International Journal on Future Revolution in Computer Science & Communication Engineering
Volume: 5 Issue: 5
ISSN: 2454-4248
54 – 58

Multivariate Missing Data Handling with Iterative Bayesian Additive Lasso (IBAL) Multiple Imputation in Multicore Environment on Cloud

Lavanya.K[1], L.S.S.Reddy[2], B. Eswara Reddy[3]
1Research Scholar,Department of Computer Science & Engineering,JNTUA College of Engineering,Ananthapuramu-515002.
lavanya.kk2005@gmail.com
2Professor, Department of Computer Science & Engineering, KLUniversity,Vaddeswaram-522502, Guntur(Dt.)
drlssreddy@kluniversity.in
3 Professor, Department of Computer Science & Engineering, JNTUA College of Engineering,Kalikiri-517234,Chittoor(Dt.)
eswarcsejntua@gmail.com

ABSTRACT: Dealing with high dimensional data of the form p>n for multivariate analysis of missingness is very complicated. It arises in many fields mainly in social science, economics and medical study; genome is an example for that where is to mention that samples are very less compared to study elements nothing but variables. The analysis is a combination of large covariate vectors with response and non-response effects of unknown functional form related to response variable of interest. Thus, there is a need for regularized regression models, with effect of smoothing parametric method to do this in this work combine regularization by incorporating different types of covariates. Although regularization approaches fits to framework but the computation high demands in high dimensional analysis they also rely on penalized estimation. The solution is to implement regularization in iteration based smoothing approaches to fit such analysis. The proposed algorithm called Iterative Bayesian Additive Lasso (IBAL) is compared with standard methods in medical analysis and produced unbiased results. The overall work done in multi core environment offered by Cloud Service called Microsoft Azure. The performance is estimated with benchmarks like Standard Error (SE), Mean Square Error (MSE), and Confidence Interval (CI).

Keywords: *Multiple Imputation, Regularized Regression, Additive Lasso, High Dimensional, and Multicore Environment.*

*****

1. Introduction

Dealing with high dimensional data of the form p>n for multivariate analysis of missingness is very complicated [7][8][9]. To make this process simple and flexible multivariate analysis is replaced with univariate analysis in iterative manner, where each variable conditionally imputed one by one to deal problem of multivariate missingness. By using simple stochastic algorithm producers randomly impute missing values one variable each at a time conditionally. Until the convergence is measured the variables are looped. Iterative imputation is a procedure applied to high dimensional demands separate framework in five stages. Firstly it focuses on background details which includes Imputation, Gibbs sampling, and Imputation in iterative approach. Second, described multivariate analysis of missing data in multicore environment. Third section introduces proposed work. Forth focus on implementation and results. Finally work is concluded.

2. Background

2.1. Multivariate Missing Data

Denoted complete data to be of size $n \times p$ where n to be considered samples and p to be set of variables, corresponding matrix to be mentioned $Y = (Y_1, ..., Y_p)$. From that considering j^{th} variable represented as $Y_j = \left(y_{1j}, ..., y_{nj}\right)^T$, it to be composed set of observed and missing values. The response related to values in k variables to be treated as vector r_k. The value will be either 0 or 1 for missing data and observed data.

After that defined observed and missing data corresponding to k variables to be mentioned as:

$$\mathbf{Y^{(k)}_{obs} = \{Y_{obs}, k = 1, ..., p\}}$$

$$\mathbf{Y^{(k)}_{mis} = \{Y_{mis}, k = 1, ..., p\}}.$$

The study continued on univariate variable Y^j composed with observed and imputed data , to be denoted as $Y^k = (Y^j{}_{obs}, Y^j{}_{mis})$ and also there is a need specifying complement and to be represented as $Y^{-k} = (Y^{-j}{}_{obs}, Y^{-j}{}_{mis})$, set of remaining variables after excluding Y^j. Here mentioned that complements related to missing and observed data and to mention in (1).

 ISSN: 2454-4248
54 – 58

$$Y^{-j}_{obs} = \{[Y_{obs}(l)], l=1,...,j-1, j+1,...,p\},$$
$$Y^{-j}_{mis} = \{[Y_{mis}(l)], l=1,...,j-1, j+1,...,p\}.$$

(1)

With all above considerations, denoted Y^j complete data with observed and missing values and Y^{-k} to be set of all values except Y^j.To model missing data number of mechanisms includes which to be mentioned one by one. First one, missing at random (MAR) where cause of missingness not depends on missing proportion but to be on observed data. Second one,Missing Completely at Random (MCAR)in which the proportion of missing factor seriously basis on values which are depends to missing data. Last one, Missing Not at Random (MNAR) [14], where missing values are depends on missing data and to be non-ignorable case .The mechanism of MCAR and MAR to be considered as ignorable case.

2.2. MI with Bayesian Modeling

The approach which is extremely popular and it is used in MI for estimating unknown regression coefficients to impute missing values using mechanism of Gibbs Sampling and data augmentation [11].General mechanism of MI includes for every data set, determine set estimators using predictive distribution for each variables and with that missing data to be drawn using posterior predictive distribution and finally generated M imputed data sets. To establish conditional distribution MI incorporated parametric approach called Bayesian inference in which missing data is derived by specifyinga joint distributions $P(Y|\theta)$ and with a prior $\pi(\theta)$.

The complete idea is shown in (2)

$$P(Y_{mis}|Y_{obs}) = \int \Theta P(Y_{mis}|XY_{obs},\theta)\ P(\theta|Y_{obs})\,d\theta \quad (2)$$

Where $P(\theta|X) \propto \Psi(\theta)P(Y|\theta)$. It is very difficult to perform such distribution directly to resolve the mechanism of Gibbs sampler or Markov chain Monte Carlo techniques were used to draw approximate samples. The study in this work used data augmentation strategy to iteratively draw θ given (Y_{obs}, Y_{mis}) and Y_{mis} given (Y_{obs}, θ).

2.2.1. Gibbs sampler

The method Gibbs sampler [11], helps to solves high dimensional problems in the form of jointdistribution but first it divides a set of complex distribution into simple conditionaldistributions.Consider, $P(Y)$ represents the complete joint distribution and be simulated with iterative draws with set of conditional distribution and include final result as mostfrequent drawn values.

Let be consider an example how distributions are derived with number of steps:

At stage of k , joint distribution is derived using $Y^{(k)} = \{Y_1^{(k)}, Y_2^{(k)}, ..., Y_n^{(k)}\}$

Similarly in stage (k+1) to be considered as $Y^{(k+1)} = \{Y_1^{(k+1)}, Y_2^{(k+1)}, ..., Y_n^{(k+1)}\}$

With the following detailed procedure can be shown with respect to all varaibales:

$$Y_1^{(k+1)} = \{Y_1|Y_2^k, Y_2^k, ..., Y_n^k]$$
$$Y_2^{(k+1)} = \{Y_2|Y_1^k, Y_3^k, ..., Y_n^k]$$
$$\vdots$$
$$Y_n^{(k+1)} = \{Y_n|Y_1^k, Y_2^k, ..., Y_{n-1}^k]$$

Similarly, the procedure is applied continuously and derived $Y^{(k+2)}$, $Y^{(k+3)}$, and so on. The process of sequence $\{Y^{(k)}, k = 1, 2, 3, \ldots, n\}$ will follow markov chain strategy.

2.2.2. Data Augmentation

Another special consideration to Gibbs sampler is two component method which is calledData augmentation. The two component sampler represented as $Y=(Y_1^{(k)}, Y_2^{(k)})$.

Considering approach components are drawn with the following way:

A. Approximate $Y_1^{(k+1)}$, with the given conditional distribution $P(Y_1|Y_2^{(k+1)})$
B. Similarly for $Y_2^{(k+1)}$, derived with distribution conditionally $P(Y_2|Y_1^{(k+1)})$.

Applied probability distributions to observed data iteratively to update parameter value.

The estimated value is used to impute missing values and to be considered as Imputation Stage, and the estimation of value done at Posterior Stage.

$$Stage_{Im} = Y_m^{(k+1)} = P(Y_m|\theta(t), Y_o)$$
$$Stage_{Post}\ \theta^{(k+1)} = P(\theta|Y_m^{(k+1)}, Y_o)$$

Consider data set $Y_{(k-1)}$ with observedand missing data at iteration (k – 1).

To derive $Y(k)$, process need to be done variable by variable.

 ISSN: 2454-4248
54 – 58

Algorithm Gibbs chain:

Step 0. Set $Y \leftarrow Y_{(k-1)}$ and update the variables of Y one at a time.

Step 1. Draw $\theta=P(\theta \mid Y_o, Y_{-1})$ and $Y_m=P(Y_m \mid Y_o, Y_{-1}, \theta)$

.

.

.

Step p. Draw $\theta=P(\theta \mid Y_o, Y_{-p})$ and $Y_m=P(Y_m \mid Y_o, Y_{-p}, \theta)$.

Observed in each step Markov chain applied toAlgorithm and derives posterior distribution to the Bayesian model and modified values of the parameters and then finally imputed data.

2.3. Multicore Environment

Multi core processors have several execution units where Scheduler allocatesthe tasks which are small programs to different cores.Example: Consider three imputation tasks namely T1, T2 and T3 and each one had subsequent workflow execution to mention as F1, F2 & F3. The Task one execute 3 functions, T2 two and T3 only one and the complete idea described in Figure 2.3. To execute each set of functionalities with respect Task demands individual core and is provisioned by Multi core environment offered by cloud.

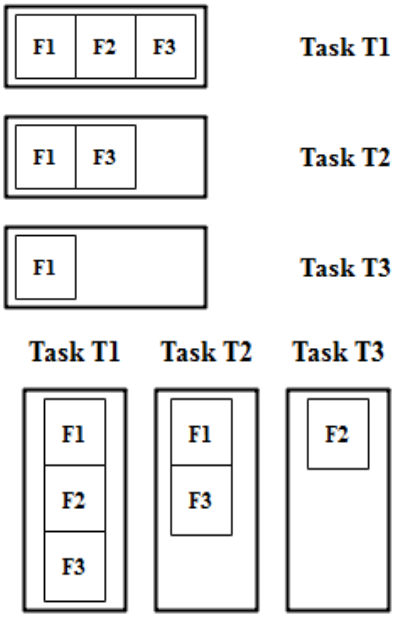

Fig.2.3: Multicore Environment

Each task has its own set of properties which could be dependent or independent of each other. They are collectively called functionality of a task. Tasks are allocated across several cores without considering the functionalities.

3. Proposed Method

3.1. Additive Lasso

The model of applying variable selection in case of n > p with Additive models [15] known to most of applications. We assume $E[y_{ij}|x_i] = h((\sum_{j=1}^{p} f_j(y_{ij})))$ and the function h applied toevery element of an unknown parameter reside in function f. Number of alternatives to approximate function f and among the smoothing is the best strategy with different cubic or regression spine introduced. The mechanism is very much suit for high dimensional data modelling.

3.2. Regularization methods

The rationale of regularization methods [12] is to penalize the loglikelihoodfunction:

$$\beta'=\arg\min \beta(||Y-X\beta||^2)+g(\beta)$$

Whereg (β) is a penality function and it is equal to $\lambda||\beta||^2$ which is leads to Ridge regression which $\lambda||\beta||^2$ leads to Lasso.

$$\beta'=\arg\min \beta(||Y-X\beta||2+\lambda||\beta||1)$$

The technique, least absolute shrinkage and selection operator (lasso) for estimation in linear models was first formulated. Based on this lasso method which adds apenalty term to the residual sum of squares and penalizes the coefficients ofcandidate predictors, shrinking some of the unimportant coefficients to zero and thus achieves variable selection where $\lambda>0$ the regularization parameter is.

3.3 Iterative Bayesian Additive Model for Multiple Imputation (IBALMI)

The method of iterative imputation, demands to determine estimation of p conditional models, $g_j(X_j|X_{-j},\theta_j)$, for $\theta_j \in \Phi_j$ with prior distributions $\pi_j(\theta_j)$ where $j=1,.....,p$. When there is no ambiguity, used g_j to be considered as specification to the conditional model for variable j.In case of imputation strategy used $g^1(X_m^1 \mid X_o^1, X^{-1}, \theta^1)$ to refer to the conditional distribution of missing data X_m^1 given (X_o^1, X^{-1}) and θ^1.

 ISSN: 2454-4248
54 – 58

Algorithm: Iterative Proposed MI with Bayesian and Additive Models

Step 1: Apply MI for X^1 at k-1 imputation variable 1

Draw θ^1 from posterior distribution $P(\theta^1 \mid X_o^1, X^{-1})$ is associated with g^1 and Π^1, Draw $g^1(X_m^1 \mid x_o^1, x^{-1})$

.

.

Step p: Apply MI for X^p at k-1 imputation variable p

Draw θ^p from posterior distribution $P(\theta^p \mid X_o^p, X^{-p})$ is associated with g^p and Π^p, Draw $g^p(X_m^p \mid x_o^p, x^{-p})$

Iterative imputation is that, where need to generate regression model for X^p given X^{-p} with p steps.The implementation which reduces the model complexity. In contrast, full Bayesian or likelihood modeling requires the more difficult task of constructing a joint model for X. Whether it is preferable to perform p easy tasks or one difficult task depends on the problem at hand. The proposed work, applied extended additive model to perform imputation which results minimized penalized least square criterion.

$$\beta^1 = \arg\min_{\beta}(\|Y - X\beta\|^2 + \sum_{j=1}^{p} \lambda_j \int_{aj}^{bj} f_j(x)^2 dx)$$

Where [aj, bj] is interval for which an estimate of fj is sought. Each function in above equation is penalized by a separate fixed smoothing parameter λj

3.5 Implementation and Results

The proposed method applied over clinical study with gene expression data with number of patients considered to be 200 and from each subject collected expression oriented features with the form of 1036 genes and are generated through microarray experiments. From that picked one of expression as outcome y, and rest are to be considered as set of predictors and results regression model to be considered as:

$$E(y \mid z_1, \ldots\ldots, z_{1035}) = \beta_0 + \beta_1 z_1 + \ldots\ldots\ldots + \beta_{1035} z_{1035}$$

To fit analysis to large sample study above mentioned model took much time it is required to apply regularized iterative method and then model to be rewritten as:

$$E(y \mid z_1, \ldots\ldots, z_{568}) = \beta_0 + \beta_1 z_1 + \ldots\ldots\ldots + \beta_{56} z_{568}$$

In original data there is no missing data and to perform imputation strategy generated missing data with logistic model to the original data set and the model is to be represented as:

$$\log it[P(\delta = 0)] = 1 + y + z_{45} - 2z_{1234} .$$

Approach is worked for estimation of β with possible imputation methods over simulation of 300sets. The performance of propose imputation method compared with standard methods with bench marks of Bias and RMSE (Root Mean Square Error) and is shown in Table 3.5.1.

Table3.5.1: Performance analysis of imputation methods in gene analysis

	Estimation Error	Standard Error	Bias	RMSE
CC	0.622	0.392	-0.232	0.455
SI	0.402	0.327	-0.012	0.328
RMI (1035 predictors)	0.412	0.331	-0.022	0.332
BLMI(1035 predictors)	0.331	0.306	0.059	0.312
IBALMI(1035 predictors)	0.408	0.303	-0.018	0.304
RMI(568 predictors)	0.388	0.3	0.002	0.300
BLMI(568 predictors)	0.35	0.305	0.04	0.307
IBALMI(568 predictors)	0.32	0.299	0.001	0.299

Consistent with the simulation results to minimize computation time, our proposed algorithm IBALMI(Iterative Bayesian Additive Lasso MI) result better performance compare to standard methods and also shown that results to be similar to complete case analysis. Later, study applied by dividing entire data sets into subsets from that compare to original complete data sub set produce least MSE and unbiased results. Compared Single Imputation (SI), RegularizedMI (RMI) the proposed algorithm Bayesian Lasso MI (BLMI) andBayesian Iterative Lasso MI (BLMI) results better performance. Moreover, it is observed that compared to complete data small predictors give better performance in terms of RMSE.The complete analysis done on multicore envirionment provisioned by cloud amazon to reduce model complexity and results shown in Table 3.5.2.As a final results multicore produces high speed compared to physical environment especially in high provisioned resources.

Table3.5.2: Multicore performance analysis of proposed method in gene analysis

Case Study	Sequential Excution Time(Sec) Ts	No.of Cores	Multi_CoreExcution Time(Sec) Tm	Speed (Ts/Tm)
q=20,M=10,p=100	5.237	1	5.40	0.97
		2	3.56	1.47
		4	1.35	3.88
		8	1.12	4.69
q=50,M=30,p=500	90	1	91.43	0.98
		2	75.34	1.19
		4	32.87	2.74
		8	27.33	3.29
q=80,M=50,p=5000	428	1	418.60	1.02
		2	287.72	1.49
		4	136.34	3.14
		8	101.60	4.21

4. Conclusion

In summary, our numerical results suggest that the Regularized methods and its extensions are better suited for

 ISSN: 2454-4248
54 – 58

MI which is an inherently iterative Bayesian multiplies imputing missing values in the presence of high-dimensional data than the other regression methods, which opens us the door to conduct imputation in the high-dimensional setting (p>n or p >> n). Compared with the existing MI approaches based on the classical regression techniques, the main advantages of the proposed methodology are as follows: (1) it is directly applicable to both low-dimensional and high-dimensional data and (2) it is a principled approach achieving simultaneous predictor selection and parameter estimation in imputation models.

References:

[1]. Aittokallio. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Briefings in Bioinformatics, 11(2):253–264, 2010.
[2]. Graham, J. W., Hofer, S. M., Piccinin, A. M. (1994), "Analysis with missing data in drug prevention research." National Institute on Drug Abuse Research Monograph 142, 13-63.
[3]. Aittokallio. Dealing with missing values in large-scale studies: microarray data imputation and beyond. Briefings in Bioinformatics, 11(2):253–264, 2010.
[4]. Little RJ, D'Agostino R, Cohen ML, et al. The prevention and treatment of missing data in clinical trials. N Engl J Med. 2012;367(14):1355–1360
[5]. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. Journal of Machine Learning Research, 99:2287–2322, 2010.
[6]. Ibrahim J, Molenberghs G. Missing data methods in longitudinal studies: A review. Test (Madr) 2009;18:1–43
[7]. Gromski, P. S., Xu, Y., Kotze, H. L., Correa, E., Ellis, D. I., Armitage, E. G., Turner, M. L., & Goodacre, R. (2014). Influence of missing values substitutes on multivariate analysis of metabolomics data. Metabolites, 4(2), 433-452.
[8]. Chiu C-C, Chan S-Y, Wang C-C, Wu W-S. Missing value imputation for microarray data: a comprehensive comparison study and a web tool. BMC Syst Biol. 2013;7(S-6):12. doi: 10.1186/1752-0509-7-S6-S12.
[9]. Stuart EA, Azur M, Frangakis C, et al. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. Am J Epidemiol. 2009;169(9):1133–1139.
[10]. Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. B. (2014). Bayesian data analysis (Vol. 2). Boca Raton, FL, USA: Chapman & Hall/CRC.
[11]. Gilks, W. R. and Wild, P. P. (1992). Adaptive rejection sampling for gibbs sampling. Appl. Statist, 41(2):337–348.
[12]. Allen and R. Tibshirani. Transposable regularized covariance models with an application to missing data imputation. Annals of Applied Statistics, 4(2):764–790, 2010.
[13]. Consentino, F. and Claeskens, G. (2011). Missing covariates in logistic regression, estimation and distribution selection. Statistical Modelling, 11(2):159–183.
[14]. Josse, J. and Husson, F. (2016). missMDA: A package for handling missing values in multivariate data analysis. Journal of Statistical Software, 70(1):1–31.
[15]. de Jong, S. van Buuren, and M. Spiess. Multiple imputation of predictor variables using generalized additive models. Communications in Statistics - Simulation and Computation, 45(3):968–985, 2014. ISSN 1532-4141