REFINING MULTIVARIATE NORMAL IMPUTATIONS TO ACCOMMODATE NON-NORMAL DATA

Juwon Song[1] and Thomas R. Belin[2]

[1]*Department of Biostatistics and Applied Mathematics, The University of Texas M.D. Anderson Cancer Center, 1515 Holcombe Blvd, Box 447, Houston, TX 77030, USA*
jwsong@mdanderson.org

[2]*Department of Biostatistics, UCLA School of Public Health, 51-267 Center for Health Sciences, Los Angeles, CA 90095*

KEY WORDS: Multiple Imputation, Multivariate Normal Distribution, Importance Sampling, SIR Algorithm

Abstract

An MCMC algorithm based on a multivariate normal distributional assumption provides the basis for widely available statistical software (such as in SAS and S-PLUS) for conducting multiple imputation. However, when data do not fit well with the multivariate normal distribution, this technique may introduce biased estimates. Here we adapt the sampling-importance-resampling (SIR) algorithm (Rubin 1987a) to perform multiple imputation by first generating imputations based on a multivariate normal distribution and then refining the values drawn in the first stage using importance resampling, making use of a more realistic distributional assumption. We first show the feasibility of the method in a simple example where missing values are missing completely at random (MCAR). We discuss the complexity of adapting the method to the more plausible situation where missingness is not MCAR but may be missing at random (MAR). We then outline some potential extensions of the SIR idea that suggest useful avenues to explore.

1. Introduction

When data are not fully observed, missing values often cause difficulty in the analysis. For incomplete data, one of the simplest way to handle missing values is complete-case analysis, analyzing only cases with all variables measured. However, it is well known that this method is inefficient even when missing data are missing completely at random (MCAR) and may result in biased estimates when missing data are missing at random (MAR) (Little and Rubin, 2002; Little, 1992). Imputation is a technique to fill in plausible values for each missing item. It is well known that single imputation, i.e., imputing a single plausible value for each missing item, predictably underestimates variances. Therefore, multiple imputation (Rubin, 1987b), a technique for imputing more than one plausible value for each missing item to reflect uncertainty about those items, has become a standard approach for handling missing data, as it is known to provide valid inferences under a much broader range of conditions.

Even though there are other parametric and nonparametric imputation techniques, multiple imputation is often conducted under a multivariate normal distributional assumption because of its conceptual simplicity and ease of application. Popular commercial statistical software such as SAS and S-PLUS provides tools for multiple imputation based on the multivariate normal distribution. However, data sets often include different types of variables, and variables are often distributed non-normally. When variables are not normally distributed, transformations are often recommended, but for some variables, there may be no transformation available to achieve normality.

Figure 1 shows an example of a highly skewed variable. For example, if we observe the number of hospital visits for the past one month, many people might report zero visits while some people report a much larger number of visits. Since almost 70% of responses are zero, no simple transformation is available to transform this variable to normality. This type of data might be analyzed using a Poisson model or using a two-part model where a logistic model is used to predict zero vs. non-zero and a companion assumption is made for the distribution of the positive values.

Another type of the non-normally distributed variable is shown in Figure 2. In questionnaires used in behavioral science research, items are often measured with Likert scales. For some of these items, participants might tend to prefer extreme choices (for example, “strongly agree” or “strongly disagree”) than a choice from the center of the distribution (for example, “neutral”), with the resulting distribution having peaks at the extremes.

Here, we examine the effect of multiple imputation based on a normal distributional assumption for non-

normally distributed variables. We then propose a refined method of multiple imputation using the transformation and the sampling-importance-resampling (SIR) algorithm. Section 2 describes the algorithm, and Section 3 shows simulation results applied to hypothetical data sets. Discussion follows with a summary and directions for further research.

Figure 1. An example of a highly skewed variable.

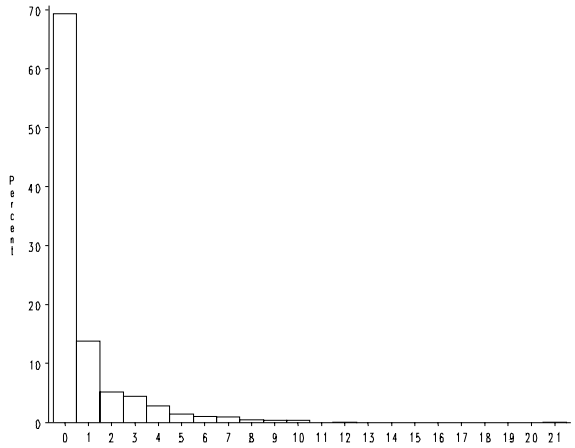
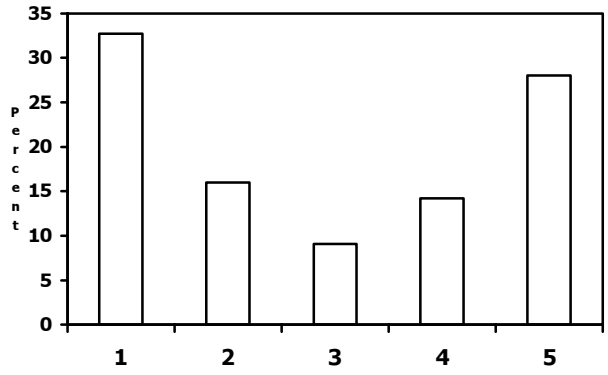


Figure 2. An example of non-normally distributed variable.



2. Algorithm

Even when data include both normally and non-normally distributed variables, multiple imputation is often conducted using a multivariate normal distributional assumption because of its ease of application. Therefore, we consider a strategy where multiple imputation would be conducted under a multivariate normal distributional assumption in a first stage, after which there would be an effort to refine the distribution of imputed values to be similar to the true underlying distribution. The algorithm consists of the following three steps:

(1) Conduct multiple imputation based on a multivariate normal model,

(2) Transform imputed values to fit the model assumption better,

(3) Apply the sampling-importance-resampling (SIR) algorithm to refine imputed values.

These steps are described in detail in following subsections.

2.1. Step 1: Multiple imputation based on a multivariate normal model

Consider a data set, Y. Following Rubin (1987b), we denote observed elements of Y as Y_{obs} and unobserved elements as Y_{mis}, so that the data can be written as $Y = (Y_{obs}; Y_{mis})$. Let's also denote $f(Y|\theta)$ the density of Y indexed by unknown parameters θ. Multiple imputation can be conducted by data augmentation (Tanner and Wong, 1987) that impute missing values using the following two steps iteratively:

With an initial value of θ, $\theta^{(0)}$, and for t =1, 2,…,

I-step: simulates missing values from its conditional predictive distribution, $Y_{mis}^{(t)} \sim P\left(Y_{mis} \middle| Y_{obs}, \theta^{(t-1)}\right)$, at the iteration t,

P-step: simulates the parameters from the complete-data posterior distribution, $\theta^{(t)} \sim P\left(\theta \middle| Y_{obs}, Y_{mis}^{(t)}\right)$, at the iteration t.

Iterating the I-step and P-step yields a stochastic sequence of missing values and parameters, which can be expected to converge in settings where data and prior assumptions are sufficient to identify model parameters.

When data follow the multivariate normal distribution, the conditional predictive distribution is also normal, and I-step can be represented as

$$Y_{mis}^{(t)} \sim N\left(b_0^{(t-1)} + \sum_{k \in O(s)} b_k^{(t-1)} Y_k, R_{Y_{mis}|Y_{obs}}^{(t-1)}\right),$$

where $b_0^{(t-1)}$, $b_k^{(t-1)}$, and $R_{Y_{mis}|Y_{obs}}^{(t-1)}$ indicate the intercept, slope, and residual variance of the regression of Y_{mis} on Y_{obs} at iteration t, and $O(s)$ indicates observed subsets of Y.

2.2. Step 2: Transformation of imputed values

Even when a variable with missing values does not follow the normal distribution, imputations based on the multivariate normal data will add a normal error

term. To show the effect of this, we examined the scenarios represented in Figures 1 and 2. Figures 3 and 4 compare the distributions of observed values with the distributions of imputed values. In Figure 3, zero values show a high peak among imputed values because values imputed with negative values were truncated to zero. Among positive values, the imputed are consistent with a bell-shaped pattern coupled with truncation, while observed values are far from normal. Similarly, Figure 4 shows that imputed values are quite uniformly distributed due to the truncation of imputed values to the minimum and maximum plausible values, and the distribution of the imputed values are again far from the distribution of the observed values.

Figure 3. Comparison of observed values and imputed values based on the multivariate normal distribution (Data from Figure 1).

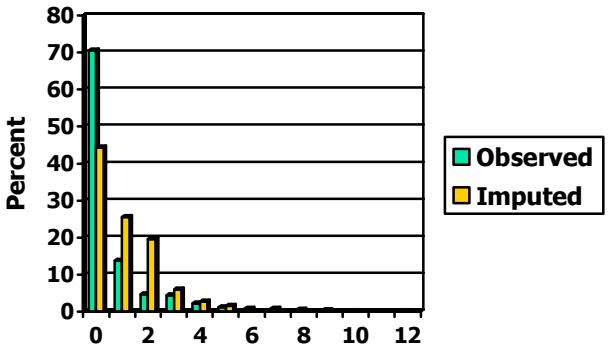
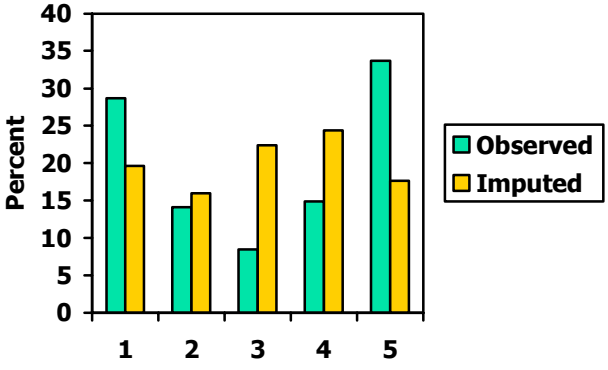


Figure 4. Comparison of observed values and imputed values based on the multivariate normal distribution (Data from Figure 2).



To overcome this lack of fit, we tried an ad-hoc transformation of the imputed values to have a similar distribution with the observed values. When missing data mechanism can be assumed to be missing completely at random (MCAR) (Little and Rubin, 2002), the transformation is simple. Since incomplete data can be considered as a random sample of complete data, the distribution of the observed data can serve as a reference distribution. When a variable with missing values follows a non-normal parametric distribution, we can (1) store the parameters of Y, $\theta^{(m)} = (\hat{\mu}^{(m)}, \hat{\sigma}^{2(m)})$, at imputation m, (2) calculate the probability of each imputed value under the normal distribution cdf with $\theta^{(m)}$, (3) calculate parameters of an appropriate distribution using these parameters, and (4) transform imputed values using an inverse cdf of the true distribution evaluated at the probability obtained at (2). For example, if a variable follow a gamma distribution, $G(\alpha,\beta)$, α and β, are approximated with $\hat{\beta} = \hat{\sigma}^{2(m)} / \hat{\mu}^{(m)}$ and $\hat{\alpha} = \hat{\mu}^{(m)} / \hat{\beta}$.

Another simpler approach incorporates an empirical distribution of Y_{obs}: (1) store the parameters of Y, $\theta^{(m)} = (\hat{\mu}^{(m)}, \hat{\sigma}^{2(m)})$, at imputation m, (2) calculate the probability of each imputed value under the normal distribution cdf with $\theta^{(m)}$, and (3) transform imputed values to follow the non-normal distribution using the inverse cdf of the observed data.

When the missing data mechanism is missing at random (MAR), the transformation can be extended to consider other covariates. Under the multivariate normal distribution assumption, I-step was

$$y_{i,mis} \sim N\left(b_0 + \sum_{k \in O_i(s)} b_k y_{ik}, R_{y_{i,mis}|y_{i,obs}}\right),$$

where $y_{i,\text{mis}}$ indicates unobserved variables in ith observation of Y, $y_{i,\text{obs}}$ indicates observed variables in ith observation of Y, and $O_i(s)$ indicates the set of observed variables in ith observation of Y. Since imputed values at each observation are draws from the multivariate normal distribution, it can be transformed to a non-normal parametric distribution using its mean (predicted values) and variances (residual terms). Multiple imputation usually outputs $\theta^{(m)} = (\hat{\mu}^{(m)}, \hat{\sigma}^{2(m)})$, but they can be easily transformed as the estimates of regression parameters of Y_{mis} on Y_{obs}, b_k s and $R_{y_{i,mis}|y_{i,obs}}$ using the sweep algorithm.

2.3. Step 3: Application of the sampling-importance-resampling (SIR) algorithm

The Sampling Importance Resampling algorithm (Rubin, 1987a) suggests using the following four steps to produce multiply imputed data sets:

Step 1: Select an approximation to the joint posterior density. When the joint posterior density of (θ, Y_{mis}), $p(\theta, Y_{mis})$, is hard to evaluate, there might exist an approximation, $h(\theta, Y_{mis})$, that can be easily evaluated.

Step 2: M values of θ and Y_{mis} are randomly drawn from the approximated joint posterior density of $h(\theta, Y_{mis})$, where M is much larger than the target number of imputations, m.

Step 3: In each M set of drawn values, the importance ratio for each $(\theta^{(j)}, Y_{mis}^{(j)})$, $j=1,\ldots, M$., can be calculated.

Step 4: Draw m values of Y_{mis} with probability proportional to the importance ratio.

Here, the importance ratio, $r(\theta^{(j)}, Y_{mis}^{(j)} \mid Y^{(j)}) \propto p(\theta^{(j)}, Y_{mis}^{(j)} \mid Y^{(j)})\, p(\theta^{(j)}) / h(\theta^{(j)}, Y_{mis}^{(j)} \mid Y^{(j)})$, indicates a ratio of the true vs. approximated posterior density of θ and Y_{mis}. When $p(Y_{mis} \mid \theta, Y)$ is tractable, importance ratios can be calculated by $r(\theta^{(j)}, Y_{mis}^{(j)} \mid Y^{(j)}) \propto p(Y^{(j)} \mid \theta^{(j)})\, p(\theta^{(j)}) / h(\theta^{(j)} \mid Y^{(j)})$, and the SIR algorithm could be streamlined.

Figure 5. Comparison of observed values and imputed values refined with the transformation / SIR algorithm (Data from Figure 1).

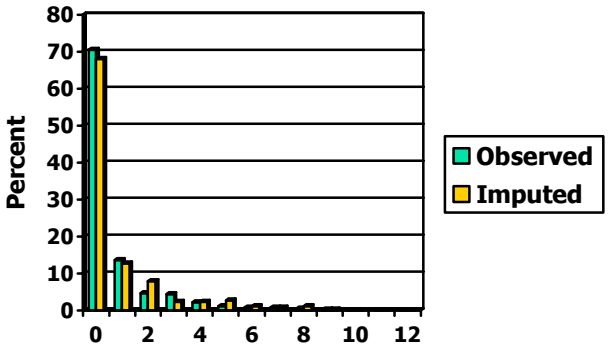
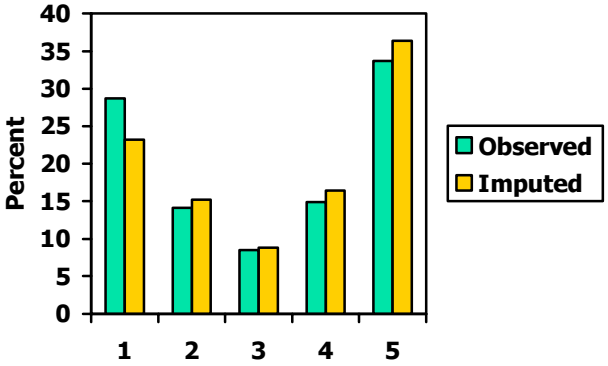


Figure 6. Comparison of observed values and imputed values refined with the transformation / SIR algorithm (Data from Figure 2).



While this algorithm has been often applied to approximate the density of θ, we consider a different situation here. We approximate the conditional density of Y_{mis} with an approximate normal distribution. In the simplest case, we streamline the process by assuming the same parameter distribution. This indicates that we are interested in the posterior change when the likelihood changes.

Figures 5 and 6 show the comparison of the distributions between observed values and imputed values after applying the SIR algorithm. The distribution of the imputed values is now very similar to the distribution of the observed values.

3. Simulation

We conducted a simulation to evaluate the performance of the proposed method. This method is compared with unadjusted multiple imputation and complete data without missing values. Three variables were generated. We denote the data as Y, and the three variables as y_1, y_2, and y_3. The first two variables were assumed to follow independent standard normal distributions: $(y_1, y_2) \sim N_2(\underline{1}, I)$, where $\underline{1}$ indicates a two dimensional vector of 1 and I is an 2×2 identity matrix. To generate a variable from a non-normal distribution, the third variable was generated from a gamma distribution with parameters related to the first two variables as follows:

$$y_3 \sim \text{Gamma}(\alpha, \beta),$$

$$\text{where } \alpha = c_1 (y_1 + y_2)^{(-2/3)} (b_0 + b_1 y_1 + b_2 y_2)^2$$

$$\text{and } \beta = c_2 (y_1 + y_2)^{(2/3)} (b_0 + b_1 y_1 + b_2 y_2)^{-1}.$$

Three different choices of α and β were considered to examine the effect of multiple imputation under various types of skewness. We chose $c_1 = 0.1$, $c_2 = 1.0$, $b_1 = 3.0$, and $b_2 = 0.1$, and the value of b_0 varied to represent different skewness. The first scenario set the value of b_0 to –1, the second scenarios to -5, and the third scenario to –7. Then, y_3 was rounded off to be an integer. Figures 7-9 shows the distribution of a simulated data set under three scenarios, respectively. (Note that Figure 9 is identical to Figure 1).

The sample size was assumed to be 1000. We deleted 25% of y_3 completely at random. The data were first imputed five times under the multivariate normal distribution assumption and multiple imputation inference was used to summarize the findings. Then, the refined procedure was applied using a transformation and the SIR algorithm. We generated one hundred imputed data sets and chose five data sets

based on the SIR algorithm. This comparison was repeated 1000 times.

Tables 1 shows the results comparing multiple imputation based on the multivariate normal distribution with the suggested refinement. To evaluate bias, the mean of y_3 from data before deletion was also included. The mean and SE indicate the average and standard error of y_3 from 1000 simulated data sets, and the coverage indicates the percentage with which the 95% confidence interval of x_3 from 1000 imputed data sets covered the true mean. When skewness was not severe (in Scenario 1), both imputation methods showed good coverage with the average close to 1 from data before deletion. When skewness got severe, bias in means was seen in multiple imputation based on the multivariate normal distribution, but the refined imputation based on the SIR algorithm showed a sample mean close to the true mean. Coverage was also low for multiple imputation based on the multivariate normal distribution, while refining the procedure with SIR produced satisfactory coverage rates. The standard error was slightly bigger in refining with SIR than multiple imputation based on the multivariate normal distribution in all three scenarios.

Table 1. The mean, standard deviation, and coverage of y_3.

Scenario	Method	Mean	SE	Coverage
1	Before deletion	1.414		
	Normal	1.472	0.064	97.8
	SIR refined	1.403	0.069	99.9
2	Before deletion	1.010		
	Normal	1.095	0.062	85.2
	SIR refined	1.001	0.069	99.8
3	Before deletion	0.816		
	Normal	0.912	0.061	76.6
	SIR refined	0.807	0.069	99.9

4. Discussion

Multiple imputation under the multivariate normal distribution assumption is easy to implement with commercial software, but it requires careful consideration of the distributional assumption and whether any transformation is needed. Refining multiple imputation based on the normal distribution assumption to incorporate non-normal data might be able to serve as a helpful tool in some cases that transformations to the normal distribution do not work well.

Figure 7. Distribution of y_3 from a simulated data set under Scenario 1.

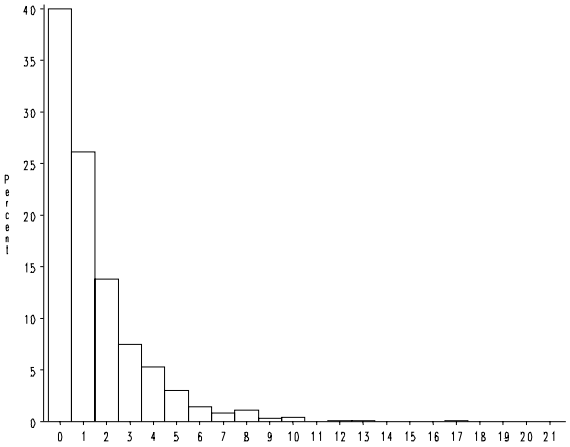


Figure 8. Distribution of y_3 from a simulated data set under Scenario 2.

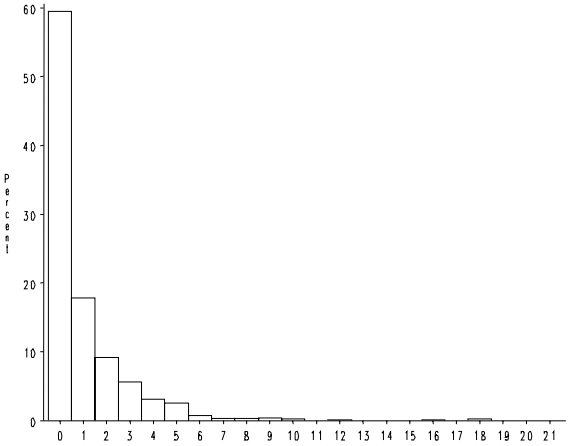
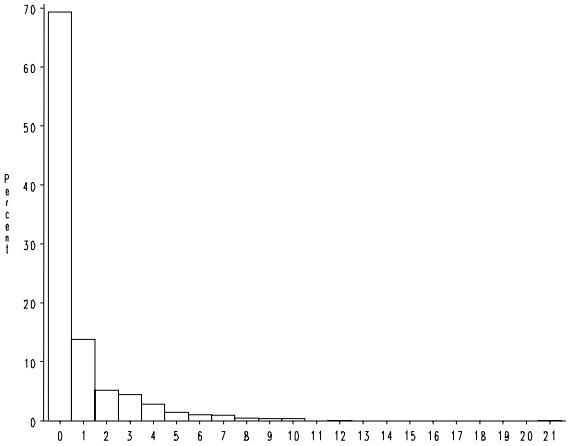


Figure 9. Distribution of y_3 from a simulated data set under Scenario 3.



When the missing data mechanism is MCAR and the variables do not follow a normal distribution, but the imputation model assumes a normal distribution, it may be that the mean is not seriously biased with moderate percentage of missing values. However, when skewness gets severe, the estimation of the mean from multiple imputation based on the multivariate normal distribution assumption might have substantial bias. Moreover, the distribution of the imputed values might be different to the distribution of the true values. Refining imputed values using the SIR algorithm would help making the distribution of imputed values similar to that of the observed values. However, as noted in the simulation, their standard errors might be slightly bigger than under a multivariate normal assumption.

When the missing data mechanism is MAR, implementation can be extended by separating the prediction and the residual term from imputed values, and adjust them using a transformation and the SIR algorithm.

On the other hand, there are alternative multiple imputation algorithms often useful to data sets with various types of variables. Some examples include hot-deck imputation and sequential regression multivariate imputation (Raghunathan et. al, 2001). Comparison of the suggested refinement with those methods would be interesting.

References

Little, R. J. A. (1992), "Regression With Missing X's: A Review," *Journal of the American Statistical Association*, 87, 1227-1237.

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data*, New York: John Wiley.

Raghunathan, T. E., Lepkowski, J. M., Hoewyk, J. V., and Solenberger, P. (2001). A Multivariate Technique for Multiply Imputing Missing Values Using a Sequence of Regression Models, *Survey Methodology*, 27, 85-95.

Rubin, D. B. (1987)[a]. Comment on "A Noniterative Sampling/Importance Resampling Alternative to the Data Augmentation Algorithm for Creating a Few Imputations When Fractions of Missing Information are Modest: The SIR Algorithm" by Tanner and Wong, *Journal of the American Statistical Association* **82**, 543-546.

Rubin, D. B. (1987)[b]. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.

Tanner, M.A., and Wong, W. H. (1987), "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of the American Statistical Association*, 82**,** 528-550.