# Preservation of Skip Patterns and Covariance Structure through Semi-Parametric Whole-Questionnaire Imputation

David Judkins[1], Tom Krenzke[1], Andrea Piesse[1], Zizhong Fan[2], and Wen-Chau Haung[1]
David Judkins, Westat, 1650 Research Blvd., Rockville, MD 20850[1]
MedImmune, Inc., One MedImmune Way, Gaithersburg, MD 20878[2]

## Abstract

Hot-deck imputation schemes are attractive because of how well they preserve complex features of marginal distributions, such as heaping of income reports at round figures, in addition to marginal means and variances. Historically, they have been less successful at preserving multivariate structure. The authors have previously reported on recent imputation methodology for preserving important features of the multivariate structure of whole-questionnaire data, including skip patterns and the strongest bivariate associations among ordered variables. This methodology is based on predictive mean matching with recursive unsupervised modeling of ordered and binary variables to be imputed. They have recently extended its capabilities to include preservation of associations among unordered variables. This paper contains a report on the methodology, including a simulation study in which the new algorithm is tested in a head-to-head competition with two other imputation approaches.

**Keywords**: Hot-deck, Predictive mean matching

## 1. General Considerations for the Design of Imputation Software for Data Publishers

The needs of a data publisher are typically different than those of a secondary analyst whose research may involve a limited number of variables and who may therefore be willing to invest substantial time and energy on maximum likelihood or optimal Bayesian estimation of model parameters. Typically, the publisher must impute all missing data at low expense to support a variety of unforeseeable analyses. We assert that the publisher will have done a good job if questionnaire skip patterns are respected (no pregnant men), other strong bivariate patterns are preserved (few Yiddish-speaking Eskimos), and essential features of all marginal distributions are preserved (first- and second-order moments, ranges, and discontinuities). Another level of achievement would be to preserve all first-order linear models under ignorable nonresponse, by which we mean that an analyst fitting a linear main-effects model for one variable in terms of some set of other variables in the data set would obtain unbiased estimates of the fixed effects in the model. We have developed a semi-parametric imputation system based on this set of performance goals. In this paper, we report on the algorithm, the results of some testing, and some comparisons with alternative approaches.

Another important goal is to be able to construct frequentist confidence intervals around post-imputation point estimates of various quantities. Although we also considered this goal when designing the system, we have not yet tested the utility of the algorithm for this purpose.

Possibly the most widely used imputation procedure is the simple hot-deck. In general, it consists of random matching observations within cells, reaching across soft boundaries when required, and never crossing hard boundaries in the search for a donor from which to obtain an imputed value. The traditional hot-deck approach imputes all missing data, preserves marginal distributions (including shapes, ranges, discontinuities, and all order moments), and is quick and inexpensive to implement if the data contain very simple or no skip patterns. However, simple hot-deck imputation does a poor job of preserving multivariate structure, strongly attenuates associations between variables, and becomes costly when complex skip patterns are involved, as discussed in Marker, Judkins, and Winglee, 2001. For these reasons, various enhancements to the hot-deck approach have been proposed (see, for example: Judkins, Mosher, and Botman, 1991; England et al., 1993; Fahimi et al., 1993; Judkins, 1997).

The use of Bayesian parametric algorithms for data imputation, e.g., IVEware (Raghunathan, Solenberger, and Hoewyk, 2002), has grown in recent years. The basic idea is to draw imputed values from a posterior predictive distribution specified by a regression model, usually with a flat or non-informative prior distribution for the parameters in the regression model. While this approach should do better than traditional hot-deck imputation at preserving multivariate structure, it too has its disadvantages. For example, Bayesian methods are often heavily reliant on normality assumptions and are not designed to cope well with unusually shaped distributions, such as heaping of reported income at round thousands. Their ability to produce imputed data that adhere to skip patterns is often limited, which can be problematic when working with survey data. Also, despite advances in computing power, substantial expense can be involved in monitoring the convergence of MCMCs.

## 2. A Semi-Parametric Algorithm

To address the complex missing patterns in survey data, Judkins (1997) proposed an iterative process called cyclic n-partition hot-decks. This semi-parametric approach cycles through sequences of hot-deck imputations, in which the most current completed data value of each survey item (either from the previous or current cycle) is considered for imputation models for each item. It is analogous to the Gibbs sampler in that covariance structures are preserved through iterations of parametric modeling, however it has the additional benefit of preserving semi-parametric distributions among survey data. The cyclic n-partition hot-deck is the underlying algorithm used in the semi-parametric approach that is evaluated in this paper.

One of the goals of our semi-parametric approach is to impute missing data for the entire questionnaire with one push of a button – after some preparatory work. The preparation includes categorizing all variables into one of the following types:

- Ordered with range restrictions, including all binary (ON);
- Ordered with a general range (RN); or
- Unordered categorical (UC).

Once the variable type is identified for each variable, the questionnaire and data values are reviewed to identify missing data patterns. This involves specifying the following for each variable: skip controllers, values of each skip controller that lead to same skip path, special missing values, inapplicable values, and special values (exceptions to general monotonicity).

The semi-parametric imputation approach is built on the hot-deck engine. The procedure begins by initially imputing all target variables (i.e., items requiring imputation) with a very simple hot-deck. Hard boundaries for each variable are defined by specified auxiliary variables and skip controllers, after collapsing on unique skip paths. After the initial hot-deck imputation, a model for each target variable is fit in terms of those variables without inapplicable or special values on the set of observations on which the target variable is neither inapplicable nor special. The model is formed on the set of observations for which the target variable is not inapplicable, not special, and not missing, using simple forward stepwise regression selection. Predicted outcome values from the final model are used to assist imputation through one of the following approaches, depending on the type of target variable:

- Predictive mean matching (ON variables);
- Adding of empirical residuals (RN variables); or
- Clustering (UC variables).

The predictive mean matching procedure for imputing ON variables first fits a linear regression model, and then uses a hot-deck with the skip controllers as hard boundaries and model-based predicted values as soft boundaries. Optionally, predicted values are coarsened prior to matching so as to facilitate meaningful multiple imputations, if desired for variance estimation purposes.

The adding of empirical residuals procedure for RN variables initially fits a linear regression model, and subsequently adds empirical residuals from hot-deck donors from the same skip path. If the residual variance is a function of the target variable, the donor pool of empirical residuals can be restricted to a cell defined by similar coarsened predicted values of the target variable.

The clustering procedure for UC variables fits a separate linear regression for each level, and subsequently conducts a k-means clustering algorithm on the vector of predicted values for each level. The algorithm is run four times, with 2, 5, 10, and 25 clusters. After the clustering algorithm is processed, hot-deck imputation is used with the skip controllers as hard boundaries and various cluster memberships as soft boundaries.

Use of the semi-parametric approach has drastically reduced the time and cost to conduct item imputation (Piesse, Judkins, and Fan, 2005). Under simulated strongly informative missing data mechanisms, the analysis of data imputed using the semi-parametric algorithm resulted in smaller biases and variances on marginal means and smaller bias in correlations than the analysis of complete case data alone (Krenzke, Judkins, and Fan, 2005). Unpublished empirical studies have shown that the semi-parametric approach preserves the correlation between two variables (including both binary and continuous variables) when non-monotone missing data patterns exist and missingness is completely at random for each variable. Still, the need for further evaluation prompted the following simulation study.

### 3. Simulated Populations

We developed four test scenarios. Three of these were designed to play to the strengths of the semi-parametric algorithm we developed. The fourth was designed to play to the strengths of one of the best known parametric imputation systems currently available, IVEware, from the University of Michigan.

In scenario #1, referred to herein as “strange pop”, there are two variables with range restrictions and/or discontinuities, as well as a very unusual dependency. It is easiest to describe the pair by construction and by graphs. Let $X \sim \mathrm{U}(-1,1)$. Let $e_Y \sim \mathrm{N}(0,1/25)$. Let

$$Y = \begin{cases} e_Y + \max\left\{-6, \min\left[6, 2\Phi^{-1}\left(\dfrac{X+1}{2}\right)\right]\right\} & \text{if } |X| > 0.5; \\ e_Y & \text{otherwise.} \end{cases}$$

Figures 1, 2, and 3 show the marginal distributions of $X$ and $Y$, and the conditional distribution of $Y$ given $X$. The following are some of the essential features of this population: $X$ is bounded by cliffs; $Y$ has a large concentration near zero, with a substantial gap either side of zero in which values are highly unlikely; the conditional distribution of $Y$ near the center range of $X$ is flat; and the conditional distribution of $Y$ near the extremes of $X$ is exponential. Clearly, we would be surprised to find a pair of variables like this in the survey setting, but this scenario was designed to demonstrate the ability of our algorithm to handle the unexpected without human intervention.

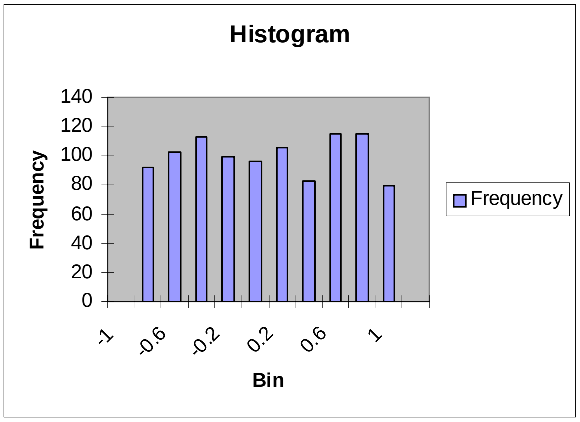

Figure 1. Marginal distribution of *X* in strange pop

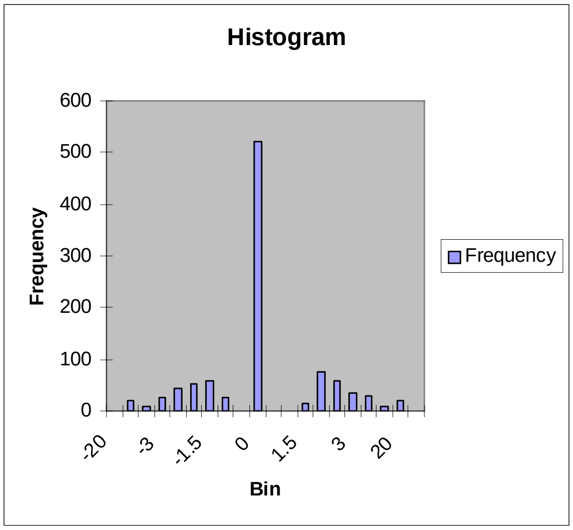

Figure 2. Marginal distribution of *Y* in strange pop

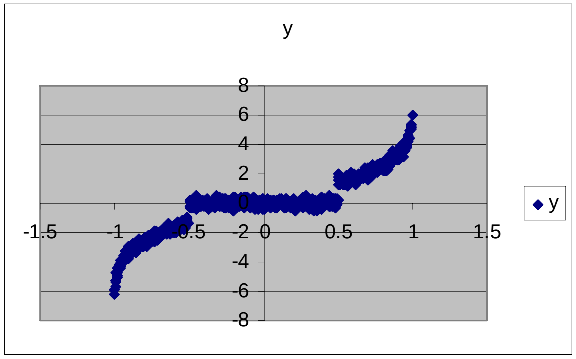

Figure 3. Conditional distribution of *Y* given *X* in strange pop

In scenario #2, there are three variables, of which one is binary ($X$) and two are unordered multinomials ($Y$ with three levels and $Z$ with four levels). The full joint distribution is given in Figure 4. The log-linear model for $X$, $Y$, and $Z$ has two-way interactions, but no three-way interaction.

| $X = 1$ | $Y = 1$ | $Y = 2$ | $Y = 3$ |
|---|---|---|---|
| $Z = 1$ | 0.0201 | 0.0177 | 0.0130 |
| $Z = 2$ | 0.0212 | 0.0216 | 0.0184 |
| $Z = 3$ | 0.0135 | 0.0216 | 0.0288 |
| $Z = 4$ | 0.0191 | 0.0264 | 0.0303 |
| $X = 2$ | $Y = 1$ | $Y = 2$ | $Y = 3$ |
| $Z = 1$ | 0.0367 | 0.0356 | 0.0214 |
| $Z = 2$ | 0.0471 | 0.0531 | 0.0370 |
| $Z = 3$ | 0.0448 | 0.0792 | 0.0866 |
| $Z = 4$ | 0.0776 | 0.1181 | 0.1113 |

Figure 4. Three-way table with two-way interactions.

Scenario #3 has the same number of cells and three-way layout as scenario #2, but the cell frequencies follow a distinct checkerboard pattern. The full joint distribution for $X$, $Y$, and $Z$ is given in Figure 5. The corresponding log-linear model involves both two- and three-way interactions.

| $X = 1$ | $Y = 1$ | $Y = 2$ | $Y = 3$ |
|---|---|---|---|
| $Z = 1$ | 0.0067 | 0.0467 | 0.0067 |
| $Z = 2$ | 0.0467 | 0.0067 | 0.0467 |
| $Z = 3$ | 0.0067 | 0.0467 | 0.0067 |
| $Z = 4$ | 0.0467 | 0.0067 | 0.0467 |
| $X = 2$ | $Y = 1$ | $Y = 2$ | $Y = 3$ |
| $Z = 1$ | 0.0367 | 0.0767 | 0.0367 |
| $Z = 2$ | 0.0767 | 0.0367 | 0.0767 |
| $Z = 3$ | 0.0367 | 0.0767 | 0.0367 |
| $Z = 4$ | 0.0767 | 0.0367 | 0.0767 |

Figure 5. Three-way table with checkerboard pattern.

Scenario #4 consists of a pair of bivariate normal variables:

$$X \sim N(60,9)$$
$$Y \sim N(120+X,1)$$

For all four test scenarios, we generated 100 data sets of 2000 observations each. An item nonresponse rate of 30 percent was then applied to each item independently across observations. The missing data mechanism was completely at random (MCAR).

## 4. Judging the Results

The imputation results are easiest to judge for scenario #4. For the pair of bivariate normal variables, we used three measures of success. One is the average value of the Kolmogorov-Smirnov statistic for comparing completed $X$ with uncensored $X$ over the 100 simulated data sets.[1] The second is a parallel test for the marginal distribution of $Y$. The third is the difference between the regression coefficient for $X$ in a linear model for $Y$ based on the uncensored data and the same coefficient from a model based on the completed data after imputation, averaged over the 100 simulated data sets.

The results are slightly more difficult to judge for scenarios #2 and #3. The parameters in log-linear models could be compared, but there are many of these. As a single measure of structure preservation, we thought it most interesting to focus on the "difference" between the uncensored and completed three-way tables. This comparison involves two tables (uncensored and completed), each with 24 cells (representing the full joint distribution of $X$, $Y$, and $Z$). As a simple and familiar looking criterion for similarity, we computed the chi-square test for equality of the two tables. However, this statistic can behave poorly with small cell sample sizes, so we also computed the simple $\ell^{24}$ distance between the two tables.

Hence, the first measure is

$$\chi^2 = \sum_{k=1}^{24} \frac{\left(n_k - n_{kc}\right)^2}{\left(n_k + n_{kc}\right)}$$

and the second is

$$\ell^{24} = \sqrt{\sum_{k=1}^{24} \left(n_k - n_{kc}\right)^2}$$

where $n_k$ is the number of uncensored observations in cell $k$, and $n_{kc}$ is the number of completed observations in cell $k$. Note that both of these measures are functions of sufficient statistics for the full log-linear model. Other functions are also possible. Also note that these criteria are not

[1] Throughout this paper, the adjective, "uncensored", refers to the simulated data before setting 30 percent to missing. The adjective, "completed", refers to the combination of simulated data that were not set to missing and imputed data.

tantamount to the properly discredited criterion of minimum mean-square prediction error at the subject level. Finally, the values of the chi-square and $\ell^{24}$ distance measures were averaged over the 100 simulated data sets.

Scenario #1 poses the greatest challenge in terms of developing criteria by which to judge the success of an imputation procedure. We used Kolmogorov-Smirnov statistics to measure the preservation of the marginal distributions of $X$ and $Y$, but the conditional distribution of $Y$ given $X$ is a complex function. A linear approximation would obviously be inadequate. The data publisher cannot predict how a secondary analyst might choose to summarize the relationship of the two variables. In an attempt to capture the general essence of the conditional distribution, we chose the invariance of a nonparametric regression of $Y$ on $X$ as one criterion of success. We first fit a cubic spline with nine knots to the uncensored data, and then to the imputed data; the results of which are shown in Figure 6. Given the graphical nature of this criterion, we only applied it to one large population with 10000 observations.

We also developed two quantitative criteria, loosely based on the Hosmer-Lemeshow graphs for the fit of logistic regression models, which could be applied to each of the 100 simulated data sets. The general idea is to compare uncensored and completed values of $Y$ within narrow bands of uncensored and completed values of $X$. The measures were constructed in several steps. First, the 2000 observations were grouped into 100 strata based on uncensored $X$. Then the mean and standard deviation of uncensored $Y$ was computed within each uncensored-$X$ band. In parallel, the 2000 observations were grouped into 100 strata based on completed $X$, and the mean and standard deviation of completed $Y$ was computed within each completed-$X$ band. Next, we computed the correlation between the two sets of conditional $Y$-means across the 100 $X$-bands, and between the two sets of conditional $Y$-standard deviations across the 100 $X$-bands. Finally, both correlations were averaged over the 100 simulated data sets.

## 5. Competing Approaches

It is often easier to evaluate a new procedure by comparing it to similar, existing methods. For this reason, we applied two other imputation approaches to each of the four simulated populations described in Section 3. As previously mentioned, one of the best known parametric imputation systems currently available is IVEware, from the University of Michigan. Thus, it seemed natural to compare the performance of this procedure to our semi-parametric approach. To further diversify the imputation approaches being compared, we also chose to test a completely nonparametric method, using sequential hot-decks (as in Fahimi et al., 1993).

Like our semi-parametric algorithm, IVEware requires the specification of various parameters. For example, variable types were declared as continuous for strange pop and the bivariate Normal pair (scenarios #1 and #4) and as categorical for the other populations (scenarios #2 and #3). For strange pop, imputed values for $X$ were restricted to the interval (-1,1). The COEF option was used to perturb model coefficients using a multivariate Normal approximation of the posterior distribution, for the bivariate Normal variables. For the other scenarios, the SIR (Sampling-Importance-Resampling) option was used to generate model coefficients from the actual posterior distribution of model parameters. For strange pop, 10 iterations were used per imputation run; 3 iterations were used for the other scenarios. A different random imputation seed was used for each of the 100 simulated data sets.

With the nonparametric approach, a dummy variable was used as the hard boundary in all hot-deck imputations. For scenarios #1 and #4, the order of sequential hot-decks in the nonparametric approach first imputed $X$ and $Y$ for observations missing data on both variables. Then observations still missing $X$ were imputed conditional on $Y$, i.e., using $Y$ as a soft boundary. Finally, observations still missing $Y$ were imputed conditional on $X$. For scenarios #2 and #3, the nonparametric approach first imputed $X$, $Y$, and $Z$ for observations missing data on all three variables. Subsequent hot-decks were applied in the following order: missing $X$ and $Z$ conditional on $Y$, missing $X$ and $Y$ conditional on $Z$; missing $Y$ and $Z$ conditional on $X$; missing $X$ conditional on $Y$ and $Z$; missing $Z$ conditional on $X$ and $Y$; and missing $Y$ conditional on $X$ and $Z$.

## 6. Results and Discussion

Figure 6 shows the nonparametric regressions of $Y$ on $X$ using only the imputed data from each of the three imputation approaches applied to strange pop. As is evident, the results of the semi-parametric procedure and the nonparametric procedure are similar. There are a very small number of outliers, but the regression lines for the semi-parametric and nonparametric procedures both closely resemble that for the uncensored data. The regression line based on the data imputed by IVEware is much less satisfactory. Unless specifically told otherwise, IVEware expects variables to be linearly related to each other, and forces them to look so after imputation.

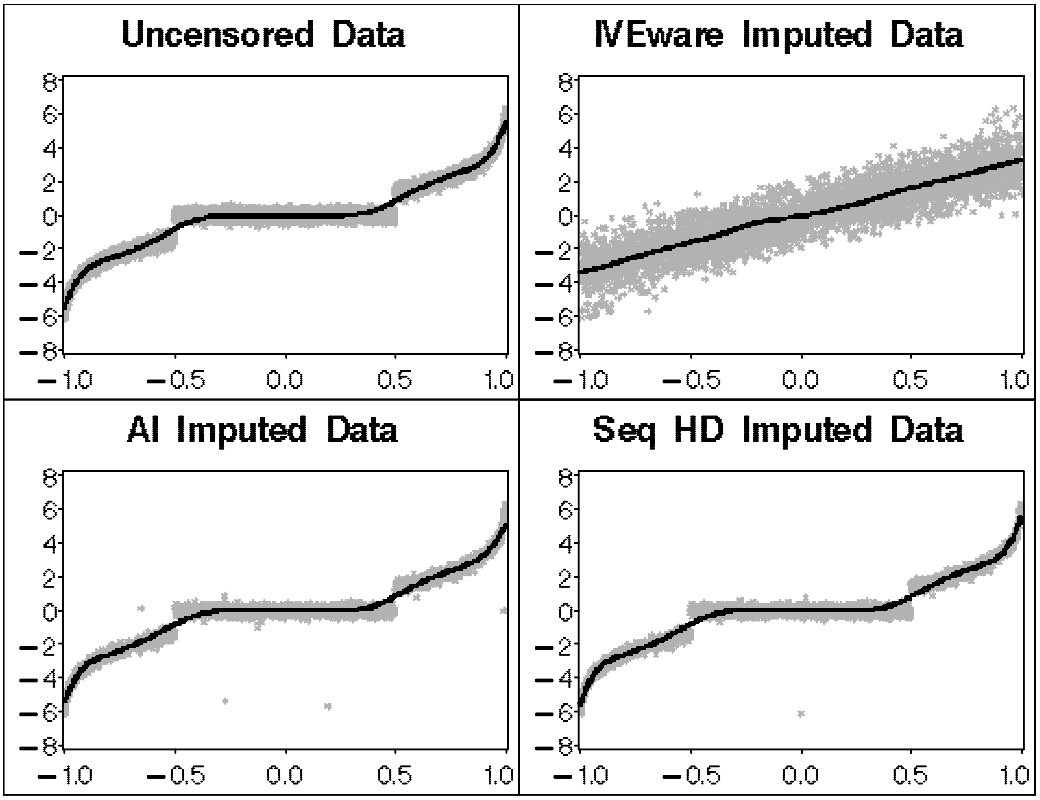

Figure 6. Four nonparametric regressions of *Y* on *X*. Upper left: uncensored data. Upper right: data imputed by IVEware. Lower left: data imputed by central algorithm of this paper (AI=AutoImpute). Lower right: Data imputed by fully nonparametric procedure (sequential hot-decks).

Table 1 shows the results for strange pop averaged over the 100 simulated data sets of 2000 observations each. The semi-parametric procedure is essentially tied with the nonparametric procedure on all four criteria. IVEware performs the worst on three of the four criteria, but the difference is small with respect to preserving the conditional mean structure of the data. IVEware captures the essential feature that *Y* increases monotonically with *X*, however, it does very poorly on preserving the marginal distributions of *X* and *Y*.

Table 2 shows the results for scenario #3. Again, it is a near tie between the semi-parametric and nonparametric procedures. IVEware clearly does not perform as well as the other two imputation approaches.

Table 1. Results for strange pop

| Criterion | Semi-para-metric (1) | IVE-ware (2) | Non-para-metric (3) | (1) vs (2) | (1) vs (3) | (2) vs (3) |
|---|---|---|---|---|---|---|
| Correlation of uncensored and completed $Y$-means across $X$-bands | 0.998 | 0.982 | 0.999 | * | | * |
| Correlation of uncensored and completed $Y$-standard deviations across $X$-bands | 0.561 | 0.319 | 0.602 | | | |
| Kolmogorov-Smirnov for $X$ | 0.036 | 0.104 | 0.037 | * | | * |
| Kolmogorv-Smirnov for $Y$ | 0.038 | 0.192 | 0.038 | * | | * |

$^*$ Difference is significant at the .05 level

Table 2. Results for scenario #3 (checkerboard)

| Criterion | Semi-para-metric (1) | IVE-ware (2) | Non-para-metric (3) | (1) vs (2) | (1) vs (3) | (2) vs (3) |
|---|---|---|---|---|---|---|
| Chi-square distance between uncensored and completed tables | 30.3 | 97.1 | 23.2 | * | | * |
| $\ell^{24}$ distance between uncensored and completed tables | 59.0 | 110.6 | 75.6 | * | | |

$^*$ Difference is significant at the .05 level

With respect to scenarios #2 and #4, all three imputation procedures performed well and achieved similar scores on the judging criteria that had been chosen. For scenario #4, we had anticipated IVEware to do better than the other procedures, but this was not the case. However, with a smaller sample size IVEware would probably have performed the best on the pair of bivariate normal variables. This is because the semi-parametric and nonparametric approaches require larger sample sizes to detect patterns among the data.

Considering all four test scenarios, the nonparametric procedure clearly produces the consistently best results. However, it is not a feasible approach for surveys with large numbers of variables due to the multitude of unique missing data patterns. Semi-parametric and parametric procedures are the only practical options for large-scale imputation by data publishers. Of course, there are many parametric procedures other than those used by IVEware. With sufficient time and energy, analysts could develop appropriate parametric models for “strange pop” and a variety of other unusual populations. An advantage of the semi-parametric approach, however, is that the level of required human involvement is minimal. As stated earlier, the only information required for model selection by the semi-parametric procedure is whether each variable is ordered with range restrictions, homoscedastic ordered with a general range, heteroscedastic ordered with a general range, or unordered. Provided that sample sizes are large enough for the important patterns among the data to appear strongly, the algorithm will take care of the rest. Of course, as with most imputation procedures, our algorithm can be defeated by high order interactions of relevant variables.

For future work, we will be focusing on post-imputation variance estimation. The semi-parametric algorithm can produce multiple imputations. We plan to research whether the use of these multiple imputations with Rubin’s formula leads to post-imputation confidence intervals with nominal or better coverage – the standard that Neyman originally proposed, that Rubin has forcefully reminded us of, and with which we agree.

## References

England, A., Hubbell, K., Judkins, D., and Ryaboy, S. (1994). Imputation of medical cost and payment data. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 406-411.

Fahimi, M., Judkins, D., Khare, M., and Ezzati-Rice, T. M. (1993). Serial imputation of NHANES III with mixed regression and hot-deck techniques. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 292-296.

Judkins, D. R. (1997). Imputing for Swiss cheese patterns of missing data. *Proceedings of Statistics Canada Symposium* 97, New Directions in Surveys and Censuses, 143-148.

Krenzke, T., Judkins, D., and Fan, Z. (2005). Vector imputation at Westat. Presented at Statistical Society of Canada Meetings, Saskatoon, Saskatchewan.

Marker, D. A., Judkins, D. R., and Winglee, M. (2001). Large-scale imputation for complex surveys, in *Survey Nonresponse*, Eds. R. M. Groves, D. A. Dillman, E. L. Eltinge, and R. J. A. Little. New York: Wiley.

Piesse, A., Judkins, D., and Fan, Z. (2005). Item imputation made easy. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, 3476-3479.

Raghunathan, Solenberger, and Hoewyk, (2002). *IVEware: Imputation and Variance Estimation Software Users Guide*. Ann Arbor: Institute for Social Research, University of Michigan.

Rubin, D. B. (1996). Multiple imputation after 19+ years. *Journal of the American Statistical Association*, 91, 473-489.