

**University of Illinois at Chicago
School of Public Health
Division of Epidemiology and Biostatistics**

*Technical report#:2006-001
February 2006*

Title: The Plausibility of Multivariate Normality Assumption
When Multiply Imputing Non-Gaussian Continuous
Outcomes: A Simulation Assessment

Author(s): Hakan Demirtas, Sally A. Freels and Recai M. Yucel

**Affiliation(s): University of Illinois at Chicago, Division of Epidemiology and
Biostatistics; University of Massachusetts at Amherst, Department of Biostatistics
and Epidemiology**

The Plausibility of Multivariate Normality Assumption When Multiply Imputing Non-Gaussian Continuous Outcomes: A Simulation Assessment

Hakan Demirtas, Sally A. Freels and Recai M. Yucel *

February 17, 2006

Abstract

Multiple imputation under the assumption of multivariate normality has emerged as a frequently used model-based approach in dealing with incomplete continuous data in recent years. Despite its simplicity and popularity, however, its plausibility has not been thoroughly evaluated via simulation. In this work, the performance of multiple imputation under a multivariate Gaussian model with unstructured covariances was examined on a broad range of simulated incomplete datasets that exhibit varying distributional characteristics such as skewness and multimodality that are not accommodated by a Gaussian model. Behavior of efficiency and accuracy measures was explored to determine the extent to which the procedure works properly. The conclusion drawn is that although the real data rarely conform with multivariate normality, imputation under the assumption of normality is a fairly reasonable tool, even when the assumption of normality is clearly violated; and the fraction of missing information is high, especially when the sample size is relatively large. While we discourage its uncritical, automatic and possibly inappropriate use, we report that its performance is better than we expected, leading us to believe that it is probably an underrated approach.

Key Words: Multivariate normality; Multiple imputation; Symmetry; Skewness; Multimodality.

1 Introduction and motivation

Missing data is the norm rather than the exception in most datasets. Determining a suitable analytical approach in the presence of incomplete observations is a major focus of scientific inquiry due to the additional complexity that arises through missing data. Incompleteness generally complicates the statistical analysis in terms of biased parameter estimates, reduced statistical power and degraded confidence intervals, and thereby may lead to false inferences (Little and Rubin, 2002).

Advances in computational statistics have produced flexible missing-data procedures with a sound statistical basis. One of these procedures involves multiple imputation (MI) (Rubin, 1987), a simulation

*Hakan Demirtas (e-mail:demirtas@uic.edu) is an Assistant Professor and Sally A. Freels is an Associate Professor of Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL 60612. Recai Yucel is an Assistant Professor of Biostatistics, Biostatistics and Epidemiology, School of Public Health and Health Sciences, 410 Arnold, University of Massachusetts-Amherst, MA 01003

technique that replaces each missing datum with a set of $m > 1$ plausible values. The m versions of complete data are then analyzed by standard complete-data methods and the results are combined into a single inferential statement using arithmetic rules to yield estimates, standard errors and p-values that formally incorporate missing data uncertainty into the modeling process. The key ideas and advantages of MI were reviewed by Rubin (1996) and Schafer (1999).

The essential step in MI is filling in the missing data by drawing from the conditional distribution of the missing data given the observed data which typically involves positing a parametric model for the data and using it to derive this conditional distribution. Joint multivariate normality (MVN) among the variables in the dataset has often been perceived as a natural assumption since the conditional distribution of the missing data given the observed data will then also be multivariate normal.

To the best of our knowledge, despite its frequent use, the plausibility of the MVN assumption for continuous responses has not been thoroughly evaluated via a broad range of simulated examples representing the situations such as multimodality, skewness, non-zero peakedness, heavy tails, and flatness of the density that clearly violate the characteristics of MVN. This article examines how MI under MVN performs under different underlying distributional properties, sample sizes and fractions of missing information.

Organization of this paper is as follows. The next section provides some key operational attributes of MI under the MVN assumption along with a brief coverage of computational routines. In Section 3, a fairly comprehensive simulation study is presented. Section 4 includes concluding remarks and discussion.

2 Imputing under multivariate normality assumption

To set the notation, the dataset is assumed to be a matrix of n rows and p columns, with rows corresponding to units and columns corresponding to variables. We denote the complete data by $Y = (Y_{obs}, Y_{mis})$, where Y_{obs} and Y_{mis} stand for the observed and missing portions of the matrix, respectively. Suppose that the distribution of Y depends on a set of unknown parameters of interest θ . Let R be the associated set of missing-value indicators. The elements of R take the values 1 or 0, indicating whether the corresponding elements of Y are observed or not, respectively. The conditional distribution of R given Y depends on the set of parameters γ . Let (y_{obs}, r) be the realized value of (Y_{obs}, R) .

The properties of missing-data methods vary depending on the manner in which data became missing; every missing-data technique makes implicit or explicit assumptions about the missing-data mechanism. Many missing-data procedures in use today assume that missing values are missing at random (MAR) (Rubin, 1976). The missing values are said to be MAR if $P(R = r | Y_{obs} = y_{obs}, Y_{mis}; \gamma) = P(R = r | Y_{obs} = y_{obs}; \gamma)$ holds for all possible γ . Under MAR, the probability distribution of the indicators of missingness may depend on the observed data but must be functionally independent of the missing data. An important special case of MAR is missing completely at random (MCAR). Under MCAR, $P(R = r | Y_{obs} = y_{obs}, Y_{mis}; \gamma) = P(R = r; \gamma)$ for all possible γ . In this case, the response probabilities are independent of both the observed and the unobserved parts of the dataset. If MAR is violated, the response probabilities depend on unobserved data; in this case, the missing values are said to be missing not at random (MNAR). MNAR situations require special care; to obtain correct inferences, one must specify a joint probability model for the complete data and the indicators of missingness.

A missing-data mechanism is said to be ignorable if (a) the missing data are MAR and (b) the parameters γ and θ are distinct (Little and Rubin, 2002). The term ignorable suggests that the missing-data mechanism can, in some sense, be ignored when performing statistical analyses. Rubin (1976) precisely explained what it means to ignore the missing-data mechanism, both from frequentist and likelihood/Bayes standpoints, and provided conditions under which ignoring the missing-data mechanism is valid for inferences about θ . In the frequentist case, ignoring the missing-data mechanism means fixing R at its realized value and using $P(Y_{obs} | R = r; \theta, \gamma)$ as a repeated-sampling distribution. That is, we pretend that Y_{obs} are the data that we had intended to collect. In the likelihood/Bayes situation, ignoring the missing-data mechanism means using $L(\theta | y_{obs}) = \int P(Y_{obs} = y_{obs}, Y_{mis}; \theta) dY_{mis}$ as the likelihood function for θ . The conditions under which these approaches are valid differ. In the likelihood/Bayes case, ignoring the missing-data mechanism is valid when θ and γ are distinct and the missing data are MAR. In the frequentist case (ad hoc approaches such as case deletion or model-based approaches such as Generalized Estimating Equations), the stronger condition of MCAR is needed. This definition of ignorability implicitly assumes that one is working within a likelihood-based or Bayesian context. The reason why the missing-data mechanism can be ignored under this condition is that the joint log-likelihood for θ and γ partitions as $l(\theta, \gamma; y_{obs}, r) = l(\theta; y_{obs}) + l(\gamma; r)$.

Information about the complete-data population parameter θ is contained fully in the first term; inferences about θ are unaffected by R , and there is no need to model $P(R = r|y, \gamma)$.

After reviewing the fundamentals of missing data, we now describe the key characteristics of multiple imputation (MI) under the MVN assumption. MI is a Monte Carlo technique (Rubin 1987, 1996) in which the missing values are replaced by a set of $m > 1$ simulated versions of them. These simulated values are drawn from a Bayesian posterior predictive distribution for the missing values given the observed values and the missingness indicators. Carrying out MI requires two sets of assumptions. First, one must propose a model for the data distribution which should be plausible and should bear some relationship to the type of analysis to be performed. The second set of assumptions pertains to type of missingness mechanism. An assumption of MAR is commonly employed for MI. However, the theory of MI does not necessarily require MAR; MI may also be performed under nonignorable models (Demirtas and Schafer, 2003; Demirtas, 2005). For the purposes of this article, we assume ignorable nonresponse.

The key idea of MI is that it treats missing data as an explicit source of random variability to be averaged over. The process of creating imputations, analyzing the imputed datasets, and combining the results is a Monte Carlo version of averaging the statistical results over the predictive distribution of the missing data, $\int P(\theta|Y) P(Y_{mis}|Y_{obs}) dY_{mis}$. In practice, a large number of multiple imputations is not required; sufficiently accurate results can often be obtained with $m \leq 10$. Once the imputations have been created, the m completed datasets may be analyzed without regard for missing data; all relevant information on nonresponse is now carried in the imputed values. Once the quantities have been estimated, the m versions of the estimates and their standard errors are combined by simple arithmetic as described by Rubin (1987).

Let y_{ij} denote an individual element of Y , $i = 1, 2, \dots, n$, $j = 1, 2, \dots, p$. The i^{th} row of Y is $y_i = (y_{i1}, y_{i2}, \dots, y_{ip})^T$. Assume that y_1, y_2, \dots, y_n are independent realizations of a random vector, denoted as (Y_1, Y_2, \dots, Y_p) , which has a multivariate normal distribution with mean vector μ and covariance matrix Σ ; that is $y_1, y_2, \dots, y_n | \theta \sim N(\mu, \Sigma)$, where $\theta = (\mu, \Sigma)$ is the unknown parameter and Σ is positive definite. The complete-data likelihood with this setting is proportional to $|\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{1}{2} \sum_{i=1}^n (y_i - \mu)^T \Sigma^{-1} (y_i - \mu)\right\}$. The maximum likelihood estimators for μ and Σ are well-known: $\hat{\mu} = \bar{y} = n^{-1} \sum_{i=1}^n y_i$ and $\hat{\Sigma} = S = n^{-1} \sum_{i=1}^n (y_i - \bar{y})(y_i - \bar{y})^T$. When imputations are created under Bayesian arguments, MI has a natural

interpretation as an approximate Bayesian inference for the quantities of interest based on the observed data. MI can be performed by first running an EM-type algorithm (Dempster, Laird and Rubin, 1977), and then by employing a data augmentation procedure (Tanner and Wong, 1987), as implemented in some software packages (e.g. Splus). The EM algorithm is useful for two reasons: it provides good starting values for the data augmentation scheme, and it gives us an idea about the convergence behavior. Data augmentation using the Bayesian paradigm has been perceived as a natural tool to create multiply imputed datasets. For further details, see Schafer (1997) and Schimert et al. (2001). Below, we give a brief description of the MI process using data augmentation.

When both μ and Σ are unknown, the conjugate class for the multivariate normal data model is the normal inverted-Wishart family. When a $p \times p$ matrix X has an inverted-Wishart density ($W^{-1}(k, \Gamma)$) with degrees of freedom parameter k and inverse-scale parameter Γ , the density is proportional to $|X|^{-\frac{(k+p+1)}{2}} \exp\{-\frac{1}{2}\text{tr}(\Gamma^{-1}X^{-1})\}$ for $k \geq p$. Bayesian inference for $\theta = (\mu, \Sigma)$ proceeds with the formulation of prior distributions: Suppose that $\mu|\Sigma \sim N(\mu_0, \tau^{-1}\Sigma)$, where the hyperparameters μ_0 and $\tau > 0$ are fixed and known; and $\Sigma \sim W^{-1}(k, \Gamma)$, where $p \leq k$ and $\Gamma > 0$. The prior density for θ is then $f(\theta) \propto |\Sigma|^{-\frac{(k+p+2)}{2}} \exp\{-\frac{1}{2}\text{tr}(\Gamma^{-1}\Sigma^{-1})\} \exp\{-\frac{\tau}{2}(\mu-\mu_0)^T \Sigma^{-1}(\mu-\mu_0)\}$, and after some algebraic manipulations the complete-data likelihood can be re-expressed as $\propto |\Sigma|^{-\frac{n}{2}} \exp\{-\frac{n}{2}\text{tr}(\Sigma^{-1}S)\} \exp\{-\frac{n}{2}(\bar{y}-\mu)^T \Sigma^{-1}(\bar{y}-\mu)\}$. Multiplying the prior and likelihood, the posterior distribution $P(\theta|Y)$ has also a normal inverted-Wishart form with new values for (τ, k, μ_0, Γ) . In other words, the complete-data posterior is normal inverted-Wishart: $\mu|\Sigma, Y \sim N(\mu_0^*, (\tau^*)^{-1}\Sigma)$; and $\Sigma|Y \sim W^{-1}(k^*, \Gamma^*)$, where the updated hyperparameters are $\tau^* = \tau + n$, $k^* = k + n$, $\mu_0^* = (\frac{n}{\tau+n})\bar{y} + (\frac{\tau}{\tau+n})\mu_0$, and $\Gamma^* = [\Gamma^{-1} + nS + (\frac{\tau n}{\tau+n})(\bar{y}-\mu_0)(\bar{y}-\mu_0)^T]^{-1}$. When no strong prior information is available about θ , one may apply Bayes' theorem with the improper prior $f(\theta) \propto |\Sigma|^{-\frac{(p+1)}{2}}$, which is the limiting form of the normal inverted-Wishart density as $\tau \rightarrow 0$, $k \rightarrow -1$ and $\Gamma^{-1} \rightarrow 0$. In our simulated examples, we used this non-informative prior reflecting a state of relative ignorance which is often colloquially expressed as "let the data talk".

Initial estimates for θ are typically obtained by the EM algorithm. Then, data augmentation scheme is implemented as follows: First, a value of missing data from the conditional predictive distribution of Y_{mis} , $Y_{mis}^{(t+1)} \sim P(Y_{mis}|Y_{obs}, \theta^{(t)})$, is drawn. Then, conditioning on $Y_{mis}^{(t+1)}$, a new value of θ from its complete-

data posterior, $\theta^{(t+1)} \sim P(\theta|Y_{obs}, Y_{mis}^{(t+1)})$ is drawn. Repeating these two steps from a starting value $\theta^{(0)}$ yields a stochastic sequence $(\theta^{(t)}, Y_{mis}^{(t)}) : t = 1, 2, \dots$ whose stationary distribution is $P(\theta, Y_{mis}|Y_{obs})$, and the subsequences $\theta^{(t)}$ and $Y_{mis}^{(t)}$ have $P(\theta|Y_{obs})$ and $P(Y_{mis}|Y_{obs})$ as their respective distributions. For a reasonably large number of iterations, the convergence to these stationary distributions is achieved. Since the complete-data likelihood is assumed to follow a multivariate normal distribution, drawing from conditional distributions above is relatively straightforward and can be performed by applying sweep operators to subsets of the vector μ and the matrix Σ .

What if these distributional assumptions are violated? Researchers rarely witness that the MVN assumption holds with the real data. Since it is impossible to formulate analytical structures and computational routines for every situation, practitioners generally rely on the MVN assumption when imputing continuous data. Hence, an assessment of viability is warranted via simulated scenarios that represent different sort of distributional violations. Describing a real phenomenon by generating the environment within which the process under consideration operates is not uncommon and is often the only feasible way of evaluation. For this reason, we devise a simulation study which we describe below.

3 A simulation study

The framework of the simulation study consists of complete data generation from several bivariate distributions, imposing missing values under ignorable missingness mechanisms, MI under the MVN assumption, parameter estimation and evaluation.

3.1 Design overview

3.1.1 Data generation

Complete data were generated using the following seven bivariate continuous distributions. For what follows, we assume that the two variables marginally follow identical densities, and f stands for the univariate probability density function. The correlations were chosen to be 0 except for the normal distribution. For its rationale, see the Discussion Section.

- *Normal distribution:* $f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$ with $E[Y] = \mu$, $V[Y] = \sigma^2$, where μ and $\sigma > 0$ are location and scale parameters, respectively. We set $\mu = 1$ and $\sigma = 1$. The correlation was

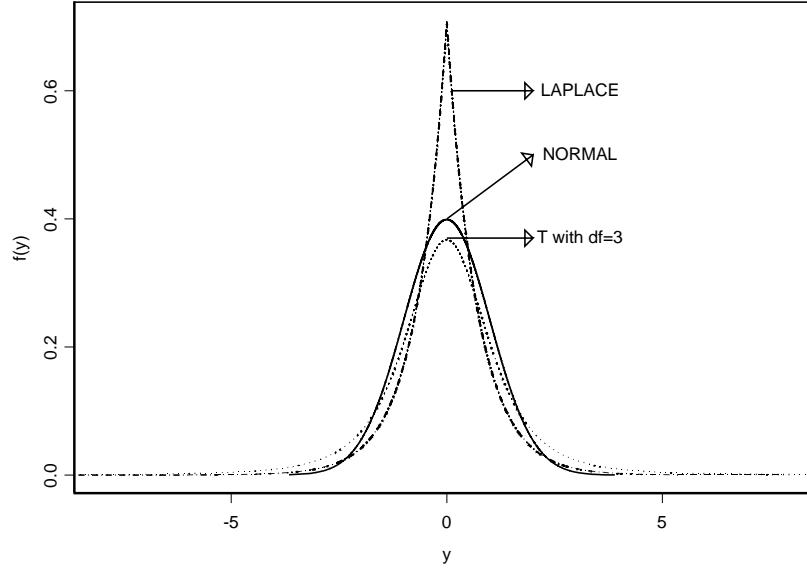


Figure 1: Density functions of Laplace($0, \sqrt{2}$), Normal($0, 1$) and $t(3, 0, 1/\sqrt{3})$ distributions. Parameter values were chosen such that $E[Y] = 0$ and $V[Y] = 1$ for all three densities.

set equal to 0.2, 0.8, and -0.5 in three simulated scenarios.

- *t distribution*: $f(y|\nu, \mu, \sigma) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} [1 + \frac{1}{\nu}(\frac{y-\mu}{\sigma})^2]^{-(\nu+1)/2}$ with $E[Y] = \mu$ for $\nu > 1$, $V[Y] = \frac{\nu}{\nu-2}\sigma^2$ for $\nu > 2$, where μ , σ and ν are the location, scale and degrees of freedom parameters, respectively. We set $\mu = 1$ and $\sigma = 1$. ν was chosen to be 3, 10, and 25 to allow for different tail behaviors with smaller ν corresponding to heavier tails.
- *Laplace (double-exponential) distribution*: $f(y|\alpha, \lambda) = \frac{\lambda}{2} \exp(-\lambda|y-\alpha|)$ with $E[Y] = \alpha$, $V[Y] = 2/\lambda^2$, where α and $\lambda > 0$ are the location and inverse scale parameters, respectively. We set $\alpha = 1$, and $\lambda = 1, 2$, and 10 . As λ gets smaller, tails get heavier.

See Figure 1 for a comparative tail behaviors of normal, t and Laplace given the same mean and variability across the distributions. All these densities are unimodal and symmetric, but t and Laplace have heavier tails than normal. ¹

- *Uniform distribution*: $f(y|a, b) = (b-a)^{-1}$, $a \leq y \leq b$ with $E[Y] = (b+a)/2$, $V[Y] = (b-a)^2/12$,

¹The parameter values used in Figure 1 are different from the ones used in simulations. This graph is created for the purpose of exposition of different tail behaviors.

where a and b are the lower and upper bounds of the support of y . We take $(0, 1)$, $(-3, 5)$ and $(2, 8)$ for (a, b) . Here, the density is flat (no mode exists).

For the next three densities, see Figure 2 for different sets of parameter values chosen. For illustration purposes, we refer to Figure 2 which is formed in a 6×3 matrix format for easier visibility and interpretability. Beta, Weibull and normal mixture densities correspond to the first, second and third columns of the matrix graph, respectively. Furthermore, plot $[u, v]$ stands for the plot in row u and column v .

- *Beta distribution:* $f(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$, $0 < y < 1$ with $E[Y] = \frac{\alpha}{\alpha+\beta}$, $V[Y] = \frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$, where $\alpha > 0$ and $\beta > 0$ are the shape parameters. Depending on the choice of parameters, the plot of the Beta density can take a variety of forms. By fixing α at 5, the values 1.5, 3, 5, 10, and 30 for β yield the densities of heavily negatively skewed (plot [1,1]), slightly negatively skewed (plot [2,1]), symmetric (plot[3,1]), mildly positively skewed (plot [4,1]), and moderately positively skewed (plot [5,1]) shapes, respectively. Finally, when $\alpha = 1$ and $\beta = 3$, the mode is at the boundary (plot [6,1]). Note that it cannot be heavily positively skewed given its nature.
- *Weibull distribution:* $f(y|\gamma, \delta) = \frac{\delta}{\gamma^\delta}y^{\delta-1}\exp(-(\frac{y}{\gamma})^\delta)$, $y > 0$ with $E[Y] = \gamma\Gamma(1 + 1/\delta)$, $V[Y] = \gamma^2[\Gamma(1 + 2/\delta) - \Gamma^2(1 + 1/\delta)]$, where $\gamma > 0$ and $\delta > 0$ are the scale and shape parameters. In a similar fashion to the Beta distribution, the Weibull density takes different graphical forms with respect to different choice of parameter values. By setting $\gamma = 1$, and $\delta = 1, 1.5, 2, 3.6, 5$, and 20, we obtain “mode at the boundary” (plot [2,1]), heavily positively skewed (plot [2,2]), slightly positively skewed (plot[3,2]), symmetric (plot [4,2]), mildly negatively skewed (plot [5,2]), and moderately negatively skewed (plot [6,2]) densities, respectively. Note that it cannot be heavily negatively skewed.
- *Mixture of univariate normal distributions:*
 $f(y|\mu_1, \mu_2, \sigma_1, \sigma_2, p) = \frac{p}{\sigma_1\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{y-\mu_1}{\sigma_1})^2) + \frac{(1-p)}{\sigma_2\sqrt{2\pi}}\exp(-\frac{1}{2}(\frac{y-\mu_2}{\sigma_2})^2)$ with $E[Y] = p\mu_1 + (1-p)\mu_2$, $V[Y] = p[\sigma_1^2 - (p-1)(\mu_1 - \mu_2)^2] - (p-1)\sigma_2^2$, where $0 < p < 1$ is the mixing parameter. Since it is a mixture, it can be unimodal or bimodal. If $(\mu_1 - \mu_2)^2 < \frac{27\sigma_1^2\sigma_2^2}{4(\sigma_1^2\sigma_2^2)}$, then the mixture is unimodal for all values of p . If $(\mu_1 - \mu_2)^2 > \frac{8\sigma_1^2\sigma_2^2}{(\sigma_1^2\sigma_2^2)}$, then there are some values for p for which the mixture is

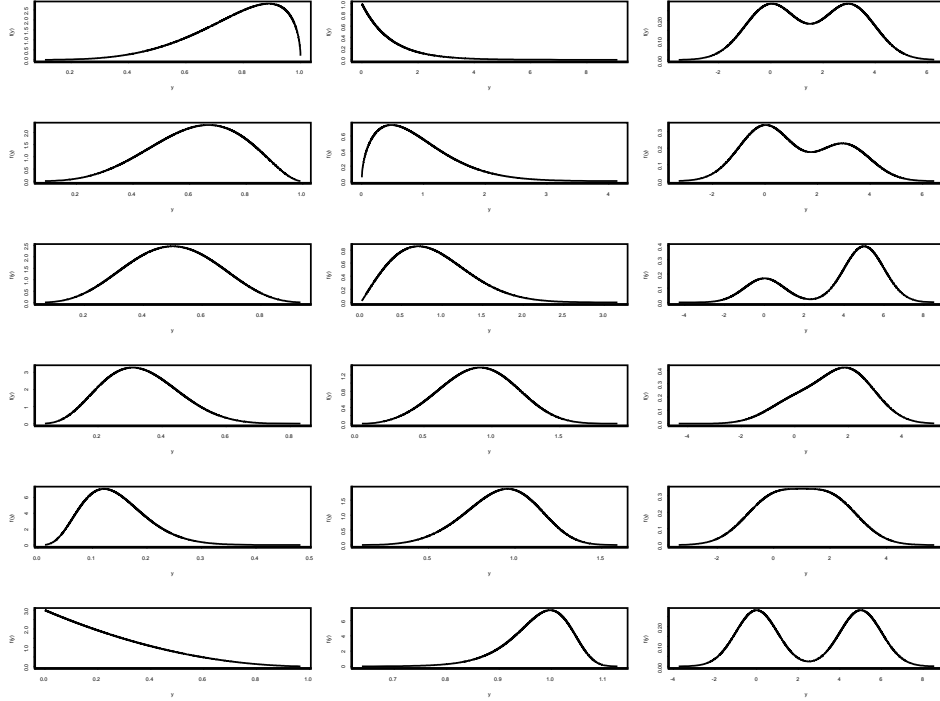


Figure 2: Density functions of Beta (first column), Weibull (second column), and Normal-mixture (third column) distributions for different set of parameters.

bimodal. Table 1 summarizes the set of assumed parameter values. The orders in Table 1 and in the third column of Figure 2 are the same.

Every distribution was chosen to address the key questions for empirically testing the validity of the MVN assumption in the imputation context. Laplace and t have heavier tails than normal and are symmetric, uniform has no mode (flat density) and is symmetric, Beta and Weibull could be symmetric or skewed depending on the choice of the parameters, and the mixture normal could be bimodal or skewed.

Table 1: Characteristics of normal mixtures for the chosen parameter values.

p	μ_1	σ_1	μ_2	σ_2	Characteristic
0.5	0	1	3	1	bimodal, balanced, close modes
0.6	0	1	3	1	bimodal, unbalanced
0.3	0	1	5	1	bimodal, unbalanced
0.3	0	1	2	1	unimodal, left-skewed
0.5	0	1	2	1	unimodal, symmetric, small curvature
0.5	0	1	5	1	bimodal, balanced, far modes

Skewness, multimodality and flat densities clearly violate the assumption of MVN and these situations should be examined for a real assessment of the impact of departures from MVN on the inferences drawn from multiply imputed datasets.

We examine two sample sizes: the number of subjects (n) in the simulated examples is 400 and 40 to assess the plausibility of the MVN assumption in large- and small-sample cases. The variables in this bivariate setting are denoted as Y_1 and Y_2 .

3.1.2 Missingness mechanism

We assume that Y_1 is always observed and Y_2 is incompletely observed. Missing values are imposed on Y_2 with missing completely at random (MCAR) and missing at random (MAR) mechanisms. Under MCAR, the mechanism that drives missingness does not depend on any variables, whereas under MAR, missingness depends on fully observed responses (Y_1 , in this case). For both sample sizes, 25% and 75% of observations in Y_2 are assumed to be missing with MCAR and MAR mechanisms. This leads to 300 or 100 observed values in the large-sample case; and 30 or 10 observed values in the small-sample case, on average, depending on the fraction of missing observations. Since the implementation of multiple imputation under the MVN assumption requires ignorability, we do not consider nonignorable missingness mechanisms in our simulations. We discuss this issue further in Section 4.

3.1.3 Parameter estimation

The mean ($\mu_2 = E[Y_2]$), variance ($\sigma_2^2 = Var[Y_2]$), correlation ($\rho = Corr(Y_1, Y_2)$) and regression (β_2 in equation $Y_2 = \beta_1 + \beta_2 Y_1$) parameters were considered. Although these quantities are obviously dependent on each other, it is illustrative to examine them separately.

3.1.4 Evaluation criteria

We created multiply imputed datasets with Splus 6.1 missing data library (Schimert et al., 2001). The procedure, which consists of complete data generation, imposing missing values, MI with data augmentation whose starting values are obtained from the EM algorithm, finding the estimates for the parameters μ_2 , σ_2^2 , ρ and β_2 , and combining them by Rubin's (1987) rules, was repeated 1000 times for each of the $2 \times 2 \times 2 \times 30 = 240$ (two sets of sample sizes, two sets of nonresponse mechanisms for each sample size, two

sets of fraction of missing values, and totally thirty different cases with seven distributions we considered) scenarios. The performance was evaluated using the following quantities that are frequently regarded as benchmark accuracy and precision measures:

Standardized bias (SB): the relative magnitude of the raw bias to the overall uncertainty in the system. If the parameter of interest is θ , the standardized bias is $100 \times \frac{E(\hat{\theta}) - \theta}{SE(\hat{\theta})}$, where SE stands for standard error. If the standardized bias exceeds 40 – 50% in a positive or negative direction, then the bias begins to have a noticeable adverse impact on efficiency, coverage and error rates (see Demirtas, 2004)).

Percentage bias (PB): the relative magnitude of the raw bias to the true value of the parameter, $100 \times \left| \frac{E(\hat{\theta}) - \theta}{\theta} \right|$. In our subjective opinion, the reasonable upper limit for the percentage bias is 5%.

Coverage rate (CR): the percentage of times that the true parameter value is covered in the confidence interval. If a procedure is working well, the actual coverage should be close to the nominal rate (i.e. Type I error rates are properly controlled). However, it is important to evaluate coverage with the other measures, because high variances can lead to higher coverage rates. We regard the performance of the interval procedure to be poor if its coverage drops below 90% (Collins et al., 2001).

Root-mean-squared error (RMSE): an integrated measure of bias and variance. It is considered to be arguably the best criterion for evaluating $\hat{\theta}$ in terms of combined accuracy and precision. $RMSE(\hat{\theta})$ is defined as $\sqrt{E_{\theta}[\hat{\theta} - \theta]^2}$.

Average width of confidence interval (AW): the distance between average lower and upper limits across 1000 confidence intervals. A high coverage rate along with narrow, calibrated confidence intervals translates into greater accuracy and higher power.

Under the above specification, SB and PB are the pure accuracy measures, AW is the pure efficiency measure, CR and RMSE are the hybrid measures. The reason we use two different bias quantities is that both have relative merits and pitfalls: SB depends on the total inherent variability which may be too small or too large, causing misleading interpretations; and PB has the assumed true value of the estimand in the denominator which similarly may take extreme values. In our experience, it is advisable to consider both accuracy benchmarks simultaneously.

3.2 Results

Due to space limitations, we only report the quantities that are relevant to the mean and variance structure. The performance of the measures of association (correlation and regression coefficients) exhibit quite similar characteristics to the mean parameter (not reported for brevity). Our findings suggest that there is little or no discernible differences in inferences between MCAR and MAR. This is hardly surprising since MI under the MVN assumption is known to operate under ignorable nonresponse; that is, as long as the reasons for missingness (Y_1 in this case) are included in the imputation process, differences should be minimal. Furthermore, in the large-sample case, differences between 25% and 75% missingness are immaterial to report. For these reasons, we tabulate the results for MCAR mechanism with 75% missingness in the large-sample case, and with 25% and 75% missingness in the small-sample case for the seven bivariate distributions.

In Tables 2-6, average estimate (AE), percentage bias (PB), standardized bias (SB), root-mean-squared error (RMSE), coverage rate (CR) and average width (AW) for $\mu_2 = \overline{Y_2}$ and $\sigma_2^2 = Var[Y_2]$ for seven distributions are tabulated. In Tables 2-6, n stands for the number of subjects in each simulated dataset, PM stands for the percentage of missing observations in Y_2 , PI stands for the parameter of interest, and TV stands for true value. The number of significant digits after the decimal point varies depending on the distribution and quantity of interest. Biases and coverage rates that are beyond acceptable limits are shown with bold characters. Moreover, the numbers may not be perfectly accurate due to rounding errors.

The results shown in Table 2 in regard to the normal distribution (the upper part) suggest that MI process yields unbiased estimates for μ_2 with good coverage rate for both large- and small-sample cases and for both rates of missing values, as one would expect. RMSE's and AW's get larger as the sample size and the fraction of missing values get smaller. In regard to σ_2^2 , the procedure seems to start suffering with $n = 40$ and 75% missingness and produces biased estimates. It is interesting to observe this, because even under a correct model specification, MI does not perform well for the variance parameter when the fraction of missing information is high in addition to having a small sample size.

The lower part of Table 2 shows the results for complete datasets that are generated by a t distribution which has heavier tails than normal distribution. Again, the estimates for the mean appear to be excellent.

However, the variance parameter is biased for $n = 40$. In addition, when degrees of freedom is small ($\nu = 3$), the coverage rates become unacceptable due to heavier tail behavior. As ν gets larger, t becomes closer to normal, and thus yields similar results.

Another example of “heavier-tail behavior” is the Laplace distribution (the upper part of Table 3). The difference between Laplace and t is that Laplace has a larger peakedness. As before, inferences for μ_2 seem to be unaffected with this distributional violation. On the other hand, the coverage rates for σ_2^2 is below 90% for all scenarios ($\lambda = 1, 2, 10$) with large biases for $n = 40$ and 75% missingness.

The uniform distribution (the lower part of Table 3) has no modes which is clearly unaccommodated by the MVN assumption. The only visible underperformance is observed in terms of bias in σ_2^2 for the small-samples in spite of good coverage rates.

The Beta (Table 4) and Weibull (Table 5) distributions have the property of possessing different shapes (symmetric, left- and right-skewed, boundary modes) depending on the choice of density parameters. For both distributions, the mean parameter performs very well for all three situations we report in Tables 4 and 5 with estimates that are unbiased, have low RMSE’s, high coverage rates, and that land on calibrated intervals. In the large-sample case, the same conclusion is valid for the variance parameter. From an accuracy standpoint, biases occur for σ_2^2 in the small-sample case. Coverage rates are within reasonable limits for both distributions except for $\delta = 1$ in Weibull which corresponds to the “mode at the boundary” scenario (plot[2,1] in Figure 2). It is noteworthy to mention that even when the distribution is heavily skewed (plot[1,1] and plot[2,2] in Figure 2), there are no biases for μ_2 with decent coverage rates in all cases; and the sample size has to be small for σ_2^2 to start not working properly.

What if the underlying density is formed through a mixture distribution? Several potential situations that may arise with a mixture distribution are tabulated and graphed in Section 3.1.1 (Table 1 and the third column of Figure 2). In the large-sample case, both μ_2 and σ_2^2 seem to have good properties (negligible bias and coverage rates that are close to nominal levels). In the small-sample case, coverage rates are satisfactory for σ_2^2 . However, when the resulting density is bimodal (Figure 2) with unbalanced modes (plot[3,2] and plot[3,3]) or is skewed (plot[3,4]), both parameters under consideration appear to be biased to varying degrees (Table 6). In addition, μ_2 has unacceptably low coverage percentages in these scenarios.

When the density is bimodal with balanced modes (plot[3,1] and plot[3,6]) or is unimodal (plot[3,5]), we do not observe biasedness and inflated Type I errors except for the variance parameter (bias only, coverage is still good) in the small-sample case.

Overall performance of MI under the MVN assumption with simulated datasets that were generated under non-Gaussian continuous distributions is better than we expected. Non-Gaussian features such as the flatness of the density, heavy tails, non-zero peakedness, skewness and multimodality do not appear to hamper the good performance of MI for the mean structure in the large-samples in the overwhelming majority of cases, even when there is a higher-than-typical rate of missingness (75%). Furthermore, the variance structure also appears to be preserved in large-samples. However, quite expectedly, the variance parameter is more sensitive to the departures from underlying distributional assumptions than the mean parameter, particularly when the sample size is small. Biases and inflation in Type I error rates occur with smaller samples, especially as the fraction of missing information gets larger.

4 Discussion

It should be noted that although transformations (e.g. Box-Cox) can move the marginal distributions closer to normality, the correlation structure in a multivariate setting also changes with transformations to an extent that leads to interpretation problems especially in the presence of missing data. The fraction of missing information hinges on how variables are related to each other among other things, and perturbations to the correlations almost certainly alter the conditional distribution of missing data given observed data, raising more questions than they solve. In the imputation context, the degree of relatedness among variables is as important as the marginal behaviors, and transformations may seriously degrade these associations. Somewhat connected to this issue, in our simulations we assumed zero correlation between the two variables in most scenarios. The rationale is that some of the missing information is anticipated to be recovered to the extent that the variables are correlated in the MI process; in assessing the performance under the MVN assumption, assuming uncorrelatedness represents a worst-case situation in the sense that incompletely observed variable (Y_2) does not receive any help from the fully observed variable (Y_1). Therefore, the results presented can be expected to change with non-zero correlation in favor of the MVN assumption

Table 2: Results for *Normal* and *t distributions*. δ refers to the correlation between Y_1 and Y_2 in Normal distribution, and ν refers to the degrees of freedom in t distribution.

Distribution	PI	n	PM	δ	TV	AE	PB	SB	RMSE	CR	AW	
NORMAL	μ_2	400	75	0.2	1	0.9940	0.6	-5.81	0.10	91.5	0.43	
	μ_2	400	75	0.5	1	0.9932	0.7	-9.62	0.07	96.7	0.31	
	μ_2	400	75	-0.8	1	1.0030	0.3	2.99	0.10	93.2	0.39	
	μ_2	40	25	0.2	1	1.0093	0.9	4.52	0.21	93.6	0.74	
	μ_2	40	25	0.5	1	1.0118	1.2	6.61	0.18	90.2	0.65	
	μ_2	40	25	-0.8	1	0.9965	0.4	-2.01	0.18	96.5	0.71	
	μ_2	40	75	0.2	1	0.9780	2.2	-6.09	0.36	91.1	1.43	
	μ_2	40	75	0.5	1	0.9937	0.6	-2.57	0.25	95.3	1.10	
	μ_2	40	75	-0.8	1	0.9495	5.0	-14.05	0.36	90.4	1.37	
	σ_2^2	400	75	0.2	1	1.0314	3.1	20.03	0.16	94.6	0.65	
	σ_2^2	400	75	0.5	1	1.0345	3.4	26.60	0.13	93.9	0.52	
	σ_2^2	400	75	-0.8	1	1.0184	1.8	12.72	0.15	95.8	0.61	
	σ_2^2	40	25	0.2	1	1.0927	9.3	30.57	0.32	93.6	1.24	
	σ_2^2	40	25	0.5	1	1.0395	3.9	13.03	0.30	92.7	1.09	
	σ_2^2	40	25	-0.8	1	1.1404	14.0	48.31	0.32	95.4	1.28	
	σ_2^2	40	75	0.2	1	1.3271	32.7	43.31	0.82	91.8	3.49	
	σ_2^2	40	75	0.5	1	1.2339	23.4	42.76	0.59	95.0	2.51	
	σ_2^2	40	75	-0.8	1	1.3946	39.5	35.11	1.19	93.3	3.55	
	T	μ_2	400	75	3	1	0.9998	0.0	-0.09	0.18	96.4	0.74
		μ_2	400	75	10	1	1.0117	1.2	10.91	0.11	94.7	0.48
		μ_2	400	75	25	1	0.9995	0.1	-0.46	0.11	96.0	0.46
μ_2		40	25	3	1	0.9877	1.2	-3.65	0.34	97.1	1.22	
μ_2		40	25	10	1	0.9976	0.2	-1.19	0.20	95.4	0.82	
μ_2		40	25	25	1	1.0093	0.9	4.48	0.20	96.0	0.79	
μ_2		40	75	3	1	0.9834	1.7	-2.83	0.58	95.8	2.44	
μ_2		40	75	10	1	1.0364	3.6	8.05	0.45	93.7	1.83	
μ_2		40	75	25	1	1.0327	3.3	8.07	0.41	93.8	1.72	
σ_2^2		400	75	3	3	2.8908	3.6	-5.88	1.85	62.6	1.77	
σ_2^2		400	75	10	1.25	1.2543	0.3	1.72	0.25	90.1	0.79	
σ_2^2		400	75	25	1.087	1.1104	2.2	13.09	0.18	93.7	0.69	
σ_2^2		40	25	3	3	3.2839	9.5	8.53	3.33	60.1	3.81	
σ_2^2		40	25	10	1.25	1.3561	8.5	20.51	0.53	90.3	1.55	
σ_2^2		40	25	25	1.087	1.1881	9.3	28.41	0.37	94.2	1.37	
σ_2^2		40	75	3	3	10.0906	236.4	9.00	78.86	78.4	55.81	
σ_2^2		40	75	10	1.25	1.9769	58.2	38.73	2.01	94.0	6.69	
σ_2^2		40	75	25	1.087	1.7584	61.8	59.97	1.30	92.5	5.03	

Table 3: Results for *Laplace* and *Uniform distributions*. λ refers to the inverse scale parameter in Laplace distribution, and a/b refers to the lower/upper limits of the range in Uniform distribution.

Distribution	PI	n	PM	λ	TV	AE	PB	SB	RMSE	CR	AW	
LAPLACE	μ_2	400	75	1	1	1.0137	1.4	9.17	0.15	93.1	0.62	
	μ_2	400	75	2	1	0.9954	0.5	-6.31	0.07	95.0	0.30	
	μ_2	400	75	10	1	0.9995	0.0	-3.32	0.02	93.9	0.06	
	μ_2	40	25	1	1	1.0063	0.6	2.24	0.28	92.1	1.03	
	μ_2	40	25	2	1	1.0082	0.8	6.18	0.13	94.6	0.51	
	μ_2	40	25	10	1	0.9979	0.2	-7.74	0.03	91.7	0.10	
	μ_2	40	75	1	1	0.9474	5.3	-11.04	0.48	93.2	2.06	
	μ_2	40	75	2	1	1.0170	1.7	6.29	0.27	90.5	1.09	
	μ_2	40	75	10	1	0.9984	0.2	-2.89	0.06	93.8	0.21	
	σ_2^2	400	75	1	2	2.0370	1.9	7.65	0.48	80.8	1.27	
	σ_2^2	400	75	2	0.5	0.4988	0.2	-1.02	0.12	82.1	0.32	
	σ_2^2	400	75	10	0.02	0.0205	2.3	9.74	0.01	80.3	0.43	
	σ_2^2	40	25	1	2	2.1551	7.8	15.96	0.98	79.7	2.49	
	σ_2^2	40	25	2	0.5	0.5233	4.7	10.63	0.22	81.1	0.60	
	σ_2^2	40	25	10	0.02	0.0205	4.3	9.00	0.01	80.5	0.02	
	σ_2^2	40	75	1	2	2.8814	44.1	39.45	2.40	88.3	8.39	
	σ_2^2	40	75	2	0.5	0.7719	54.4	41.54	0.71	88.8	2.25	
	σ_2^2	40	75	10	0.02	0.0330	65.0	23.02	0.06	88.4	0.09	
	UNIFORM	μ_2	400	75	0/1	0.5	0.5012	0.2	4.07	0.03	95.0	0.13
		μ_2	400	75	-3/5	1	0.9922	0.8	-2.59	0.30	94.6	1.27
		μ_2	400	75	2/8	5	5.0214	0.4	12.79	0.17	96.8	0.77
μ_2		40	25	0/1	0.1	0.4986	0.3	-2.52	0.06	94.3	0.21	
μ_2		40	25	-3/5	1	1.0253	2.5	4.65	0.54	94.1	2.16	
μ_2		40	25	2/8	5	5.0047	0.1	1.43	0.33	94.7	1.32	
μ_2		40	75	0/1	0.1	0.5125	2.5	12.74	0.10	95.1	0.47	
μ_2		40	75	-3/5	1	1.0491	4.9	19.10	1.06	92.2	4.71	
μ_2		40	75	2/8	5	5.0047	0.2	1.49	0.65	95.9	2.98	
σ_2^2		400	75	0/1	0.083	0.0859	3.1	33.00	0.01	95.8	0.10	
σ_2^2		400	75	-3/5	5.333	5.5214	3.5	19.59	0.40	94.3	4.24	
σ_2^2		400	75	2/8	3	3.0982	3.3	33.06	0.31	96.1	1.94	
σ_2^2		40	25	0/1	0.083	0.0890	6.8	41.74	0.02	95.3	0.10	
σ_2^2		40	25	-3/5	5.333	5.5982	4.9	28.74	1.29	95.2	4.32	
σ_2^2		40	25	2/8	3	3.2869	9.6	50.82	0.63	96.4	1.95	
σ_2^2		40	75	0/1	0.083	0.1201	44.1	72.24	0.06	95.8	0.33	
σ_2^2		40	75	-3/5	5.333	9.6110	80.2	62.56	8.05	95.8	36.94	
σ_2^2		40	75	2/8	3	4.8951	63.2	38.73	5.23	94.9	15.61	

Table 4: Results for *Beta distribution*. α and β refer to the shape parameters.

PI	n	PM	α	β	TV	AE	PB	SB	RMSE	CR	AW
μ_2	400	75	5	1.5	0.7692	0.7693	0.0	0.26	0.0164	92.0	0.0673
μ_2	400	75	5	3	0.6250	0.6250	0.0	-0.02	0.0149	97.0	0.0705
μ_2	400	75	5	5	0.5000	0.5016	0.3	11.27	0.0143	95.9	0.0661
μ_2	400	75	5	10	0.3333	0.3336	0.1	1.67	0.0135	93.0	0.0516
μ_2	400	75	5	30	0.1429	0.1428	0.0	-0.87	0.0058	94.9	0.0266
μ_2	400	75	1	3	0.2500	0.2501	0.1	0.59	0.0232	89.8	0.0832
μ_2	40	25	5	1.5	0.7692	0.7679	0.2	-4.41	0.0297	96.2	0.1148
μ_2	40	25	5	3	0.6250	0.6267	0.3	5.88	0.0291	96.1	0.1183
μ_2	40	25	5	5	0.5000	0.5053	1.1	20.22	0.0268	95.2	0.1094
μ_2	40	25	5	10	0.3333	0.3321	0.4	-5.62	0.0224	94.8	0.0893
μ_2	40	25	5	30	0.1429	0.1422	0.5	-6.26	0.0108	95.5	0.0435
μ_2	40	25	1	3	0.2500	0.2519	0.8	5.69	0.0339	94.7	0.1429
μ_2	40	75	5	1.5	0.7692	0.7642	0.7	-8.37	0.0601	96.3	0.2427
μ_2	40	75	5	3	0.6250	0.6277	0.4	4.12	0.0659	93.6	0.2647
μ_2	40	75	5	5	0.5000	0.4996	0.1	-0.87	0.0505	95.5	0.2368
μ_2	40	75	5	10	0.3333	0.3381	1.4	12.01	0.0402	94.3	0.1876
μ_2	40	75	5	30	0.1429	0.1470	2.9	17.92	0.0236	91.9	0.0931
μ_2	40	75	1	3	0.2500	0.2468	1.3	-4.46	0.0725	93.1	0.3103
$\sigma_{\mu_2}^2$	400	75	5	1.5	0.0237	0.0245	3.4	20.44	0.0040	91.4	0.0157
$\sigma_{\mu_2}^2$	400	75	5	3	0.0260	0.0270	3.8	26.19	0.0039	96.5	0.0172
$\sigma_{\mu_2}^2$	400	75	5	5	0.0227	0.0231	1.7	12.92	0.0030	96.2	0.0143
$\sigma_{\mu_2}^2$	400	75	5	10	0.0139	0.0142	2.6	18.54	0.0019	95.3	0.0089
$\sigma_{\mu_2}^2$	400	75	5	30	0.0034	0.0035	4.2	23.41	0.0006	94.7	0.0023
$\sigma_{\mu_2}^2$	400	75	1	3	0.0375	0.0385	2.6	17.14	0.0058	94.1	0.0242
$\sigma_{\mu_2}^2$	40	25	5	1.5	0.0237	0.0262	10.8	32.97	0.0081	94.9	0.0298
$\sigma_{\mu_2}^2$	40	25	5	3	0.0260	0.0281	8.1	27.52	0.0079	93.0	0.0322
$\sigma_{\mu_2}^2$	40	25	5	5	0.0227	0.0238	4.7	17.01	0.0063	95.8	0.0269
$\sigma_{\mu_2}^2$	40	25	5	10	0.0139	0.0151	8.9	31.11	0.0041	95.4	0.0174
$\sigma_{\mu_2}^2$	40	25	5	30	0.0034	0.0038	10.4	32.58	0.0011	95.3	0.0043
$\sigma_{\mu_2}^2$	40	25	1	3	0.0375	0.0407	8.4	29.84	0.0110	93.8	0.0465
$\sigma_{\mu_2}^2$	40	75	5	1.5	0.0237	0.0364	53.9	45.05	0.0310	94.0	0.1103
$\sigma_{\mu_2}^2$	40	75	5	3	0.0260	0.0402	54.6	55.68	0.0291	95.8	0.1192
$\sigma_{\mu_2}^2$	40	75	5	5	0.0227	0.0330	45.1	60.27	0.0198	96.5	0.0971
$\sigma_{\mu_2}^2$	40	75	5	10	0.0139	0.0213	53.7	49.31	0.0168	94.4	0.0628
$\sigma_{\mu_2}^2$	40	75	5	30	0.0034	0.0052	53.2	50.62	0.0040	94.8	0.0149
$\sigma_{\mu_2}^2$	40	75	1	3	0.0375	0.0568	51.6	47.08	0.0453	91.9	0.1723

Table 5: Results for *Weibull distribution*. δ refers to the shape parameter.

PI	n	PM	δ	TV	AE	PB	SB	RMSE	CR	AW
μ_2	400	75	1	1.0000	1.0131	1.3	8.18	0.1602	95.4	0.6457
μ_2	400	75	1.5	0.9027	0.9065	0.4	5.57	0.0675	94.2	0.2627
μ_2	400	75	2	0.8862	0.8757	1.2	-21.66	0.0498	95.0	0.2018
μ_2	400	75	3.6	0.9011	0.9013	0.1	2.09	0.0550	91.6	0.1211
μ_2	400	75	5	0.9182	0.9154	0.3	-13.44	0.0210	96.5	0.0911
μ_2	400	75	20	0.9735	0.9737	0.0	3.06	0.0067	90.8	0.0263
μ_2	40	25	1	1.0000	1.0183	1.8	9.98	0.1842	93.1	0.7381
μ_2	40	25	1.5	0.9027	0.9069	0.5	3.92	0.1069	95.7	0.4532
μ_2	40	25	2	0.8862	0.8891	0.3	3.22	0.0887	94.0	0.3494
μ_2	40	25	3.6	0.9011	0.9023	0.1	2.09	0.550	91.7	0.2046
μ_2	40	25	5	0.9182	0.9144	0.4	-10.18	0.0369	96.3	0.1545
μ_2	40	25	20	0.9735	0.9738	0.0	2.35	0.0113	93.5	0.0430
μ_2	40	75	1	1.0000	1.0497	5.0	11.84	0.4558	92.9	1.6683
μ_2	40	75	1.5	0.9027	0.9214	2.1	7.70	0.2421	90.0	0.9961
μ_2	40	75	2	0.8862	0.8860	0.0	-0.10	0.1772	93.4	0.7597
μ_2	40	75	3.6	0.9011	0.8959	0.6	-5.41	0.0965	93.1	0.4029
μ_2	40	75	5	0.9182	0.9210	0.3	3.59	0.0799	90.1	0.3182
μ_2	40	75	20	0.9735	0.9706	0.3	-11.72	0.0248	92.5	0.0996
$\sigma_{\sigma_2}^2$	400	75	1	1.0000	1.0062	0.6	2.37	0.2616	76.7	0.6417
$\sigma_{\sigma_2}^2$	400	75	1.5	0.3757	0.3840	2.2	11.09	0.0751	90.3	0.2426
$\sigma_{\sigma_2}^2$	400	75	2	0.2146	0.2167	17.1	5.72	0.0370	90.1	0.1385
$\sigma_{\sigma_2}^2$	400	75	3.6	0.0773	0.0799	3.4	24.05	0.0113	96.0	0.0500
$\sigma_{\sigma_2}^2$	400	75	5	0.0442	0.0449	1.6	10.13	0.0068	93.5	0.0275
$\sigma_{\sigma_2}^2$	400	75	20	0.0036	0.0037	1.5	7.55	0.0007	90.2	0.0023
$\sigma_{\sigma_2}^2$	40	25	1	1.0000	1.1170	11.7	20.89	0.5708	79.2	1.2736
$\sigma_{\sigma_2}^2$	40	25	1.5	0.3757	0.4100	9.1	24.9	0.1416	90.3	0.4712
$\sigma_{\sigma_2}^2$	40	25	2	0.2146	0.2356	27.3	30.62	0.0716	95.1	0.2665
$\sigma_{\sigma_2}^2$	40	25	3.6	0.0773	0.0822	6.3	23.16	0.0216	95.7	0.0941
$\sigma_{\sigma_2}^2$	40	25	5	0.0442	0.0469	6.0	22.95	0.0118	97.0	0.0536
$\sigma_{\sigma_2}^2$	40	25	20	0.0036	0.0038	3.9	11.29	0.0012	91.9	0.0044
$\sigma_{\sigma_2}^2$	40	75	1	1.0000	1.9244	92.4	38.63	2.5596	92.0	6.2061
$\sigma_{\sigma_2}^2$	40	75	1.5	0.3757	0.6357	69.2	46.33	0.6172	93.1	1.9072
$\sigma_{\sigma_2}^2$	40	75	2	0.2146	0.3341	80.5	49.92	0.2669	93.6	0.9234
$\sigma_{\sigma_2}^2$	40	75	3.6	0.0773	0.1060	37.1	54.91	0.0594	95.4	0.2641
$\sigma_{\sigma_2}^2$	40	75	5	0.0442	0.0611	38.2	39.91	0.0455	93.5	0.1723
$\sigma_{\sigma_2}^2$	40	75	20	0.0036	0.0062	69.2	30.92	0.0085	92.9	0.0211

Table 6: Results for *Normal-mixture distribution*. $(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ refers to the vector of the mixing proportion, mean and variance parameters in the mixture.

PI	n	PM	$(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$	TV	AE	PB	SB	RMSE	CR	AW
μ_2	400	75	(0.5,0,1,3,1)	1.5	1.5012	0.1	0.71	0.1723	96.5	0.7789
μ_2	400	75	(0.6,0,1,3,1)	1.2	1.1965	0.3	-1.94	0.1801	95.2	0.7813
μ_2	400	75	(0.3,0,1,5,1)	3.5	3.5017	0.0	0.65	0.2626	95.7	1.0842
μ_2	400	75	(0.3,0,1,2,1)	1.4	1.3986	0.5	-4.54	0.1399	94.9	0.5971
μ_2	400	75	(0.5,0,1,2,1)	1.0	1.0020	0.2	1.37	0.1459	95.8	0.6310
μ_2	400	75	(0.5,0,1,5,1)	2.5	2.4787	0.9	-7.34	0.2905	94.2	1.1712
μ_2	40	25	(0.5,0,1,3,1)	1.5	1.4788	1.4	-6.33	0.3341	93.7	1.3579
μ_2	40	25	(0.6,0,1,3,1)	1.2	1.8047	50.4	182.33	0.6893	59.6	1.3425
μ_2	40	25	(0.3,0,1,5,1)	3.5	1.4768	57.8	-397.22	2.0860	4.8	1.8485
μ_2	40	25	(0.3,0,1,2,1)	1.4	0.5839	58.3	-314.24	0.8563	17.8	1.0066
μ_2	40	25	(0.5,0,1,2,1)	1.0	0.9838	1.6	-6.07	0.2660	93.8	1.0423
μ_2	40	25	(0.5,0,1,5,1)	2.5	2.5387	1.5	7.90	0.4894	93.1	1.9876
μ_2	40	75	(0.5,0,1,3,1)	1.5	1.4535	3.1	-6.26	0.7410	94.4	3.0890
μ_2	40	75	(0.6,0,1,3,1)	1.2	1.8806	56.7	105.94	0.9348	80.6	2.8635
μ_2	40	75	(0.3,0,1,5,1)	3.5	1.5849	54.7	-207.56	2.1247	48.5	3.9240
μ_2	40	75	(0.3,0,1,2,1)	1.4	0.6254	55.3	-149.45	0.9313	62.3	2.1289
μ_2	40	75	(0.5,0,1,2,1)	1.0	0.9589	4.1	-8.99	0.4571	94.5	2.1045
μ_2	40	75	(0.5,0,1,5,1)	2.5	2.5458	1.8	4.28	1.0686	90.1	4.2106
σ_2^2	400	75	(0.5,0,1,3,1)	3.25	3.3000	1.5	14.17	0.3552	94.5	2.1521
σ_2^2	400	75	(0.6,0,1,3,1)	3.16	3.2665	3.4	26.82	0.4102	96.9	2.0317
σ_2^2	400	75	(0.3,0,1,5,1)	6.25	6.4532	3.3	26.92	0.7797	93.7	4.0851
σ_2^2	400	75	(0.3,0,1,2,1)	1.84	1.8867	2.5	18.28	0.2593	95.3	1.1857
σ_2^2	400	75	(0.5,0,1,2,1)	2.00	2.0704	3.5	25.92	0.2797	94.7	1.3246
σ_2^2	400	75	(0.5,0,1,5,1)	7.25	7.4428	2.7	30.60	0.6574	95.8	4.7273
σ_2^2	40	25	(0.5,0,1,3,1)	3.25	3.5739	10.0	47.00	0.7600	95.9	4.1433
σ_2^2	40	25	(0.6,0,1,3,1)	3.16	3.4990	10.7	46.56	0.8015	93.8	4.0290
σ_2^2	40	25	(0.3,0,1,5,1)	6.25	6.6623	6.6	29.19	1.4678	96.5	7.6097
σ_2^2	40	25	(0.3,0,1,2,1)	1.84	1.9773	7.5	25.84	0.5473	95.0	2.2773
σ_2^2	40	25	(0.5,0,1,2,1)	2.00	2.1191	6.0	24.11	0.5069	94.6	2.4227
σ_2^2	40	25	(0.5,0,1,5,1)	7.25	7.8019	7.6	51.48	1.2033	95.8	8.7430
σ_2^2	40	75	(0.5,0,1,3,1)	3.25	5.2983	63.0	55.64	4.2044	96.0	15.7478
σ_2^2	40	75	(0.6,0,1,3,1)	3.16	4.6554	47.3	60.51	2.8832	95.7	13.4347
σ_2^2	40	75	(0.3,0,1,5,1)	6.25	9.1524	46.4	58.28	5.7532	94.5	25.5332
σ_2^2	40	75	(0.3,0,1,2,1)	1.84	2.5920	40.9	48.71	1.7137	94.9	6.8619
σ_2^2	40	75	(0.5,0,1,2,1)	2.00	2.7350	36.7	58.02	1.4617	96.2	7.1710
σ_2^2	40	75	(0.5,0,1,5,1)	7.25	10.4258	43.8	73.17	5.3694	95.1	29.1722

given this conservative position we have taken.

The main motivation of our work was to evaluate the effects of misspecification of the imputation model as MVN. As the popularity of inference by multiple imputation increases, new methods for creating multiple imputations assuming more diverse class of distributions are emerging as well (He, 2005). While the new methods may be more appropriate in this endeavor, they are not as commonly implemented as the current method based on the MVN assumption (e.g. SAS PROC MI).

As noted in Section 2 the computational algorithm for creating multiple imputations relies on Bayesian arguments and users of this method must choose hyperparameters (τ, k, μ_0, Γ) for the underlying prior distributions. In this paper, we chose such priors to be mainly “non-informative” with a goal of minimal subjective influence on the “final” inferences. By changing this practice to allow more informative priors, improved inferences can be drawn. Several strategies could be employed for determining the hyperparameters such as using data to obtain prior “guesses” on the parameters (Schafer and Yucel, 2002; Demirtas, 2005), or via previous studies. It is important, however, to be cautious, when imposing stronger prior beliefs, to carry out a sensitivity analyses by comparing inferences under different prior distributions. In applications with sparse or limited data, some aspects of the parameters may not be well estimated. In such cases, use of ridge-like priors (Schafer, 1997), tends to stabilize the computational procedures.

One may argue that the generalizability of the simulation results is doubtful given the countless other situations that can potentially be encountered in real life. This argument has certain validity, however, the purpose of this article is limited to evaluating how MI under MVN set-up performs with respect to incomplete multivariate non-Gaussian outcomes that exhibit varying distributional properties as well as different levels sample sizes and fractions of missing information. Although our simulated datasets are only a small portion of what may arise in applications and they may be insufficiently complex compared to the real datasets, we believe that a simulation assessment and evaluation based on a fairly comprehensive study that includes many non-Gaussian features is insightful.

Nonignorable modeling is a separate issue and beyond the scope of this manuscript. Imputing under the MVN assumption explicitly requires ignorability given the current state of the research. Furthermore, we do not attach any negative connotations to other MI models for continuous outcomes. While our results

suggest that this approach is reasonable for estimating the means and associations in most cases, it may not necessarily perform better than the other MI models in every setting. However, MI under the Gaussian model is expected to continue to draw attention from practitioners who typically prefer a readily available, approximate solution with decent properties rather than the one that is more efficient, but problem-specific and complicated to implement. While we discourage its uncritical, automatic and possibly inappropriate use, we feel that that it is a rather plausible option for imputing non-Gaussian outcomes given the simplicity of implementation and/or widespread availability of software.

5 References

- Collins LM, Schafer JL, Kam CH. (2001), A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6, 330–351.
- Demirtas H, Schafer JL. (2003), On the performance of random-coefficient pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 22, 2553–2575.
- Demirtas H. (2004), Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica*, 58, 466–482.
- Demirtas H. (2005), Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 24, 2345–2363.
- Dempster AP, Laird NM, Rubin DB. (1977), Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of Royal Statistical Society, Series B*, 39, 1–38.
- He Y. (2005), Multiple Imputation for Continuous Non-normal Missing Data. Unpublished dissertation. Department of Biostatistics , University of Michigan, Ann Arbor, MI.
- Little RJA, Rubin DB. (2002), *Statistical Analysis with Missing Data*. Second Edition. New York: Wiley.
- Rubin DB. (1976), Inference and missing data. *Biometrika*, 63, 581–592.
- Rubin DB. (1987), *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

- Rubin DB. (1996), Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91, 473–520.
- Schafer JL. (1997), *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Schafer JL. (1999), Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8, 3–15.
- Schafer JL, Yucel RM. (2002), Computational strategies for multivariate linear mixed-effects models with missing values. *Journal of Computational and Graphical Statistics*, 11, 437–457.
- Schimert J, Schafer JL, Hesterberg T, Fraley C, Clarkson DB. (2001), *Analyzing Data with Missing Values in S-plus*. Seattle, WA: Data Analysis Products Division, Insightful Corp.
- Tanner MA, Wong WH. (1987), The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association*, 82, 528–540.