

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

*Technical report#:2007-003  
May 2007*

Title: Imputation under the generalized lambda distribution

**Authors: Hakan Demirtas**

**Affiliation: University of Illinois at Chicago, Division of Epidemiology and Biostatistics.**

# Imputation under the generalized lambda distribution

Hakan Demirtas\*

May 22, 2007

## Abstract

Although the normality assumption has been regarded as a mathematical convenience for inferential purposes due to its nice distributional properties, there has been a growing interest regarding generalized classes of distributions that span a much broader spectrum in terms of symmetry and peakedness behavior. In this respect, the generalized lambda distribution (GLD) represents a viable choice. In this article, we conduct multiple imputation for univariate continuous data under the GLD to explore the extent to which this procedure works properly; and we make comparisons with normal imputation models via widely accepted accuracy and precision measures using simulated data that exhibit different distributional features as characterized by competing specifications of the third and fourth moments. Furthermore, we present a simulation study that is designed around a psychiatric trials data. Multiple imputation under the GLD that cover most of the feasible area in the skewness-elongation plane appears to have substantial potential of capturing real missing-data trends that can be encountered in biopharmaceutical practice.

**Key Words:** Multiple imputation; Normality; Symmetry; Skewness; Kurtosis

## 1 Introduction

The normality assumption is unequivocally one of the most extensively used and studied phenomena in statistics. Although real data rarely conform with normality, it has been regarded as a convenient assumption in model formation, estimation and testing due to its well-understood distributional properties. Despite its popularity, it only represents a single point in the skewness-elongation plane; and general classes of continuous distributions that span a broader spectrum in terms of symmetry and peakedness behavior (Genton, 2004)

---

\*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.

have received increased interest among statisticians. In this article, we describe multiple imputation (MI) under the generalized lambda distribution (GLD) that can accommodate a much wider range of distributional features. The salient aspects of the GLD are elaborated in Section 2.

MI is a stochastic simulation technique that involves filling-in missing data with  $m > 1$  plausible values through a predictive distribution (Rubin, 2004; Little and Rubin, 2002). Once  $m$  versions of the completed data sets are obtained, one can proceed with analyzing them with standard complete-data methods, and consolidating the results into a single inferential summary. As a result, with MI, uncertainty due to missing data is formally taken into account in the modeling process. Other key advantages of MI are reviewed by Schafer (1997, 1999). For an extensive bibliography, see Rubin (1996) and for a software review see Horton and Lipsitz (2001). The fundamental step in parametric MI is filling in the missing data by drawing from the conditional distribution of the missing data given the observed data. This usually entails positing a model for the data and using it to derive this conditional distribution. For continuous data, multivariate normality among the variables has been perceived as a natural assumption since the conditional distribution of the missing data given the observed data is then also multivariate normal. Recently, extending the practice of MI from normality to more general classes of densities has begun to receive attention (Liu, 1995; He and Raghunathan, 2006).

Considering the restrictive nature of the normality assumption, employing a distributional setup that spans a wider range of symmetry-peakedness behavior in the imputation process may provide a reasonable way to handle non-Gaussian continuous data. In this regard, the GLD approach can be thought as a sensible alternative because of the ability of accommodating a variety of distributional shapes depending on the choice of parameter values. Here, we explore the relative advantages of conducting imputation inferences under this more flexible approach via two limited simulation experiments that includes some univariate data generation mechanisms that may be encountered in practice. The rationale is to assess the feasibility of this technique as a possible impetus for extensions to the multivariate case, and to gauge its generalizability potential for creating imputations under a multivariate extension of the GLD. Given that imputation under non-normal densities is a recently emerging notion, which has potential in many research areas, it is important to

evaluate its performance in terms of commonly accepted bias and precision measures.

The organization of the rest of this paper is as follows. In Section 2, we describe essential aspects of the GLD. In Section 3, we present a simulation design based on hypothetical data and give a simple algorithm to create multiply imputed data sets under the GLD, with an examination of relative improvements over Gaussian imputation on incomplete data sets that exhibit different distributional characteristics. Subsequently, we explore the behavior of efficiency and accuracy measures to determine the extent to which the GLD works properly. Another simulation study that is devised around a real psychiatric trials data follows in Section 4. Section 5 includes concluding remarks, discussion and future directions.

## 2 Overview of the generalized lambda distribution

The generalized lambda density is a class of distributions used for parameter estimation, fitting distributions to data, or in simulation studies that primarily involve univariate data generation (Ramberg and Schmeiser, 1972, 1974; Ramberg et al., 1979). The univariate GLD is attractive because its pdf and inverse distribution function are known and its associated algorithm for data generation can be implemented with relative ease. Simulated values,  $x$ , can be drawn by the inverse cdf method  $x = \lambda_1 + [p^{\lambda_3} - (1-p)^{\lambda_4}] / \lambda_2$ , where  $\lambda_1$  and  $\lambda_2$  are location and scale parameters, respectively,  $\lambda_3$  and  $\lambda_4$  are shape parameters, and  $p \sim U[0, 1]$ . The right hand side of the equation is the inverse cdf of the GLD. Although the cdf does not exist in closed form, this is not a problem in practice since the same is true for the normal distribution. Estimation of  $\lambda$ 's can be carried out by percentile matching, moment matching, maximum likelihood, and pseudo least squares methods (Tarsitano, 2005). Our preliminary work suggested that matching moments yield very accurate estimates. For this reason, we predicate the estimation process upon moment matching. Ramberg and Schmeiser (1974) showed that the  $k^{th}$  moment of the GLD, when it exists, is given by  $E(X^k) = \lambda_2^{-k} \sum_{i=0}^k \binom{k}{i} (-1)^i \beta(\lambda_3(k-i) + 1, \lambda_4 i + 1)$  for  $\lambda_1 = 0$ , where  $\beta$  denotes the beta function. Ramberg et al. (1979) gave details of estimation and model-fitting procedures for this four-parameter probability distribution. It involves solving a set of equations that are formed through the first four moments that are capable of accommodating a wide variety of curve shapes.

Ramberg and Schmeiser (1974) derived the following expression for the mean ( $\mu$ ), the variance ( $\sigma^2$ ), and the third ( $\mu_3 = E(X - \mu)^3$ ) and fourth ( $\mu_4 = E(X - \mu)^4$ ) moments about the mean for this distribution:

$$\begin{aligned}\mu &= \lambda_1 + \frac{A}{\lambda_2} \\ \sigma^2 &= \frac{(B-A^2)}{\lambda_2^2} \\ \mu_3 &= \frac{(C-3AB+2A^3)}{\lambda_2^3} \\ \mu_4 &= \frac{(D-4AC+6A^2B-3A^4)}{\lambda_2^4}, \text{ where}\end{aligned}$$

$$\begin{aligned}A &= \frac{1}{1+\lambda_3} - \frac{1}{1+\lambda_4} \\ B &= \frac{1}{1+2\lambda_3} + \frac{1}{1+2\lambda_4} - 2\beta(1 + \lambda_3, 1 + \lambda_4) \\ C &= \frac{1}{1+3\lambda_3} - \frac{1}{1+3\lambda_4} - 3\beta(1 + 2\lambda_3, 1 + \lambda_4) + 3\beta(1 + \lambda_3, 1 + 2\lambda_4) \\ D &= \frac{1}{1+4\lambda_3} + \frac{1}{1+4\lambda_4} - 4\beta(1 + 3\lambda_3, 1 + \lambda_4) - 4\beta(1 + \lambda_3, 1 + 3\lambda_4) + 6\beta(1 + 2\lambda_3, 1 + 2\lambda_4)\end{aligned}$$

The skewness and kurtosis, as given by  $\alpha_3 = \frac{\mu_3}{\sigma^3}$  and,  $\alpha_4 = \frac{\mu_4}{\sigma^4}$ <sup>1</sup> are functions of  $\lambda_3$  and  $\lambda_4$ , but do not depend on  $\lambda_1$  and  $\lambda_2$ . One can easily compute the empirical moments  $\mu^*$ ,  $\sigma^{*2}$ ,  $\mu_3^*$ , and  $\mu_4^*$  from a given data set; then find the values of  $\lambda_3$  and  $\lambda_4$  which best solve the two simultaneous nonlinear equations  $\alpha_3(\lambda_3, \lambda_4) = \alpha_3^*$  and  $\alpha_4(\lambda_3, \lambda_4) = \alpha_4^*$ . Solving these equations can be accomplished by the Newton-Raphson method, or any other plausible root-finding or non-linear optimization routine. Calculating  $\lambda_1$  and  $\lambda_2$  is straightforward, with a caveat that the sign of  $\lambda_2$  should be the same as the sign of  $\lambda_3$  and  $\lambda_4$ . The specification of any feasible moment structure (not every combination is possible within the general constraint  $\alpha_4 > \alpha_3^2 + 1$ ) that translates to corresponding values of  $\lambda$ 's, enables us to generate random numbers via the inverse cdf method, which is equally applicable in the context of creating multiply imputed data sets.

After reviewing the fundamentals of the GLD, we describe a simulation study that includes the way one creates multiply imputed data sets under the GLD with competing moment structures in the next section.

---

<sup>1</sup>Note that  $\alpha_4$  is not normalized by subtracting 3 –which is the kurtosis of the normal distribution–, i.e. it represents kurtosis proper, not kurtosis excess.

### 3 Simulation design and imputation algorithm

*Complete data generation:* We generated complete data sets via Fleishman’s power polynomials. Fleishman (1978) argued that real-life distributions of variables are typically characterized by their first four moments. He presented a moment-matching procedure that simulates non-normal distributions often used in Monte Carlo studies. It is based on the polynomial transformation,  $Y = a + bZ + cZ^2 + dZ^3$ , where  $Z$  follows a standard normal distribution, and  $Y$  is standardized (zero mean and unit variance). The distribution of  $Y$  depends on the constants  $a, b, c$  and  $d$ , whose values were tabulated for selected values of skewness and kurtosis. This procedure of expressing any given variable by the sum of linear combinations of powers of a standard normal variate is capable of covering a wide area in the skewness-elongation plane. The reason we use Fleishman polynomials is that it covers about the same area in the skewness-elongation plane as the GLD does. Another advantage of the use of the power method is that it allows us to make a genuine assessment as to how the GLD method performs in the imputation context because of the fact that the power method and the GLD represent radically different distributional forms. The nine skewness-kurtosis pairs that were used in the simulated examples are given in Table 1. The number of observations,  $n$ , in the complete data set was chosen to be 100, 500, and 1000.

**Table 1 goes here**

*Missingness mechanism:* Missing values were assumed to be missing completely at random (MCAR). This missingness mechanism generally too simplistic for real-life applications. However, the purpose is not to conduct a sensitivity analysis with respect to the mechanism that leads to the observed data. Rather, the current paper is motivated by how tenably MI inferences can be conducted with the GLD. The nonresponse rate was chosen to be 25%, 50%, and 75%.

*Imputation algorithm:* We assume two imputation models for comparison purposes for each of the incomplete data sets generated. The first one is the normal model, where we create imputations following the standard approach of using a Bayesian predictive model of the missing data given the observed data (Schafer, 1997). For the other (the GLD), instead of adopting a Bayesian approach, we account for the parameter uncer-

tainty by obtaining nonparametric bootstrap samples (He and Raghunathan, 2006) that anchor the subsequent estimation procedure for the parameters of the GLD. Denoting the data  $Y = (Y_{obs}, Y_{mis}) = (y_1, y_2, \dots, y_{n_1}, y_{n_1+1}, \dots, y_n)^T$ , of which the first  $n_1$  elements are observed ( $Y_{obs} = (y_1, y_2, \dots, y_{n_1})^T$ ), and the remaining  $n - n_1$  elements are missing ( $Y_{mis} = (y_{n_1+1}, \dots, y_n)^T$ ), the imputation algorithm is as follows:

1. Center and scale the data so that mean is zero and variance is one.<sup>2</sup> Let the transformed data be  $Y_{obs}^*$ .
2. Draw a nonparametric bootstrap sample of size  $n_1$  from  $Y_{obs}^*$ .
3. Estimate the model parameters  $(\lambda_1, \lambda_2, \lambda_3, \lambda_4)$  by any feasible optimization routine (e.g. the Newton-Raphson algorithm).
4. Simulate independent variates from these distributions for every missing data point in  $Y_{mis}^*$ .
5. Back transform the filled-in data and the transformed observed data to the original scale.
6. Repeat steps 2-5 independently  $m = 10$  times.

*Parameters of interest:* We compared the relative performances of normal and GLD imputations on five quantiles ( $5^{th}$ ,  $25^{th}$ ,  $50^{th}$ ,  $75^{th}$ , and  $95^{th}$ ) that are known to be sensitive to model misspecification.

*Evaluation criteria:* The simulation experiment was repeated  $N = 1000$  times for each of the  $9 \times 3 \times 3 = 81$  scenarios (combinations of complete data distributions, data set sizes, and missingness rates, respectively). Obviously,  $N$  could have been chosen to be larger; however, with too many replications, the bias could turn out to be significant when it is actually not. Evaluation is conducted based on three quantities: a) *Standardized bias (SB)* is the relative magnitude of the raw bias to the overall uncertainty in the system. If

---

<sup>2</sup>This standardization is not really necessary for the subsequent estimation of the four  $\lambda$  coefficients. It is done in order to identify any possible computational and implementation-related mistakes by comparing our estimates and the tabulated values in Ramberg et al. (1979) that are based on centered and scaled data.

the parameter of interest is  $\theta$ , 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 50% in a positive or negative direction, then the bias begins to have a noticeable adverse impact on efficiency, coverage and error rates (Demirtas, 2004). b) *Coverage rate (CR)* is 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). We regard the performance of the interval procedure to be poor if its coverage drops below 90% (Collins et al., 2001). c) *Root-mean-square error (RMSE)* is 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  $\sqrt{E_{\theta}[(\hat{\theta} - \theta)^2]}$ . Under this specification,  $SB$  is a pure accuracy measure,  $CR$  and  $RMSE$  are the hybrid measures of accuracy and precision. For more detailed discussion on this evaluation system, see Demirtas (2005a, 2005b), Demirtas (2007), Demirtas and Hedeker (2007), Demirtas, Freels and Yucel (2007), and Demirtas et al. (2007). Software implementation was done in R language (2007).

### 3.1 Results

Since the results across different sample sizes and nonresponse rates yielded little or no discernible differences, we present the results for  $n = 1000$  and 50% nonresponse rate due to space limitations. In Tables 2, 3, and 4, we tabulate the average estimate ( $AE$ ),  $SB$ ,  $RMSE$ , and  $CR$  for the five quantiles under consideration across 1000 simulation replicates for both the normal imputation model and the proposed GLD imputation. In these tables, the bias and coverage quantities that do not fall into the acceptable ranges ( $> 50\%$  or  $< -50\%$  for  $SB$ , and  $< 90.0$  for  $CR$ ) are denoted with bold characters. The number of significant digits varies depending on the measures. A close examination of Tables 2-4 reveals that the GLD method outperforms the normal imputation method in terms of all three evaluation criteria to varying degrees across all scenarios except when both skewness and kurtosis are 0 and 3, respectively. In this case, which corresponds to normal underlying data, the performances are comparable. The GLD method yields remarkable results with no exceptions as indicated by negligible biases and high coverage rates. In other words, the performance of the GLD approach turns out to be superior *compared to* that of the

MI normal model in terms of commonly accepted accuracy and precision quantities in an overwhelming majority of cases, with decent properties *in every evaluation metric we considered*. This gives us hope in regard to the more general multivariate case which we briefly discuss in Section 5.

**Table 2, 3 and 4 go here**

## 4 Simulations devised around real data

Our second simulation study pertains to a psychiatric trial data from the National Institute of Mental Health Schizophrenia Collaborative Study with 437 subjects. The outcome of interest, severity of illness, was measured on an ordinal scale ranging from 1 (normal) to 7 (extremely ill), which we treat as continuous since intermediate values were obtained due to multiple raters. Measurements were planned for four time points, but missing values occurred primarily due to drop-out; about 25% of measurements were missing at the end of the study. This data set was previously analyzed by Hedeker and Gibbons (1997), Demirtas and Schafer (2003), Demirtas (2005b) from the pattern-mixture modeling standpoint. For the purposes of this work, we calculated the empirical third and moments for measurements (after centering and scaling) at each time point ( $Y_1, Y_2, Y_3, Y_4$ ) separately and computed the corresponding coefficients of Fleishman's polynomials by which we created data sets with 437 subjects. At the next step, we imposed missing values assuming 50% nonresponse rate which is substantially higher than the observed nonresponse rate, taking a conservative position. The rest of the simulation design was identical to the one presented in Section 3; the imputation algorithm, parameters of interest, evaluation criteria, and the number of replications (1000) were the same. The results for  $Y_1, Y_2, Y_3$ , and  $Y_4$  are tabulated in Table 5 whose format is very similar to Tables 2, 3, and 4. The essential conclusions remained unchanged, demonstrating the superior performance of the GLD imputation to the normal imputation in all evaluation criteria we examined in the form of lower biases and *RMSE*'s, and higher coverage rates.

**Table 5 goes here**

## 5 Discussion

We strongly believe that a close connection between random number generation and MI can be established (see Demirtas and Hedeker, 2007 for an example). In this paper, we have adapted methods developed in the random number generation literature to the context of MI. A major imputation principle is not to distort the marginal distributions and associations between observed and imputed variables; and random number generation is conducted via specified distributional properties. Observed data trends can be assumed to be applicable to the whole data set, and missing portions can be filled in with numbers that belong to the same distributional mechanism which, in a sense, is the operational logic for random number generation. This assertion clearly assumes ignorable nonresponse (once we have taken into account what we have observed, there remains no dependence on what we have not observed), where missingness fully depends on the observed quantities in the system in the sense of Rubin (1976). Although this may be considered a limitation, it might serve as a milestone for nonignorable extensions.

In our view, the promising results in the univariate case substantiate the natural continuation of this imputation method under the multivariate case. As articulated in Headrick and Mugdadi (2006), in terms of multivariate data generation, it has been demonstrated that the GLD have computational difficulties associated with 1) having to take several steps to overcome the problem of generating biased correlation coefficients, 2) having access to commercial software packages and ensuring the accuracy of numerical solutions to complicated integrals. Headrick and Mugdadi (2006) developed a methodology to simulate multivariate non-normal distributions from the GLD. We feel that making connections from random generation domain to MI domain using a multivariate version of the GLD is an exciting idea which will be taken up in future work.

There are a few limitations that need to be addressed. First, while we recognize that real incomplete data often include many variables, our focus was on univariate data. We view this as a potential building block for more realistic situations. The behavior of the third and fourth moments typically requires more modeling flexibility in terms of the area covered in the symmetry-elongation plane as well as the association among variables. This work serves as an initial feasibility study for assessing the generalizability potential to the multivariate settings. On a related note, although the power approach is capable of

picking some data trends that are unlikely to be captured by a normal model, it does not cover the entire symmetry-elongation plane. Nevertheless, considering the relative gains presented in this paper, it provides an indication that the multivariate version can lead to further improvements. Furthermore, as mentioned in Section 3, the assumed missingness mechanism (MCAR) is generally too simplistic for real-life applications. However, our purpose was not to conduct a sensitivity analysis with respect to the mechanism that leads to the observed data. Rather, the current paper was motivated by how tenably MI inferences can be conducted with the GLD. Finally, our simulation setup needed to be in manageable limits and does not span every imaginable scenario that may arise in practice. However, in our opinion, it is sufficiently comprehensive to demonstrate the superiority of the GLD imputation model in most cases.

The assumption of multivariate normality along with the Bayesian paradigm has often been regarded as a statistically defensible way of creating multiply imputed data sets for continuous data. While it is a convenient assumption and it has been shown to work well in some settings (e.g., with a large number of subjects), it is constructive to move the practice of MI to other distributions that cover a broader range of the third and fourth moments. In an attempt to go beyond the realm of normality to adequately model distributional properties that are not accommodated by a Gaussian model, there has been a growing interest in non-normal distributions. This work was motivated by the premise that the MI framework may be amenable to the GLD method, and the results presented in this paper seem to support this conjecture.

## REFERENCES

Collins, L.M., Schafer, J.L., Kam, C.H. (2001). A comparison of inclusive and restrictive strategies in modern missing data procedures. *Psychological Methods*, 6:330–351.

Demirtas, H. (2004). Simulation-driven inferences for multiply imputed longitudinal datasets. *Statistica Neerlandica*, 58:466–482.

Demirtas, H. (2005a). Bayesian analysis of hierarchical pattern-mixture models for clinical trials data with attrition and comparisons to commonly used ad-hoc and model-based approaches. *Journal of Biopharmaceutical Statistics*, 25:383–402.

Demirtas, H. (2005b). Multiple imputation under Bayesianly smoothed pattern-mixture models for non-ignorable drop-out. *Statistics in Medicine*, 24: 2345–2363.

Demirtas, H. (2007). Practical advice on how to impute continuous data when the ultimate interest centers on dichotomized outcomes through pre-specified thresholds. *Communications in Statistics-Simulation and Computation* (in press).

Demirtas, H., Arguelles, L.M., Chung, H., Hedeker, D. (2007). On the performance of bias-reduction techniques for variance estimation in approximate Bayesian bootstrap imputation. *Computational Statistics and Data Analysis*, 51:4064–4068.

Demirtas, H., Freels, S.A., Yucel, R.M. (2007). Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: A simulation assessment. *Journal of Statistical Computation and Simulation* (in press).

Demirtas, H., Hedeker, D. (2007). Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. *Statistics in Medicine*, 26:782–799.

Demirtas, H., Schafer, J.L. (2003). Gaussianization-based quasi-imputation and expansion strategies for incomplete correlated binary responses. *Statistics in Medicine*, 22:2253–2575.

Fleishman, A.I. (1978). A method for simulating non-normal distributions. *Psychometrika*, 43:521–532.

Genton, M.G. (Ed) (2004). *Skew-Elliptical Distributions and Their Applications: A Journey Beyond Normality*. Boca Raton, FL: Chapman and Hall/CRC.

He, Y., Raghunathan, T.E. (2006). Tukey’s gh distribution for multiple imputation. *The American Statistician*, 60:251–256.

Headrick, T.C., Mugdadi, A. (2006). On simulating multivariate non-normal distributions from the generalized lambda distribution. *Computational Statistics and Data Analysis*, 50:3343–3353.

Hedeker, D., Gibbons, R.D. (1997). Application of random effects pattern-mixture models for missing data in longitudinal studies. *Psychological Methods*, 2:64–78.

Horton, J.H., Lipsitz, S.R. (2001). Multiple imputation in practice: Comparison of software packages for regression models with missing variables. *The American Statistician*, 55:244–254.

- Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*. Second Edition. New York: Wiley.
- Liu, C. (1995). Missing data imputation using the multivariate t distribution. *Journal of Multivariate Analysis*, 53:139–158.
- R Development Core Team. (2007). *R: A language and environment for statistical computing, Version 2.4.1*. Vienna, Austria. URL: <http://www.r-project.org>.
- Ramberg, J.S., Dudewicz, E.J., Tadikamalla, P.R., Mykytka, E.F. (1979). A probability distribution and its uses in fitting data. *Technometrics*, 21:201–214.
- Ramberg, J.S., Schmeiser, B.W. (1972). An approximate method for generating symmetric random variables. *Communications of the ACM*, 15:987–990.
- Ramberg, J.S., Schmeiser, B.W. (1974). An approximate method for generating asymmetric random variables. *Communications of the ACM*, 17:78–82.
- Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 21:581–592.
- Rubin, D.B. (1996). Multiple imputation after 18+ years (with discussion). *Journal of the American Statistical Association*, 91:473–520.
- Rubin, D.B. (2004). *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley Classic Library.
- Schafer, J.L. (1997). *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall.
- Schafer, J.L. (1999). Multiple imputation: a primer. *Statistical Methods in Medical Research*, 8:3–15.
- Tarsitano, A. (2005). Estimation of the generalized lambda distribution parameters for grouped data. *Communications in Statistics—Theory and Methods*, 34:1689–1709.

Table 1: The skewness-kurtosis specifications with the prefixes meso, lepto, and platy stand for  $\alpha_4 = 3$ ,  $> 3$ , and  $< 3$  kurtosis, respectively.

Scenario	Skewness	Kurtosis	Property
1	0	3	symmetric-mesokurtic
2	0	6.75	symmetric-leptokurtic
3	0	2	symmetric-platykurtic
4	0.5	3	right skewed-mesokurtic
5	-0.5	3	left skewed-mesokurtic
6	0.25	6	right skewed-leptokurtic
7	-0.25	6	left skewed-leptokurtic
8	0.5	2.75	right skewed-platykurtic
9	-0.5	2.75	left skewed-platykurtic

Table 2: The performance of imputation inferences under GLD and normal model for the five quantiles. *AE*, *SB*, *RMSE*, and *CR* stand for the average estimate, standardized bias, root-mean-square error, and coverage rate, respectively, for the three skewness-kurtosis specifications. The number of simulation replicates,  $N$  is 1000; the length of the complete data vector,  $n$  is 1000; and missingness rate is 50%.

MI model	Skewness	Kurtosis	Quantile	AE	SB	RMSE	CR
<i>GLD</i>	0	3	5	0.0499486	-0.89	0.0058	97.4
			25	0.2495334	-3.67	0.0127	96.3
			50	0.4993954	-4.09	0.0148	96.2
			75	0.7498074	-1.58	0.0122	96.1
			95	0.9498895	-1.82	0.0061	96.9
<i>NORMAL</i>	0	3	5	0.0499825	-0.32	0.0054	97.5
			25	0.2497824	-1.80	0.0121	97.1
			50	0.4995701	-3.18	0.0135	97.6
			75	0.7495653	-3.65	0.0119	96.9
			95	0.9499117	-1.58	0.0057	97.2
<i>GLD</i>	0	6.75	5	0.0500296	0.53	0.0056	97.6
			25	0.2557933	41.25	0.0152	94.6
			50	0.4999347	-0.36	0.0179	95.6
			75	0.7434274	-44.17	0.0162	93.7
			95	0.9499712	-0.51	0.0057	97.5
<i>NORMAL</i>	0	6.75	5	0.0538806	<b>56.68</b>	0.0079	93.7
			25	0.2740849	<b>194.41</b>	0.0271	<b>63.1</b>
			50	0.5002284	1.70	0.0134	97.8
			75	0.7255957	<b>-186.83</b>	0.0277	<b>61.4</b>
			95	0.9461708	<b>-53.07</b>	0.0082	92.0
<i>GLD</i>	0	2	5	0.0469125	-47.44	0.0072	90.8
			25	0.2518849	15.10	0.0126	96.4
			50	0.4997525	-1.74	0.0142	96.6
			75	0.7482853	-13.39	0.0129	95.5
			95	0.9531540	49.65	0.0070	91.9
<i>NORMAL</i>	0	2	5	0.0504027	7.50	0.0054	98.3
			25	0.2314196	<b>-158.15</b>	0.0219	<b>72.4</b>
			50	0.4998634	-0.96	0.0143	96.3
			75	0.7688809	<b>156.84</b>	0.0224	<b>69.7</b>
			95	0.9495874	-7.78	0.0053	98.0

Table 3: The performance of imputation inferences under GLD and normal model for the five quantiles. *AE*, *SB*, *RMSE*, and *CR* stand for the average estimate, standardized bias, root-mean-square error, and coverage rate, respectively, for the three skewness-kurtosis specifications. The number of simulation replicates,  $N$  is 1000; the length of the complete data vector,  $n$  is 1000; and missingness rate is 50%.

MI model	Skewness	Kurtosis	Quantile	AE	SB	RMSE	CR
<i>GLD</i>	0.5	0	5	0.0477437	-47.02	0.0066	92.1
			25	0.2480900	-14.91	0.0129	95.8
			50	0.5043638	29.84	0.0152	95.8
			75	0.7507316	5.95	0.0123	96.4
			95	0.9483501	-26.66	0.0064	97.3
<i>NORMAL</i>	0.5	3	5	0.0598935	<b>175.00</b>	0.0114	<b>83.3</b>
			25	0.2385586	<b>-102.37</b>	0.0160	90.6
			50	0.4813793	<b>-138.31</b>	0.0230	<b>80.9</b>
			75	0.7449949	-40.11	0.0134	95.1
			95	0.9564973	<b>118.76</b>	0.0085	<b>79.9</b>
<i>GLD</i>	-0.5	3	5	0.0512576	20.05	0.0064	96.4
			25	0.2489546	-8.21	0.0127	96.4
			50	0.4954918	-29.80	0.0158	95.1
			75	0.7514441	11.11	0.0131	95.2
			95	0.9522544	48.54	0.0064	92.7
<i>NORMAL</i>	-0.5	0	5	0.0431758	<b>-115.66</b>	0.0089	<b>76.4</b>
			25	0.2543131	33.88	0.0134	95.8
			50	0.5180933	<b>130.73</b>	0.0228	<b>81.3</b>
			75	0.7608400	<b>97.72</b>	0.0155	<b>90.7</b>
			95	0.9401312	<b>-182.66</b>	0.0112	<b>84.5</b>
<i>GLD</i>	0.25	6	5	0.0517760	23.83	0.0065	98.2
			25	0.2488873	-8.91	0.0125	97.8
			50	0.4973669	12.34	0.0207	91.6
			75	0.7491826	-18.41	0.0174	93.3
			95	0.9515947	30.44	0.0064	93.5
<i>NORMAL</i>	0.25	6	5	0.0560095	<b>93.24</b>	0.0088	92.3
			25	0.2677680	<b>148.17</b>	0.0214	<b>78.6</b>
			50	0.4942188	-43.63	0.0145	96.1
			75	0.7265580	<b>-188.46</b>	0.0265	<b>66.0</b>
			95	0.9493758	-9.41	0.0067	94.4

Table 4: The performance of imputation inferences under GLD and normal model for the five quantiles. *AE*, *SB*, *RMSE*, and *CR* stand for the average estimate, standardized bias, root-mean-square error, and coverage rate, respectively, for the three skewness-kurtosis specifications. The number of simulation replicates,  $N$  is 1000; the length of the complete data vector,  $n$  is 1000; and missingness rate is 50%.

MI model	Skewness	Kurtosis	Quantile	AE	SB	RMSE	CR
<i>GLD</i>	-0.25	6	5	0.0491049	-11.68	0.0066	92.5
			25	0.2526175	18.83	0.0179	91.4
			50	0.5026924	15.91	0.0214	90.0
			75	0.7505555	4.21	0.0132	96.0
			95	0.9479410	-24.74	0.0069	96.3
<i>NORMAL</i>	-0.25	6	5	0.0502629	4.04	0.0065	95.0
			25	0.2725973	<b>173.44</b>	0.0261	<b>65.3</b>
			50	0.5052684	38.02	0.0148	96.3
			75	0.7314660	<b>-149.33</b>	0.0223	<b>76.0</b>
			95	0.9438255	<b>-90.52</b>	0.0092	<b>89.4</b>
<i>GLD</i>	0.5	2.75	5	0.0485412	-19.12	0.0073	90.0
			25	0.2481446	-14.26	0.0131	95.8
			50	0.5049860	34.38	0.0153	95.1
			75	0.7508067	6.54	0.0124	96.5
			95	0.9486500	-20.94	0.0066	95.9
<i>NORMAL</i>	0.5	2.75	5	0.0612920	<b>206.13</b>	0.0125	<b>77.9</b>
			25	0.2345472	<b>-141.35</b>	0.0189	<b>81.5</b>
			50	0.4799831	<b>-147.46</b>	0.0242	<b>78.2</b>
			75	0.7491222	-6.98	0.0126	96.2
			95	0.9575381	<b>133.95</b>	0.0094	<b>73.5</b>
<i>GLD</i>	-0.5	2.75	5	0.0517619	27.63	0.0066	95.8
			25	0.2502842	2.20	0.0129	96.0
			50	0.4964806	-23.62	0.0153	95.8
			75	0.7525221	18.81	0.0136	95.1
			95	0.9529195	31.70	0.0067	92.3
<i>NORMAL</i>	-0.5	2.75	5	0.0380095	<b>-128.41</b>	0.0090	<b>76.6</b>
			25	0.25233793	18.22	0.0133	95.1
			50	0.5215843	<b>153.05</b>	0.0258	<b>74.3</b>
			75	0.7661733	<b>140.82</b>	0.0198	<b>79.0</b>
			95	0.9384262	<b>-212.58</b>	0.0128	<b>76.8</b>

Table 5: Results for the real data example.

Variable	MI model	Skewness	Kurtosis	Quantile	AE	SB	RMSE	CR
$Y_1$	<i>GLD</i>	-0.4631312	3.078634	5	0.0509135	15.66	0.0059	97.1
				25	0.2477042	-18.49	0.0126	96.8
				50	0.4944317	-37.45	0.0159	95.3
				75	0.7501686	1.34	0.0125	95.9
				95	0.9523221	39.56	0.0063	94.4
$Y_1$	<i>NORMAL</i>	-0.4631312	3.078634	5	0.0435438	<b>-120.07</b>	0.0084	<b>78.5</b>
				25	0.2539572	32.06	0.0129	97.4
				50	0.5148616	<b>109.58</b>	0.0201	<b>86.7</b>
				75	0.7578036	<b>70.40</b>	0.0135	94.2
				95	0.9409185	<b>-162.10</b>	0.0107	<b>86.8</b>
$Y_2$	<i>GLD</i>	-0.5510303	2.959411	5	0.0513895	22.4	0.0063	96.3
				25	0.2488242	-9.68	0.0122	97.0
				50	0.4955295	-33.12	0.0142	94.8
				75	0.7526314	20.98	0.0128	94.5
				95	0.9519672	36.94	0.0065	93.3
$Y_2$	<i>NORMAL</i>	-0.5510303	2.959411	5	0.0424717	<b>-135.59</b>	0.0093	<b>73.6</b>
				25	0.2543171	33.69	0.0135	95.4
				50	0.5214366	<b>163.34</b>	0.0251	<b>76.5</b>
				75	0.7639952	<b>130.91</b>	0.0176	<b>86.7</b>
				95	0.9382803	<b>-227.73</b>	0.0128	<b>77.8</b>
$Y_3$	<i>GLD</i>	-0.3990908	2.259374	5	0.0502502	4.17	0.0060	96.6
				25	0.2507719	6.20	0.0125	96.7
				50	0.4968942	-21.71	0.0146	97.1
				75	0.7516106	12.07	0.0134	94.9
				95	0.9529402	45.71	0.0071	91.7
$Y_3$	<i>NORMAL</i>	-0.3990908	2.259374	5	0.0428000	<b>-145.87</b>	0.0087	<b>78.9</b>
				25	0.2419406	<b>-66.63</b>	0.0145	91.6
				50	0.5209430	<b>155.74</b>	0.0249	<b>74.6</b>
				75	0.7738696	<b>218.42</b>	0.0262	<b>53.3</b>
				95	0.9356460	<b>-248.35</b>	0.0154	<b>60.2</b>
$Y_4$	<i>GLD</i>	0.2144061	2.059205	5	0.0481439	-30.70	0.0074	90.2
				25	0.2501604	1.33	0.0120	96.9
				50	0.5021968	15.58	0.0142	96.8
				75	0.7481528	-14.46	0.0129	97.1
				95	0.9519811	29.85	0.0069	93.9
$Y_4$	<i>NORMAL</i>	0.2144061	2.059205	5	0.0572933	<b>138.15</b>	0.0090	91.2
				25	0.2271167	<b>-214.00</b>	0.0252	<b>57.3</b>
				50	0.4881166	<b>-86.33</b>	0.0182	90.7
				75	0.7641291	<b>117.63</b>	0.0186	<b>82.1</b>
				95	0.9544118	<b>78.33</b>	0.0071	<b>88.4</b>