UNITED NATIONS
ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Budapest, Hungary, 14-16 September 2015)

Topic (v): Emerging methods and data revolution

Multiple Ratio Imputation by the EMB Algorithm

Prepared by Masayoshi Takahashi,[1] National Statistics Center, Japan

I. Introduction

1. Missing data problems are ubiquitous in many fields, including official statistics, where one of the common treatments of missing data is ratio imputation (de Waal *et al.*, 2011; Thompson & Washington, 2012; Office for National Statistics, 2014). On the other hand, multiple imputation has been the recommended practice from statisticians (Rubin, 1987; Little & Rubin, 2002). Among statisticians, multiple imputation is known to be the gold standard of treating missing data (Baraldi & Enders, 2010; Cheema, 2014). While ratio imputation is often employed to deal with missing values in practice, the literature is devoid of multiple ratio imputation, leading to a gap between theory and practice. This paper proposes a novel application of the Expectation-Maximization with Bootstrapping (EMB) algorithm to ratio imputation, where multiply-imputed values will be created for each missing value. The objective of this paper is to present the mechanism of multiple ratio imputation and to assess the performance compared to traditional imputation methods. For this purpose, Monte Carlo simulation is applied to the newly-developed *R*-function for multiple ratio imputation. A small application to the 2012 Japanese Economic Census data is also presented to illustrate the usefulness of multiple ratio imputation. Also, this research implemented multiple ratio imputation by the Expectation-Maximization with Bootstrapping (EMB) algorithm in the *R* statistical environment (to be released soon).

II. Assumptions of Missing Mechanisms

2. Suppose that $\mathbf{D}$ is an $n \times p$ dataset, where n is the number of observations and p is the number of variables. Also, let $\mathbf{R}$ be a response indicator matrix, whose dimension is the same as $\mathbf{D}$. Whenever $\mathbf{D}$ is observed $\mathbf{R} = 1$, and whenever $\mathbf{D}$ is not observed $\mathbf{R} = 0$. Note, however, that *R* in Italics refers to the *R* statistical environment in this paper. Furthermore, $\mathbf{D_{obs}}$ refers to the observed part of data, and $\mathbf{D_{mis}}$ refers to the missing part of data, i.e., $\mathbf{D} = \{\mathbf{D_{obs}}, \mathbf{D_{mis}}\}$. The first assumption is Missing Completely At Random (MCAR), which is $P(\mathbf{R}|\mathbf{D}) = P(\mathbf{R})$. The second assumption is Missing At Random (MAR), which is $P(\mathbf{R}|\mathbf{D}) = P(\mathbf{R}|\mathbf{D_{obs}})$. The third assumption is Non-Ignorable (NI), which is $P(\mathbf{R}|\mathbf{D}) \neq P(\mathbf{R}|\mathbf{D_{obs}})$.

III. Existing Algorithms and Software for Multiple Imputation

3. Before moving on to the discussion of multiple ratio imputation, this section is a concise review of the existing multiple imputation algorithms and software programs. As of today, there are three major algorithms for multiple imputation.

[1] The author wishes to thank Dr. Manabu Iwasaki (Seikei University), Dr. Michiko Watanabe (Keio University), Dr. Takayuki Abe (Keio University), Dr. Tetsuto Himeno (Seikei University), Mr. Nobuyuki Sakashita (National Statistics Center), and Ms. Kazumi Wada (National Statistics Center) for their valuable comments on earlier versions of this paper. However, any remaining errors are the author's responsibility. Also, note that the views and opinions expressed in this paper are the author's own, not necessarily those of the institution. The analyses in this paper were conducted using *R* 3.1.0.

4. The first traditional algorithm is based on Markov chain Monte Carlo (MCMC). This is the original version of Rubin's (1978, 1987) multiple imputation. *R*-Package Norm currently implements this version of multiple imputation (Schafer, 1997; Fox, 2015). A commercial software program using the MCMC algorithm is SAS Proc MI (SAS, 2011). The second major algorithm is called Fully Conditional Specification (FCS), also known as chained equations by van Buuren (2012). *R*-Package MICE currently implements this version of multiple imputation (van Buuren & Groothuis-Oudshoorn, 2011; van Buuren & Groothuis-Oudshoorn, 2015). Other commercial software programs using the FCS algorithm are SPSS Missing Values (SPSS, 2009) and SOLAS (Statistical Solutions, 2011). The FCS algorithm is known to be flexible. The third relatively new algorithm is the EMB algorithm by Honaker & King (2010). *R*-Package Amelia II currently implements this version of multiple imputation (Honaker *et al.*, 2011; Honaker *et al.*, 2015). The EMB algorithm is known to be computationally efficient.

5. Assessing the superiority among the different multiple imputation algorithms is beyond the scope of the current study. Suffice it to say that, if the underlying distribution can be approximated by a multivariate normal distribution with the MAR condition, all of the three algorithms essentially give the same answers (Takahashi, 2014). As for the performance of the EMB algorithm, Honaker & King (2010) contend that the estimates of population parameters in bootstrap resamples can be appropriately used instead of random draws from the posterior. In fact, Rubin (1987) argues that the approximately Bayesian bootstrap method is proper imputation because it incorporates between-imputation variability. Also, Little & Rubin (2002) assure that the substitution of Maximum Likelihood Estimates (MLEs) from bootstrap resamples is proper because the MLEs from the bootstrap resamples are asymptotically identical to a sample drawn from the posterior distribution. Therefore, multiple imputation by the EMB algorithm can be considered to be proper imputation in Rubin's sense (1987).

6. Also, according to van Buuren (2012), the bootstrap method is computationally efficient because there is no need to make a draw from the χ^2 distribution, unlike the other traditional algorithms of multiple imputation. This means that it is not necessary to resort to the Cholesky decomposition (a.k.a. the Cholesky factorization), the property of which is that if $\mathbf{A}$ is a symmetric positive definite matrix, i.e., $\mathbf{A} = \mathbf{A^T}$, then there is a matrix $\mathbf{L}$ such that $\mathbf{A} = \mathbf{LL^T}$, which means that $\mathbf{A}$ can be factored into $\mathbf{LL^T}$, where $\mathbf{L}$ is a lower triangular matrix with positive diagonal elements (Leon, 2006).

7. Nonetheless, *R*-Package Amelia II does not allow us to estimate the ratio imputation model. In fact, none of the existing multiple imputation software programs mentioned above has an option to perform multiple ratio imputation. This paper contributes to the literature by applying the EMB algorithm to ratio imputation; thus, the new multiple ratio imputation is born.

IV. Single Ratio Imputation

8. Suppose that the population model is equation (1), where the ratio $\bar{Y}_1/\bar{Y}_2$ is generally a consistent but biased estimator of ω, except for some special cases, and the mean of ε_i is 0 with unknown variance. However, as the sample size increases, this bias tends to be negligible. Also, the distribution of the ratio estimate is known to be asymptotically normal (Cochran, 1977, p.153).

$$Y_{i1} = \omega Y_{i2} + \varepsilon_i \tag{1}$$

9. Under the following special case, the ratio estimator is unbiased, where ε_i is independent of Y_{i2} with the mean of 0 and the unknown variance of $Y_{i2}\sigma^2$ (Cochran, 1977, p.158; Shao, 2000, p.79; Liang *et al.*, 2008, p.2).

10. Ratio imputation takes the form of a simple regression model without an intercept, in which the slope coefficient is calculated by the ratio between the means of two variables (de Waal *et al.*, 2011). In other words, the ratio imputation model is equation (2), where $\hat{\omega} = \bar{Y}_{1,obs}/\bar{Y}_{2,obs}$. Also, by adding a disturbance term, ratio imputation can be made stochastic as in equation (3) (Hu *et al.*, 2001).

$$\hat{Y}_{i1} = \hat{\omega} Y_{i2} \tag{2}$$
$$\hat{Y}_{i1} = \hat{\omega} Y_{i2} + \hat{u}_i \tag{3}$$

V. Theory of Multiple Ratio Imputation

11. As the literature has demonstrated, if the missing mechanism is MAR, imputation can ameliorate the bias due to missingness (Little & Rubin, 2002; de Waal *et al.*, 2011). Caution is that imputed values are not the complete reproduction of the true values, and that the goal of imputation is generally not to replicate the truth for each missing value, but to make it possible to have a valid statistical inference. For this purpose, it is necessary to evaluate the error due to missingness, for which Rubin (1978, 1987) proposed multiple imputation as a solution. Indeed, Baraldi & Enders (2010) and Cheema (2014) demonstrate that multiple imputation is superior to listwise deletion, mean imputation, and single regression imputation. Furthermore, Leite & Beretvas (2010) contend that multiple imputation is robust to violations of continuous variables and the normality assumption. Thus, multiple imputation is the gold standard of treating missing data. The current study extends the utility of ratio imputation by transforming it to multiple imputation by way of the EMB algorithm described in this section.

12. Multiple imputation in theory is to randomly draw several imputed values from the distribution of missing data. However, missing data are by definition unobserved; as a result, the true distribution of missing data is always unknown. A solution to this problem is to estimate the posterior distribution of missing data based on observed data, and to make a random draw of imputed values. As seen in the previous section, the value of ω is estimated by $\hat{\omega} = \bar{Y}_{1,obs}/\bar{Y}_{2,obs}$. Therefore, in order to create multiple ratio imputation, the mean vector is what needs to be randomly drawn from the posterior distribution of missing data given observed data.

13. Honaker & King (2010) suggested the use of the EMB algorithm for the purpose of drawing the mean vector and the variance-covariance matrix from the posterior density. Honaker *et al.* (2011) presented a general-purpose multiple imputation software program called Amelia II, which is a computationally efficient and highly reliable multiple imputation program. Nevertheless, as seen above, Amelia II does not allow us to estimate the ratio imputation model. Thus, this paper applies the EMB algorithm to ratio imputation to create multiple ratio imputation.

14. In the rest of this section, let us review the bootstrap method and the Expectation-Maximization (EM) algorithm, in order to illustrate how the EMB algorithm works for the purpose of generating multiple ratio imputation. For this purpose, this paper uses the following example data in Table 1.

Table 1. Example Data (Simulated Weekly Income in U.S. Dollars)

ID	Income0	Income1	Income2
1	543	543	514
2	272	272	243
3	**797**	**NA**	597
4	239	239	264
5	415	415	350
6	371	371	346
7	**650**	**NA**	545
8	495	495	475
9	553	553	564
10	**710**	**NA**	558

Note. Income0 is the true complete variable. Income1 is the observed incomplete Variable at time *t* with NA = missing. Income2 is the auxiliary variable at *t* – 1.

A. Nonparametric Bootstrap

15. The first step for multiple ratio imputation is to randomly draw vectors of means from an appropriate posterior distribution to account for the estimation uncertainty. The EMB algorithm replaces the complex process of random draws from the posterior by nonparametric bootstrapping, which is a general resampling method, where samples are taken from the original sample data. The nonparametric bootstrap uses the existing sample data (size = *n*) as the pseudo-population and draws resamples (size = *n*) with replacement *M* times (Horowitz, 2001). For example, if data $Y_1, \ldots, Y_n$ are independently and

identically distributed from an unknown distribution F, this distribution is estimated by $\hat{F}(y)$, which is the empirical distribution F_n defined in equation (4), where $I(Y)$ is the indicator function of the set Y.

$$F_n(y) = \frac{1}{n}\sum_{i=1}^{n} I(Y_i \leq y) \tag{4}$$

16. Based on equation (4), bootstrap resamples are generated. The distribution $\hat{F}$ can be any estimator in order to generate the bootstrap resamples of F based on $Y_1, \ldots, Y_n$. A nonparametric estimator of F is the empirical distribution F_n defined by equation (4) (Shao & Tu, 1995; DeGroot & Schervish, 2002). Table 2 is an example of bootstrap data.

Table 2. Bootstrap Data (M = 2)

Incomplete Data		Bootstrap 1		Bootstrap 2	
Income1	Income2	IncomeB11	IncomeB12	IncomeB21	IncomeB22
543	514	NA	545	495	475
272	243	272	243	272	243
NA	597	239	264	371	346
239	264	NA	597	415	350
415	350	272	243	NA	597
371	346	553	564	543	514
NA	545	272	243	272	243
495	475	495	475	NA	545
553	564	553	564	371	346
NA	558	272	243	NA	545

Note. NA represents missing values.

17. When incomplete data are bootstrapped, the chance is that each bootstrap resample is also incomplete. Therefore, the information from incomplete bootstrap resamples is biased and inefficient. The EM algorithm refines bootstrap estimates in the next section.

B. EM Algorithm

18. The EM algorithm first assumes a certain distribution and tentative starting values for the mean and the variance-covariance. Using these starting values, an expected value of model likelihood is calculated, the likelihood is maximized, model parameters are estimated that maximize these expected values, and then the distribution is updated. The expectation and the maximization steps are repeated until the values converge, whose properties are known to be an MLE (Schafer, 1997; Iwasaki, 2002; Do & Batzoglou, 2008). Formally, the EM algorithm can be summarized as follows. Starting from an initial value θ_0, repeat the following two steps:

E-step: $Q(\theta|\theta_t) = \int l(\theta|Y)\, P(Y_{mis}|Y_{obs}; \theta_t) dY_{mis}$, where $l(\theta|Y)$ is log likelihood.
M-step: Maximize $\theta_{t+1} = \arg\max_{\theta} Q(\theta|\theta_t)$ with respect to θ.
Under certain conditions, it is proven that $\theta_t \to \hat{\theta}$ $(t \to \infty)$.

19. The values in Table 2 are incomplete. If the EM algorithm is used to refine these values, the EM mean for incomeB11 is 405.741 and the EM mean for incomeB12 is 398.100; also, the EM mean for incomeB21 is 450.912 and the EM mean for incomeB22 is 420.400. Using these values, the ratio will be estimated as 1.019 and 1.072, respectively. Thus, in this small example, the estimated ratio is 1.046 on average, ranging from 1.019 to 1.072. This variation captures the estimation uncertainty due to missingness, which is called the between-imputation variance (Little & Rubin, 2002). Obviously, real applications require a much larger value of M (Graham *et al.*, 2007; Bodner, 2008).

C. Application of the EMB Algorithm to Multiple Ratio Imputation

20. The multiple ratio imputation model is defined by equation (5), where tilde means that these values are drawn from an appropriate posterior distribution of missing data. In other words, $\tilde{\omega}$ is a vector of ratios drawn from the appropriate posterior taking estimation uncertainty into account and $\tilde{u}_i$ is the disturbance term taking fundamental uncertainty into account (King *et al.*, 2001).

$$\begin{aligned} \tilde{Y}_{i1} &= \tilde{\omega} Y_{i2} + \tilde{u}_i, \text{where} \\ \tilde{\omega} &= \frac{\tilde{\bar{Y}}_1}{\tilde{\bar{Y}}_2} \end{aligned} \tag{5}$$

21. Table 3 presents the result of multiple ratio imputation, where $M = 2$, using the same example data as in Table 1. The model is $\widetilde{Income}_1 = \tilde{\omega} \times Income_2 + \tilde{u}_i$, where the mean of $\tilde{\omega}$ is 1.050 with the standard deviation of 0.048, ranging from 0.903 to 1.342 if $M = 100$. This variation captures the stability of the imputation model, which serves as a diagnostic method for imputation, because the simulation standard error (essentially, between-imputation variance) can be appropriately used for assessing the likeliness of the simulation estimator being close to the true parameter of interest (DeGroot & Schervish, 2002). Note that, in Table 3, the values of Imputation1 and Imputation2 for ID 3, 7, and 10 change over columns Imputation1 to Imputation2, because the values in these rows are imputed values. Also, note that the values in the other rows do no change over columns, because these are observed values.

Table 3. Multiple Ratio Imputation Data ($M = 2$)

ID	Income1	Income2	Imputation1	Imputation2
1	543	514	543.000	543.000
2	272	243	272.000	272.000
3	**NA**	597	**620.917**	**662.732**
4	239	264	239.000	239.000
5	415	350	415.000	415.000
6	371	346	371.000	371.000
7	**NA**	545	**571.100**	**600.655**
8	495	475	495.000	495.000
9	553	564	553.000	553.000
10	**NA**	558	**597.406**	**637.115**

22. Just as in regular multiple imputation (Little & Rubin, 2002), the estimates by multiple ratio imputation can be combined as follows. Let $\hat{\theta}_m$ be an estimate based on the m-th multiply-imputed dataset. The combined point estimate $\bar{\theta}_M$ is equation (6). The variance of the combined point estimate consists of two parts. Let v_m be the estimate of the variance of $\hat{\theta}_m$, $\text{var}(\hat{\theta}_m)$, let $\bar{W}_M$ be the average of within-imputation variance, let $\bar{B}_M$ be the average of between-imputation variance, and let T_M be the total variance of $\bar{\theta}_M$. Then, the total variance of $\bar{\theta}_M$ is equation (7), where $(1 + 1/M)$ is an adjustment factor because M is not infinite. If M is infinite, $\lim_{M\to\infty} \left(1 + \frac{1}{M}\right) \tilde{v}_M = \tilde{v}_M$. In short, the variance of $\bar{\theta}_M$ takes into account within-imputation variance and between-imputation variance.

$$\bar{\theta}_M = \frac{1}{M} \sum_{m=1}^{M} \hat{\theta}_m \tag{6}$$

$$T_M = \bar{W}_M + \left(1 + \frac{1}{M}\right) \bar{B}_M = \frac{1}{M} \sum_{m=1}^{M} v_m + \left(1 + \frac{1}{M}\right) \left[\frac{1}{M-1} \sum_{m=1}^{M} \left(\hat{\theta}_m - \bar{\theta}_M\right)^2 \right] \tag{7}$$

VI. Monte Carlo Evidence

23. Using a total of the 45,000 simulated datasets with various characteristics, this paper compares the Relative Root Mean Square Errors (RRMSE) of the estimators for the mean, the standard deviation, and the *t*-statistics in regression across different missing data handling techniques. The data used in this section are a modified version of the simulated data used by King *et al.* (2001). The Monte Carlo experiments here are based on 1,000 iterations, each of which is a random draw from the following multivariate normal distribution: Variables y1 and y2 are normally distributed with the mean vector (6, 10) and the standard deviation vector (1, 1), where the correlation between y1 and y2 is set to 0.6. Each set of these 1,000 data is further repeated for n = 50, n =100, n =200, n =500, and n =1,000; thus, there are 5,000 datasets of five different data sizes. Our simulated data assume that the population model is equation (8).

$$Y_{i1} = \omega Y_{i2} + \varepsilon_i, where$$

$$\omega = \frac{\bar{Y}_1}{\bar{Y}_2} = 0.6, \varepsilon_i \sim N(0, 0.64) \tag{8}$$

24. Furthermore, each of these 5,000 datasets is made incomplete using the three data generation processes of MCAR, MAR, and NI as in Table 4, with the average missing rates of 15%, 25%, and 35%. These missing rates approximately cover the range from 10% to 40% missingness. Note that Variable y1 is the target incomplete variable for imputation, Variable y2 is completely observed in all of the situations to be used as the auxiliary variable, and Variable r in Table 4 is 1,000 sets of continuous uniform random numbers ranging from 0 to 1 for the missingness mechanism.

Table 4: Missingness Mechanisms and Missing Rates

MCAR	Missingness of y1 is a function of r. 15%: y1 is missing if r > 0.85. 25%: y1 is missing if r > 0.75. 35%: y1 is missing if r > 0.65.
MAR	Missingness of y1 is a function of y2 and r. 15%: y1 is missing if y2 > 10 and r > 0.7. 25%: y1 is missing if y2 > 10 and r > 0.5. 35%: y1 is missing if y2 > 10 and r > 0.3.
NI	Missingness of y1 is a function of y1 and r. 15%: y1 is missing if y1 > 6 and r > 0.7. 25%: y1 is missing if y1 > 6 and r > 0.5. 35%: y1 is missing if y1 > 6 and r > 0.3.

25. Therefore, there is a total of 45,000 datasets, i.e., 1,000 datasets multiplied by five sample sizes, three missing mechanisms, and three missing rates. The overall performance can be captured by the Mean Square Error (MSE), which is defined as equation (9), where θ is the truth and $\hat{\theta}$ is an estimator. The MSE measures the dispersion around the true value of the parameter, suggesting that an estimator with the smallest MSE is the best of a competing set of estimators (Gujarati, 2003).

$$MSE(\hat{\theta}) = E(\hat{\theta} - \theta)^2 \tag{9}$$

26. Following Di Zio & Guarnera (2013), this study uses the Relative Root Mean Square Error (RRMSE), which is defined as equation (10), where θ is the truth, $\hat{\theta}$ is an estimator, and T is the number of trials. For example, θ in the following analyses is the mean, the standard deviation, and the *t*-statistic based on complete data. $\hat{\theta}$ is the estimated quantity based on imputed data.

$$RRMSE(\hat{\theta}) = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\left(\frac{\hat{\theta} - \theta}{\theta}\right)^2} \tag{10}$$

27. In the following analyses, the multiple ratio imputation model sets the number of multiply-imputed datasets (M) to 100, based on the recent findings in the multiple imputation literature (Graham *et al.*, 2007; Bodner, 2008).

VII. Summary of the Monte Carlo Results

A. RRMSE Comparison of the Mean

28. The standard recommendation (de Waal *et al.*, 2011) is that if the goal is to calculate a point estimate, the choice is deterministic single ratio imputation. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is as good as that of deterministic single ratio imputation, which is known to be a preferred method for the estimation of the mean. If multiple ratio imputation equally performs well compared to deterministic single ratio imputation, this means that multiple ratio imputation attains the highest performance in estimating the mean.

29. In all of the 45 patterns, deterministic ratio imputation and multiple imputation both outperform listwise deletion. Even when the missing mechanism is MCAR, the results by imputation are always better than those of listwise deletion. Between the ratio imputation methods, deterministic ratio imputation slightly performs better than multiple ratio imputation in 32 out of the 45 patterns with 13 ties. However, 43 out of the 45 patterns are within a 0.01-point difference in terms of the RRMSE. Thus, there are no significant differences between deterministic ratio imputation and multiple ratio imputation.

30. Furthermore, this difference is expected to disappear as M approaches infinity. In general, under the situations where the model is correctly specified and the assumption of MAR is satisfied, both single imputation and multiple imputation ($M = \infty$) would be unbiased and agree on the point estimation (Donders *et al.*, 2006). The results in the Monte Carlo experiments assure that this general relationship also applies to the relationship between single ratio imputation and multiple ratio imputation. Therefore, on average, multiple ratio imputation can be expected to give essentially the same answers as to the estimation of the mean, compared to deterministic ratio imputation.

B. RRMSE Comparison of the Standard Deviation

31. The standard recommendation (de Waal *et al.*, 2011) is that if the goal is to estimate the variation of data, the choice is stochastic single ratio imputation. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is as good as that of stochastic ratio imputation, which is known to be a preferred method to estimate the standard deviation.

32. In all of the 45 patterns, multiple ratio imputation always outperforms listwise deletion. Even when the missing mechanism is MCAR, the results by multiple ratio imputation are always better than those of listwise deletion. In contrast, stochastic ratio imputation outperforms listwise deletion in only 20 out of the 45 patterns. Especially, when the missing mechanism is MCAR, listwise deletion often outperforms stochastic ratio imputation in 14 out of the 15 patterns, although the difference is minimal. This implies that when missing data are suspected to be MCAR, there is a chance that using stochastic ratio imputation may make the situation worse than simply using listwise deletion.

33. Between the ratio imputation methods, multiple ratio imputation often performs better than stochastic ratio imputation, 43 out of the 45 patterns. Therefore, this study contends that multiple ratio imputation is the preferred method for the estimation of the standard deviation. The results in the Monte Carlo experiments imply that, regardless of missing mechanisms, multiple ratio imputation should be used for the purpose of estimating the standard deviation.

C. RRMSE Comparison of the *t*-Statistics

34. The standard recommendation (van Buuren, 2012; Hughes *et al.*, 2014) is that if the goal is to obtain valid inferences with standard errors, the choice is multiple imputation which is a superior variance-estimation method. Thus, the main purpose of this comparison is to show that the performance of multiple ratio imputation is better than that of regular multiple imputation in terms of estimating the *t*-statistics when the true model is equation (8). The comparison of the *t*-statistics in regression is appropriate, because it is the quantity of interest for many applied researchers in disputing whether an independent variable has some impact on a dependent variable.

35. When the true model is equation (8), it is expected that multiple ratio imputation is more accurate and efficient than regular multiple imputation. The comparison of multiple ratio imputation and Amelia II is appropriate, because the algorithm is the same EMB under the same platform of the *R* statistical environment. In all of the 45 patterns, regular multiple imputation and multiple ratio imputation both outperform listwise deletion. Furthermore, multiple ratio imputation always outperforms regular multiple imputation under the condition where the true population model is equation (8). According to Cheema (2014, p.58), comparisons of *t* and *F* statistics are fair because the complete sample and the imputed sample are identical in all respects including power, except for the fact that no values were missing in the complete sample while some values were missing in the imputed values. Therefore, the difference in the observed value of statistics are caused by the difference between imputed values and their true counterparts.

36. Therefore, multiple ratio imputation adds an important option for the tool kit of imputing and analyzing the mean, the standard deviation, and the *t*-statistics. If the true model is equation (8), multiple ratio imputation is at least as good as and in many cases better than the other traditional imputation methods for the three quantities of interest, regardless of the missingness mechanisms. To be fair, this paper never claims that multiple ratio imputation is always superior to regular multiple imputation. If the true model is not equation (8), the superiority shown in this section is not guaranteed.

VIII. Empirical Example

37. This section presents a small example that demonstrates how switching from the existing methods can change substantive conclusions. Note that the results presented in this section are analyzed by the National Statistics Center of Japan, using the dataset of the 2012 Japanese Economic Census for Business Activity (Statistics Bureau of Japan, 2012). Also note that the views and opinions expressed in the analysis using this dataset are the author's own, not necessarily those of the institution. The data in this section focus on the case of Division I (Wholesale and Retail Trade) in the prefecture of Tokyo. The target variable for imputation is turnover, and the quantity of interest is the mean of turnover. Variable cost is used as a proxy variable for turnover, because turnover is expected to increase proportionately to cost. Using the number of employees, the units are classified in this study; in other words, our data focus on the establishments and enterprises with the number of employees equal to 1. For comparisons, this section has the following two models. Equation (11) is deterministic single ratio imputation, and equation (12) is multiple ratio imputation.

$$\widehat{Turnover}_i = \hat{\omega} Cost_i \tag{11}$$

$$\widetilde{Turnover}_i = \tilde{\omega} Cost_i + \tilde{u}_i \tag{12}$$

38. Table 5 presents the result of analysis. The listwise deletion estimate is 3569.12. In contrast, the point estimate by deterministic ratio imputation is 3526.73. The listwise deletion estimate is 1.012 times as large as the estimate by deterministic ratio imputation. The point estimate by multiple ratio imputation is 3526.69, and it is 1.000 times as large as the estimate by deterministic ratio imputation. Although the point estimates are almost equal between single and multiple ratio imputation, the multiple ratio imputation model has an additional row for BISD (4.74). With this information, the imputer/analyst is approximately 95% confident that the true mean value of complete turnover is somewhere between 3517.21 and 3536.16, after taking the error due to missingness into account. In fact, the imputer/analyst can be confident that the estimate by ratio imputation (3526.7) is meaningfully different from the listwise deletion estimate (3569.1) which is outside the confidence interval, a finding unascertainable with the traditional single ratio imputation method.

Table 5. Mean of Turnover (Division I)

	Listwise Deletion	Deterministic Ratio Imputation	Multiple Ratio Imputation
Mean	3569.12	3526.73	3526.69
BISD	NA	NA	4.74
CI (95%)	NA	NA	3517.21, 3536.16

Note. NA means Not-Applicable. $M = 100$ for multiple ratio imputation.
The number of observations is 3,811. Units are a million yen.

39. Apparently, the difference between the listwise deletion estimate and the estimate by ratio imputation looks small; however, note that the unit is a million yen, which is approximately 8,000 U.S. dollars ($1 = 125 yen). Furthermore, this is the mean among 3,811 establishments and enterprises. Therefore, the difference in the total amount of turnover between the two estimates is 161.55 billion yen, or $1.29 billion, which is quite large.

IX. Conclusion

40. This paper proposed a novel application of the EMB algorithm to ratio imputation, and presented the mechanism and the usefulness of multiple ratio imputation. For this purpose, Monte Carlo simulation was applied to the newly-developed *R*-function for multiple ratio imputation. An application to the 2012 Japanese Economic Census data was also presented to illustrate the usefulness of multiple ratio imputation. This research showed that the fit of multiple ratio imputation was at least as good as or in many cases better than that of single ratio imputation and regular multiple imputation if the assumption holds. Specifically, for the purpose of estimating the mean, the performance of deterministic ratio imputation and multiple ratio imputation are essentially identical, with multiple ratio imputation having additional information on estimation uncertainty. In order to estimate the standard deviation, multiple ratio imputation outperforms stochastic ratio imputation. If the goal is to estimate the *t*-statistics in regression, multiple ratio imputation clearly outperforms regular multiple imputation when the assumptions of the ratio model are satisfied. These findings are important because researchers are recommended to use different ways of imputation, depending on the type of statistical analyses, meaning that there are no one-size-fit-for-all imputation methods (Poston & Conde, 2014). Thus, multiple ratio imputation will be a valuable addition for treating missing data problems, which will expand the choice of missing data treatments.

References

[1] Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. *Journal of School Psychology*, 48(1), 5-37.

[2] Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling*, 15, 651-675.

[3] Cheema, J. R. (2014). Some general guidelines for choosing missing data handling methods in educational research. *Journal of Modern Applied Statistical Methods*, 13(2), 53-75.

[4] Cochran, W. G. (1977). *Sampling Techniques*, 3rd edition. New York, NY: John Wiley & Sons.

[5] DeGroot, M H., & Schervish, M. J. (2002). *Probability and Statistics*, 3rd edition. Boston, MA: Addison-Wesley.

[6] de Waal, T., Pannekoek, J., & Scholtus, S. (2011). *Handbook of Statistical Data Editing and Imputation*. Hoboken, NJ: John Wiley & Sons.

[7] Di Zio, M., & Guarnera, U. (2013). Contamination model for selective editing. *Journal of Official Statistics*, 29 (4), 539-555.

[8] Do, C. B., & Batzoglou, S. (2008). What is the expectation maximization algorithm? *Nature Biotechnology*, 26 (8), 897-899.

[9] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*, 59, 1087-1091.

[10] Fox, J. (2015). *Package 'Norm'* [Computer software]. Retrieved from: http://cran.r-project.org/web/packages/norm/norm.pdf

[11] Graham, J. W., Olchowski, A. E., & Gilreath, T. D. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, 8(3), 206-213.

[12] Gujarati, D. N. (2003). *Basic econometrics*, 4th edition. New York, NY: McGraw-Hill.

[13] Honaker, J., & King, G. (2010). What to do about missing values in time series cross-section data. *American Journal of Political Science*, 54(2), 561-581.
[14] Honaker, J., King, G., & Blackwell, M. (2011). Amelia II: a program for missing data. *Journal of Statistical Software*, 45(7), 1-47.
[15] Honaker, J., King, G., & Blackwell, M. (2015). *Package 'Amelia'* [Computer software]. Retrieved from: http://cran.r-project.org/web/packages/Amelia/Amelia.pdf
[16] Horowitz, J. L. (2001). The bootstrap. In J. J. Heckman & E. Leamer (Eds), *Handbook of Econometrics* (pp.3160-3228), vol.5. Amsterdam: Elsevier.
[17] Hu, M., Salvucci, S., & Lee, R. (2001). *A Study of Imputation Algorithms.* Working Paper No. 2001–17. U.S. Department of Education. National Center for Education Statistics. Retrieved from: http://nces.ed.gov/pubs2001/200117.pdf
[18] Hughes, R. A., Sterne, J. A. C., & Tilling, K. (2014). Comparison of imputation variance estimators. *Statistical Methods in Medical Research*, forthcoming.
[19] Iwasaki, M. (2002). *Fukanzen Data no Toukei Kaiseki* (*Foundations of Incomplete Data Analysis*). Tokyo: EconomistSha Publications, Inc.
[20] King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: an alternative algorithm for multiple imputation. *American Political Science Review*, 95(1), 49-69.
[21] Leite, W., & Beretvas, S. (2010). The performance of multiple imputation for Likert-type items with missing data. *Journal of Modern Applied Statistical Methods*, 9(1), 64-74.
[22] Leon, S. J. (2006). *Linear Algebra with Applications*, 7th edition. Upper Saddle River, NJ: Pearson/Prentice Hall.
[23] Liang, H., Su, H., & Zou, G. (2008). Confidence intervals for a common mean with missing data with applications in AIDS study. *Computational Statistics & Data Analysis*, 53(2), 546-553.
[24] Little, R. J. A., & Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, second edition. Hoboken, NJ: John Wiley & Sons.
[25] Office for National Statistics. (2014). Change to imputation method used for the turnover question in monthly business surveys. *Guidance and methodology: retail sales*. Retrieved from: http://www.ons.gov.uk/ons/guide-method/method-quality/specific/economy/retail-sales/index.html
[26] Poston, D., & Conde, E. (2014). Missing data and the statistical modeling of adolescent pregnancy. *Journal of Modern Applied Statistical Methods*, 13(2), 464-478.
[27] Rubin, D. B. (1978). Multiple imputations in sample surveys: a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34.
[28] Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley & Sons.
[29] SAS Institute Inc. (2011). *SAS/STAT 9.3 User's Guide*. Retrieved from: http://support.sas.com/documentation/cdl/en/statug/63962/HTML/default/viewer.htm
[30] Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Boca Raton, FL: Chapman & Hall/CRC.
[31] Shao, J. (2000). Cold deck and ratio imputation. *Survey Methodology*, 26(1), 79-85.
[32] Shao, J., & Tu, D. (1995). *The Jackknife and Bootstrap*. New York, NY: Springer.
[33] SPSS Inc. (2009). *PASW Missing Values 18*. Retrieved from: http://www.unt.edu/rss/class/Jon/SPSS_SC/Manuals/v18/PASW Missing Values 18.pdf
[34] Statistical Solutions. (2011). *SOLAS Version 4.0 Imputation User Manual*. Retrieved from: http://www.solasmissingdata.com/wp-content/uploads/2011/05/Solas-4-Manual.pdf
[35] Statistics Bureau of Japan. (2012). *Economic Census for Business Activity*. Retrieved from: http://www.stat.go.jp/english/data/e-census/2012/index.htm
[36] Takahashi, M. (2014). An assessment of automatic editing via the contamination model and multiple imputation. *Proceedings of the Work Session on Statistical Data Editing*, United Nations Economic Commission for Europe, 1-10.
[37] Thompson, K. J., & Washington, K. T. (2012). A response propensity based evaluation of the treatment of unit nonresponse for selected business surveys. *Federal Committee on Statistical Methodology 2012 Research Conference*. Retrieved from: https://fcsm.sites.usa.gov/files/2014/05/Thompson_2012FCSM_III-B.pdf
[38] van Buuren, S. (2012). *Flexible Imputation of Missing Data*. Boca Raton, FL: Chapman & Hall/CRC.
[39] van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: multivariate imputation by chained equations in R. *Journal of Statistical Software*, 45(3), 1-67.
[40] van Buuren, S., & Groothuis-Oudshoorn, K. (2015). *Package 'mice'* [Computer software]. Retrieved from: http://cran.r-project.org/web/packages/mice/mice.pdf