MCMC impute missing values and Bayesian variable selection for logistic regression model to predict Pima Indian Diabetes

Gongli Li¹*, Yueze Liu², Han Li², Ruikuan Yao² and Chenyang Li²

¹College of Statistics and Big data, Henan university of Economics and Law, Zhengzhou, China
²School of Information Science & Engineering Lanzhou University, Lanzhou, China

*Corresponding author: u6141461@anu.edu.au

Abstract. Diabetes mellitus is a metabolic disease that causes high blood sugar. The risk factor of diabetes can be reduced significantly by early precise prediction. Lots of literatures published for diabetes prediction, but nearly all of them proposed frequentist Machine learning algorithm to classify and build models. Besides, their data preprocessing methods are not professional. In this literature, we are proposing MCMC filling missing values and Bayesian variable selection for logistic regression model to classify diabetes patients. It shows great performance (AUC = 0.884, sensitivity =0.805, specificity= 0.875).

Keywords: MCMC, missing value imputation, Bayesian machine learning, Pima Indian Diabetes

1. Introduction
Diabetes is a progressive, chronic disease characterized by elevated levels of blood glucose, and diabetes could lead to complications in many parts of the body and increase the risk of dying prematurely [1]. The research [2] in 2017 demonstrates that 451 million people are living with diabetes in the world, it will rise to 693 million by 2045. Diabetes can be controlled and prevented by early accurate prediction. Recently, there are plenty of literatures proposed and published for diabetes prediction, such as random forest [3] and Gaussian process classifier [4]. However, nearly all of them proposed frequentist Machine learning algorithm to build models, and their dealing missing data methods, such as omitting missing values and filling missing values by attribute’s mean or median, lost lots of information. In this literature, MCMC filling missing values and Bayesian variable selection for logistic regression model are proposed to improve data quality and build prediction model.
Table 1. PID attributes summary table

| Attributes          | Description                                                | Mean±SD     |
|---------------------|-------------------------------------------------------------|-------------|
| Pregnant            | Number of pregnancies                                      | 3.85±3.4    |
| Glucose             | Plasma glucose concentration at 2 hours                    | 121.7±30.5  |
| Pressure            | diastolic blood pressure (mm Hg)                           | 72.4±12.4   |
| SkinThick.          | Triceps skin fold thickness (mm)                            | 29.15±10.5  |
| Insulin             | 2-hour serum insulin (μU/ml)                                | 155.6±118.8 |
| BMI                 | Body Mass index \((\text{weight (kg)} / \text{Height (inches)}^2)\) | 32.5±6.9    |
| Pedigree            | Diabetes pedigree function \(\sum K_i (88 - ADM_i + 20 \sum K_j (ALC_j - 14) + 50_j)\) | 0.47±0.33   |
| Age                 | Age of an individual                                        | 33.24±11.76 |

2. Data and method

2.1. Data description
The Bayesian model analysis and predict based on public source Pima Indians Diabetes dataset. The dataset consists 768 female potential diabetic patients (268 diabetic patients, and 500 non-diabetic patients) from the Pima Indian population near Phoenix, Arizona [5]. The Fig. 1 shows the density plots for different attributes with group diabetic and non-diabetic, and the number of zero values of each attribute in the original dataset. It is impossible for Blood pressure, Insulin, Glucose, Skin Thickness and BMI equal zero, so these values are labelled as missing values. The data summary table 1 shows attribute description with mean and standard deviation excluded the missing values.

![Figure 1. Density plots for all attributes.](image-url)
2.2. **Proposed framework**

The Fig. 2 shows the proposed framework of the literature. Firstly, preprocess the PID dataset to improve data quality. Secondly, split the data to testing set and training set. Use training set to build the Bayesian model, and testing set evaluates the performance.

2.2.1. **Data pre-process.** The first step of the proposed framework is data pre-process. It includes remove outliers (Q), MCMC impute missing values (R) and standardization (S).

The outlier is an observation that lies an abnormal distance from other observations in a random sample from a population. Identify and reject outliers in the literature can be written as in (1).

\[
Q(x) = \begin{cases} 
  x, & \text{if } x \in [Q_1 - 1.5 \times IQR, Q_3 + 1.5 \times IQR] \\
  \text{reject, otherwise} 
\end{cases} \tag{1}
\]

MCMC impute missing values [6] was proposed on 1978 for imputing non-response in sample surveys, and the mathematical formulation for the simulation can be formulated as (2).

\[
P(\theta|X_{\text{obs}}) = \int p(\theta, X_{\text{mis}}|X_{\text{obs}})dX_{\text{mis}} = \int p(\theta|X_{\text{mis}}, X_{\text{obs}})p(X_{\text{mis}}|X_{\text{obs}})dX_{\text{mis}} \tag{2}
\]

Where the missing values \(X_{\text{mis}}\) can be treated as unknown parameters, so the posterior distribution of \(X_{\text{mis}}\) can be simulated by drawing the \(X_{\text{mis}}^{(d)}\) from their posterior predictive distribution \(p(X_{\text{mis}}|X_{\text{obs}})\) to complete the data set, and then drawing \(\theta\) from the completed data posterior distribution \(p(\theta|X_{\text{mis}}, X_{\text{obs}})\). After \(S\) time iteration, the missing values can be imputed by the mean of \(X_{\text{mis}}^{(d)}\) draws as (3).

\[
R(X_{\text{mis}}) = E(X_{\text{mis}}|X_{\text{obs}}, \theta, \Sigma) = \frac{\sum_{i=1}^{S} X_{\text{mis}}^{(i)}}{S} \tag{3}
\]
Table 2. Algorithm 1

**Algorithm 1** Bayesian multivariate normal model with missing values

| Input: | Prior: $\theta \sim \text{MVN}(\mu_0, A_0)$, $\Sigma \sim \text{Inverse-Wishart}(v_0, S_0^2)$. |
|-------|--------------------------------------------------|
|       | Data: $X_{\text{obs}} = [x_{ij}|x_{ij} \text{ not missing}]$, $X_{\text{mis}} = [x_{ij}|x_{ij} \text{ missing}]$. |
|       | starting values $(\theta^{(0)}, \Sigma^{(0)})$, and iteration times $S$. |
| Output: | $((\theta^{(1)}, \ldots, \theta^{(S)}), (\Sigma^{(1)}, \ldots, \Sigma^{(S)}), (x_{\text{mis}}^{(1)}, \ldots, x_{\text{mis}}^{(S)}))$ |

for iteration $i = 1, 2, \ldots, S$ do

1. Sampling $x_{\text{mis}}^{(i+1)}$ from $X_{\text{mis}} | X_{\text{obs}}, \theta^{(i)}, \Sigma^{(i)} \sim p(X_{\text{mis}} | X_{\text{obs}}, \theta^{(i)}, \Sigma^{(i)})$
2. Sampling $\theta^{(i+1)}$ from $p(\theta | X_{\text{obs}}, X_{\text{mis}}^{(i)}, \Sigma^{(i)})$
3. Sampling $\Sigma^{(i+1)}$ from $p(\Sigma | X_{\text{obs}}, x_{\text{mis}}^{(i+1)}, \theta^{(i+1)})$

end for

In this literature, Bayesian multivariate normal model with missing values is used for imputation. Gibbs sampler algorithm can be applied to the model, because the full conditional distribution of each parameter can be derived. The algorithm of Bayesian multivariate normal model with missing values by Gibbs sampler is given in table 2 (algorithm 1).

Z-score normalization as (4) is used for achieving standard normal distribution with zero mean and unit variance.

$$S(x) = \frac{x - \mu}{\sigma}$$

(4)

2.2.2. **Bayesian variable selection for logistic regression model.** Metropolis algorithm for the Bayesian logistic regression variable selection model is introduced in this section. The (5) shows the framework of Logistic regression variable selection that $\gamma = [y_0, y_1, \ldots, y_p]$ and $\beta = [\beta_0, \beta_1, \ldots, \beta_p]$ are unknown parameters, where $y_j \in \{0,1\}$ and $\beta_j \in \mathbb{R}$. Each value of $\theta = (\gamma, \beta)$ presents different models.

$$P(Y = 1 | x_0, y, \beta) = \frac{\exp(\theta_0)}{1 + \exp(\theta_0)}$$

where $\theta_0 = \beta_0 y_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

(5)

The table 3 (algorithm 2) shows the detail of the Bayesian variable selection for logistic regression model sampling process. Update $\gamma$ and $\beta$ are two steps for every iteration. Metropolis algorithm is applied for update $\beta$. In the update $\gamma$ step, the Metropolis sampler searches through the model space for $\gamma = (y_0, \ldots, y_p)$ with higher posterior probabilities. For current model selected coefficient $\gamma^{(s)} = (y_0^{(s)}, \ldots, y_p^{(s)})$, generating a new value for $y_j^{(s+1)}$ in $\gamma^{(s+1)} = (y_0^{(s+1)}, \ldots, y_p^{(s+1)})$ is sampled from $p(\gamma^{(s+1)} | y, x, y_j^{(s)})$, where $y_j^{(s)}$ means the values of $\gamma^{(s)}$ exclude the regressor $j$. Therefore, the conditional odds can be defined as (6).

$$\text{odds}_j = \frac{p(Y = 1 | y, x, y_j^{(s)})}{p(Y = 0 | y, x, y_j^{(s)})} = \frac{P(Y = 1)}{P(Y = 0)} \times \frac{P(y_j = 1 | y, x, y_j^{(s)})}{P(y_j = 0 | y, x, y_j^{(s)})}$$

(6)
Table 3. Algorithm 2

Algorithm 2 Bayesian variable selection for logistics regression

Input:
Prior: identical and independent priors $\beta_j \sim \text{Normal}(\mu_j, \delta_j^2)$, where $j = \{0, \ldots, p\}$.
Proposal distribution: $\beta^* | \beta^{(s)} \sim \text{MVN}(\beta^{(s)}, \Sigma)$

Starting model selected coefficient $\gamma^{(0)} = (\gamma_0^{(0)}, \ldots, \gamma_p^{(0)})^T$

Iteration time $S$ and starting value $\beta^{(0)}$.

Output: 
$\{ (\beta_0^{(1)}, \ldots, \beta_p^{(1)}), \ldots, (\beta_0^{(S)}, \ldots, \beta_p^{(S)}) \}$ for iteration $i = 1, 2, \ldots, S$

for iteration $i = 1, 2, \ldots, S$ do
  update $\gamma$
    1. Set $\gamma = \gamma^{(0)}$.
    2. For $j \in \{1, \ldots, p\}$ in random order, replace $\gamma_j$ sampling from $p(\gamma_j^{(s+1)} | y, X, \gamma_{-j}^{(s+1)})$.
    3. Set $\gamma^{(s+1)} = \gamma$

  update $\beta$
    4. Sample $\beta^* \sim \text{MVN}(\beta^{(s)}, \Sigma)$
    5. Acceptance ratio: $r = \prod_{j=1}^p \text{dnorm}(\beta_j^*; \mu_j, \delta_j^2) \prod_{i=1}^n \text{dbern}(y_i; \beta_{\gamma_i}^{(s+1)}; X_{\gamma_i}^{(s+1)})$
      $\prod_{j=1}^p \text{dnorm}(\beta_j; \mu_j, \delta_j^2) \prod_{i=1}^n \text{dbern}(y_i; \beta_{\gamma_i}^{(s)}; X_{\gamma_i}^{(s)})$
      Where $\beta$ and $X$ mean only include the column where $\gamma_j = 1$.
    6. Let: $\beta^{s+1} = \begin{cases} \beta^*, & \text{with probability } \min(r, 1) \\ \beta^{(s)}, & \text{with probability } \min(r, 1) \end{cases}$

end for

3. Results and discussion

3.1. Pre-processing results
There are 84 outliers based on the outlier detection (1). After remove the outliers, skewness and kurtosis problems can be solved.

Figure 3. Multivariate normal distribution of the attributes with missing values
The attributes with missing values (Glucose, BloodPressure, SkinThickness, sqrt(Insulin), BMI) follow multivariate normal distribution on Fig.3, so Bayesian multivariate normal model with missing values could be applied here to impute missing values. The weakly informative prior parameter setting and thinning the sequence (save every 10th sample) are applied. Iteration times $S=50000$. After the iteration, the missing values can be imputed by the mean of 5000 thinned drawings $X_{mis}$.

3.2. **Bayesian logistic regression variable selection result**

After remove 84 outliers in pre-processed step, there are 684 observations. Random select 184 observations as testing set and the other 500 as training set. For the input values in the algorithm 2, iteration times $S=300000$ and thinning the sequence (save every 50th sample) are applied. Assume weakly information prior and input values, where $\beta^{(0)}$ and $\Sigma$ is the coefficient and unscaled covariance matrix of logistic regression by MLE method, and the starting model selected coefficient $\gamma_j^{(0)}$ is randomly generated by Bern($p=0.5$).

**Figure 4.** Trace of MCMC chains for $\beta_j \times \gamma_j$

**Figure 5.** Posterior density for $\beta_j \gamma_j$
Fig. 4 shows the trace of MCMC chains for $\beta_j^{(6)} \times \gamma_j^{(6)}$ is stationary for all $j={0,...,8}$. Fig. 5 shows the posterior density of $\beta_3 \gamma_3$, $\beta_4 \gamma_4$, $\beta_5 \gamma_5$, $\beta_8 \gamma_8$ are located around 0. The posterior density of $\beta_1 \gamma_1$, $\beta_6 \gamma_6$ and $\beta_7 \gamma_7$ are bimodal, and the posterior density of $\beta_0 \gamma_0$, $\beta_2 \gamma_2$ are followed normal distribution. The ROC and best cutoff confusion matrix of the Bayesian logistic regression variable selection model are shown in Fig. 6. On the ROC plot, the area under curve is 0.884 and the optimal cutoff to classify the test observations as diabetic patients is $p > 0.2925$. The confusion matrix plot shows detail of the optimal point. The accuracy is 0.826 with sensitivity =0.805 and specificity=0.875.

![Figure 6. ROC and best cutoff confusion matrix of the Bayesian logistic regression variable selection model](image)

3.3. Results comparison

| Authors and published year | Impute missing value technique | Outlier rejection technique | Feature reduction technique | classifier | performance |
|-----------------------------|-------------------------------|-------------------------------|-----------------------------|------------|-------------|
| S. Bashir et al. (2016) [7]  | KNN impute                    | extreme studentized deviate   | -                           | HM-BagMoov | AUC: 0.926; sensitivity: 0.787; specificity: 0.926 |
| M. Maniruzzaman et al. (2017) [4] | Median                       | -                             | gaussian process classifier | AUC: 0.918; sensitivity: 0.943; specificity: 0.633 |
| M. Maniruzzaman et al. (2018) [8] | Group median                 | Median                        | -                           | Random forest | AUC: 0.93; sensitivity: 0.966; specificity: 0.797 |
| Q. Wang et al. (2019) [3]   | NB method                     | -                             | -                           | Random forest | AUC: 0.928; sensitivity: 0.85; specificity: 0.63 |
| K. Hasan (2020) et al. [9]   | mean method                   | IQR                           | Correlation                 | Ensembling of AB and XB | AUC: 0.95; sensitivity: 0.789; specificity: 0.934 |
| Li (2014) [10]              | -                             | -                             | -                           | Ensembling of SVM, ANN & NB | AUC: 0.53; sensitivity: 0.858; specificity: 0.868 |
| A.K. Dewangan et al. (2015) [11] | Manual                       | -                             | Manual                      | Ensembling of MLP & NB | AUC: 0.7; sensitivity: 0.641; specificity: 0.909 |
| S.P. Chatrati et al. 2020 [12] | -                             | -                             | -                           | Discriminant Analysis | AUC: 0.7; sensitivity: 0.720; specificity: 0.67 |
| Our proposed (2020)         | MCMC impute                   | IQR                           | -                           | variable selection of Bayesian logistic regression | AUC: 0.884; sensitivity: 0.805; specificity: 0.875 |

*Bayesian variable selection for logistics regression
From the table 4, it is clear that the performance of methods with preprocess is better than that others without any preprocess, because there are 763 missing values and some outliers. The framework proposed in [9] and [4] used attribute’s mean or median to impute missing values is basic way and may loss lots of information. The [7] used KNN impute missing values, the weakness is imputed values may far away from the central tendency of the population distribution. MCMC impute missing values is the safe and proper way.

4. Conclusion and future work
In this literature, MCMC filling missing values and building the Bayesian model to predict diabetes in PID dataset is different with the other proposed models. In the future, diagnostic and analysis other medical data in Bayesian way will be applied.

Acknowledgments
Corresponding author is the first author Gongli Li.

References
[1] World Health Organization. Global report on diabetes[J]. 2016.
[2] Cho N H, Shaw J E, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045[J]. Diabetes research and clinical practice, 2018, 138: 271-281.
[3] Wang Q, Cao W, Guo J, et al. DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values[J]. IEEE Access, 2019, 7: 102232-102238.
[4] Cho N H, Shaw J E, Karuranga S, et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045[J]. Diabetes research and clinical practice, 2018, 138: 271-281.
[5] Rubin D B. Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse[C]//Proceedings of the survey research methods section of the American Statistical Association. American Statistical Association, 1978, 1: 20-34.
[6] Bashir S, Qamar U, Khan F H. IntelliHealth: a medical decision support application using a novel weighted multi-layer classifier ensemble framework[J]. Journal of biomedical informatics, 2016, 59: 185-200.
[7] Maniruzzaman M, Rahman M J, Al-MehediHasan M, et al. Accurate diabetes risk stratification using machine learning: role of missing value and outliers[J]. Journal of medical systems, 2018, 42(5): 92.
[8] Wang Q, Cao W, Guo J, et al. DMP_MI: An effective diabetes mellitus classification algorithm on imbalanced data with missing values[J]. IEEE Access, 2019, 7: 102232-102238.
[9] Hasan M K, Alam M A, Das D, et al. Diabetes Prediction Using Ensembling of Different Machine Learning Classifiers[J]. IEEE Access, 2020, 8: 76516-76531.
[10] Li L. Diagnosis of diabetes using a weight-adjusted voting approach[C]//2014 IEEE International Conference on Bioinformatics and Bioengineering. IEEE, 2014: 320-324.
[11] kumar Dewangan A, Agrawal P. Classification of diabetes mellitus using machine learning techniques[J]. International Journal of Engineering and Applied Sciences, 2015, 2(5).
[12] Chatrati S P, Hossain G, Goyal A, et al. Smart home health monitoring system for predicting type 2 diabetes and hypertension[J]. Journal of King Saud University-Computer and Information Sciences, 2020.