Hot Deck imputation for multivariate missing data

Jae Kwang Kim[1] Wayne Fuller[2]

[1 2] Department of Statistics, Iowa State University, Ames, IA, USA
[1] Corresponding author: Jae Kwang Kim, e-mail: jkim@iastate.edu

Abstact

Fractional hot deck imputation, considered in Fuller and Kim (2005), is extended to multivariate missing data. The joint distribution of the study items is nonparametrically estimated using a discrete approximation, where the discrete transformation also serves to define imputation cells. The procedure first estimates the probabilities for the cells and then imputes real observations for missing items. Calibration weighting is used to reduce the imputation variance. Replication variance estimation is discussed.

KEY WORDS: EM algorithm, Imputation cell, Replication variance estimation.

1. Introduction

Item nonresponse occurs when a sampled unit provides some information but fails to respond to all items. Imputation, the substitution of values for missing data, is a very popular technique for handling item nonresponse. Hot deck imputation is an imputation procedure in which the value assigned for a missing item is taken from respondents in the current sample.Haziza (2009) and Andridge and Little (2010) provide comprehensive overviews of hot deck imputation methods in survey sampling.

Fractional hot deck imputation was proposed by Kalton and Kish (1984) as a way of achieving efficient hot deck imputation. Kim and Fuller (2004) and Fuller and Kim (2005) provided a rigorous treatment of fractional hot deck imputation and discussed variance estimation. However, their approach is not directly applicable to multivariate missing data. Hot deck imputation for multivariate missing data with arbitrary missing pattern is a notoriously difficult problem because it is difficult to preserve the covariance structure in the imputed data. Judkins et al. (2007) proposed an iterative hot deck imputation procedure that is closely related to the data augmentation algorithm of Tanner and Wong (1987). Judkins et al. (2007) did not provide variance estimators. Other non-hot-deck imputation procedures for multivariate missing data include the multiple imputation approach of Raghunathan et al (2001) and parametric fractional imputation of Kim (2011). The approaches of Judkins et al. (2007) and Raghunathan et al. (2001) are based on conditionally specified models and the imputation methods from the conditionally specified model are subject to the model compatibility problem (Chen, 2011). Conditional models for different missing patterns calculated directly from the observed patterns may not be compatible with each other. The parametric fractional imputation of Kim (2011) used the joint density to create imputed values, but the imputed values are artificial in that they may not be observed values from the sample.

In this paper, we discuss extension of fractional hot deck imputation to multivariate missing data such that the covariance structures are well preserved and variance

estimation is relatively easy. The proposed method is easy to understand and easier to implement than existing methods. Refinement of the procedure is required.

2. Proposed method

2.1 Transformation to integer-valued observations

To discuss multivariate fractional hot deck imputation, suppose that we have a vector of K items, where $\mathbf{Y_i} = (\mathrm{Y}_{1i}, \mathrm{Y}_{2i}, \cdots, \mathrm{Y}_{Ki})$ is the vector for the i-th individual. We write $(\mathrm{Y}_{i,obs}, \mathrm{Y}_{i,mis})$ to denote the (observed, missing) part of $\mathbf{Y}$ in unit i. Because missing patterns can be different for different units, it is understood that $mis = mis(i) \subset \{1,2,\cdots,K\}$. Ideally, the imputed values for $\mathrm{Y}_{i,mis}$ are generated from the conditional distribution $\mathrm{f}(\mathrm{y}_{i,mis}|\mathrm{y}_{i,obs})$. Computation of such predictionrequires model specification and parameter estimation for the specified model. Instead of specifying the conditional distribution, we will consider the joint distribution without specifying a fully parametric joint model.

The basic step in the proposed imputation is temporary replacement of the original data by a discrete approximation. Each continuous variable is transformed into a discrete variable by dividing the range into a small finite number of segments. Let $\tilde{Y}_{ki}$denote the discrete version of Y_{ki}.One simple way of computing the discrete version is to divide the range into groups of equal length. Note that, if Y_k is observed then $\tilde{Y}_k$ is observed.

2.2 Estimated probabilities

Once the $\tilde{Y}_k$ are constructed, we use the observed value $\tilde{Y}_k$ to compute the joint probability of $(\tilde{Y}_1, \tilde{Y}_2, \cdots, \tilde{Y}_k)$. The computation of the joint probability can be performed using a modified version of the EM algorithm. To facilitate discussion, let K=3. Let each variable have five categories. In practice there will be some categorical variables and those variables can have different number of categories. Let the population fraction falling in category *rst* be π_{rst}. We assume that all conditional probabilities are the same for the observed data as for missing data. That is, for example, $\mathrm{P}(\tilde{y}_{i1} = k_1, \tilde{y}_{i2} = k_2|\tilde{y}_{i3} = k_3)$ is the same for observed data as for data with (y_{i1}, y_{i2})missing. Thus, it is possible to estimate the π_{rst} and the conditional probabilities from the sample observations. We estimate positive probabilities for those cells with positive observed probabilities. For a particular missing configuration the conditional probability for each category of each missing variable can be computed. For example the probability that a unit with Y_3 missing falls in category t is

$$\pi_{rs|t} = \frac{\pi_{rst}}{\sum_{jk} \pi_{jkt}}$$

For an unequal probability sample, the conditional probabilities are computed using the sampling weights.

In the EM algorithm, the E-step is essentially the same as applying fully efficient fractional imputation (FEFI) of Fuller and Kim (2005) using all possible combinations of imputed values. The imputed values for $\tilde{Y}_{i,mis}$ are taken from the support of $\tilde{Y}_{mis(i)}$ where the support of $\tilde{Y}_{mis(i)}$ is equal to that for the respondents. Let $\tilde{y}^*_{j,mis(i)}$, $(\mathrm{j} = 1, \cdots, M_i)$ be the set of possible values of $\tilde{Y}_{i,mis(i)}$ in the sample. Once the realized values $\tilde{Y}_{i,mis(i)}$ are

imputed, we can use the idea of EM by weighting (Ibrahim, 1990) to compute the fractional weights. The fractional weight assigned to $\tilde{Y}_{i,mis} = \tilde{y}^*_{j,mis(i)}$ at the *t*-th EM iteration is

$$\tilde{w}^*_{ij(t)} = \frac{\tilde{\pi}_t(\tilde{y}_{i,obs}, \tilde{y}^*_{j,mis(i)})}{\sum_j \tilde{\pi}_t(\tilde{y}_{i,obs}, \tilde{y}^*_{j,mis(i)})} \qquad (1)$$

where $\tilde{\pi}_{\mathrm{t}}(\tilde{\mathrm{y}}_1, \cdots, \tilde{\mathrm{y}}_{\mathrm{k}})$ is the current value of the joint probabilities $\mathrm{P}(\tilde{Y}_1 = \tilde{y}_1, \cdots, \tilde{Y}_K = \tilde{y}_{\mathrm{K}})$ evaluated at the *t*-th EM algorithm. If unit *i* does not have any missing data, then $\tilde{w}^*_{ij(t)} = 1$. Computing fractional weights in (1) corresponds to E-step of the EM algorithm. Once the FEFI is constructed as above, the M-step is to update the joint probabilities by using the weighted average of the FEFI data using fractional weights. That is,

$$\tilde{\pi}_{\mathrm{t+1}}(\tilde{\mathrm{y}}_1, \cdots, \tilde{\mathrm{y}}_{\mathrm{K}}) = (\textstyle\sum_{\mathrm{i}\in\mathrm{A}} \mathrm{w_i})^{-1} \sum_{\mathrm{i}\in\mathrm{A}} \sum_{\mathrm{j}=1}^{\mathrm{M_i}} \mathrm{w_i}\, \tilde{w}^*_{ij(t)} \mathrm{I}\{\tilde{\mathrm{y}}^*_{\mathrm{ij1}}, \cdots, \tilde{\mathrm{y}}^*_{\mathrm{ijK}} = \tilde{y}_{\mathrm{K}}\} \qquad (2)$$

where $\mathrm{w_i}$ is the sampling weight for unit *i* and $\tilde{\mathrm{y}}^*_{\mathrm{ijk}}$ is the *j*-th imputed value for $\tilde{Y}_{\mathrm{ik}}$. If $\tilde{Y}_{\mathrm{ik}}$is observed, then $\tilde{\mathrm{y}}^*_{\mathrm{ijk}}$ is the observed value.

Given the probabilities a "fully efficient" estimator of the mean vector can be computed. Let $\bar{\mathrm{y}}_{\mathrm{j,rst}}$ be the mean for variable *j* in cell *rst*. Then the estimated mean of y is

$$\hat{\boldsymbol{\mu}}_{\mathbf{j}} = \sum_{\mathbf{r}} \sum_{\mathbf{s}} \sum_{\mathbf{t}} \bar{\mathbf{y}}_{\mathbf{j,rst}}\, \hat{\boldsymbol{\pi}}_{\mathbf{rst}}. \qquad (3)$$

2.3 Imputation

Let $\tilde{\mathrm{y}}_{\mathrm{ji}}$ denote the cell value (1,2,3,4,5) of variable $\mathrm{Y_j}$ for individual *i.* Let $\tilde{\mathbf{y}}_{\mathrm{i}}$ denote the **y**-vector for individual *i*. Let $\boldsymbol{\delta}_{\mathrm{c}}$ be the indicator for missingness for observation c, where the *j*-th element of $\boldsymbol{\delta}_{\mathrm{c}}$ is zero if $\tilde{\mathbf{y}}_{\mathrm{jc}}$ is missing. Let $\tilde{\mathbf{y}}_{\mathbf{i}} \cdot \boldsymbol{\delta}_{\mathbf{c}}$ be the element by element product. An observation that has no missing is a potential donor for observation c. Assume 5 values are to be imputed for each missing value. If there are 5 or more observations with no missing and

$$\tilde{\mathbf{y}}_{\mathbf{t}} \cdot \boldsymbol{\delta}_{\mathbf{c}} = \tilde{\mathbf{y}}_{\mathbf{i}} \cdot \boldsymbol{\delta}_{\mathbf{c}}$$

then these observations form the donor set. If not, an additional operation is required to define the donor set.

Given $\boldsymbol{\delta}_{\mathrm{c}}$ the conditional probabilities for each possible vector of imputed values can be calculated. A set of five donor cells is selected with probability proportional to the conditional probability. Then a donor is selected from each of the selected cells.

Given the imputed values the imputed mean for y_j is

$$\bar{y}_{j,imp} = \left(\sum_i w_i\right)^{-1} \sum_i w_i y_{ji} =: \sum_i w_i^* y_{ji,imp}\,.$$

Let $\bar{y}_{j,fif}$ be the estimator of (3). Note that

$$\bar{y}_{j,imp} - \bar{y}_{j,fif} = \sum_{i \in A_{j,imp}} w_i^* (y_{ji,imp} - \bar{y}_{j,fif})$$

where

$$\bar{y}_{j,fif} = \sum_{i \in A_{j,obs}} w_i^* y_i + \sum_{i \in A_{j,imp}} w_i^* \bar{y}_{j,fif}$$

and $y_{ji,imp}$ is the mean of imputed values for y_j for element i. Compute the variance of $\bar{y}_{j,imp} - \bar{y}_{j,fif}$ under the assumption that the y_{ji} are a simple random sample. If

$$\sum_j \left[\hat{V}(\bar{y}_{j,imp} - \bar{y}_{j,fif})\right]^{-1} \left(\bar{y}_{j,imp} - \bar{y}_{j,fif}\right)^2$$

exceeds, say, the 10 % point of a chi-squared distribution with k degrees of freedom the imputed values are rejected and a new set of donors selected. There are alternative rejection criteria that can be considered.

2.4 Fractional weighting

Given a set of donors that satisfy the restrictions, the regression procedure can be used to modify the weights to give the fully efficient estimator.

In some cases, the number of fractional imputations for unit i, M_i, can be quite large. In this case, computing all the joint probabilities may not be feasible. For example, if there are 100 items in the survey then computing all the joint probability is practically impossible. Because the joint probabilities are used to approximate the imputation model (the conditional model of missing data given the observed data), we have only to specify the imputation models (nonparametrically). The first step is to investigate the missing data patterns and then specify the set of variables needed for the imputation model. For example, suppose that we have $(Y_1, Y_2, \cdots, Y_{10})$ and only four missing patterns: (1) All observed, (2) Only Y_2 missing, (3) Only Y_1 missing, (4) Both $Y_1$and Y_2 missing. In this case, we may have only to specify the set of variables that are related with (Y_1, Y_2). Thus, for example, $(Y_1, Y_2, Y_3, Y_5, Y_{10})$ might be used to perform the FEFI method. The reduced model approach is appealing but there is a danger of having incompatible models when we specify different conditional models for many different missing patterns. Tools for model incompatibility may need to be developed. In some cases, we can use a log-linear model to fit a sparse model for the joint probability. For example, we can fit a sparse model for $\log\{\pi_{ijk}\}$ to obtain smoothed fractional weights. Such an approach can be used to construct a reduced imputation model.

Once the imputation model is finalized and the imputation size M_i is large, we can use the PPS sampling to obtain a reduced set of imputed values. Regression weighting can be used to preserve the marginal fractional weights. In the PPS sampling, the size measure is proportional to the original fractional weights after accounting for certainty selection. That is, we first select the candidates with $w_{ij}^* > 1/M$ with certainty.

2.5 Fractional imputation

We now discuss how to perform fractional imputation for $y_{i,mis}$ using fractional imputation for $\tilde{y}_{i,mis}$. Given the value of $\tilde{y}^*_{ij,mis} = \tilde{y}_{j,mis}$, $(\mathrm{j} = 1,2,\cdots,\mathrm{M})$, we can perform a single hot deck imputation from the donor set of observed units with the same value of $\tilde{y}_{j,mis(i)}$. Here, the value of $\tilde{y}_{j,mis(i)}$ can be used as an imputation cell. If at least one donor is identified from the set of fully responding units with $\tilde{y}_{k,mis(i)} = \tilde{y}^*_{ij,mis(i)}$, we can use the donor to obtain imputed values $\tilde{y}^*_{ij,mis} = y_{k,mis(i)}$ for missing $y_{i,mis}$. If such a donor is not identified from the set of fully responding units, then we use hot deck imputation marginally using the marginal values of $\tilde{y}_{j,mis(i)}$ (for each item separately). The discrete version preserves most of the correlation structure in Y_k, and the marginal hot deck imputation will perform well if there is no systematic variation within categories of $\tilde{Y}$.

3. Variance estimation

We now consider variance estimation for the proposed fractional imputation estimator using a replication method.
The replication variance estimator of $\hat{\theta}_n$ takes the form of

$$\widehat{\boldsymbol{V}}_{\boldsymbol{rep}}(\,\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}) = \sum_{\boldsymbol{k=1}}^{\boldsymbol{L}} \boldsymbol{c_k}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}^{(\boldsymbol{k})} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}\right)^{\mathbf{2}} \tag{4}$$

where L is the number of replicates, c_k is the replication factor associated with replication k, and $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}^{(\boldsymbol{k})}$ is the k th replicate of $\hat{\theta}_n$. If $\hat{\theta}_n = \sum_{i=1}^n y_i/n$, then we can write If $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}^{(\boldsymbol{k})} = \sum_{i=1}^n w_i^{(k)} y_i$ for some replication weights $w_1^{(k)}, w_2^{(k)}, \cdots, w_n^{(k)}$. For example, in the jackknife method, we have L=n, $c_k = (n-1)/n$ and

$$w_i^{(k)} = \begin{cases} (n-1)^{-1} & \text{if } i \neq k \\ 0 & \text{if } i = k \end{cases}$$

If we use the above jackknife method for If $\hat{\theta}_n = \sum_{i=1}^n y_i/n$, the resulting jackknife estimator in (4) is algebraically equivalent to $n^{-1}(n-1)^{-1}\sum_{i=1}^n (y_i - \bar{y}_n)^2$.

The replication method for fractional imputation consists of computing a replicated version of joint probability $\tilde{\pi}$, denoted by $\tilde{\pi}^{(k)}$, and then computing replicated fractional weights. For fractional weights of the form (1), we can use

$$\widetilde{\boldsymbol{w}}_{\boldsymbol{ij}}^{*(\boldsymbol{k})} = \frac{\widetilde{\boldsymbol{\pi}}^{(\boldsymbol{k})}(\widetilde{\boldsymbol{y}}_{\boldsymbol{i,obs}}\,,\,\widetilde{\boldsymbol{y}}^*_{\boldsymbol{j,mis(i)}})}{\sum_{\boldsymbol{j}} \widetilde{\boldsymbol{\pi}}^{(\boldsymbol{k})}(\widetilde{\boldsymbol{y}}_{\boldsymbol{i,obs}}\,,\,\widetilde{\boldsymbol{y}}^*_{\boldsymbol{j,mis(i)}})} \tag{5}$$

to obtain initial replication fractional weights. If the initial fractional weights are modified by a regression weighting procedure, then the replicated fractional weight can be constructed similarly, using (5) as the initial fractional weights.

References

Andridge, R.R. and Little, R.J.A. (2010). "A review of hot deck imputation for survey nonresponse". *International Statistical Review*, **78**, 40-64.

Chen, H.Y. (2010). "Compatibility of conditionally specified models". *Statistics and Probability Letters*, 80, 670-677.

Fuller, W.A. and Kim, J.K. (2005). "Hot deck imputation for the response model". *Survey Methodology*, **31**,139-149.

Haziza, D. (2009). "Imputation and inference in the presence of missing data". In *Handbook of Statistics*, Volume **29**, *Sample Surveys*: *Theory Methods and Inference*, Edited by C.R. Rao and D. Pfeffermann, 215-246.

Ibrahim, J. G. (1990). "Incomplete data in generalized linear models". *Journal of the American Statistical Association*, **85**, 765-769.

Judkins,D., Krenzke, T., Piesse, A., Fan,Z., and Haung, W.C. (2007). "Preservation of skip patterns and covariate structure through semi-parametric whole questionnaire imputation". In *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 3211-3218.

Kalton, G. and Kish, L. (1984). "Some efficient random imputation methods". *Communications in Statistics: Series A*, **13**, 1919-1939.

Kim, J.K. (2011). "Parametric fractional imputation for missing data analysis", *Biometrika*, **98**, 119--132.

Kim, J.K. and Fuller, W.A. (2004). "Fractional hot deck imputation". *Biometrika*, **91**, 559-578.

Tanner, M.A. and Wong, W.H. (1987). "The calculation of posterior distribution by data augmentation". *Journal of the American Statistical Association*, **82**, 528-540.