The Wild Bootstrap Resampling in Regression Imputation Algorithm with a Gaussian Mixture Model

Aisyah Mat Jasin [1], Daniel Neagu[1] and Attila Csenki [1]

[1] University of Bradford, Bradford BD7 1DP, UK
A.MatJasin@bradford.ac.uk
D.Neagu@bradford.ac.uk
A.Csenki@bradford.ac.uk

Abstract. Unsupervised learning of finite Gaussian mixture model (FGMM) is used to learn the distribution of population data. This paper proposes the use of the wild bootstrapping to create the variability of the imputed data in single missing data imputation. We compare the performance and accuracy of the proposed method in single imputation and multiple imputation from the R-package Amelia II using RMSE, R-squared, MAE and MAPE. The proposed method shows better performance when compared with the multiple imputation (MI) which is indeed known as the golden method of missing data imputation techniques.

Keywords: Missing Data Imputation, Gaussian Mixture Model, Bootstrap.

1 Introduction

Missing data can occur in data records for various reasons, such as: data entry errors, system failures, or respondents who avoid answering questions within a survey. Various methods have been proposed to deal with the missing data problem. The standard technique is discarding observations or variables that contain missing values. The deletion method is inappropriate when the missing proportions are high, resulting in inefficient parameter estimates, and estimated results tend to be underestimated. To deal with these issues, imputation methods can be used to substitute missing values with plausible values. For example, the single mean imputation consists of replacing the missing values with the mean, median or mode value. However, this simple approach produces biased analysis results. The multiple imputation method introduced in [1] is a complex approach where missing data are filled-in by drawing multiple sets of complete data that contain different plausible values. This method is complicated and computationally expensive [2], especially for large data sets because execution processes are implemented through three phases in several iterations. The improved version of the single imputation technique such as conditional mean imputation, which incorporates the statistical and machine learning methods with multivariate Gaussian mixture models (GMM)[3] have gained interest in many years[4].

The conditional mean imputation (also known as ordinary least square, OLS) or regression imputation can preserve the data distribution, according to Di Zio [5]. The

conventional OLS $\hat{y}_i = \beta_0 + \sum_{j=1}^{J} \beta_j x_j + \varepsilon_i$ implementation requires the use of random error ε_i which can be obtained in two ways [6]: 1) draw a random error with underlying assumption that it is independent and identically distributed, that follows a Gaussian distribution with zero mean and finite variance; 2) draw a random error with replacement from the empirical distribution of the estimated residuals $\varepsilon_i = y_i - \hat{y}_i$ [7]. Problems can occur in the random error and residual ε_i in method (1) that will create the sparsity problem whereas the random ε_i generation will be either too large or too small although the normality distribution assumption is met. The sparsity of data in method (2) will be inconsistent if the data distribution has different clusterings and each cluster consists of a different density. The sparsity of data creates some problems such as increases in the variance between the imputed and original data.

The conditional mean imputation proposed in [5] does not consider adding the residuals. Although this method may preserve the data distribution, it will underestimate the variability, introduce the bias on imputed data and the result of imputed data will be highly inaccurate. The additional steps are required to improve data sparsity in the random error ε_i generated in the OLS to obtain a better predicted missing value.

The main objective of this study is to investigate the random error and employ the wild bootstrap [8] [9] on the missing data prediction using regression imputation on the Gaussian mixture model. The wild bootstrap is used to improve the variance in heteroscedasticity issue when the data variance is not homoscedastic [8][9]. Further details about the wild bootstrap approach are discussed in the next section that introduces the modelling framework.

In this paper, we employ the wild bootstrap to the single imputation technique in missing value prediction, since the GMM framework is flexible to learn multimodal data distribution. We combine the GMM model with the proposed missing data prediction method. We also employ the wild bootstrap to investigate the effect of the sparsity of imputed data in a different mixture data distribution case. Thus, we would like to show that the performance of single imputation may perform well, and as good as the implementation of MI. We assume that the data is missing data at random (MAR).

This paper is organized as follows: in Section 2, we present the Gaussian mixture model framework and the proposed regression imputation with wild bootstrap technique. In Section 3 we discuss the experimental evaluation and experimental results. Section 4 concludes the paper and identifies further directions for research and study.

2 Modelling Framework

GMM is a powerful probabilistic model used in predicting specifically in data clustering [5]. This model is flexible to learn from different data distributions by fitting the probability density function (PDF) to represent different clusters [3]. The well-known

strategy for finding the Maximum Likelihood (ML) parameter estimation uses the Expectation-Maximization (EM) algorithm [10]. GMM applications to missing data problems have been studied extensively for example in [4][5][11].

2.1 Definitions

Suppose the data set $\mathbf{X}$ having N units of independent and identically distributed (i.i.d) data points with p -column vectors can be written as follows:

Observation (index)	X_1	X_2	..	X_l	..	X_p
1	x_{11}^O	x_{12}^O	..	x_{1l}^O	..	x_{1p}^O
..	:	:	..	..	..	:
n_1	$x_{n_1 1}^O$	$x_{n_1 2}^O$	..	$x_{n_1 l}^O$	..	$x_{n_1 p}^O$
n_1+1	$x_{(n_1+1)1}^M$	$x_{(n_1+1)2}^M$	..	NA	..	$x_{(n_1+1)p}^M$
:	:	:	..		..	:
N	x_{N1}^M	x_{N2}^M	..	NA	..	x_{Np}^M

Figure 1: A sample data set with missing values

Figure 1 illustrates a data set that contains missing values (highlighted with *NA* in the relevant cells). Let $\mathbf{X} = \{X_1, X_2, ..., X_p\}$ be the random variable of the $N \times p$ data matrix. In the imputation process, Rao and Shao [12] suggested to create a set of respondents $\mathbf{X}^O$ and a set of non-respondents $\mathbf{X}^M$ separately. The variable $\mathbf{X}^O$ denotes the $n_1 \times p$ matrix where n_1 is the size of observed data while $\mathbf{X}^M$ denote the $n_0 \times p$ matrix where $n_0 = N - n_1$ is the number of missing values that occur in $\mathbf{x}_l$. Let $\mathbf{x}_l$ of size $n_1 \times 1$ vector contain observed data and n_0 be the size of missing values in $\mathbf{x}_l$.

2.2 Multivariate Gaussian Mixture Model

The Maximum Likelihood (ML) is an approach to estimate the parameters of the distribution from multivariate GMM using the Expectation-Maximization EM algorithm [10]. The data in GMM are distributed by different k Gaussian components and estimated as follows:

$$f(\mathbf{x};\boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k f(\mathbf{x} \mid \boldsymbol{\theta}_k) \qquad (1)$$

where $f(\mathbf{x} \mid \boldsymbol{\theta}_k)$ is the density of p-variate Gaussian distribution with the k component. The vector $\boldsymbol{\Phi}$ contains the full set of parameters in the mixture model $\boldsymbol{\Phi} = (\pi_1, ..., \pi_K; \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_K)$, where $\boldsymbol{\theta}_k$ is the vector of unknown parameters of mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$.

The mixing coefficients (or weights) π_k for the k^{th} component must satisfy the conditions $0 < \pi_k < 1$, and $\sum_{k=1}^{K} \pi_k = 1$. The GMM is a dynamic model where it is not required to specify any column vector to be an input or output particularly.

2.3 The General EM Algorithm

The EM algorithm is a statistical tool to find the maximum likelihood estimates of the set parameters such as mean, variances, covariances and regression coefficients of a model. The optimisation algorithm introduced by Dempster, Laird, and Rubin [10] starts with an initial estimate of $\boldsymbol{\Phi}$ and iteratively executes the process until it satisfies the convergence criteria. The iterative process has two steps known as the E-step and the M-step. The E-step computes the probability membership τ_{ik} for all data points x_i of mixture component k. The M-step will update the value of the parameter $\boldsymbol{\Phi}$ with respect to the k Gaussian component. Let denote q as an iteration counter, the expected values of the posterior distribution are computed by:

$$\hat{\tau}_{ik}^{(q)} = \frac{\hat{\pi}_k f(\mathbf{x}_i^o \mid \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}}_k)}{\sum_{j=1}^{K} \hat{\pi}_j f(\mathbf{x}_i^o \mid \hat{\boldsymbol{\mu}}_j, \hat{\boldsymbol{\Sigma}}_j)} \tag{2}$$

In the M-step, we use the expected values in the posterior distribution (2) to re-estimate the means, covariances and mixing coefficients. The new set of parameters $\boldsymbol{\Phi}^{(q+1)}$ are updated as follows:

$$\hat{\pi}_k^{(q+1)} = \frac{N_k}{N} \text{ for } k=1,...,K, \tag{3}$$

$$\hat{\mu}_k^{(q+1)} = \frac{1}{N_k} \sum_{i=1}^{N} \tau_{ik} \hat{\mathbf{x}}_{ik} \tag{4}$$

$$\hat{\boldsymbol{\Sigma}}_k^{(q+1)} = \frac{1}{N_k} \sum_{i=1}^{N} \tau_{ik} [(\hat{\mathbf{x}}_{ik} - \boldsymbol{\mu}_k)(\hat{\mathbf{x}}_{ik} - \boldsymbol{\mu}_k)^T + \hat{\boldsymbol{\Sigma}}_{ik}^{MM}] \tag{5}$$

The algorithm then iterates the E-step and M-step until convergence is achieved.

2.4 The Least Square Method

The conditional mean imputation is also known as regression imputation [13]. The imputed values are regressed from independent variables $\mathbf{X}_p$. Let consider the following linear regression model:

$$x_{il} = \beta_0 + \beta_1 x_i + \varepsilon_i , \quad i = 1, 2, ..., n \tag{6}$$

where the response variable x_{il} is predicted from regression coefficients β_0 and β_1 with random error $\varepsilon_i \sim N(0, \sigma^2)$ i.i.d. and uncorrelated. The matrix development of equation (6) is presented as follows:

$$\mathbf{x}_l = \begin{bmatrix} x_{1l} \\ \vdots \\ x_{Nl} \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{N1} & \dots & x_{Np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_N \end{bmatrix}$$

In general, $\mathbf{x}_l$ is an $N \times 1$ vector of the dependent variable contains missing values, $\mathbf{X}$ is a $N \times p$ matrix of observed variables, $\boldsymbol{\beta}$ is a $p \times 1$ vector of the regression coefficients and $\boldsymbol{\varepsilon}$ is a $N \times 1$ vector of random errors. The general least square estimator of $\boldsymbol{\beta}$ based on observed values is:

$$\hat{\boldsymbol{\beta}}^O = \left(\mathbf{X}^T \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{X}_l \tag{7}$$

In the presence of missing data, the imputed values are obtained by the conditional mean imputation technique which corresponds to imputed values generated from a set of regression equation calculated in (7) as discussed in [14][13]. There are two ways to generate the random error component ε_i . The random error component ε_i can be generated either with $\varepsilon_i \sim N(0, \sigma^2)$ or residual.

2.5 Fundamentals of the Bootstrap Method

The bootstrap non-parametric resampling technique was proposed by Efron [15] for estimating a standard error , confidence interval in various types of distributions. This method was extended in [16] and [17] to generate the random error ε_i in the regression model. Let $\mathbf{X} = \left\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_{n_1}\right\}$ is a random sample from p-variate normal distribution K where n_1 refers to the size of observed data $\mathbf{X}^O$ as shown in Figure 1. Let $\mathbf{X}^{(b_k)}$ denote the bootstrap resampled data generated by sampling with replacement from the original dataset $\mathbf{X}_k$ where b indicates the counter $b = 1, ..., B$ of drawing samples of bootstrap and k refers to the current Gaussian component. In this study, the resampling and parameter estimation are implemented on the observed data $\mathbf{X}_k^O$ where the superscript O refers to observed data.

2.6 The Wild Bootstrap

Wu [8] introduced the wild bootstrap to deal with the heteroscedasticity issue. Later, a better approximation of the wild bootstrap was proposed by Liu [9]. The wild bootstrap is based on the modification of the bootstrap residual approach of the least square estimation. Wu [8] improved the resampling residual with replacement in bootstrap by drawing a value of t_i^* that follow a standard normal distribution with zero mean and unit variance:

$$x_{il}^b = x_i^T \hat{\beta} + t_i^* \frac{\hat{\varepsilon}_i}{\sqrt{1-w_i}} \tag{8}$$

where $w_i = x_i^T (\mathbf{X}^T \mathbf{X}) x_i$. However, the error variance $t_i^* \hat{\varepsilon}_i$ are inconsistent. Therefore, authors in [18] proposed to compute t_i^* by drawing a sample a_i with replacement:

$$t_i^* = a_i = \frac{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}_i}{\sqrt{n_{1k}^{-1} \sum_{i=1}^{n_{1k}} (\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2}} \tag{9}$$

where $\bar{\hat{\varepsilon}} = n_{1k}^{-1} \sum_{i=1}^{n_{1k}} \hat{\varepsilon}_i$.

The second wild bootstrap technique employed in this study is the Liu's bootstrap [9]. Liu [9] proposed t_i^* in Wu [8] by resampling a set of central residual with zero mean and unit variance that has third central moments equal to one. Liu proposed two procedures to draw random numbers t_i^*. However, we consider the second procedure as it is appropriate for normal distribution. Liu's bootstrap is conducted by drawing random numbers:

$$t_i = D_1 D_2 - E(D_1) E(D_2) \tag{10}$$

where D_1 and D_2 are random i.i.d that follows normal distribution with means $0.5*(\sqrt{17/6} + \sqrt{1/6})$ and $0.5*(\sqrt{17/6} - \sqrt{1/6})$ respectively, and variance 0.5.

2.7 The Non-Parametric Wild Bootstrap Applied in Missing Data Imputation

The bootstrap procedure based on the resample approach in the GMM is described in the following steps:

1. Initiate the set of parameters $\mathbf{\Phi}$ with K-means algorithm.
2. Compute the residual for each Gaussian component:
    a. Fit Gaussian mixture model using the parameter values from the step 1.
    b. Compute the residual: $\hat{\varepsilon}_k = \mathbf{X}_{lk} \hat{\boldsymbol{\beta}}_k$ where *k* is the Gaussian component *k* = 1,..,*K*.
3. For b = 1,.., B
    a. Draw a vector $\hat{\varepsilon}_k$of n_{1k} i.i.d sample with a simple random sampling with replacement. The vector $\hat{\varepsilon}_k$ is generated from step 2b with respect to the

option of the Wu's [8] or Liu's[9] bootstrap procedure as discussed in the Section 2.6.

b. Fit Gaussian mixture model using the parameter values from the step 1.
c. In the E-step,
   i. Compute the posterior probabilities vector τ_{ik} in equation (2) on the observed data.
d. In the M-step,
   i. Impute the missing values of size n_{0k} using a linear regression model (6) based on OLS estimator $\hat{\boldsymbol{\beta}}^{(O_k)}$ in (12):
   $$x_{il} = \hat{\beta}_0^{(O_k)} + \hat{\beta}_1^{(O_k)} x_i + t_i^* \hat{\varepsilon}_i \ / \sqrt{1 - w_i}$$
   where the residual t_i^* taken from the step 3a.
   ii. Update the new parameter $\boldsymbol{\Phi}$ for each component in GMM as shown in (3), (4) and (5).

3 Experiments and Discussion of Results

In this section, the numerical results are presented on real and simulated datasets.

3.1 The Non-Parametric Wild Bootstrap Applied in Missing Data Imputation

Dataset: We applied various evaluation criteria on one real dataset and one artificial dataset with two variables and two Gaussian classes. The first case study is the Old Faithful Geyser dataset [19]. This dataset contains 272 records on the waiting time between geyser eruptions (waiting) and the duration of eruptions (eruptions) in Yellowstone National Park, USA.

For the artificial case study, the values are randomly sampled with 1000 observations of two Gaussian classes with different position mean values and positive-negative correlation. Data are drawn with normal distribution using the following parameters:

$$\pi_1 = 0.5\,,\ \pi_2 = 0.5$$

$$\mu_1 = (4,2)^{'}\,,\ \mu_2 = (-2,6)^{'}$$

$$\Sigma_1 = \begin{pmatrix} 1 & -0.7 \\ -0.7 & 1 \end{pmatrix}, \Sigma_1 = \begin{pmatrix} 3 & 0.9 \\ 0.9 & 3 \end{pmatrix}$$

Software: the proposed method in these experiments were conducted using Matlab version 2017a. The proposed method is compared with multiple imputation available in the R-package Amelia II. The comparisons are conducted based on the artificial missing data generated with different missing data percentages (MDP): 5%, 10%, 15% and 20%.

Imputation implementation: the missing data are imputed based on the regression imputation. Prior to the imputation process, the K-means algorithm is used to determine

initial parameter values of mixing proportion π_k, mean μ_k and covariance matrix Σ_k in GMM. The stopping criteria is based on a selected threshold where the different iterations were less than 10^{-6}.

Evaluation criteria: These experiments are designed to measure the performance and prediction accuracy between predicted and actual values. RMSE computes the deviation between predicted and actual values that employed by most missing data imputation studies. The greater the deviation means the greater variance between them. Therefore, the lower value shows better performance:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{11}$$

MAPE was used to measure the average relative error of the imputation accuracy:

$$MAPE = \frac{100}{N} \times \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{12}$$

MAE was used to measure the average error of each different in imputation:

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N} \left| y_i - \hat{y}_i \right| \tag{13}$$

R-squared values were used to describe the variance in goodness-of-fit for the regression models between observed data and the expected values of the dependent variable. The range of R-squared is between 0 and 1:

$$R^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{14}$$

3.2 Experimental results

In this study, we compare the imputation accuracy using MAPE and MAE whilst measuring the performance using RMSE and R-Squared of three methods: single regression imputation combined with Wu's and Liu's wild bootstrap and MI. The better results are highlighted in bold font.

Table 1 summarizes the performance and prediction accuracy of the three methods on the Old Faithful Geyser dataset while Table 2 shows the result estimation on the random data generation. The result of the proposed methods in RMSE shows better performance and significantly different between the MI with the proposed Wu's and Liu's method in all MDP proportions. This is shown in the 5% MDP, Wu and Liu method yielded 7.8225 and 7.8879 respectively while MI gained 9.8719. It is also found in 10%, 15% and 20% MDP where the Wu's and Liu's method have outperformed the MI where the result of Wu's shows 7.0955, 6.6819 and 6.7349 while Liu shows 7.8746, 7.0150 and 7.2354 in RMSE. In contrast, the MI obtained 8.4187, 8.7004 and 8.9103 higher than Wu's and Liu's method in 10%, 15% and 20% MDP respectively.

The R-squared values are used to quantify the overall model performance of variance in response variable explained by the independent variables. The larger the R-squared means the more variability is explained by the linear regression model. The result of R-squared presented in Table 1 showed that the proposed method gives the best performance with 0.6338% for 5% of MDP proportion followed by 0.6836, 0.7683 and 0.7127 for Wu, while Liu's obtained 0.6894, 0.6869 and 0.7050 for the 10%, 15% and 20% of MDP respectively on the Faithful data set. The R-squared obtained by the proposed method in the random generation data in the Table 2 showed less than 0.6% for all MDP percentages. In contrast, the MI in Amelia gives a lower variance than the proposed method in all MDP proportions with R-squared ranging from 0.03 to 0.2.

The imputation accuracy is measured based on the average relative error between predicted missing data and the original data using mean absolute percentage error (MAPE) and mean absolute error (MAE).

The result of MAE in the Table 1 showed that the Wu's and Liu's methods are consistently outperformed the MI method on the Old Faithful Geyser dataset. In contrast, in the Table 2, the Liu's method offered consistent and better accuracy than MI method. Meanwhile the Wu's method showed inconsistent improvement in the measure of average error magnitude to MI method on the random data generation.

As can be observed from the MAPE values obtained in Table 1, the proposed method of Wu's and Liu's performed better imputation on the Old Faithful Geyser data set. Meanwhile, by observing the MAPE values gained in the Table 2, Liu' method showed consistent to defeat the MI method compared to Wu's method.

A plot of the result shown in Figure 1 compare the outcome between multiple imputation technique in r-package Amelia II and the proposed methods.

Table 1. The MAPE, MAE, R-square and RMSE estimates on the Old Faithful Geyser dataset

		MAPE	MAE	R square	RMSE
5%	Amelia	0.6220	6.5928	0.4960	9.8719
	Wu	**0.2758**	**2.6453**	**0.6338**	**7.8225**
	Liu	**0.1713**	**1.7281**	**0.4212**	**7.8879**
10%	Amelia	0.0827	1.5636	0.6450	8.4187
	Wu	**0.0379**	**0.6959**	**0.6836**	**7.0955**
	Liu	**0.0190**	**0.3594**	**0.6894**	**7.8746**
15%	Amelia	0.0346	1.0379	0.5184	8.7004
	Wu	**0.0012**	**0.0342**	**0.7683**	**6.6819**
	Liu	**0.0167**	**0.5018**	**0.6869**	**7.0150**
20%	Amelia	0.0432	1.6841	0.4959	8.9103
	Wu	**0.0054**	**0.2144**	**0.7127**	**6.7349**
	Liu	**0.0104**	**0.3994**	**0.7050**	**7.2354**

Table 2. The MAPE, MAE, R-square and RMSE estimates on the randomly generated data

		MAPE	MAE	R square	RMSE
5%	Amelia	1.5798	0.7748	0.2439	2.0230
	Wu	**0.0469**	**0.1055**	**0.2593**	**1.8642**
	Liu	**0.0721**	**0.1604**	**0.3754**	**1.8066**
10%	Amelia	0.1259	0.1477	0.2206	2.2926
	Wu	0.1344	0.5811	0.3977	2.1557
	Liu	**0.0089**	**0.0377**	**0.5803**	**1.6889**
15%	Amelia	0.1674	0.3105	0.0272	2.2817
	Wu	**0.0272**	**0.1812**	**0.1316**	**2.3463**
	Liu	**0.0203**	**0.1282**	**0.5197**	**1.7214**
20%	Amelia	0.0501	0.1159	0.1206	2.7461
	Wu	**0.0435**	0.3528	**0.1954**	**2.1858**
	Liu	**0.0127**	**0.1070**	**0.4586**	**1.8340**

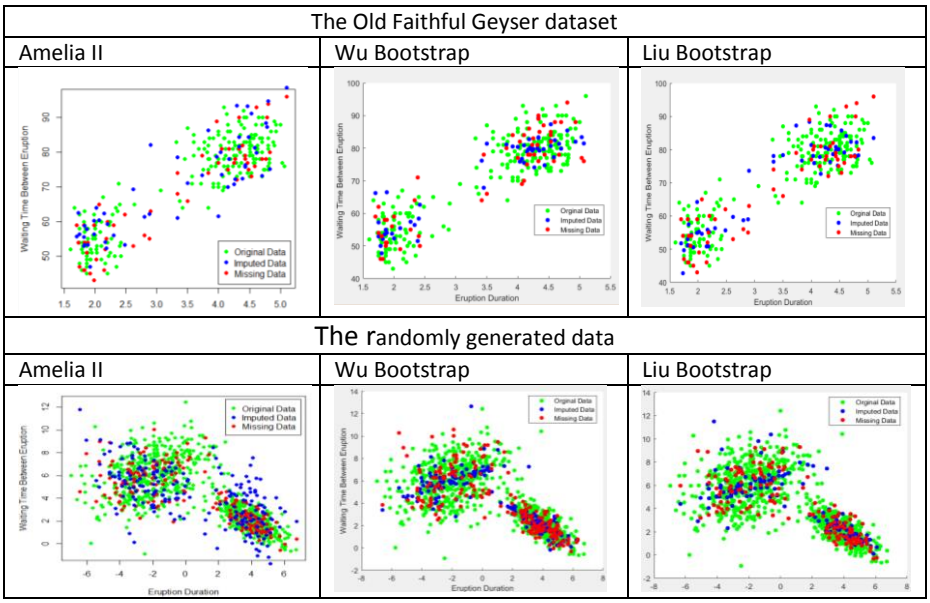

Fig. 1. The scatter plot of two datasets using R Amelia II and the proposed methods

4 Conclusions

In this paper, we proposed a method for single imputation that incorporates wild bootstrap in order to create the variability of imputed data as for example Multiple Imputation (MI) does. The MI is indeed known to be the preferred method in handling missing data problems over the years compared to the single imputation methods.

The imputation process in MI involves several steps while single imputation has simpler implementation compared to MI. The missing data in MI are imputed for M times with different plausible values and combine appropriately in the analysis stage. The sparsity of imputed data is a matter of concern because it will reflect the variance and measurement error between predicted and original data. Thus, the main purpose of this comparison is to show that the performance of single imputation in the Gaussian mixture model may perform well and as good as the implementation of MI.

The performance of this method is measured by the RMSE, R-squared, MAE, and MAPE. Based on the results, we summarize that the single missing data imputation combined with the wild bootstrap is preferrable over the MI technique for the data containing several Gaussian distributions. Furthermore, the imputation process on the Gaussian mixture model could be relevant to preserve the originality of data distribution.

Since this study is implemented on bivariate data with two Gaussian components, in the future work we will focus on multivariate data with multiple Gaussian components.

References

[1] D. B. Rubin, “Multiple imputations in sample surveys - A phenomenological Bayesian approach to nonresponse,” in *Proceedings of the Survey Research Methods Section of the American Statistical Association*, pp. 20–34(1978).

[2] J. Honaker, G. King, and M. Blackwell, “Amelia II: A program for missing data,” *J. Stat. Softw.* 45(7), 47(2006).

[3] G. McLachlan and D. Peel, *Finite Mixture Models*. Wiley, New York(2000).

[4] Z. Ghahramani and M. I. Jordan, “Supervised Learning from Incomplete Data via an EM Approach,” in *Advances in Neural Information Processing Systems*.6,120–127(1994).

[5] M. Di Zio, G. Ugo, and O. Luzi, “Imputation through finite Gaussian mixture models,” *Comput. Stat. Data Anal.* 51(11), 5305–5316(2007).

[6] M. Paik, “Fractional Imputation,”(Unpublished Doctoral Thesis),Iowa State University,USA (2009).

[7] M. S. Srivastava and M. Dolatabadi, “Multiple imputation and other resampling schemes for imputing missing observations,” *J. Multivar. Anal.* 100(9),1919–1937 (2009).

[8] C. F. J. Wu, “Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis,” *Ann. Stat.* 14(4), 1261–1295(1986).

[9] R. Y. Liu, “Bootstrap procedures under some non-i.i.d. models,” *Ann. Stat.*16(4),1696–1708(1988).

[10] A. P. Dempster, A. P. Dempster, N. M. Laird, N. M. Laird, D. B. Rubin, and D. B. Rubin, “Maximum likelihood from incomplete data via the EM algorithm,” *J. R. Stat. Soc. Ser. B*, 39(1),1–38(1977).

[11] E. Eirola, A. Lendasse, V. Vandewalle, and C. Biernacki, “Mixture of Gaussians for distance estimation with missing data,” *Neurocomputing*, 131, 32–42(2014).

[12] J. S. J.N.K. Rao, “Jackknife Variance Estimation with Survey Data Under Hot Deck Imputation,” *Biometrika Trust*, 79(4),811–822(1992).

[13] C. K. Enders, *Applied missing data analysis*.The Guilford Press,New York(2010).

[14] C. R. Harvey, “The specification of conditional expectations,” *J. Empir. Financ.*, 8(5), 573–637(2001).

[15] B. Efron, “Bootstrap Methods : Another Look at the Jackknife,” *Ann. Stat.*7(1),1–26 (1979).

[16] E. Hardle, W. and Mammen, “Comparing Nonparametric versus Parametric Regression Fits,” *Ann. Stat.*21,1926–1947(1993).

[17] M. Zadkarami, “Bootsrapping: A Nonparametric Approach to Identify the Effect of Sparsity of Data in the Binary Regression Models,” *J. Appl. Sci.*8(17), 2991–2997 (2008).

[18] F. Cribari-Neto and S. G. Zarkos, “Bootstrap methods for heteroskedastic regression models: evidence on estimation and testing,” *Econom. Rev.*18(2), 211–228(1999).

[19] A. Azzalini and A. W. Bowman, “A Look at Some Data on the Old Faithful Geyser,” *J. R. Stat. Soc.*, 39(3),357–365(1990).

Appendix A: The notation list.

$\mathbf{X}$	An entire random sample of size N and p-column.
$\mathbf{X}^O$	The observed values of random vector $\mathbf{X}$
$\mathbf{X}^M$	The p-feature vectors $\mathbf{X}$ contains missing values occur in $\mathbf{X}_l$
n_1	The number of observed data in $\mathbf{X}^O$
n_0	The number of missing data in $\mathbf{X}^M$
K	Total number Gaussian of components
$\boldsymbol{\theta}_k$	Parameter theta that consists of parameter mean vector $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$
$\boldsymbol{\mu}_k$	The mean vector
$\boldsymbol{\Sigma}_k$	The covariance matrix
π	Mixing proportion of the current Gaussian component
$0 \le \pi_k \le 1,$	The probability of mixing coefficient must be between 0 and 1.
$\sum_{k=1}^{K} \pi_k = 1$	The sum of mixing coefficient of each component must be equal to one
$\Phi = (\pi_1,...,\pi_K;\boldsymbol{\theta}_1,...,\boldsymbol{\theta}_K)$	The vector Φ containing the set of parameters π_K and $\boldsymbol{\theta}_K$
$f(\mathbf{x};\Phi)$	The mixture density containing all the parameters of mixture model
$f(\mathbf{x} \mid \theta_k)$	The mixture of density function of vector $\mathbf{X}$ conditioned on parameter estimation theta.
$f(\mathbf{x};\Phi) = \sum_{k=1}^{K} \pi_k f(\mathbf{x} \mid \theta_k)$	The probability density function governed by the set of parameters mixing coefficient and theta with K-component mixture density
τ_{ik}	The posterior probability or responsibility for each data point that belongs to the k^{th} component
β_0	Beta 0 is represented as the intercept of regression coefficient
β_1	Beta 1 is represented as the slope of regression coefficient

$x_{il} = \beta_0 + \beta_1 x_i + \varepsilon_i$	The ordinary leas square model (OLS)
$\hat{\boldsymbol{\beta}}^O = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{X}_l$	The least square estimator of $\hat{\boldsymbol{\beta}}^O$
$\varepsilon_i \sim N(0,\sigma^2)$	The random error component follows he normal distribution with mean 0 and variance
$\mathbf{X}^{(b_k)}$	The sample data after bootstrapping
$\mathbf{X}_k^O$	The observed data based on the k current Gaussian component
$b = 1,\ldots,B$	b is a counter value for bootstrap iterative process until B times
$\boldsymbol{t}_i^*$	The non-parametric bootstrap resampled residual $\hat{\varepsilon}_i = x_{il} - \hat{x}_{il}$
$w_i = x_i^T(\mathbf{X}^T\mathbf{X})x_i$	The leverage is the i^{th} diagonal element of Hat Matrix
$a_i = \frac{\hat{\varepsilon}_i - \bar{\hat{\varepsilon}}_i}{\sqrt{n_{1k}^{-1}\sum_{i=1}^{n_{1k}}(\hat{\varepsilon}_i - \bar{\hat{\varepsilon}})^2}}$	The non-parametric bootstrap resampled residual proposed to improve the random draw of Wu's algorithm
D_1 and D_2	The i.i.d that follows normal distribution
$\mathbf{X}_{n1_k}^O$	The observed data of n_1 size and k Gaussian component
$\hat{\varepsilon}_i = x_{il} - \hat{x}_{il}$	The estimated residual fitted by OLS model
q	A counter in EM algorithm iteration