablility vary with the applications and hence models required for different applications also differ which is as a challenging task for the analysts [4]. The missing data needs to be filled with high precision by applying some statistical techniques. Even highly sophisticated techniques cannot always predict the exact missing values. But they try in their own way to estimate them more accurately as far as possible.

*For correspondence

Researchers have introduced various parametric, semi-parametric and non-parametric models to deal with NMAR patterns which occur more common in real life applications. Likelihood based and Bayesian models are highly appreciated for treating NMAR data. Since Bayesian principles are prescriptive in nature, inferences can be easily made using posterior distribution when model and prior distribution are properly given. Even in complex problem space, Bayesian model outperforms likelihood methods in estimation of extreme parameters. Furthermore, it is strong enough to make inferences based on the assumed model. Genetic Algorithms are also proved to be better in finding optimal solutions in NP hard problems. These advantageous features serve as the main motivation behind combining Bayesian principles with Genetic Algorithm in the proposed methodology, Bayesian Genetic Algorithm (BAGEL) for estimating non-ignorable NMAR data. The proposed BAGEL is tested on different real datasets in estimating missing values of both continuous and discrete attributes. The results from BAGEL are compared with that of the existing techniques and it is observed that BAGEL outperforms other methods at different missing rates from 5% to 50%.

The rest of the paper is organized as follows. Section 2 discusses the literatures that have dealt with NMAR values. Section 3 introduces the proposed methodology and fitness functions that are used for continuous and discrete attributes. Section 4 shows the implementation details and analyzes the results obtained and section 5 concludes the paper highlighting the striking features of the proposed methodology.

2. Related work

In most of the missing scenarios, the data cannot be NMAR alone, but is accompanied by MAR pattern. Molenberghs and Kenward [5] have experimented on this dependency and it is observed that NMAR data primarily depends on unobserved responses and on observed ones with some low probabilities. They cannot be directly imputed by applying statistical techniques like MAR or MCAR. Models have to be created explicitly in order to incorporate complete state of the subject under study [6]. Based on outcome of that model and other observed responses, the missing values can be estimated [7]. Various models introduced by researchers for imputing NMAR data are discussed in detail in [8].

Selection and pattern mixture models are commonly followed in non-ignorable scenarios. Let dropouts and random coefficients are denoted by r and u respectively. θ represents the variables included in the outcome model and x and y represent the vector of covariates and the outcome vectors respectively. Selection models include many models like Covariate dependent dropout model, Random dropout model, Non-ignorable dropout model and Non-ignorable random coefficient dropout model which vary in the calculations. They are all based on the below factorization given in the following equation.

$$f(y, r, u|x, \theta) = f(y|u, x, \theta)f(u|x, \theta)f(r|y, u, x, \theta). \quad (1)$$

Selection models predict the missing values using either outcome variable or latent variables. When missing outcome variable is involved, the missingness is said to be outcome-dependent. They augment the complete data with missing data model. Parametric regression is commonly used as the predictor in this model. But, when the missing variables are continuous and when the model parameters are incorrectly specified, the entire estimate will be biased and hence utmost care is necessary in defining the parameters.

Pattern mixture models vary slightly from selection models in posterior estimation as given by the following factorization in Eq. (2). They also include many models and a detailed study about them is given in [9, 10].

$$f(y, r, u|x, \theta) = f(r|x, \theta)f(u|r, x, \theta)f(y|r, u, x, \theta). \quad (2)$$

When latent variables are used, the missingness is said to be random-coefficient where shared parameter models are used [11]. In the shared parameter model, observed values and the missing indicators are considered to be independent conditioned on a set of shared parameters γ.

$$f(y, r, u|x, \theta) = \int f(y|u, \gamma, x, \theta)f(r|u, \gamma, x, \theta)f(\gamma)\mathrm{d}\gamma. \quad (3)$$

Most of the literatures have relied on Maximum Likelihood (ML) methods for parameter estimation and conventional likelihood procedures for inferences [12]. Even though efficient methods are available for estimation, selection of models in complex problems is still a difficult task. Kang *et al* [13] have found that for NMAR data, ML methods are better than complete case analysis. In general, ML methods work well in NMAR environment than with MAR pattern. But, their performance degrades in complex Growth Mixture Models (GMM) with incomplete outcomes and outlier values.

Propensity score based estimations are also popular for NMAR scenarios. In order to estimate the Propensity Scores (PS), Riddles [14] have introduced a maximum likelihood based method. The calibration condition is considered as auxiliary information and using the GMM principles, the NMAR inference accuracy was improved. Jiang *et al* [15] have used propensity score adjustment for regression models with non-ignorable missing values and covariates.

Empirical likelihood methods are also commonly used in NMAR patterns. Pseudo Empirical likelihood (PEL) method is used in [16]. Zhao *et al* [17] have applied empirical likelihood principles to estimate the mean functionals with non-ignorable missing response data when the inverse probability weighted methods with and without auxiliary information are used. For NMAR values, Niu *et al* [18] have implemented empirical likelihood based method using linear regression with confidence intervals. Tang *et al*

[19] also added to this contribution by using empirical likelihood for dealing with NMAR cases. But, standard likelihood methods are computationally expensive and in order to reduce the complexity, composite marginal likelihood methods are used in [20].

Mean functions are estimated for NMAR data using a semi-parametric method in [21]. It is shown that in case of given or estimated tilting parameter, $\sqrt{}$n-consistency is derived. Wang *et al* [22] have introduced Generalized Method of Moment (GMM) and provided guidelines for identification of suitable non-parametric models for the given problem. Since non-parametric methods are more flexible with less computational complexity, they are commonly used in treating incomplete values. But, the only difficulty in their implementation for NMAR pattern is that the behavior of missing outcomes is actually not known in most of the real applications [23].

Some other methods proposed by researchers for handling non-ignorable missing outcomes are Calibration Weighting [24, 25], Generalized Linear Sample model [4], Double Sampling method [26], and Correlated random effects model [27]. Pfeffermann *et al* [28] discusses the literatures which use probability weighting to estimate these values. Liao [29] have discussed various statistical methods for NMAR pattern in quality of life data. Kim and Shao [30] provides a detailed description on the methods used for inferencing NMAR values. Lu and Zhang [31] have discussed robust growth mixture models for dealing with non-ignorable missing outcomes. Kang *et al* [13] have introduced new NMAR models for masked clinical trials with missing entries. The guidelines regarding collection of dataset with continuous non-ignorable missing values are given in [32].

Since the actual reasons for dropout are not known in real applications, subjective theory and models are used in the estimation process. Hence, sensitivity analysis is required to validate the study results of models [5]. Sensitivity analysis provides valuable information regarding the model used for non-ignorability. It is a computationally expensive process which consumes more time. In order to reduce the computational complexity in repeatedly fitting non-linear relationship of the parameters into the model, Xie *et al* [33] proposed a semi-parametric index based model which is quick and robust in adjusting standard parameter estimates of non-ignorable missingness. Selection model augments the complete data with the missing data model. Parametric regression is commonly used as the predictor in this model. When the missing variables are continuous and the model parameters are incorrectly specified, the entire estimate will be biased and hence utmost attention is needed in this context. It relaxes the linearity assumption for response probability in sensitivity analysis and avoids incorrect fitting of complicated semiparametric joint selection models. Yin and Shi [34] have also added to this research by performing simulation based sensitivity analysis.

2.1 *Optimization techniques in estimation of missing values*

Optimization techniques like Genetic Algorithms, Ant Colony Optimization, and Particle Swarm Optimization are found to be successful in finding the best solution among *n* possible solutions for a given problem and are commonly used for handling data preprocessing problems in many literatures. Even though Genetic Algorithm is the oldest among them, because of its flexibility and efficiency, even today it is one of the appreciating methods in solving combinatorial problems. In [35], GA is combined with simulated annealing for addressing ignorable missing data problems. Azadeh *et al* [36] have compared the performance of GA, PSO and Artificial Neural Networks both empirically and also statistically. It is found that GA is better than others with lower error percentage and better Pearson correlation coefficients. Duma [37] combines Genetic Algorithm with multi-layered Artificial Immune System (GA-AIS) to treat MAR, MCAR and NMAR missing patterns. Genetic Algorithms are effectively used for handling ignorable missing values in [38, 39]. For handling ignorable incomplete values in heterogeneous attributes, Lobato *et al* [40] have successfully implemented multi-objective genetic algorithm.

2.2 *Bayesian models in imputation process*

Bayesian models are capable of efficiently incorporating prior knowledge from input parameters into the missing value imputation process [41–43]. They are more preferred than simple methods like mean/mode imputation, complete case analysis, etc. Naïve Bayes (NB) uses simple conditional probability and can itself be used for imputation of missing values. Full Bayesian methods practically require tools for simulating the distribution (e.g., Monte Carlo Markov Chain and Gibbs sampler). To overcome this problem for imputing categorical data, Epifanio [44] uses Constraint Fixed Point approach for Pseudo Bayes (PB) inferential framework in which Manifest and Belief Carrier models are proposed. In [45], Approximate Bayesian Bootstrap (ABB) together with Multiple Imputation (MI) is implemented to make it suitable for NMAR. Non-Parametric Bayesian based MI is also used in [46].

The probability distribution of parameters used in the Bayesian model provides prior knowledge. The mean and variance of these distributions reflects accuracy level of the prior knowledge incorporated into the model. Bayesian methods impute the missing values as part of the overall imputation model [47]. Unlike EM algorithm which uses knowledge from the samples, Bayesian methods create and use the models for imputation which adds more flexibility to the process. Since Bayesian approach explicitly models the relationship between different levels of analysis, it is more suitable for multi-level statistical analysis. Even for

complex problems with large sample size, Bayesian methods support additional features than ML methods with almost equal asymptotical complexity [48]. They are found to be superior than covariance based methods for treating discrete missing entries. Uncertainty in small samples can be better captured using Bayesian posterior distribution than traditional ML methods. Estimates like posterior variance and credibility intervals are easily obtained from the Bayesian draws. All these factors encourage researchers to prefer Bayesian methods than standard ML methods [49].

For analyzing dropouts in latent growth models, Bayesian principle is used in [50]. Using the Bayesian principles, Mason *et al* [51] constructed a base model of interest, a submodel to estimate missing values in the covariates and then a submodel to impute the actual missing values. The models are implemented in the UK Millennium Cohort Study dataset and the sensitivity analysis also showed the robustness of the model. Enders *et al* [9] have implemented Bayesian approach for estimation of mediation effects with missing values in 'n' manifest variables with the same performance as that of ML methods. Janicki *et al* [52] have used full Bayesian model for analyzing non-ignorable categorical missing values. They showed that even when there is uncertainty about the ignorability and the corresponding model, Bayesian averaging principles can be successfully applied. Allen [53] have used Bayesian hierarchical selection model using logistic regression for treating missing values in the process of monitoring academic growth. Zhu *et al* [54] have performed Bayesian sensitivity analysis for models used in treating missing values. Moreover, the robustness of full Bayesian method is clearly explained in [2].

Many researchers have thus utilized the advantageous features of Bayesian method for NMAR data. It considers all possible values that can be substituted for missing values. Prior information that is available can be effectively used for analysis and restrictions can be easily implemented on parameters. The final accuracy of imputed solution can be improved if Bayesian methods are integrated with other optimization techniques and this idea motivates to develop Bayesian Genetic Algorithm where Bayesian method takes help from the Genetic Algorithm in evolving better solutions over generations.

3. Materials and methods

3.1 *BAGEL: The proposed methodology*

Bayesian method and Genetic Algorithms are chosen for the proposed methodology because both are simple to understand and implement. The basic Genetic Algorithm starts with encoding of the chromosome. It is an important step in any GA implementation and hence great care has to be taken to include the required parameters as genes (genotype) in the chromosome (Phenotype). As a next step, suitable chromosomes have to be seeded into the initial population using appropriate strategies. Fitness estimation is the next operation which is the brain of GA. Objective (fitness) function is defined by the user and the fitness values obtained determine the capability of chromosomes in generating new offspring in the evolution process. Based on the fitness values, better chromosomes (individuals) are selected for the crossover operation. Crossover is being done to exchange the characteristics (genes) among parents to produce offspring (children) with new characteristics. The new chromosomes are expected to be in identifying the optimal solution. Mutation is then performed with very low probability where values in the randomly selected gene positions are altered.

The pseudocode for BAGEL is given below.

Bayesian Genetic Algorithm (BAGEL)

Input: Dataset, Missing attribute and the corresponding covariates
n: Number of covariates
n_chrm: Number of chromosomes in the population
Output: Imputed value

Begin
1. Encode the chromosomes with missing attribute and n covariates
2. Initialize the population with good chromosomes
2.1 Extract better chromosomes using Bayesian model factorization given in Eq. (5)
2.2 Sort them in descending order based on their factorization values
2.3 Seed the top 'n_chrm' chromosomes into the population
3. Fitness evaluation
If the missing attribute is discrete, estimate Bayesian based fitness using Eq. (6)
If the missing attribute is continuous, estimate Bayesian based fitness using Eq. (7)
4. Genetic Operations
4.1 Select the best parents from the population using rank selection
4.2 Perform one point crossover to exchange genes among chromosomes
4.3 Perform mutation to alter some genes
5. Iterate the steps 3 and 4 until the termination condition is reached
6. Return the chromosome with highest fitness value as the solution (value to be substituted in the missing hole)
End

BAGEL differs from the basic GA procedure by applying Bayesian technique in fitness value estimation. The traditional Bayes formula given in Eq. (4) is applicable for discrete attributes.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}. \tag{4}$$

It cannot be directly applied if the attributes are continuous. The missing data may be either continuous or discrete and the covariates may also be of mixed type. To apply Bayesian technique for estimating fitness values of such mixed types, this formula needs to be redefined. For NMAR data, covariates can also be included in model creation and fitness estimation.

The steps included BAGEL are given below.

Step 1: Encoding of chromosomes

The missing attributes and the corresponding covariates need to be first identified. The covariates are selected based on chi square and correlation coefficient values for discrete and continuous attributes respectively. BAGEL chromosome contains missing attribute and n covariates required for the analysis. The structure of chromosome used in BAGEL is given below in figure 1.

Covariate 1	Covariate 2	…..	Covariate n	Missing target attribute

Figure 1. Encoding of BAGEL chromosome.

Step 2: Model creation

Consider a finite dataset $D = \{1,\ldots,N\}$ with values $\{y_i, x_i\}$ for each record $i = 1, \ldots, N$, where y stands for outcome variable and x represents the vector of covariates. The value of response variable $R_i = 1$ if the ith record is complete and 0 if value of the outcome variable in the record is missing. The subset S refers to the selected samples where $R_i = 1$ (i.e., unit $i \in D$ does not have missing values). It serves as the first filter for selecting individuals of the initial population since first part of the numerator and the denominator given in Eq. (5) works on S. Following factorization can be done for all individuals and the corresponding values are recorded.

$$f_p(y_i|x_i, R_i) = \frac{P(R_i = 1|x_i, y_i) f_D(y_i|x_i)}{P(R_i = 1|x_i)} \quad \text{for} \quad i = 1, 2, \ldots N, \tag{5}$$

where $f_p(y_i|x_i, R_i)$ is the conditional probability density function (pdf) of y_i whose $R_i = 1$ ($i \in S$). The pdf values can be obtained by using appropriate distribution which suits the dataset. Even though pdf that is being applied may vary from problem to problem, normal distribution commonly holds for many cases. In all our experiments, normal distribution is assumed. If there is only one variable under analysis, univariate normal distribution is used, bivariate distribution for two variables and multivariate distribution for multiple variables. The above formula given in Eq. (5) can be applied irrespective of the type of missing and covariate attributes. The individuals are then sorted in descending order based on their pdf values. Those with high values which top the list are seeded into the initial population. Thus, the above pdf $f_p(y_i|\, x_i,\, R_i)$ serves as the model which holds main responsibility of introducing required solutions into the population.

Step 3: Fitness estimation

Fitness values are then estimated for all chromosomes in the population using Bayesian formula. If the missing attribute is discrete, probability can be calculated based on the Bayesian formula itself given in Eq. (6) where X_{miss} refers to the missing attribute; Y represents the covariates and θ includes the model parameters. Since, the missing attribute is discrete, simple probability formula given below can better classify the records. θ and Y may contain both discrete and continuous attributes. Hence during their analysis, probability density function (pdf) is used if the involved attribute is continuous.

$$P(X_{miss}|Y, \theta) = \frac{P(Y, \theta|X_{\text{miss}}).P(X_{\text{miss}})}{P(Y, \theta)}. \tag{6}$$

If the value to be imputed is continuous, the probability density function replaces probability shown in Eq. (7).

$$f(X_{miss}|Y, \theta) = \frac{f(Y, \theta|X_{\text{miss}}).f(X_{\text{miss}})}{\int s f(s) f(Y, \theta|X = s)\mathrm{d}s}, \tag{7}$$

where S represents the set of samples in the population which are chosen from the distribution. The standard distribution is assumed in all the cases.

Step 4: Genetic operations

By applying suitable selection mechanism, parents for crossover are then selected. Suitable crossover techniques like one point, two point or uniform crossover can be implemented to generate new offspring. On the new ones generated, mutation is being performed where random or some specific genes are modified or swapped to prevent the chromosomes from premature convergence.

Step 5: Termination

The steps from fitness estimation to mutation are repeated until termination criterion is reached. The algorithm can be run for either fixed number of generations or until same values remain in the population for n consecutive generations.

4. Results and discussion

BAGEL is implemented on real datasets to evaluate its accuracy in estimation of missing values. The population size plays a major role in the estimation process. If number of chromosomes in the population is too low, the required chromosomes for analysis may be missed out. Also if it is too high, GA may run for a long time thereby increasing the computational cost. Hence, appropriate population size should be maintained. In our experiments, the accuracy of BAGEL is estimated with varied population sizes. The other genetic parameters chosen for the experiments are:

Encoding scheme:	Real value encoding
Selection mechanism:	Rank selection
Crossover mechanism:	One point crossover
Mutation mechanism:	Swap mutation
Elitism:	10%
Crossover probability:	0.90
Mutation probability:	0.10

The bias cannot be estimated without knowing original values of the missing values. Hence after introducing missing values manually, BAGEL is implemented to impute these missing values and its accuracy is then evaluated against their original values.

For discrete attributes, the classification accuracy is calculated as

$$Classification\,accuracy\,\% = \frac{\text{No of values correctly classified}}{\text{Total number of missing values}}. \quad (8)$$

As given in Eq. (9), Root Mean Square Error (RMSE) is calculated for continuous attributes to evaluate the bias percentage where x_i and y_i refers to the original and imputed values respectively for i varying from 1 to n.

$$RMSE = \frac{1}{n}\sqrt{\sum_{j=1}^{n}(x_i - y_i)^2}. \quad (9)$$

4.1 *Discrete attributes*

BAGEL is implemented on Pima Indians diabetes and adult datasets from UCI repository and a real dataset collected from students in an engineering college to evaluate its performance. The student dataset is collected from 300 students with the attributes given in table 1.

4.1a *Experimental analysis:* BAGEL is tested with population size of 40, 50, 60 and 70 and better performance is seen with population size of 70 (due to space constraints, the detailed results are not given). The results of BAGEL are compared with that of ABB [45], Bayesian based MI [46] and GA-AIS [37]. Since GA is stochastic in nature, BAGEL is executed for 30 runs and the error% from all the runs are averaged and reported. Classification error % of all these methods in imputing missing discrete values at different missing rates like 5%, 10%, 20%, 30%, 40% and 50% in adult, Pima Indians diabetes and student datasets are given in figures 2, 3 and 4 respectively. In the student dataset, assume that female students have missed out values for the gender attribute. Based on the assumption, gender of female students is manually deleted at different proportions. Hence, missingness depends upon the missing values themselves, it falls under NMAR category.

The number of generations required for convergence of BAGEL in the three datasets is given in table 2. It is observed that average of 19.39 generations is required for BAGEL convergence.

Table 1. Description of student dataset.

S. no	Attribute name	Attribute type
1	Student name	String
2	Age	Continuous
3	Gender	Discrete
4	Height	Continuous
5	Weight	Continuous
6	Family income	Continuous
7	Monthly expenses	Continuous
8	Spectacles	Discrete

Even at 50% missingness, error % of BAGEL is only around 6%. BAGEL can be applied for discrete attributes with multiple classes. It is facilitated in BAGEL by including all the classes for discrete attributes in the chromosome structure.

Table 3 shows comparison of BAGEL's classification accuracy with that of other methods. BAGEL shows large improvement in performance when compared to ABB. When compared with MI and GA-AIS, a marginal increase of 0.56% and 0.45% is seen respectively. The striking feature of BAGEL over the compared Bayesian based MI method is that BAGEL does not demand expertise from the

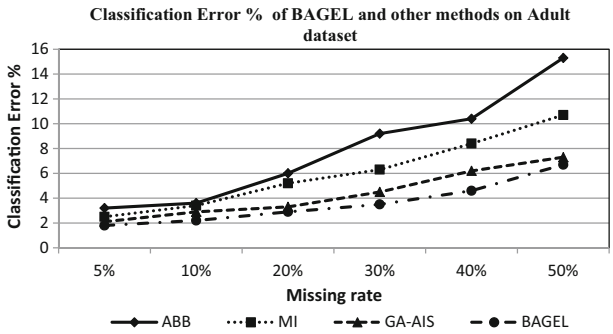

Figure 2. Classification error % of BAGEL and other methods on adult dataset.

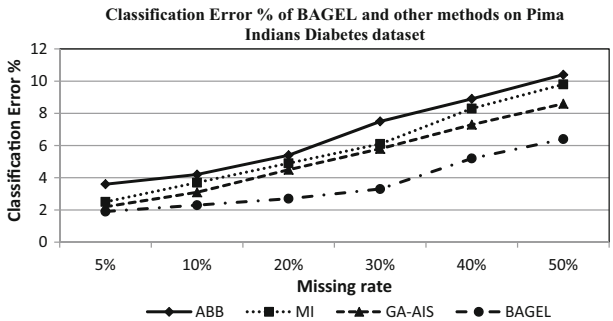

Figure 3. Classification error % of BAGEL and other methods on Pima Indians diabetes dataset.

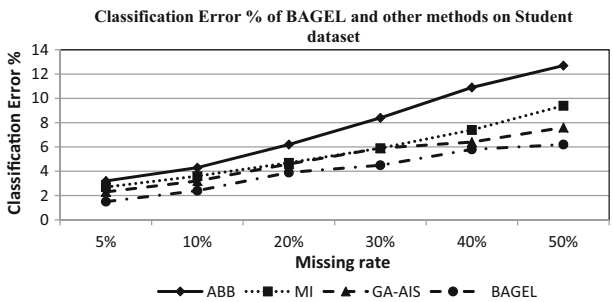

Figure 4. Classification Error % of BAGEL and other methods on student dataset.

Table 2. BAGEL convergence velocity for estimation of NMAR discrete attribute.

	Number of generations taken for convergence		
Missing rate (%)	Adult	Pima Indians diabetes	Student
5	10	9	10
10	13	12	14
20	17	17	16
30	20	22	21
40	24	26	26
50	29	31	32
Average number of generations			19.39

Table 3. Difference in performance of BAGEL and other methods for discrete attributes.

Methods	Difference in classification accuracy (%)
BAGEL ~ ABB	3.78
BAGEL ~ MI	0.56
BAGEL ~ GA-AIS	2.45

Table 4. Chi square values for adult dataset.

Missing rate (%)	γ^2 value calculated	Table value
5	1.25	3.84
10	1.56	
20	0.48	
30	0.36	
40	0.79	
50	0.21	

user as MI does. Since both GA-AIS and BAGEL use principles of GA, only slight difference is noted between their performances.

4.1b *Statistical analysis*

Chi square test is used to validate the performance of BAGEL in imputing missing NMAR discrete attributes statistically. In all the experiments, the significance level (α) is set as 0.05 and the hypotheses used are:

Null hypothesis

H0: There is no significant difference between observed and expected values

Alternate hypothesis

H1: There is a significant difference between observed and expected values

(i) Adult dataset

In this dataset, the target attribute has two classes and hence $n = 2$. The degree of freedom is thus $2 - 1 = 1$.

Table 5. Chi square values for Pima Indians diabetes dataset.

Missing rate (%)	γ^2 value calculated	Table value
5	3.7	3.84
10	2.83	
20	1.92	
30	1.984	
40	3.05	
50	2.82	

Table 6. Chi square values for student dataset.

Missing rate (%)	γ^2 value calculated	Table value
5	1.4	3.84
10	2.12	
20	0.97	
30	1.94	
40	2.35	
50	2.82	

From table 4, it is observed that at all missing rates ranging from 5% to 50%, the calculated γ^2 is less than the table value for the given degree of freedom (1). Hence the null hypothesis is accepted which states that there is no significant difference between the observed and expected values.

(ii) Pima Indians diabetes dataset

The number of classes in the target attribute is 2 in this dataset. Degree of freedom is thus $2 - 1 = 1$. From table 5, it is observed that the calculated γ^2 is less than the table value (3.84) at all missing rates. The null hypothesis is therefore accepted and it is concluded that there is no significant difference between the expected and the values imputed by BAGEL.

(iii) Student dataset

The number of classes in the target attribute (gender) is 2 (male and female). Degree of freedom is thus $2 - 1 = 1$. From table 6, it is observed that the calculated γ^2 is less than the table value (3.84) at all missing rates and the null hypothesis is therefore accepted.

4.2 *Continuous attributes*

For demonstrating performance of BAGEL in estimating missing continuous NMAR values, two UCI repository datasets (Indian liver patient, housing) and the student dataset used for discrete attribute are used.

4.2a *Experimental analysis*

BAGEL is implemented with population size of 70 and its RMSE% is compared against that of Bayesian based MI [46], EL [18] and GA-AIS [37]. The results for Indian liver

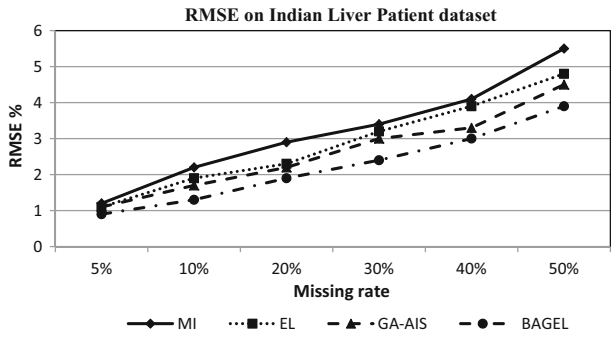

Figure 5. RMSE of BAGEL and other methods on Indian liver patient dataset.

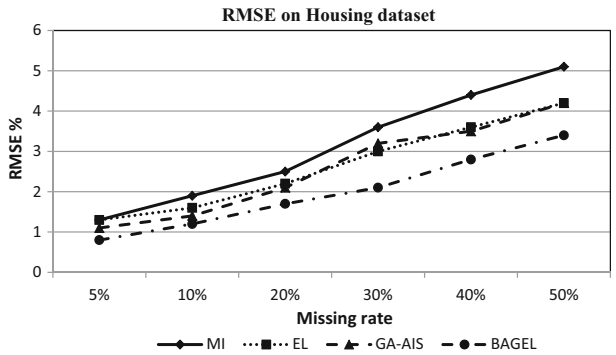

Figure 6. RMSE of BAGEL and other methods on housing dataset and student dataset.

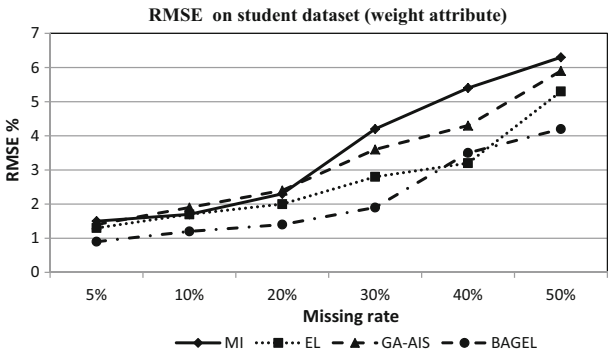

Figure 7. RMSE of BAGEL and other methods for weight attribute.

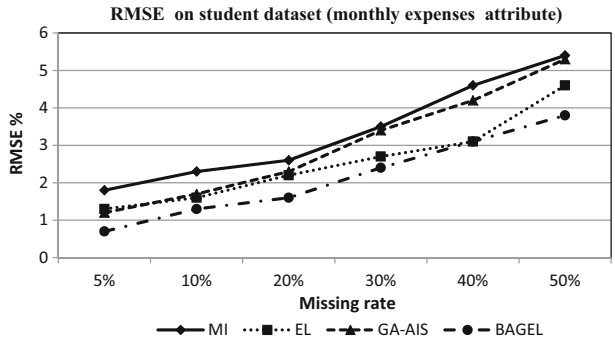

Figure 8. RMSE of BAGEL and other methods for monthly expenses attribute.

patient and housing datasets are given in figures 5 and 6 respectively.

In the dataset given in table 1, some students fail to provide details of their weight and monthly expenses. The overweight students and those who have high family income and monthly expenses do not provide their values. Since they are optional fields, some students have not given the values and hence this missingness comes under NMAR category.

Bayesian model given in Eq. (5) is used to select individuals for the initial population. For both weight and monthly expenses attribute, the RMSE% obtained from BAGEL is compared with that of existing methods.

Case 1 Weight attribute

RMSE% is calculated for the attributes weight and monthly expenses at different missing rates and at different population sizes. Population size of 70 shows better performance and hence RMSE% of BAGEL with population size of 70 is compared against that of MI, EL and GA-AIS as depicted in figure 7. Substituting mean of the available values produces more biased results. When more records are missing, the accuracy obviously decreases for all the methods. Even then, the results confirm superior performance of BAGEL with maximum of 5% of RMSE at 50% missing rate which is highly appreciable. Whenever better chromosomes are identified using fitness function, GA explores capability of that chromosome in attaining the final solution and hence its performance is improved at all missing rates.

Case 2 Monthly expenses attribute

Since population size of 70 shows better performance in monthly expenses attribute also, RMSE of BAGEL with population size of 70 is compared against that of MI, EL and GA-AIS in figure 8. At 5% missingness, BAGEL produced only 0.95% error. Similarly at all missing rates, value of RMSE is considerably lower. Even at 50% missingness, the imputation accuracy is about 95%.

For this attribute also, BAGEL shows better performance than other three methods. Even though the distribution of values for monthly expenses attribute is high, BAGEL has less RMSE% at all missing rates. Specifically below 20% missing rate, RMSE is less than 2%. This proves the combined effort made by Bayesian model, Bayesian fitness estimation and GA in imputing the missing values.

From table 7, convergence velocity of BAGEL in estimation of missing values in Indian liver patient, housing and student (weight and monthly expenses attribute) datasets can be observed. Average of 20.33 generations is

Table 7. BAGEL convergence velocity for continuous attributes.

	Number of generations taken for convergence			
Missing rate (%)	Indian liver patient	Housing	Weight	Monthly expenses
5	10	9	9	10
10	12	14	11	12
20	16	17	15	16
30	19	21	21	23
40	27	30	28	28
50	32	35	37	36
Average number of generations				20.33

Table 8. Difference in performance of BAGEL and other methods for continuous attributes.

Methods	Difference in RMSE (%)
BAGEL ~ MI	9.56
BAGEL ~ EL	2.15
BAGEL ~ GA-AIS	0.87

required for BAGEL in imputing non-ignorable missing values.

Table 8 shows the difference in RMSE% between BAGEL and other algorithms. It is found that performance of BAGEL is superior to others by imputing the missing values more accurately.

4.2b *Statistical analysis*

The correlation coefficient (r) calculated between actual and imputed values in Indian liver patient and housing datasets is given in table 9.

The standard deviation of the original student dataset without any missing values and the one which is calculated after imputing the missing values using BAGEL ($\sigma 2$) is given in table 10. In the original dataset, the value of standard deviation ($\sigma 1$) is 8.188. The standard deviation ($\sigma 2$) values in the new dataset with imputed values for missing ones are very close to $\sigma 1$. But the difference in σ decreases when the population size is increased. Hence when population size increases, the accuracy of the imputed values also improves but with some extra cost of execution time needed for processing large population. It is to be noted that at 5% missingness, the average $\sigma 2$ value that is obtained using BAGEL is 8.071 which is highly appreciable. The difference between standard deviation in original and BAGEL imputed values at 5% (8.188–8.071) is 0.117. But when the percentage of missingness increases, the difference in standard deviation values also increases due to the fact that only less number of required samples is available for analysis. For example, at 50% missingness, the deviation (8.188−7.457) is 0.731.

For the attribute monthly expenses, value of σ in the complete dataset without missing values is 482.62. The minimum and maximum values for the attribute are 150 and 750 respectively. The missing values are imputed using BAGEL and σ obtained after filling the missing values with new values in the dataset at different missing rates and population sizes are reported in table 11. The difference in values between $\sigma 1$ and $\sigma 2$ is low ($504.03 - 482.62 = 21.41$) at 5% missingness which ensures that the imputed values are very near to the original values. At 50% missingness, the deviation ($545.18 - 482.62 = 62.56$) is comparatively high due to less number of appropriate individuals available for analysis.

The correlation coefficient calculated between actual and estimated values of weight and monthly expenses attribute in the student dataset using different methods are given in table 12. For all the four methods, it is to be noted that the value of r decreases when missing rate increases. Compared to MI, EL and GA-AIS, BAGEL achieves r value close to 1. It indicates that the correlation between actual values and the values imputed by using BAGEL are highly correlated.

5. Conclusion

Non-ignorable missing values greatly affect the accuracy of analysis results and hence need to be treated with efficient techniques. Since most of the literatures have concentrated on either discrete or continuous attributes, the issue of handling heterogeneous missing values remains unfixed. This paper has addressed this issue by proposing a novel and simple method (BAGEL) for handling non-ignorable missing values in the datasets with both discrete and continuous attributes. The imputation accuracy depends on the input chromosomes fed into the initial population. Realizing this importance, BAGEL has introduced a simple Bayesian based model for seeding the initial population with good chromosomes. This in turn effectively reduces the computational time by preventing entry of unfit chromosomes into the population. The Bayesian fitness function effectively evaluates the chromosomes with simple calculations. The hybridization of Bayesian and Genetic Algorithm principles greatly reduces the computational complexity. The implementation results show that in

Table 9. Correlation coefficient of BAGEL and other methods on Indian liver patient and housing datasets.

	Indian liver patient				Housing			
Missing rate (%)	MI	EL	GA-AIS	BAGEL	MI	EL	GA-AIS	BAGEL
5	0.85	0.91	0.92	0.97	0.87	0.92	0.94	0.96
10	0.82	0.88	0.89	0.95	0.83	0.90	0.89	0.93
20	0.79	0.83	0.87	0.91	0.78	0.87	0.88	0.91
30	0.74	0.80	0.84	0.88	0.73	0.84	0.85	0.88
40	0.70	0.76	0.8	0.85	0.69	0.83	0.82	0.84
50	0.66	0.74	0.78	0.80	0.66	0.80	0.79	0.80

Table 10. σ in original vs imputed dataset for weight attribute.

		$\sigma 2$ (Imputed dataset)					
$\sigma 1$ (Original dataset)	Population size	5%	10%	20%	30%	40%	50%
8.188	40	7.995	7.791	7.497	7.319	7.278	7.224
	50	8.059	7.835	7.61	7.460	7.338	7.467
	60	8.097	7.859	7.6	7.556	7.573	7.537
	70	8.134	7.904	7.821	7.625	7.668	7.601
	Avg ($\sigma 2$)	8.071	7.847	7.632	7.561	7.464	7.457

Table 11. σ in original vs imputed dataset for monthly expenses attribute.

		$\sigma 2$ (Imputed dataset)					
$\sigma 1$ (Original dataset)	Population size	5%	10%	20%	30%	40%	50%
482.62	40	518.36	522.23	522.23	531.68	548.26	556.32
	50	509.26	519.26	526.98	529.33	537.65	554.98
	60	498.22	512.56	518.69	527.65	532.26	539.87
	70	490.26	499.56	512.26	520.37	524.23	529.56
	Avg ($\sigma 2$)	504.03	513.41	521.54	527.26	535.60	545.18

Table 12. Correlation coefficient of BAGEL and other methods on student dataset.

	Weight attribute				Monthly expenses attribute			
Missing rate (%)	MI	EL	GA-AIS	BAGEL	MI	EL	GA-AIS	BAGEL
5	0.80	0.93	0.94	0.96	0.81	0.92	0.94	0.95
10	0.78	0.88	0.92	0.94	0.77	0.88	0.90	0.92
20	0.74	0.84	0.87	0.92	0.76	0.85	0.87	0.90
30	0.69	0.80	0.83	0.89	0.70	0.82	0.86	0.87
40	0.64	0.77	0.81	0.86	0.68	0.77	0.82	0.84
50	0.60	0.75	0.79	0.82	0.60	0.75	0.79	0.81

treating both discrete and continuous attributes, BAGEL results in better performance than MI, ABB, EL and GA-AIS methods in terms of imputation accuracy. The convergence velocities of BAGEL in all the datasets at different missing rates are also found to be encouraging. The statistical tests are also conducted to validate the empirical results. Chi square values of BAGEL at missing rates from 5% to 50% show that the actual and imputed values are highly relevant. For continuous attributes too, the original and imputed values of BAGEL are highly correlated with r value close to 1. It reduces the burden for analysts in searching appropriate model suitable for their dataset. BAGEL with integrated model (Bayesian model) and optimization technique (Genetic Algorithm) serves as the unified novel approach which is suitable for handling NMAR condition. In future, Bayesian models can be integrated with other optimization algorithms to further enhance the imputation performance.

References

[1] Belin T R 2009 Missing data: What a little can do, and what researchers can do in response. *Am. J. Ophthal.* 148(6): 820–822

[2] Zhang Z and Wang L 2012 A note on the robustness of a full Bayesian method for non-ignorable missing data analysis. Brazil. *J. Prob. Stat.* 26(3): 244–264

[3] Wang S, Jiao H and Xiang Y 2013 The effect of nonignorable missing data in computerized adaptive test on item fit statistics for polytomous item response models. *Annual meeting of the National Council on Measurement in Education.* April 27–30, 2013, San Francisco, CA

[4] Pfeffermann D and Sikov N 2011 Imputation and estimation under nonignorable nonresponse in household surveys with missing covariate information. *J. Offic. Stat.* 27(2): 181–209

[5] Molenberghs G and Kenward M G 2007 *Missing data in clinical studies.* West Sussex, England: John Wiley

[6] Molenberghs G 2009 Incomplete data in clinical studies: Analysis, sensitivity and sensitivity analysis. *Drug Inform. J.* 43(4): 409–429

[7] Pfeffermann D 2011 Modelling of complex survey data: Why model? Why is it a problem? How can we approach it?. *Surv. Method* 37(2): 115–136

[8] Xie H 2010 Adjusting for nonignorable missingness when estimating generalized additive models. *Biomet. J.* 52(2): 186–200

[9] Enders C K, Fairchild A J and MacKinnon D P 2013 A Bayesian approach for estimating mediation effects with missing data. *Multivar. Behav. Res.* 48(3): 340–369

[10] Muthen B, Asparouhov T, Hunter A and Leuchter A 2011 Growth modeling with non-ignorable dropout: Alternative analyses of the STAR*D antidepressant trial. *Psychol. Methods*16(1): 16–33

[11] Feldman B J and Rabe-Hesketh S R 2012 Modeling achievement trajectories when attrition is informative. *J. Educ. Behav. Stat.* 37(6): 703–736

[12] Song W, Yao W and Xing Y 2014 Robust mixture regression model fitting by Laplace distribution. *Comput. Stat. Data Anal.* 71: 128–137

[13] Kang S, Little R J and Kaciroti N 2015 Missing not at random models for masked clinical trials with dropouts. *Clin. Trials* 12(2):139–148

[14] Riddles M K 2013 *Propensity score adjusted method for missing data*. PhD thesis, Iowa State University

[15] Jiang D, Zhao P and Tang N 2016 A propensity score adjustment method for regression models with nonignorable missing covariates. *Comput. Stat. Data Anal.* 94: 98–119

[16] Fang F, Hong Q and Shao J 2010 Empirical likelihood estimation for samples with nonignorable nonresponse. *Stat. Sinica* 20: 263–280

[17] Zhao H, Zhao P Y and Tang N S 2013 Empirical likelihood inference for mean functionals with nonignorably missing response data. *Comput. Stat. Data Anal.* 66(10): 101–116

[18] Niu C, Guo X, Xu W and Zhu L 2014 Empirical likelihood inference in linear regression with non-ignorable missing response. *Comput. Stat. Data Anal.* 79: 91–112

[19] Tang N S, Zhao P Y and Zhu HT 2014 Empirical likelihood for estimating equations with nonignorably missing data. *Stat. Sinica* 24: 723–747

[20] Varin C, Reid N and Firth D 2011 An overview of composite likelihood methods. *Stat. Sinica* 21: 5–42

[21] Kim J K and Yu C L 2011 A semiparametric estimation of mean functionals with nonignorable missing data. *J. Am. Statist. Assoc.* 106(493): 157–165

[22] Wang S, Shao J and Kim J K 2014 An instrument variable approach for identification and estimation with nonignorable nonresponse. *Stat. Sinica* 24: 1097–1116

[23] Miao W, Ding P and Geng Z 2015 Identifiability of normal and normal mixture models with nonignorable missing data. arXiv:1509.03860

[24] Kim J K 2009 Calibration estimation using empirical likelihood in survey sampling. *Stat. Sinica* 19(1): 145–157

[25] Kott P S 2009 Calibration weighting: Combining probability samples and linear prediction models. In: D Pfeffermann and C R Rao (Eds.) *Handbook of statistics 29B; Sample surveys: Inference and analysis*. Amsterdam: North Holland, 55–82

[26] Aronow P M, Gerber A S, Green D P and Kern H 2013 *Double sampling for missing outcome data in randomized experiments*. Typescript, Yale University

[27] Karl A T, Yang Y and Lohr S L 2013 A correlated random effects model for nonignorable missing data in value-added assessment of teacher effects. *J. Educ. Behav. Stat.* 38(6): 557–603

[28] Pfeffermann D and Sverchkov M 2009 Inference under informative sampling. In: D Pfeffermann and C R Rao (Eds.) *Handbook of Statistics 29B; Sample Surveys: Inference and Analysis*. Amsterdam: North Holland, 455–487

[29] Liao K 2012 *Statistical methods for non-ignorable missing data with applications to quality-of-life data*. PhD thesis, University of Pennsylvania

[30] Kim J K and Shao J 2013 *Statistical methods for handling incomplete data*. Chapman &Hall/CRC

[31] Lu Z and Zhang Z 2014 Robust growth mixture models with non-ignorable missingness: Models, estimation, selection, and application. *Comput. Stat. Data Anal.*71: 220–240

[32] Paiva T and Reiter J P 2015 *Stop or continue data collection: A nonignorable missing data approach for continuous variables*. arXiv: 2015. : 1511.02189

[33] Xie H, Qian Y and Qu L M 2011 A semiparametric approach for analyzing nonignorable missing data. *Stat. Sinica* 21: 1881–1899

[34] Yin P and Shi J Q 2015 Simulation based sensitivity analysis for non-ignorable missing data. arxiv:1501.05788

[35] Nelwamondo F V and Marwala T 2008 Techniques for handling missing data: Applications to online condition monitoring. *Int. J. Innov. Comp., Inform. Cont.* 4(6): 1507–1526

[36] Azadeh S M, Asadzadeh R, Jafari-Marandi S, Nazari-Shirkouhi G, Khoshkhou B, Talebi S and Naghavi A 2013 Optimum estimation of missing values in randomized complete block design by genetic algorithm. *Knowl. Based Syst.* 37(1): 37–47

[37] Duma M 2013 Partial imputation of unseen records to improve classification using a hybrid multi-layered artificial immune system and genetic algorithm. *Appl. Soft Comp.* 13(12): 4461–4480

[38] DeviPriya R and Kuppuswami S 2014 Drawing inferences from clinical studies with missing values using genetic algorithm. *Int. J. Bioinf. Res. Appl.* 10(6): 613–627

[39] DeviPriya R and Kuppuswami S 2015 A novel approach for imputation of missing continuous attribute values in databases using genetic algorithm. *Int. J. Inform. Tech. Manag.* 14(2/3):185–200

[40] Lobato F, Sales C, Araujo I, Tadaiesky V, Diaa L, Ramos L and Santana A 2015 Multi objective genetic algorithm for missing data imputation. *Pattern Recogn. Lett.* 68(P1): 126–131

[41] Celeux G, Forbes F, Robert C and Titterington D 2006 Deviance information criteria for missing data models. *Bayes. Anal.* 1(4): 651–674

[42] Kruschke J K, Aguinis H and Joo H 2012 The time has come: Bayesian methods for data analysis in the organizational sciences. *Organiz. Res. Methods* 15(4): 722–752

[43] Lu Z L, Zhang Z and Lubke G 2011 Bayesian inference for growth mixture models with latent class dependent missing data. *Multivar. Behav. Res*. 46(4): 567–597

[44] Epifanio G D 2006 *A Pseudo Bayes approach for non-ignorable non-response in categorical survey data*. Dip. Economi, Finanza e Stat., Technical Report, Univ. di Perugia

[45] Siddique J and Belin T R 2008 Using an approximate Bayesian bootstrap to multiply impute nonignorable missing data. *Comput. Stat. Data Anal.* 53(2): 405–415

[46] Si Y 2012 *Non-parametric Bayesian methods for multiple imputation of large scale incomplete categorical data in panel studies*. PhD thesis, Duke University

[47] Asparouhov T and Muthen B 2010 *Bayesian analysis of latent variable models using MPlus*. Version 4. http://www.statmodel.com

[48] Lunn D, Jackson C, Best N, Thomas A and Spiegelhalter D 2013 *The BUGS Book – A practical introduction to Bayesian analysis*. Boca Raton, FL: CRC Press

[49] Little R 2011 Calibrated Bayes, for statistics in general, and missing data in particular. *Stat. Sci*. 26(2): 162–174

[50] Tanaka D and Kanazawa Y 2010 Bayesian analysis of the latent growth model with dropout. *Discussion paper series,* Department of Social Systems and Management, University of Tsukuba

[51] Mason A, Richardson S, Plewis I and Best N 2012 Strategy for modelling nonrandom missing data mechanisms in observational studies using Bayesian methods. *J. Offic. Stat.* 28(2): 279–302

[52] Janicki R and Malec D 2013 A Bayesian model averaging approach to analyzing categorical data with nonignorable nonresponse. *Comput. Stat. Data Anal.* 57: 600–614

[53] Allen J 2015 A Bayesian Hierarchical selection model for academic growth with missing data. *ACT Working Paper Series,* WP-2015-04

[54] Zhu H, Ibrahim J G and Tang N 2014 Bayesian sensitivity analysis of statistical models with missing data. *Stat. Sinica* 24(2):871–896