

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

*Technical report#:2006-002
April 2006*

Title: Practical advice on how to impute continuous data
when the ultimate interest centers on dichotomized
outcomes through pre-specified thresholds

Author(s): Hakan Demirtas

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

Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds

Hakan Demirtas *

April 6, 2006

Abstract

Multiple imputation under the multivariate normality assumption has often been regarded as a viable model-based approach in dealing with incomplete continuous data in the last two decades. A situation where the measurements are taken on a continuous scale with an ultimate interest in dichotomized versions through discipline-specific thresholds is not uncommon in applied research, especially in medical and social sciences. In practice, researchers generally tend to impute missing values for continuous outcomes under a Gaussian imputation model, and then dichotomize them via commonly-accepted cut-off points. An alternative strategy is creating multiply imputed data sets after dichotomization under a log-linear imputation model that uses a saturated multinomial structure. In this work, the performances of the two imputation methods were examined on a fairly wide range of simulated incomplete data sets that exhibit varying distributional characteristics such as skewness and multimodality. Behavior of efficiency and accuracy measures was explored to determine the extent to which the procedures work properly. The conclusion drawn is that dichotomization before carrying out a log-linear imputation should be the preferred approach except for a few special cases. I recommend that researchers use the atypical second strategy whenever the interest centers on binary quantities that are obtained through underlying continuous measurements. A possible explanation is that erratic/idiosyncratic aspects that are not accommodated by a Gaussian model are probably transformed into better-behaving discrete trends in this particular missing-data setting. This premise outweighs the assertion that continuous variables inherently carry more information, leading to a counter-intuitive, but potentially useful result for practitioners.

Key Words: Multivariate normality; Multiple imputation; Log-linear models; Skewness; Multimodality.

1 Introduction and motivation

Missing data is the norm rather than the exception in most data sets. 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).

*Hakan Demirtas (e-mail:demirtas@uic.edu) is an Assistant Professor of Biostatistics, Division of Epidemiology and Biostatistics (MC923), University of Illinois at Chicago, 1603 West Taylor Street, Chicago, IL, 60612.

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 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 (1997, 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 usually involves positing a parametric model for the data and using it to derive this conditional distribution. For continuous data, joint multivariate normality among the variables 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.

When all the variables are categorical, a log-linear imputation model is typically used (Schafer, 1997). If the sample size is assumed fixed, the set of cell frequencies in a contingency table has a multinomial distribution. If there are no restrictions on the parameters other than they are true probabilities, then the model is said to be saturated. Log-linear models are a flexible class of models for specifying possible dependencies among variables. With complete data, using a Dirichlet prior distribution for the saturated model leads to a conjugate analysis. The posterior distribution is again Dirichlet with updated parameters involving the data and prior parameters.

In applied research, it is not uncommon that a set of continuous measurements or observations are obtained with an ultimate interest in dichotomized versions through pre-specified threshold points. One well-known example is the obesity status (obese, not obese) based on body mass index. In medicine, blood pressure and cholesterol levels are often categorized. In fact, many diagnostic procedures require some sort of classification system or an underlying continuous variable.

The research question that motivates this study is how to conduct multiple imputation inference in regard to the order of imputation and dichotomization. When incomplete continuous data are collected, should one impute continuous outcomes under the normality assumption and consequently dichotomize

the completed data, or is it better to dichotomize the outcomes before carrying out MI under a log-linear model? This article examines the plausibility of both approaches via a broad range of simulated examples representing the situations such as multimodality, skewness, non-zero peakedness, heavy tails, and flatness of the continuous densities. Common sense suggests that the first approach is more plausible on the grounds that continuous data naturally carry more information. However, the results of the simulation study do not appear to support this conjecture.

Organization of this paper is as follows. The next section provides some key operational attributes of MI under normal and log-linear models along with a brief coverage of computational routines. In Section 3, a fairly comprehensive simulation study that spans different underlying distributional properties of the original outcomes is presented. Section 4 includes practical suggestions, concluding remarks and discussion.

2 Fundamentals of missing data and multiple imputation

To set the notation, the data set 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 data set. 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).

After reviewing the fundamentals of missing data, I now describe the key characteristics of multiple imputation (MI). 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, ignorable nonresponse is assumed.

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 data sets, 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 data sets 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).

2.1 Imputing continuous data under normal models

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 missing data library, SAS procedure PROC MI). 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 data sets. For further details, see Schafer (1997) and Schimert et al. (2001). Below, a brief description of the MI process using data augmentation is given.

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 $|X|^{-\left(\frac{k+p+1}{2}\right)} \exp\left\{-\frac{1}{2} \text{tr}(\Gamma^{-1} X^{-1})\right\}$ 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 $f(\theta) \propto |\Sigma|^{-\left(\frac{k+p+2}{2}\right)} \exp\left\{-\frac{1}{2} \text{tr}(\Gamma^{-1} \Sigma^{-1})\right\} \exp\left\{-\frac{\tau}{2} (\mu - \mu_0)^T \Sigma^{-1} (\mu - \mu_0)\right\}$, and after some algebraic manipulations the complete-data likelihood can be re-expressed as $\propto |\Sigma|^{-\frac{n}{2}} \exp\left\{-\frac{n}{2} \text{tr}(\Sigma^{-1} S)\right\} \exp\left\{-\frac{n}{2} (\bar{y} - \mu)^T \Sigma^{-1} (\bar{y} - \mu)\right\}$. 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

$\tau^* = \tau + n$, $k^* = k + n$, $\mu_0^* = \left(\frac{n}{\tau+n}\right)\bar{y} + \left(\frac{\tau}{\tau+n}\right)\mu_0$, and $\Gamma^* = \left[\Gamma^{-1} + nS + \left(\frac{\tau n}{\tau+n}\right)(\bar{y} - \mu_0)(\bar{y} - \mu_0)^T\right]^{-1}$. When no strong prior information is available about θ , one may apply Bayes' theorem with the improper prior $f(\theta) \propto |\Sigma|^{-\left(\frac{p+1}{2}\right)}$, which is the limiting form of the normal inverted-Wishart density as $\tau \rightarrow 0$, $k \rightarrow -1$ and $\Gamma^{-1} \rightarrow 0$. In the simulated examples, this noninformative prior was used to reflect a state of relative ignorance which is often bluntly 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 .

2.2 Imputing binary data under saturated multinomial models

The saturated multinomial is more general than the multivariate normal in the sense that it allows for three-way or higher associations among the variables, whereas the multivariate normal captures simple associations only. However, the extra generality does not come for free. In many applications, some of the high-order associations may be poorly estimated.

Let Y_1, Y_2, \dots, Y_p denote a set of categorical variables. Suppose that Y_j takes possible values $1, 2, \dots, d_j$ for $j = 1, 2, \dots, p$. Here, levels $1, 2, \dots, n_j$ are nominal or unordered categories. If values of Y_1, Y_2, \dots, Y_p are measured for a sample of n units, then the complete data can be expressed as an $n \times p$ data matrix Y . If the sample units are independently and identically distributed (iid), then without loss of information one can reduce Y to a contingency table with D cells, where $D = \prod_{j=1}^p d_j$ is the number of distinct combinations of the levels of Y_1, Y_2, \dots, Y_p . Indexing the cells of the contingency table by a single subscript $d = 1, 2, \dots, D$,

and letting x_d be the number of sample units that fall into cell d , $x = (x_1, x_2, \dots, x_D)$ denotes the entire set of cell frequencies or counts. If the sample units are iid and the sample size $n = \sum_{d=1}^D x_d$ is regarded fixed, then x has a multinomial distribution. $x|\theta \sim M(n, \theta)$ indicates that x is multinomial with index n and parameter $\theta = (\theta_1, \theta_2, \dots, \theta_D)$, where θ_d is the probability that a unit falls into cell d . The probability distribution for x is given by $P(x|\theta) = \frac{n!}{x_1!x_2!\dots x_D!}\theta_1^{x_1}\theta_2^{x_2}\dots\theta_D^{x_D}$. The cell probabilities must satisfy $\sum_{d=1}^D \theta_d = 1$, therefore the multinomial has $D - 1$ free parameters. Such a model is said to be saturated, because it includes the maximum number of free parameters ($D - 1$). The multinomial distribution has the convenient property that after collapsing or partitioning the contingency table, the resulting distribution is also multinomial.

A common way to conduct Bayesian inference with a multinomial model is to choose a parametric family of prior distributions whose density has the same functional form as the likelihood $L(\theta|Y) \propto \theta_d^{x_d}$. Suppose that $\theta = (\theta_1, \theta_2, \dots, \theta_D)$ is a vector of random variables with the property that $\theta_d \geq 0$ for $d = 1, 2, \dots, D$ and $\sum_{d=1}^D \theta_d = 1$. Then θ is said to have a Dirichlet distribution, $\theta|\alpha \sim D(\alpha)$ with parameter $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_D)$ if its density is $P(\theta|\alpha) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\dots\Gamma(\alpha_D)}\theta_1^{\alpha_1-1}\theta_2^{\alpha_2-1}\dots\theta_D^{\alpha_D-1}$, where $\alpha_0 = \sum_{d=1}^D \alpha_d$. Multiplying the Dirichlet density by the multinomial likelihood produces a Dirichlet posterior with updated parameters α^* , where $\alpha_j^* = \alpha_j + x_j$ for $j = 1, 2, \dots, D$.

Creating multiply imputed data sets using log-linear models with multinomial likelihood and Dirichlet prior can be carried out with EM and data augmentation algorithms. Computational routines are roughly based on properties of multinomial and Dirichlet distributions, and grouping the observations with respect to their missingness patterns. Choosing the prior hyperparameters is an important step. Schafer (1997) discusses noninformative, flattening and data-dependent values for the hyperparameters. In this work, I choose to work with noninformative priors. When little prior information is available about θ , it may be plausible to take α 's equal to a common value, say c . However, there is no unique choice for c that represents a state of ignorance about θ . Without strong prior information, a reasonable point estimate for θ would be the ML estimate, $\hat{\theta} = (x_1/n, x_2/n, \dots, x_D/n)$. This is the posterior mean of θ under the improper prior with $c = 0$. It is also the posterior mode under the uniform prior with $c = 1$. Furthermore, Jeffrey's invariance principle leads to the choice of $c = 0.5$ which is the default value in Splus missing data library (Schimert et al., 2001). For this reason, we use this prior in the simulations described in

Section 3. When there are constraints that are imposed by a log-linear model, the likelihood follows a product multinomial distribution. In such cases, the constrained Dirichlet as the prior distribution needs to be adopted. Schafer's (1997) book includes a wealth of examples drawn from real-life studies along with comprehensive coverage of computational and algorithmic details in regard to imputation under the log-linear model; interested readers should consult with Schafer (1997) for deeper practical, computational and conceptual issues.

For the purposes of this paper, I focus on binary variables. In the next subsection, the question that motivates this work is re-visited.

2.3 Imputation and dichotomization: which one should come first?

The real data rarely conform with multivariate normality. However, if the immediate goal is to create multiply imputed data sets, which in turn to be dichotomized, the normal MI model may be tempting. On the other hand, erratic aspects of continuous densities may be eliminated to some extent when dichotomization is performed before carrying out MI under a log-linear model. The price to be paid may be thought of losing information with continuous-binary conversion before the imputation. Should one employ IMVND (Impute under Multivariate Normality then Dichotomize) or DILLM (Dichotomize then Impute under a Log-linear Model) strategies? Although intuition favors IMVND, it is worthwhile to design a simulation study that anchors a comparative view. An assessment of viability is warranted under both approaches via simulated scenarios that represent different sorts of distributional features of the continuous variables.

Describing a real phenomenon by generating an environment within which the process under consideration operates is not uncommon and is often the only feasible way of evaluation. For this reason, I devise a simulation study which I describe below in an attempt to answer the primary research question under consideration.

3 A simulation study

The framework of the simulation study consists of complete data generation from seven bivariate distributions, imposing missing values under ignorable missingness mechanisms, MI under IMVND and DILLM, 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, it is assumed 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:* $f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2\sigma^2}(y - \mu)^2\right)$, where μ and $\sigma > 0$ are location and scale parameters, respectively. We set $\mu = 1$ and $\sigma = 1$. The correlation was set equal to 0.2.
- *t distribution:* $f(y|\nu, \mu, \sigma) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} \left[1 + \frac{1}{\nu}\left(\frac{y-\mu}{\sigma}\right)^2\right]^{-(\nu+1)/2}$, 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, since smaller ν corresponds to heavier tails.
- *Laplace (double-exponential) distribution:* $f(y|\alpha, \lambda) = \frac{\lambda}{2} \exp(-\lambda|y - \alpha|)$, where α and $\lambda > 0$ are the location and inverse scale parameters, respectively. We set $\alpha = 1$, and $\lambda = 1$, 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.

- *Uniform distribution:* $f(y|a, b) = (b - a)^{-1}$, $a \leq y \leq b$ with $E[Y] = (b + a)/2$, where a and b are the lower and upper bounds of the support of y . The pair $(0, 1)$ is chosen 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 4×3 matrix format for easier visibility and interpretability. Beta and Weibull densities correspond to the first and second rows of the matrix graph, respectively; Normal-mixture densities are shown in the third and fourth row. Furthermore, $\text{plot}[u, v]$ stands for the plot in row u and column v .

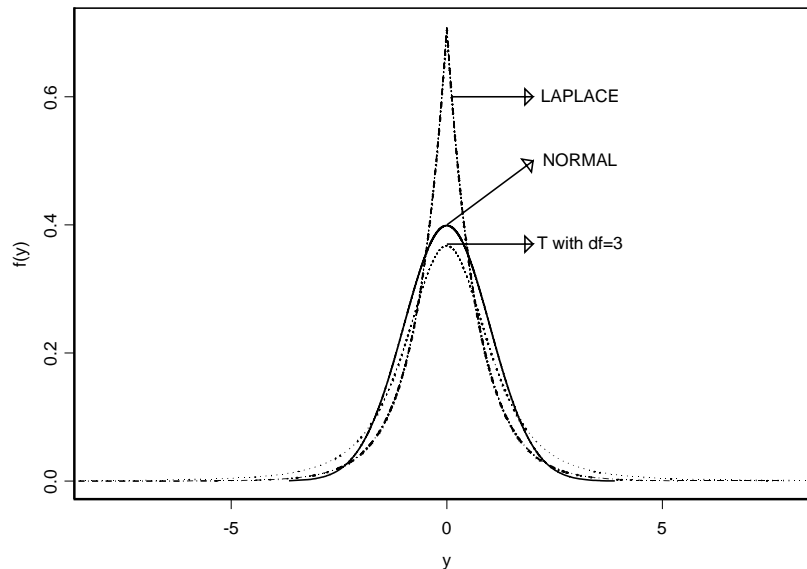


Figure 1: Density functions of Laplace($0, \sqrt{2}$), Normal($0, 1$) and $t(3, 0, 1/\sqrt{3})$ distributions. Parameters were chosen such that $E[Y] = 0$ and $Var[Y] = 1$ for all three densities. Note that the parameter values are different from the ones used in simulations. This graph was created for the purpose of exposition of different tail behaviors.

- Beta distribution:* $f(y|\alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}y^{\alpha-1}(1-y)^{\beta-1}$, where $0 < y < 1$, $\alpha > 0$ 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 5, 30, and 1.5 for β yield the densities of symmetric (plot [1,1]), moderately positively skewed (plot [1,2]), and heavily negatively skewed (plot [1,3]) shapes, respectively. Note that it cannot be heavily positively skewed.
- Weibull distribution:* $f(y|\gamma, \delta) = \frac{\delta}{\gamma^\delta}y^{\delta-1}exp(-(\frac{y}{\gamma})^\delta)$, where $y > 0$, $\gamma > 0$ 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 $\delta = 1.5, 3.6$, and 20, we obtain 'heavily positively skewed (plot [2,1]), symmetric (plot [2,2]), and moderately negatively skewed (plot [2,3]) densities, respectively. Note that it cannot be heavily negatively skewed.

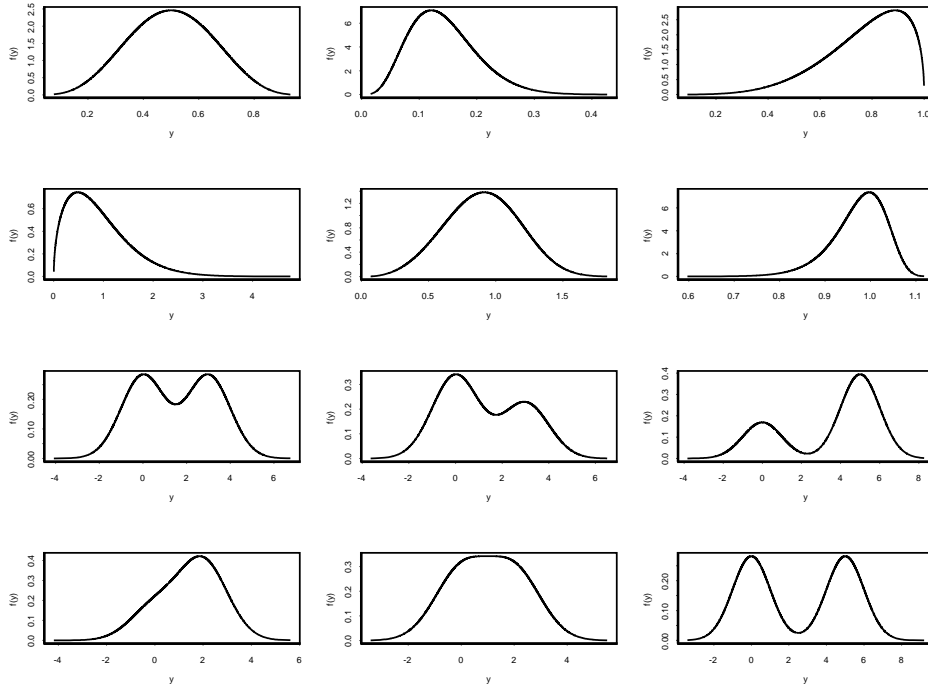


Figure 2: Density functions of Beta (first row), Weibull (second row), and Normal-mixture (third and fourth row) distributions for different sets of parameters.

- *Mixture of univariate normal distributions:*

$f(y|\mu_1, \mu_2, \sigma_1, \sigma_2, p) = \frac{p}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_1}{\sigma_1}\right)^2\right) + \frac{(1-p)}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu_2}{\sigma_2}\right)^2\right)$, 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 bimodal. Table 1 summarizes the set of assumed parameter values. The first three rows in Table 1 and the third row of Figure 2, and the last three rows in Table 1 and the fourth row of Figure 2 are in the same order.

Every distribution was chosen to address the key questions for empirically testing the performance of IMVND and DILLM strategies. 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. Skewness, multimodality and flat densities clearly violate the normality assumption and these situations should be examined for a

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

real assessment of the impact of departures from normality on the inferences drawn from two competing imputation models.

The number of subjects (n) in the simulated examples is 400. 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. This assumption does not have any impact on conclusions drawn in this work and can easily be relaxed. 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). Specifically, logit of the missingness probability in Y_2 is taken as a linear function of Y_1 . 75% of observations in Y_2 are assumed to be missing with MCAR and MAR mechanisms. This leads to 100 observed values on average. Since the implementation of multiple imputation under both models requires ignorability, nonignorable missingness mechanisms were not considered in the simulations. This issue is discussed further in Section 4.

3.1.3 Parameter estimation

The parameter of interest is the proportion of 1's ($p_2 = E[Y_2]$) under both approaches. The marginal expectation of binary variables (which also determines the variability of \hat{p}_2) is arguably the most commonly used quantity of interest in practice. Other important quantities such as odds ratios, regression and correlation coefficients could have been examined in a more complicated setting. However, simpler answers are necessary building-blocks for more complex ones.

3.1.4 Evaluation criteria

I created multiply imputed data sets with Splus 6.1 missing data library (Schimert et al., 2001). The procedure, which consists of complete data generation, imposing missing values, MI under IMVND and DILLM with data augmentation whose starting values were obtained from the EM algorithm, finding the estimates for the parameters p_2 , and combining them by Rubin's (1987) rules, was repeated 1000 times for each of the $2 \times 5 \times 16 = 160$ (two sets of nonresponse mechanisms, five different cut-off points for Bernoulli proportions, and sixteen different versions of seven distributions under consideration) scenarios. To make a real comparison, identical incomplete data sets were used for IMVND and DILLM for each of the 1000 replicates in the simulation. The relative performances were 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, $100 \times \frac{E(\hat{\theta}) - \theta}{\theta}$. A reasonable upper limit for the percentage bias can be taken as 5% in either direction.

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 my limited experience, it is advisable to consider both accuracy benchmarks simultaneously.

3.2 Results

The findings suggest that there is little or no discernible differences in inferences between MCAR and MAR. MI under normal and log-linear models is known to operate under ignorable nonresponse; that is, as long as the reasons for missingness (Y_1 under MAR) are included in the imputation process, differences are expected to be minimal. For this reason along with space restrictions, I only report the results for MCAR mechanism.

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 p_2 are tabulated for seven distributions. In Tables 2-6, *IM* stands for imputation method (IMVND or DILLM), “Cut-off” represents the threshold that leads to the proportion of 1’s whose expected values are shown under the column of TV (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. Table 3 to 6 include an additional column of parameters, pertinent to the defining characteristics of the densities.

Table 2 includes four densities. For the normal distribution, biases and coverages are within reasonable limits for both imputation models with a slight advantage of IMVND, as one would expect. For the t and Laplace distributions that are symmetric with heavier tails than normal, severe biases occur under IMVND for all cut-offs except when cut-off corresponds to 0.5 which is the mean and mode. Coverage rates are unacceptably low at the extremes (when TV is 0.1 or 0.9) probably owing to heavier tails. For the Uniform density that has a flat shape, DILLM gains a definite advantage over IMVND in terms of

accuracy measures, again except for $TV = 0.5$. The big picture in Table 2 is that underlying density should correctly be specified in order IMVND to perform well. When the density is symmetric and flat, IMVND performs satisfactorily only when the cut-off corresponds to the mean. If the density is unimodal with heavier tails in addition to being symmetric, coverage rates also suffer at the tails under IMVND, whereas DILLM does not seem to be sensitive to these complications in positive sense.

Skewness is another key issue that needs to be addressed. The Beta and Weibull distributions can be formed to yield symmetric and skewed shapes depending on the choice of parameter values. Results for the Beta distribution are tabulated in Table 2. When the density is symmetric ($\alpha = \beta = 5$), IMVND performs slightly better than DILLM in general. When it is positively skewed ($\alpha = 5, \beta = 30$) or negatively skewed ($\alpha = 5, \beta = 1.5$), DILLM outperforms IMVND with a significant margin for most cut-offs from bias and coverage standpoints. Only exception occurs when the cut-off is close to the mode of the distribution (positively skewed case with cut-off that leads to the true value of 0.1, see Figure 2). A similar message emerges for the Weibull distribution (see Table 4). In the symmetric case ($\gamma = 1, \delta = 3.6$), IMVND has a minor advantage over DILLM. However, when symmetry is perturbed in either direction (positive skewness [$\gamma = 1, \delta = 1.5$]), negative skewness [$\gamma = 1, \delta = 20$]), coverage percentages and biases suggest a definitive edge in favor of DILLM. Again, only visible exception is negatively skewed shape with a true value of 0.90 that happens to coincide with the mode. The bottom line is that symmetric densities that have a similar tail behavior to that of the normal give marginally better results under IMVND; and if skewness is present, DILLM gains an overwhelming advantage except when the selected cut-off points that anchor resulting Bernoulli proportions are very close the continuous modes.

Bimodality issues were explored through a mixture of univariate normal distributions (see Figure 2 and Table 1). Results for the Normal-mixture densities are shown in Tables 5 and 6. Across the six scenarios with this particular distributional form, DILLM appears to be the clear winner with one serious exception of unimodal, symmetric density (plot[4,2] in Figure 2) where the performances are compatible. Other situations that deliver fairly similar biases and coverages are symmetric shapes (plot[3,1] and plot[4,3]) with a true value of 0.5, and cases where mode(s) is(are) in close proximity with cut-offs culminating in the specified true values of marginal means (the proportions 0.1 and 0.9 in plot[3,1], 0.6 in plot[3,2], and

0.9 in plot[4, 1]).

One common trend across Tables 2-6 is that average widths (and not surprisingly RMSE's) are smaller under IMVND in most cases. However, narrower intervals are good only when the coverage is correct. In most applications, there is a trade-off between bias and precision. Generally, one cannot reduce the bias and variability simultaneously, because they typically move in opposite directions. For further discussion of this issue, see the next section.

The overall conclusion is that when continuous measurements are obtained with an ultimate interest in dichotomized versions of them, DILLM should be the preferred strategy over IMVND, except for two situations. 1) Underlying densities are a close approximation to multivariate normality in terms of modality, symmetry and tail behavior. 2) The cut-off point coincides with the mode. Although I do not have a cogent mathematical argument to explain this phenomenon, a conceptual reasoning might be that the erratic aspects of the continuous densities are probably eliminated through collapsing to some extent, when dichotomization is performed before carrying out MI under a log-linear model.

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 the simulations zero correlation between the two variables was assumed 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 both approaches, assuming uncorrelatedness represents a worst-case situation in the sense that incompletely observed variable (Y_2) literally does not receive any help from the fully observed variable (Y_1). Furthermore,

Table 2: Results for *Normal*, *t*, *Laplace* and *Uniform* distributions.

Distribution	IM	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
NORMAL	IMVND	-0.28155	0.1	0.10188	1.88	8.22	0.0229	96.0	0.115
	DILLM			0.10315	3.16	10.88	0.0291	95.1	0.130
	IMVND	0.74665	0.4	0.39561	-1.10	-10.48	0.0420	95.9	0.184
	DILLM			0.39782	-0.54	-4.21	0.0516	94.3	0.211
	IMVND	1.00000	0.5	0.49661	-0.68	-8.40	0.0404	95.5	0.191
	DILLM			0.49334	-1.33	-12.85	0.0521	94.8	0.213
	IMVND	1.25335	0.6	0.59905	-0.16	-2.53	0.0377	95.2	0.187
	DILLM			0.59745	-0.42	-5.76	0.0442	95.5	0.206
	IMVND	2.28155	0.9	0.89851	-0.17	-6.38	0.0233	96.2	0.112
	DILLM			0.89497	-0.56	-17.35	0.0293	95.3	0.129
Distribution	IM	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
T	IMVND	-0.63774	0.1	0.14246	42.46	97.45	0.0608	78.9	0.135
	DILLM			0.10561	5.61	17.48	0.0325	93.7	0.130
	IMVND	0.72333	0.4	0.42279	5.70	56.00	0.0465	95.4	0.190
	DILLM			0.40297	0.74	6.25	0.0475	94.5	0.210
	IMVND	1.00000	0.5	0.49829	-0.34	-4.27	0.0398	95.2	0.190
	DILLM			0.49757	-0.48	-5.08	0.0477	95.8	0.211
	IMVND	1.27667	0.6	0.57410	-4.32	-60.37	0.0500	92.9	0.188
	DILLM			0.59928	-0.12	-1.35	0.0531	91.1	0.206
	IMVND	2.63774	0.9	0.85508	-4.99	-103.02	0.0625	75.6	0.134
	DILLM			0.89308	-0.77	-22.54	0.0314	95.9	0.133
Distribution	IM	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
LAPLACE	IMVND	-0.60944	0.1	0.12056	20.56	63.10	0.0385	88.9	0.126
	DILLM			0.10541	5.41	16.85	0.0325	94.7	0.133
	IMVND	0.77686	0.4	0.42865	7.16	64.89	0.0525	92.9	0.185
	DILLM			0.40430	1.07	8.07	0.0533	93.3	0.209
	IMVND	1.00000	0.5	0.49912	-0.18	-2.24	0.0391	95.5	0.188
	DILLM			0.50222	0.44	4.39	0.0505	93.0	0.216
	IMVND	1.22314	0.6	0.57749	-3.75	-55.52	0.0463	93.3	0.187
	DILLM			0.60311	0.52	5.80	0.0536	92.6	0.210
	IMVND	2.60944	0.9	0.87706	-2.55	-67.76	0.0408	89.4	0.129
	DILLM			0.89642	-0.40	-11.02	0.0326	92.1	0.132
Distribution	IM	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
UNIFORM	IMVND	0.1	0.1	0.08854	-11.46	-57.83	0.0228	94.5	0.106
	DILLM			0.10379	3.79	12.62	0.0302	95.7	0.130
	IMVND	0.4	0.4	0.37683	-5.79	-53.52	0.0490	90.5	0.177
	DILLM			0.40480	1.20	9.19	0.0523	91.4	0.208
	IMVND	0.5	0.5	0.50282	0.56	6.36	0.0443	93.8	0.189
	DILLM			0.50307	0.61	5.60	0.0547	94.7	0.217
	IMVND	0.6	0.6	0.62609	4.35	65.15	0.0477	91.4	0.187
	DILLM			0.59762	-0.40	-4.72	0.0502	94.8	0.205
	IMVND	0.9	0.9	0.91122	1.25	56.89	0.0227	96.0	0.105
	DILLM			0.89548	-0.50	-15.43	0.0296	95.2	0.132

Table 3: Results for *Beta* distribution with $\alpha = 5$ and $\beta = 5, 30,$ and 1.5 .

IM	β	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
IMVND	5	0.30088	0.1	0.09861	-1.39	-5.97	0.0232	96.1	0.112
DILLM				0.10543	5.43	17.17	0.0320	92.5	0.128
IMVND	5	0.45896	0.4	0.40178	0.44	4.14	0.0428	95.2	0.184
DILLM				0.40754	1.89	14.48	0.0525	93.9	0.210
IMVND	5	0.50000	0.5	0.49733	-0.53	-6.59	0.0404	95.5	0.189
DILLM				0.49796	-0.41	-4.24	0.0481	94.9	0.211
IMVND	5	0.54104	0.6	0.60672	1.12	14.91	0.0455	94.6	0.188
DILLM				0.60451	0.75	8.58	0.0527	92.8	0.210
IMVND	5	0.69912	0.9	0.90212	0.24	9.86	0.0216	95.8	0.111
DILLM				0.89363	-0.71	-22.72	0.0287	95.6	0.131
IMVND	30	0.07342	0.1	0.10187	1.87	6.37	0.0294	91.3	0.115
DILLM				0.11339	13.39	41.79	0.0346	95.6	0.135
IMVND	30	0.12210	0.4	0.43242	8.11	72.63	0.0551	88.1	0.186
DILLM				0.40206	0.52	3.89	0.0528	93.0	0.211
IMVND	30	0.13615	0.5	0.53036	6.07	78.70	0.0490	92.9	0.187
DILLM				0.49549	-0.90	-8.89	0.0508	93.8	0.213
IMVND	30	0.15106	0.6	0.63158	5.26	86.32	0.0483	92.5	0.183
DILLM				0.60312	0.52	6.75	0.0461	95.1	0.203
IMVND	30	0.22186	0.9	0.88493	-1.68	-63.11	0.0282	96.0	0.122
DILLM				0.89379	-0.68	-19.55	0.0323	95.4	0.130
IMVND	1.5	0.55071	0.1	0.12174	21.74	101.90	0.0304	95.7	0.129
DILLM				0.10599	5.99	20.28	0.0300	94.6	0.130
IMVND	1.5	0.75539	0.4	0.35571	-11.07	-121.43	0.0573	83.2	0.181
DILLM				0.40914	2.29	17.50	0.0529	94.1	0.211
IMVND	1.5	0.79907	0.5	0.44641	-10.72	-133.51	0.0669	81.9	0.192
DILLM				0.49994	-0.01	-12.85	0.0523	94.9	0.214
IMVND	1.5	0.83761	0.6	0.55397	-7.67	-104.52	0.0636	84.6	0.189
DILLM				0.59660	-0.57	-6.62	0.0514	94.0	0.209
IMVND	1.5	0.94559	0.9	0.91419	1.58	49.91	0.0317	80.4	0.101
DILLM				0.89296	-0.78	-21.46	0.0335	92.5	0.131

Table 4: Results for *Weibull* distribution with $\gamma = 1$, and $\delta = 1.5, 3.6,$ and 20 .

IM	δ	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
IMVND	1.5	0.22308	0.1	0.09362	-6.38	-19.30	0.0336	83.2	0.105
DILLM				0.10834	8.34	26.08	0.0330	95.5	0.130
IMVND	1.5	0.63902	0.4	0.45520	13.80	131.66	0.0693	79.5	0.187
DILLM				0.40591	1.48	11.72	0.0507	93.6	0.211
IMVND	1.5	0.78322	0.5	0.55998	12.00	156.73	0.0711	81.4	0.185
DILLM				0.50430	0.86	8.41	0.0511	93.9	0.208
IMVND	1.5	0.94338	0.6	0.64390	7.32	131.36	0.0551	89.5	0.183
DILLM				0.59586	-0.69	-8.56	0.0484	95.0	0.214
IMVND	1.5	1.74372	0.9	0.87178	-3.14	-131.61	0.0354	95.3	0.131
DILLM				0.89525	-0.53	-15.06	0.0318	95.1	0.128
IMVND	3.6	0.53521	0.1	0.10294	2.94	8.22	0.0254	95.4	0.115
DILLM				0.10675	6.75	22.93	0.0301	94.9	0.131
IMVND	3.6	0.82978	0.4	0.39344	-1.64	-15.28	0.0433	95.5	0.190
DILLM				0.39734	-0.67	-5.36	0.0495	95.2	0.205
IMVND	3.6	0.90320	0.5	0.49029	-1.94	-21.82	0.0454	96.1	0.190
DILLM				0.49331	-1.34	-12.77	0.0527	94.5	0.216
IMVND	3.6	0.97601	0.6	0.59812	-0.31	-4.56	0.0411	95.7	0.186
DILLM				0.59511	-0.82	-9.80	0.0500	93.1	0.209
IMVND	3.6	1.26071	0.9	0.90218	0.24	9.20	0.0237	96.4	0.112
DILLM				0.89447	-0.61	-19.34	0.0291	95.5	0.126
IMVND	20	0.89358	0.1	0.12125	21.25	85.71	0.0326	95.6	0.125
DILLM				0.10665	6.65	21.67	0.0313	94.6	0.132
IMVND	20	0.96697	0.4	0.36701	-8.25	-87.03	0.0502	88.9	0.181
DILLM				0.39665	-0.84	-6.91	0.0485	95.2	0.211
IMVND	20	0.98184	0.5	0.46198	-7.61	-91.44	0.0563	90.0	0.192
DILLM				0.50316	0.63	6.18	0.0510	94.8	0.211
IMVND	20	0.99564	0.6	0.55206	-7.99	-112.10	0.0642	86.4	0.189
DILLM				0.59510	-0.82	-9.26	0.0530	93.1	0.207
IMVND	20	1.04258	0.9	0.90227	0.25	7.85	0.0290	90.1	0.109
DILLM				0.89401	-0.67	-21.34	0.0286	95.2	0.131

Table 5: Results for *Normal-mixture* densities shown in the third row of Figure 2. $(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$ refers to the vector of the mixing proportion, mean and variance parameters in the mixture.

IM	$(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
IMVND	(0.5,0,1,3,1)	-0.84275	0.1	0.10056	0.56	2.41	0.0280	95.2	0.125
DILLM				0.10566	5.66	17.59	0.0376	94.1	0.147
IMVND	(0.5,0,1,3,1)	0.79254	0.4	0.36414	-8.97	-127.94	0.0470	91.8	0.173
DILLM				0.39740	-0.65	-6.88	0.0428	95.5	0.196
IMVND	(0.5,0,1,3,1)	1.50000	0.5	0.49888	-0.23	-3.37	0.0382	96.0	0.177
DILLM				0.49552	-0.90	-10.95	0.0460	95.3	0.186
IMVND	(0.5,0,1,3,1)	2.20619	0.6	0.63433	5.72	94.44	0.0541	87.1	0.177
DILLM				0.59739	-0.44	-5.96	0.0487	92.6	0.198
IMVND	(0.5,0,1,3,1)	3.84511	0.9	0.89799	-0.22	-8.54	0.0286	95.1	0.124
DILLM				0.89398	-0.67	-18.23	0.0385	94.7	0.146
IMVND	(0.6,0,1,3,1)	-0.67625	0.1	0.10903	9.03	40.89	0.0288	95.5	0.131
DILLM				0.10212	2.12	6.98	0.0353	92.8	0.150
IMVND	(0.6,0,1,3,1)	1.39396	0.4	0.35567	-11.08	-130.32	0.0608	80.5	0.175
DILLM				0.39995	-0.01	-0.11	0.0495	94.9	0.218
IMVND	(0.6,0,1,3,1)	2.07845	0.5	0.45842	-8.32	-118.72	0.0593	84.6	0.178
DILLM				0.49715	-0.57	-6.23	0.0508	93.9	0.200
IMVND	(0.6,0,1,3,1)	2.57504	0.6	0.59537	-0.77	-13.58	0.0394	95.5	0.177
DILLM				0.60154	0.26	3.97	0.0437	95.4	0.181
IMVND	(0.6,0,1,3,1)	3.96639	0.9	0.91133	1.26	59.88	0.0270	93.4	0.115
DILLM				0.89864	-0.15	-4.51	0.0350	93.8	0.149
IMVND	(0.3,0,1,5,1)	-1.06822	0.1	0.08402	-15.98	-79.65	0.0306	84.2	0.106
DILLM				0.10465	4.65	14.70	0.0369	92.6	0.145
IMVND	(0.3,0,1,5,1)	0.17801	0.4	0.49404	23.51	295.29	0.0993	29.3	0.173
DILLM				0.40384	0.96	9.01	0.0437	93.5	0.196
IMVND	(0.3,0,1,5,1)	0.56580	0.5	0.57807	15.61	224.80	0.0904	49.4	0.173
DILLM				0.50370	0.74	7.53	0.0541	95.8	0.220
IMVND	(0.3,0,1,5,1)	1.06404	0.6	0.64840	8.07	139.91	0.0644	79.1	0.175
DILLM				0.59926	-0.12	-1.41	0.0575	93.6	0.222
IMVND	(0.3,0,1,5,1)	5.42914	0.9	0.86394	-4.01	165.39	0.0471	96.0	0.140
DILLM				0.89417	-0.65	-19.86	0.0348	95.2	0.151

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

IM	$(p, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$	Cut-off	TV	AE	PB	SB	RMSE	CR	AW
IMVND	(0.3,0,1,2,1)	-1.06985	0.1	0.08964	-10.36	43.71	0.0308	88.1	0.120
DILLM				0.10467	4.67	15.71	0.0350	95.3	0.146
IMVND	(0.3,0,1,2,1)	0.14400	0.4	0.42162	5.40	53.72	0.0506	93.6	0.197
DILLM				0.40773	1.93	15.33	0.0558	94.0	0.218
IMVND	(0.3,0,1,2,1)	0.48505	0.5	0.52269	4.54	58.20	0.0500	95.2	0.197
DILLM				0.50038	0.08	0.77	0.0543	93.2	0.215
IMVND	(0.3,0,1,2,1)	0.85196	0.6	0.62724	2.87	48.53	0.0444	95.5	0.197
DILLM				0.59740	-0.43	-5.29	0.0542	95.3	0.227
IMVND	(0.3,0,1,2,1)	2.47104	0.9	0.89239	-0.85	-32.79	0.0294	95.7	0.136
DILLM				0.89436	-0.63	-18.41	0.0361	95.2	0.150
IMVND	(0.5,0,1,2,1)	-0.85007	0.1	0.10135	1.35	8.22	0.0268	96.4	0.127
DILLM				0.10598	5.98	20.54	0.0346	95.9	0.151
IMVND	(0.5,0,1,2,1)	0.58662	0.4	0.38950	-2.62	-28.82	0.0428	95.5	0.192
DILLM				0.40020	0.05	0.43	0.0507	96.0	0.217
IMVND	(0.5,0,1,2,1)	1.00000	0.5	0.49882	-0.24	-3.03	0.0440	95.3	0.194
DILLM				0.50001	0.00	0.03	0.0510	96.1	0.216
IMVND	(0.5,0,1,2,1)	1.41159	0.6	0.61233	2.06	32.60	0.0447	95.8	0.195
DILLM				0.60024	0.04	0.48	0.0543	93.8	0.217
IMVND	(0.5,0,1,2,1)	2.84628	0.9	0.90051	0.06	2.20	0.0280	95.4	0.127
DILLM				0.89488	-0.57	-15.84	0.0377	92.1	0.149
IMVND	(0.5,0,1,5,1)	-0.84078	0.1	0.11034	10.34	45.55	0.0299	95.3	0.125
DILLM				0.10187	1.87	5.78	0.0373	90.4	0.144
IMVND	(0.5,0,1,5,1)	0.84134	0.4	0.32621	-18.45	-243.53	0.0847	43.8	0.154
DILLM				0.40084	0.21	1.93	0.0485	94.6	0.189
IMVND	(0.5,0,1,5,1)	2.50000	0.5	0.49949	-0.10	-1.91	0.0316	95.4	0.149
DILLM				0.49877	-0.25	-3.91	0.0364	95.1	0.144
IMVND	(0.5,0,1,5,1)	4.15832	0.6	0.67819	13.03	250.20	0.0892	41.2	0.156
DILLM				0.60621	1.03	14.96	0.0468	95.8	0.194
IMVND	(0.5,0,1,5,1)	5.84182	0.9	0.88745	-1.39	-54.44	0.0312	95.3	0.125
DILLM				0.89786	-0.24	-7.56	0.0333	94.5	0.149

how continuous and binary correlations after the dichotomization are related is a fuzzy notion and merits further study as far as the comparative performance of IMVND and DILLM.

As noted in Section 2, the computational algorithm for creating multiple imputations relies on Bayesian arguments and users of this method must choose hyperparameters for the underlying prior distributions. In this paper, I chose such priors to be noninformative 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. Using noninformative priors in both imputation models is believed to lead to a fair and impartial comparison.

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 IMVND and DILLM set-ups performs with respect to incomplete continuous outcomes that exhibit varying distributional properties. Even though the simulated data sets are only a small portion of what may arise in applications and they are not sufficiently complex compared to the real data sets, I believe that a simulation assessment and evaluation based on a fairly broad study that includes many non-Gaussian features is insightful.

Although the bias measures strongly favor DILLM, average widths and RMSE’s are smaller in IMVND compared to DILLM. As mentioned in Section 3, narrower confidence intervals are good, but only subject to the correct coverage rate which is direct function of the bias. When Type I errors are not properly controlled as measured by the coverage rate, it means that the procedure is not working well. Do we err on the side of bias or variability? In the less-than-perfect world we live in, if one is forced to choose between the two of them, the “less bias” option is more attractive. The main reason is that variance reduction techniques are more general, better-developed and better-accepted while bias reduction techniques are

typically problem-specific. Another reason is that large variability implies a higher chance of capturing the truth in the confidence intervals. In other words, a reasonable degree of inflation in the variation is acceptable if the reward is unbiasedness. Finally, if the estimates are biased, what sense does it make to have small variability? Perhaps, I am placing unduly heavy weight on the coverage rate, but I reckon that this is the right way of evaluating the plausibility of a given procedure. Needless to say: "got to have a small variation" platitude is not really valid here.

There are a few other points that deserve discussion. First, imputing under the normal and log-linear models explicitly requires ignorability given the current state of the research. Nonignorable modeling is a separate issue and beyond the scope of this manuscript. The reason is that imposing missing values by ignorable missingness mechanisms under which both normal and log-linear imputation models are known to operate is essential to make a genuine comparison. Second, no negative connotations are attached to other MI models (Van Buuren et al., 1999; Raghunathan et al., 2001) which may potentially outperform IMVND and DILLM in some settings. Third, thresholds that govern the distribution of dichotomous outcomes may be controversial and different thresholds may cause major changes in the inferences. However, since this equally applies to both approaches, it is inconsequential. Fourth, no theoretical justifications are provided as to why and how DILLM should be the preferred strategy when the ultimate interest is about the dichotomized outcomes. This study is limited to a practical domain and the goal is giving practical advice to applied researchers. Given sufficient effort, a more theoretical reasoning can be developed. Finally, a comprehensive account on deep properties of the two imputation techniques is not intended to be given in this article. There are many twists and angles, especially in MI under a log-linear model such as structural zeroes and insufficient amount of data for supporting arbitrarily complex associations among the variables. Again, the purpose is limited to simple recommendations as to how one should proceed when faced with this specific missing-data problem.

I conclude with a re-iteration of the primary message of this study: When continuous measurements are obtained with an ultimate interest in dichotomized versions of them, dichotomization before carrying out a log-linear imputation should be the preferred strategy over more intuitive approach of imputing under a Gaussian model before dichotomization, except for the situations where underlying densities are

a close approximation to multivariate normality in terms of modality, symmetry and tail behavior; and the situations where cut-offs coincide with the mode. A possible explanation is that erratic/idiosyncratic aspects that are not accommodated by a Gaussian model are probably transformed into better-behaving discrete trends in this particular missing-data setting. This premise outweighs the assertion that continuous variables inherently carry more information, leading to a counter-intuitive, but potentially useful result for practitioners. A comparison for relative performances of the two imputation approaches with ordinal(ized) outcomes is a natural continuation and will be taken up in future work.

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.
- Little RJA, Rubin DB. (2002), *Statistical Analysis with Missing Data*. Second Edition. New York: Wiley.
- Raghunathan TE, Lepkowski JM, van Hoewyk J, Solenberger P. (2001), A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27, 85–95.
- 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.
- Van Buuren S, Boshuizen HC, Knook L. (1999), Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in Medicine*, 18, 681–694.