

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

*Technical report#:2006-003
May 2006*

Title: A simple suggestion on imputing continuous data when
the eventual interest pertains to ordinalized outcomes via
threshold concept

Author(s): Hakan Demirtas

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

A simple suggestion on imputing continuous data when the eventual interest pertains to ordinalized outcomes via threshold concept

Hakan Demirtas *

April 12, 2006

Abstract

Multiple imputation under the multivariate normality assumption has often been considered a workable model-based approach in dealing with incomplete continuous data. A situation where the measurements are taken on a continuous scale with an eventual interest in ordinalized versions via threshold concept is commonly encountered in applied research, especially in medical and social sciences. In practice, researchers ordinarily impute missing values for continuous outcomes under a Gaussian imputation model, and then ordinalize them via pre-specified cutoff points. An alternate strategy is creating multiply imputed data sets after ordinalization 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 broad range of simulated incomplete data sets that exhibit varying distributional characteristics such as skewness and multimodality. Behavior of efficiency and accuracy measures was investigated to determine the degree to which the procedures work appositely. The conclusion drawn is that ordinalization before carrying out a log-linear imputation should be the preferred procedure except for a few special cases. It is recommended that researchers use the less common second strategy whenever the interest centers on ordinal quantities that are obtained through underlying continuous measurements. This postulate is probably due to the transformation of non-Gaussian features into better-behaving categorical trends in this particular missing-data environment. This premise preponderates the factual argument that continuous variables intrinsically convey more information, leading to a counter-intuitive, but potentially beneficial result for practitioners.

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

1 Introduction and motivation

Missing data is the rule rather than the aberration in statistical practice. Determining an appropriate analytical strategy in the absence of completeness is a consequential focus of scientific exploration on account of the extra intricacy that arises through missing data. Missing values generally sophisticate 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).

Improvements in computational statistics have produced flexible missing-data procedures with a sensible statistical element. One of these procedures involves multiple imputation (MI) (Rubin, 1987), a simulation technique that replaces each missing datum with a set plausible values. The versions of complete data

*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.

are then analyzed by standard complete-data methods and the results are combined into a single illative summary using arithmetic rules to yield estimates, standard errors and p-values that formally consolidate missing data uncertainty into the modeling process. The key ideas and benefits of MI were reviewed by Rubin (1996) and Schafer (1997, 1999).

The fundamental 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 entails propounding 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 is then also 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 ordinalized versions through pre-specified threshold points. In medicine, blood pressure and cholesterol levels are often categorized. Other examples include size of a tumor, progression of a disease, etc. In fact, many diagnostic procedures in health sciences 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 ordinalization. When incomplete continuous data are collected, should one impute continuous outcomes under the normality assumption and consequently ordinalize the completed data, or is it better to ordinalize the outcomes before carrying out MI under a log-linear model? This article examines the credibility of both approaches via a broad range of simulated examples typifying 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 inherently 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 simulation study that spans different underlying distributional properties of the original outcomes is presented. Section 4 includes concluding remarks and discussion.

2 Overview of missing data and multiple imputation

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). Under MAR, missingness is related to the observed data, but conditionally independent of the missing responses. A special case of MAR is missing completely at random (MCAR). Under MCAR, nonresponse is independent of observed and unobserved data; a weaker version of the MCAR assumption allows dependence on fully observed covariates. If the response probabilities depend on unobserved data; in this case, the missing values are said to be missing not at random (MNAR). A missing-data mechanism is said to be ignorable if the missing data are MAR or MCAR, together with a minor technical condition called distinctness (Little and Rubin, 2002).

Multiple imputation (MI) is a Monte Carlo technique (Rubin 1987, 1996) in which the missing values are replaced by a set of 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. In practice, a large number of multiple imputations is not required; sufficiently accurate results can often be obtained with several imputations. Once the imputations have been created, the 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 several 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 $|X|^{-\frac{(k+p+1)}{2}} \exp\{-\frac{1}{2}\text{tr}(\Gamma^{-1}X^{-1})\}$ 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 the simulated examples, this noninformative prior was used to reflect a state of relative ignorance.

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 ordinal 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, we 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. Interested readers should consult with Schafer (1997) for deeper practical, computational and conceptual issues.

In the next subsection, the question that motivates this work is re-visited.

2.3 Imputation and ordinalization: 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 ordinalized, the normal MI model may be tempting. On the other hand, erratic aspects of continuous densities may be eliminated to some degree when ordinalization is performed before carrying out MI under a log-linear model. The price to be paid may be thought of losing information with continuous-ordinal conversion before the imputation. Should one employ IMVNO (Impute under Multivariate Normality then Ordinalize) or OILLM (Ordinalize then Impute under a Log-linear Model) strategies? Although intuition favors IMVNO, 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 which consists of imperfect proxies of what is believed to be underlying truth 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 IMVNO and OILLM, 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:* $f(y|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{\sigma^2}(y - \mu)^2\right)$, 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.

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, negatively skewed
0.5	0	1	2	1	unimodal, balanced, small curvature
0.5	0	1	5	1	bimodal, balanced, far modes

- *t distribution*: $f(y|\nu, \mu, \sigma) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)\sqrt{\nu\pi}\sigma} [1 + \frac{1}{\nu}(\frac{y-\mu}{\sigma})^2]^{-(\nu+1)/2}$, 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.
- *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).
- *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, positively skewed, and negatively skewed shapes, respectively.
- *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 positively skewed, symmetric, and negatively skewed densities, respectively.
- *Mixture of univariate normal distributions*:
 $f(y|\mu_1, \mu_2, \sigma_1, \sigma_2, p) = \frac{p}{\sigma_1\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{y-\mu_1}{\sigma_1})^2) + \frac{(1-p)}{\sigma_2\sqrt{2\pi}} \exp(-\frac{1}{2}(\frac{y-\mu_2}{\sigma_2})^2)$, 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. (In the presence of bimodality, “balanced” is used rather than “symmetric” which may be vague.)

Every distribution was chosen to address the key questions for empirically testing the performance of IMVNO and OILLM 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 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 are 100, 500, and 1000. 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 25, 125, and 250 observed values on average, given different values of N . 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 number of ordinal categories was chosen to be four (1, 2, 3, 4). The relative proportions of ordinal categories were determined by the following five sets of cutoffs through percentiles of the respective distributions: (0.25, 0.25, 0.25, 0.25), (0.40, 0.30, 0.20, 0.10), (0.10, 0.20, 0.30, 0.40), (0.40, 0.10, 0.10, 0.40), and (0.10, 0.40, 0.40, 0.10). The percentiles that lead to these sets of proportions were computed either analytically or by simulation, depending on the distribution. The parameters of interest are the marginal proportions of observed categories in Y_2 ($p_{21}, p_{22}, p_{23}, p_{24}$) under both approaches. Other 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 IMVNO and OILLM with data augmentation whose starting values were obtained from the EM algorithm, finding the estimates for the parameters, and combining them by Rubin's (1987) rules, was repeated 1000 times for each of the $3 \times 2 \times 5 \times 16 = 480$ (three sets of sample sizes, two sets of nonresponse mechanisms, five different cutoff points for ordinal proportions, and sixteen different versions of seven distributions under

consideration) scenarios. To make a real comparison, identical incomplete data sets were used for IMVNO and OILLM 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).

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. For these reasons, it is advisable to consider both accuracy benchmarks simultaneously.

3.2 Results

In Table 2, the results for $N = 500$ were tabulated since the differences across sample sizes were little or indiscernible. Furthermore, due to space limitations, aggregated versions of the results were reported rather than full ones.¹ There are four parameters and five cutoff vectors, leading to 20 combinations. The SB column refers to the number of times where standardized bias under IMVNO or OILLM is smaller than the other across the 20 combinations. For example, $x - y$ under the SB column means that in x cases, the standardized bias under IMVNO is smaller than that under OILLM, whereas in y cases, the opposite is true. The PB column stands for the number of cases (out of 20) where the absolute percentage bias is smaller than 0.05, for IMVNO and OILLM, respectively. Similarly, the CR column represents the number of cases where the coverage rate is greater than 90% under both approaches. The assumed missingness mechanisms and the key distributional properties were also included. Other evaluation criteria such as root mean-squared error and average widths were examined, but were not reported for brevity.

Across the table, no significant differences between MCAR and MAR mechanisms were found. For the normal distribution, biases give advantage to IMVNO, as one would expect, with comparable coverage rates. For the t and Laplace distributions which have heavier tails than normal, OILLM gains a significant edge across all metrics. For the Beta and Weibull distributions which could exhibit varying distributional

¹Full results are available upon request.

characteristics with respect to symmetry/skewness depending on the choice of parameters, IMVNO performs better when the marginal shape is symmetric with similar tail behavior to normal in terms of bias, whereas OILMM's performance is superior when the skewness exists in either direction; the coverage rates under these two distributions are either similar or in favor of OILMM. For the normal mixtures, OILLM is the clear winner across all three comparison columns with the exception that the coverage rates are comparable when the density is unimodal and balanced.

It should be noted that missing values were imposed based on continuous measurements. In a sense, OILLM seems to beat IMVNO in its own game from the missingness-mechanism standpoint (with a few exceptions) since the discrete data model and the mechanism that leads to the incomplete data are not truly compatible. At this point, a natural question arises: How would OILLM perform when the mechanism that drives the nonresponse is based on the ordinalized data, not the continuous data, compared to OILLM that was implemented before? Imposing missing values using ordinalized data did not yield any sizable differences under MCAR. However, further improvements were observed under MAR.

The overall conclusion is that when continuous measurements are obtained with an eventual interest in ordinalized versions of them, OILLM should be the preferred strategy over IMVNO, except for the situations where the underlying densities are a close approximation to multivariate normality in terms of modality, symmetry and tail behavior. Although a cogent mathematical argument is not provided to explain this phenomenon, a conceptual reasoning could be that the erratic aspects of the continuous densities are probably eliminated through collapsing to some extent, when ordinalization is performed before carrying out MI under a log-linear model.

4 Discussion

It should be noted that although transformations 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, how continuous and ordinal correlations after the ordinalization are related is a fuzzy notion and merits further study as far as the comparative performance of IMVNO and OILLM.

Table 2: Aggregated results for $N = 500$ for seven distributions across 20 marginal proportion-cutoff combinations.

Distribution	Mechanism	Property	SB	$PB < 0.05$	$CR > 0.90$	
Normal	MCAR	Symmetric	16 – 4	20 – 14	20 – 20	
	MAR		16 – 4	19 – 10	20 – 20	
T	MCAR	Symmetric-heavier tails	3 – 17	2 – 15	6 – 20	
	MAR		4 – 16	4 – 14	6 – 20	
Laplace	MCAR	Symmetric-heavier tails	0 – 20	9 – 16	11 – 20	
	MAR		3 – 17	1 – 10	12 – 19	
Uniform	MCAR	Flat density	3 – 17	4 – 16	13 – 20	
	MAR		4 – 16	4 – 14	15 – 20	
Beta	MCAR	Negatively skewed	3 – 17	2 – 14	10 – 20	
	MAR		2 – 18	2 – 14	10 – 20	
	MCAR	Symmetric	12 – 8	18 – 15	20 – 20	
	MAR		12 – 8	19 – 14	20 – 20	
	MCAR	Positively skewed	5 – 15	4 – 13	16 – 20	
	MAR		5 – 15	6 – 13	17 – 20	
Weibull	MCAR	Positively skewed	3 – 17	2 – 15	9 – 20	
	MAR		3 – 17	4 – 14	10 – 20	
	MCAR	Symmetric	17 – 3	20 – 14	20 – 20	
	MAR		15 – 5	20 – 15	20 – 20	
	MCAR	Negatively skewed	3 – 17	4 – 14	11 – 20	
	MAR		4 – 16	6 – 13	12 – 20	
	Normal-mixture	MCAR	Bimodal-balanced-close modes	5 – 15	6 – 14	8 – 20
		MAR		6 – 14	4 – 13	13 – 19
MCAR		Bimodal-unbalanced-close modes	1 – 19	3 – 15	11 – 20	
MAR			5 – 15	5 – 13	13 – 20	
MCAR		Bimodal-unbalanced-far modes	0 – 20	1 – 15	3 – 20	
MAR			1 – 19	1 – 17	2 – 20	
MCAR		Unimodal-negatively skewed	2 – 28	9 – 14	19 – 20	
MAR			4 – 16	11 – 16	20 – 20	
MCAR		Unimodal-balanced-small curvature	5 – 15	10 – 14	20 – 20	
MAR			10 – 10	11 – 14	20 – 20	
MCAR		Bimodal-balanced-far modes	0 – 20	2 – 16	8 – 20	
MAR			1 – 19	2 – 14	7 – 20	

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 IMVNO and OILLM set-ups performs with respect to incomplete continuous outcomes that exhibit varying distributional properties. Even though the simulated data sets are only a tiny 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 study that includes many non-Gaussian features is insightful. Obviously, a much more comprehensive simulation study that spans a broader range of situations (more variables, parameters, etc.) could have been designed. However, the scope of this article is defined as giving simple advice for this particular missing-data problem which is encountered in practice quite often, especially in medical and social sciences.

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 IMVNO and OILLM 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 OILLM should be the preferred strategy when the ultimate interest is about the ordinalized outcomes. Again, 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 of 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 ordinalized versions of them, ordinalization before carrying out a log-linear imputation should be the preferred strategy over more intuitive approach of imputing under a Gaussian model before ordinalization, except for the situations where underlying densities are a close approximation to multivariate normality in terms of modality, symmetry and tail behavior. The reason could be 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 preponderates the factual argument that continuous variables intrinsically convey more information, leading to a counter-intuitive, but potentially beneficial result for practitioners.

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.
- 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.