Appendix

Regularity Assumptions

Regularity assumptions necessary for the calculations given in Section 2 may take many forms. The regularity conditions suggested here are not supposed to be technically exhaustive but merely suggestive of what may be needed:

1. "Standard regularity assumptions" on the complete and observed data models, so that the MLE is asymptotically linear and efficient in both models.
2. The draws from the posterior are asymptotically normal, i.e. given the observed data
$$\sqrt{n}\left(\tilde{\theta}_{nj} - \hat{\theta}_n\right) \xrightarrow{\mathcal{D}} N\left(0, I(\theta_0)^{-1}\right).$$
Often this follows from the same regularity conditions used to obtain the asymptotic distribution of the observed data MLE.
3. With $\tilde{X}_i \sim \mathcal{L}_{\theta_n}(X_i|Y_i = y_i)$,
$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} s_{\tilde{X}_i|y_i}(\theta_n) \xrightarrow{\mathcal{D}} N\left(0, EI_Y(\theta_0)^{-1}\right)$$
for any sequence θ_n with $\sqrt{n}(\theta_n - \theta_0)$ bounded and (almost) every sequence of y_is. In practice, a Lindeberg condition should be verified.
4. Uniform laws of large numbers hold for
$$\frac{1}{n}\sum_{i=1}^{n} D_\theta s_{\tilde{X}_i|y_i}(\theta) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} D_\theta s_{Y_i}(\theta)$$
on a compact set of θ with θ_0 as an inner point; here $\tilde{X}_i \sim \mathcal{L}_\theta(X_i|Y_i = y_i)$. In practice this could be verified from additional smoothness (a Lipschitz condition) or more general empirical process techniques.

[*Received May 2001, accepted June 2002*]

Discussion: Efficiency and Self-efficiency With Multiple Imputation Inference

Xiao-Li Meng[1] and Martin Romero[2]

[1]Harvard University, Cambridge, MA 02138, USA. E-mail: meng@stat.harvard.edu [2]Instituto Mexicano del Petróleo, Eje Central Lázaro Cárdenas 152, México D.F., C.P. 07730 México

Summary

By closely examining the examples provided in Nielsen (2003), this paper further explores the relationship between self-efficiency (Meng, 1994) and the validity of Rubin's multiple imputation (RMI) variance combining rule. The RMI variance combining rule is based on the common assumption/intuition that the efficiency of our estimators decreases when we have less data. However, there are estimation procedures

that will do the opposite, that is, they can produce more efficient estimators with less data. Self-efficiency is a theoretical formulation for excluding such procedures. When a user, typically unaware of the hidden self-inefficiency of his choice, adopts a self-inefficient complete-data estimation procedure to conduct an RMI inference, the theoretical validity of his inference becomes a complex issue, as we demonstrate. We also propose a diagnostic tool for assessing potential self-inefficiency and the bias in the RMI variance estimator, at the outset of RMI inference, by constructing a convenient proxy to the RMI point estimator.

Key words: Congeniality; Incomplete data; Missing data; Variance decomposition.

1 Introduction

The main purpose of this paper is to investigate Nielsen's (2003) "counterexamples" to Rubin's multiple imputation (RMI) variance combining rule. Before doing so, we would also like to take this opportunity to comment on Nielsen's two main statements made in his abstract, namely, (1) Bayesian multiple imputation does not guarantee proper multiple imputation, and (2) Bayesian multiple imputation is inefficient even when it is proper. Anyone who only reads these statements, and who has not previously been involved in the research on RMI, is likely to be under the impression that the culprit here is "Bayesian". Logically, statement (1) is obviously correct: simply being a Bayesian does not automatically make one more "proper", in both its general and specific sense (as defined in Rubin, 1987), than a frequentist. An erroneous, or even just naive Bayesian analysis, be it for imputation or for any other purpose, can do just as much damage as any other erroneous method.

In the same spirit, the user's complete-data estimation *procedure* cannot be completely arbitrary. One way to regulate the users' procedures is to require them to be *self-efficient* (Meng, 1994). This requirement is weaker than being fully efficient (e.g., MLE), because it merely excludes any complete-data procedure that *can improve upon itself by applying the same procedure to a part of the same data*. After discussing Nielsen's statement (2), as it is a more transparent issue, we explore the intriguing relationship between the self-efficiency and the validity of RMI inference in all Nielsen's examples.

2 Theoretical Efficiency versus Practical Efficiency

Nielsen's key point underlying his statement (2) is neither targeted at Bayesian nor RMI per se. It is more a general comment about statistical inference based on simulation, in particular, Markov chain Monte Carlo (MCMC). Theoretically, such an inference is never fully efficient, in both statistical and computational senses, compared to the same inference based on exact analytic calculation. But, as Nielsen emphasizes, such a comparison is "hardly fair", since if the analytic calculation is feasible, then there is no real need to conduct simulation. However, in the context of public-use databases, for which RMI was originally designed, even if it is possible for imputers to do everything analytically (which is never the case), they still need to resort to simulation in order to produce multiple imputations, for the obvious reason that the imputers are not the ones who will conduct the subsequent, typically many, statistical analyses using the imputed databases.

Nielsen's point that one should use all k draws within each of the m independent Markov chains instead of just the final m independent draws (the last draw from each chain) reemphasizes sound advice that many advocated and followed in the literature (e.g., Gelman & Rubin, 1992). In general, one key exception to this rule is when the subsequent function evaluations are much more expensive than making the draws, in which case one would purposely ignore some of the dependent draws to achieve practical independence (known as "thinning", as in Meng & Schilling, 2002). Such cost considerations are quite common in statistical practice; for example, they underlie the reason for adopting cluster sampling instead of unit sampling, even if the former is typically statistically less efficient.

For similar practical considerations, when RMI is needed, there are at least two reasons why it can be wise to sacrifice a little bit of theoretical efficiency for practical efficiency, at least for now. First, such km imputations will no longer be conditionally independent (given the observed data), thereby making subsequent analysis more complicated. Although it is possible to derive more general combining rules for specific dependent structures (e.g., as with nested multiple imputation; see Shen, 2000 and Rubin, 2003), currently it is not clear whether it is possible to have a single set of combining rules for dealing with arbitrarily (conditionally) dependent multiply-imputed data sets. This, of course, is a topic worthy of research, but in order for such rules to be relevant for practical purposes, we first need to assess the feasibility of producing multiply-imputed datasets with large km. If km is small to moderate (e.g., 5–50), it would be more efficient, again from a practical point of view, to just produce km (conditionally) independent multiple imputations.

Second, at least in the context of public-use databases, it is not possible to publish all average estimates, averaging over all km draws, that would cover the entire range of the possible estimators in the subsequent analyses. Even when the imputer and analyst are the same actor, the cost of the complete-data analysis can be such that to repeat it km times and then average is not practically desirable when km is large. This might sound strange, as one may wonder how can an investigator be able to make km draws, which involves a full Bayesian prediction machinery, yet not able to conduct km Bayesian analyses with the km complete data sets? However, one must keep in mind that outside the domain of public-use databases, RMI is most useful when a hybrid inferential method is needed. In his discussion of Meng (1994), Zaslavsky (1994) succinctly summarizes this "hybrid" perspective:

> *"Because it may be so difficult to specify fully a Bayesian analysis, in many problems the best strategy can be to use a model-based Bayesian inference for the part that requires it, in particular the imputation of missing data, and to use frequentist methods, relying on estimates of means and variances and on approximate normality, for the rest of the inference. Multiple imputation is a device for such a combined approach, ... This strategy may engender uncongeniality of the analytic methods used in different parts of the inference, even though each is appropriate for its part of the inferential task, and even in cases in which the same organization carries out both parts of the analysis. Nonetheless, the mixed strategy is desirable when it is the most tractable valid approach."*

For such a hybrid approach, the cost of the two parts can be very different. Often, the imputation is the more expensive part, which limits the number of imputations one can create. Sometimes the complete-data analysis is the more expensive part, be it parametric, semi-parametric, or non-parametric. This was the case, for example, in Tu, Meng & Pagano (1993), who used RMI to deal with the delay in death reporting to US's Centers for Disease Control (CDC). The multiply-imputed datasets were then used for an AIDS survival analysis using Cox regression model, with the resulting estimates combined via Rubin's rules to reach an inferential conclusion. The imputation turned out to be relatively straightforward, based on a (conditional) negative binomial model for the number of delayed reports, and the posterior distribution for the model parameters was well approximated by a multivariate normal distribution thanks to the size of the CDC data (nearly 90,000 cases). However, the large sample size makes it a non-trivial task, relative to the effort of producing multiple imputations, to conduct, for example, 1,000 Cox regressions. Nor were such a large number of replications needed. Tu, Meng & Pagano (1993) reported results using $m = 10$ and $m = 50$, and the results were practically the same, taking into account the simulation variation. The rationale for adopting RMI for this AIDS survival analysis was further discussed in the rejoinder of Meng (1994).

Zaslavsky (1994) documented three complex situations where such hybrid approaches (not necessarily involving MI) are useful. Multiple imputation is by no means the only approach for handling

complex incomplete-data problems, but it is certainly one of the more feasible and flexible due to its separation of the modeling for missing data and of the complete-data analysis, regardless of how many actors are involved in making the final inference. A possible price one pays for this feasibility is some theoretical inefficiency, either statistical or computational. But, as Zaslavsky concluded, such a strategy is desirable when it is the *most tractable valid* approach, not necessarily the most efficient approach in theory. See Rubin (1996) for more discussions on theoretical efficiency versus practical efficiency, especially regarding avoiding unwise efforts made towards unachievable theoretical optimality when facing practical constraints.

3 Full Efficiency versus Self-efficiency

The conclusion of the previous section, of course, does not imply, in any way, that one should settle for an approach simply because it is feasible. Indeed, a major reason for the abuse of any popular statistical method, such as the bootstrap, is because the procedure is deceptively easy to use, especially when it is a part of a commercial software. The user, of course, should be concerned with the validity of any method being used. Facing the possibility of uncongeniality, which is a key feature of multi-party inferences, the user should also be informed of what types of validity the method was designed for. This might sound a bit rhetorical as many investigators would naturally equate "validity" with "the right answer", which is typically perceived to be unique. But as statisticians know, with real-life statistical inference there is never a single correct answer. This point is worth emphasizing because it reminds us that the concept of validity is not absolute, and just because one method is valid does not automatically imply invalidity or inferiority of all other methods.

In the context of RMI, a great example of this fact is that although a standard confidence approach is valid, in the sense of delivering its promised nominal level of coverage (e.g., 95%), it can actually be inadmissible because the corresponding interval approach from RMI delivers shorter intervals, yet with at least the same nominal coverage (Meng, 1994; Rubin, 1996). This seemingly paradoxical phenomenon is due to the fact that the interval from RMI can be more efficient, albeit it could be made even more so if additional quantities, such as those cited by Nielsen's equation (11), were available to the user. Rubin (1996) provided an overview of this concept of *confidence-validity*, as Neyman (1934) originally conceived "confidence", and its importance in the context of RMI. Meng (1994) derived a set of conditions that not only guarantees the confidence-validity of RMI interval estimators but its superiority over the user's standard interval estimator without using RMI (even assuming the user's analysis is free of nonresponse bias).

In particular, Meng (1994) introduced the notion of *self-efficiency*. This was based on the following observation. Since Rubin's variance combining rule relies on the following decomposition,

$$\textit{Total Variance} = \textit{Within-imputation Variance} + \textit{Between-imputation Variance}, \tag{1}$$

it is clear that, for this decomposition to be useful, the estimator from RMI cannot be more efficient than the user's complete-data estimator. This could happen, however, if the user's (complete-data) estimation procedure can actually produce a more efficient estimator when applied to the incomplete data than applied to the complete data. This is because the RMI estimator is an adjusted version of user's complete-data procedure applied to the incomplete data, adjusted for the non-response bias that was modeled by the imputer. As such, the RMI estimator it is often highly correlated with the (hypothetical) user's incomplete-data estimator without imputation, and thus the inherited properties of the user's estimation procedure may carry over to the RMI estimator. This is one reason that the congeniality concept in Meng (1994) was defined with respect to the user's *estimation procedure*, not just the user's *complete-data estimator*. A more important and somewhat subtle reason for this "insistence", as Nielsen noted in his Section 4, is that the frequentist operating characteristics of any incomplete-data analysis, be it RMI or otherwise, must depend on the missing data mechanism

(MDM), that is, the process that leads to the observed data pattern. Clearly, the user's *complete-data estimator* alone does not and cannot capture the MDM, but user's *estimation procedure* is driven by the MDM, and thus it is a more relevant ingredient for formulating congeniality.

Estimation procedures that can provide more efficient estimators with less data may appear to be counterintuitive, but they do exist, even in simple examples. One such example was discussed in Meng (2001), using an example of Robins & Wang (2000), in the context of linear regression under heteroscedasticity. Because the standard least-square estimator (LSE) does not properly weight each observation with heteroscedastic errors, it is possible for the LSE estimator based on a partial data set to be more efficient than the complete-data LSE, when those data points with larger variances are missing. The self-efficient requirement, in a nutshell, eliminates such procedures. In other words, the LSE estimator may not be self-efficient with respect to a selection process that depends on the covariate (i.e., missing at random (MAR), but not missing completely at random (MCAR), as defined in Rubin (1976)). LSE estimators are typically used because of their robustness properties. However, to what degree such considerations contradict self-efficiency, which by itself also appears to be a desirable requirement, is a topic that requires much more research.

Mathematically, *self-efficiency* is defined in the following way:

Definition. (Meng, 1994): Let W_c be a data set, and let W_o be a subset of W_c created by a selection mechanism. A statistical estimation procedure $\hat{\theta}(\cdot)$ for θ is said to be *self-efficient* (with respect to the selection mechanism) if there is no $\lambda \in (-\infty, +\infty)$ such that the mean-squared error of $\lambda\hat{\theta}(W_o) + (1 - \lambda)\hat{\theta}(W_c)$ is less than that of $\hat{\theta}(W_c)$.

Note that implicit in this definition is the assumption that the estimation procedure, $\hat{\theta}(\cdot)$, is well-defined for any subset of W_c that is selected by the *specified* missing-data mechanism; $\hat{\theta}(\cdot)$ is well understood as a convention, but its rigorous mathematical formulation is beyond the scope of this paper because of the potentially arbitrary dimension and structure of W_o. Also note that this definition does not exclude the possibility that $\hat{\theta}(\cdot)$ is self-efficient for some values of θ but not for others, when the mean-squared error (MSE) or asymptotically the variance of $\hat{\theta}(\cdot)$ depends on θ. We also emphasize that at the first sight the requirement that $\hat{\theta}(W_c)$ is no less efficient than any $\lambda\hat{\theta}(W_o) + (1 - \lambda)\hat{\theta}(W_c)$ may seem to be too strong for formulating our intuitive desire that $\hat{\theta}(\cdot)$, as an estimation procedure, should not produce a more efficient estimator with less data, which might appear to be adequately formulated by only requiring $\hat{\theta}(W_c)$ be no less efficient than $\hat{\theta}(W_o)$. In actuality, the linear combination formulation is the more sensible, though a bit more subtle, theoretical formulation of our desire. This is because any $\hat{\theta}(W_c)$ can be rewritten as $\lambda\hat{\theta}(W_c) + (1 - \lambda)\hat{\theta}(W_c)$ for any λ.

Consider the following scenario. Suppose both of us (Meng and Romero) are provided with the same complete-data set W_c, and we both adopt the same procedure $\hat{\theta}(\cdot)$. A third person, say Nielsen, is not provided with any data but only our two estimates, and he has the task of combining them to form a single estimate. Although there are many ways of combining two estimates, the linear combination is typically the easiest, and in fact the only combining method needed asymptotically when variance is used as the measure of efficiency (by a Taylor-expansion argument). Note that in order to maintain the consistency when linearly combining two consistent estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ for the same θ, the linear weights in $\lambda_1\hat{\theta}_1 + \lambda_2\hat{\theta}_2$ must satisfy $\lambda_1 + \lambda_2 = 1$, but neither of them is required to be between 0 and 1 (as we normally think of weights). Obviously, when we both apply $\hat{\theta}(\cdot)$ to W_c, Nielsen will end up with the same $\hat{\theta}(W_c)$ regardless of his choice of the linear weight λ. However, if $\hat{\theta}(\cdot)$ is not self-efficient, as defined above, then Nielsen can achieve a more efficient combined estimator than $\hat{\theta}(W_c)$ by asking one of us to throw away some of the data, that is, by applying $\hat{\theta}(\cdot)$ to W_o instead of W_c. Clearly, this is exactly what we try to avoid when we want to eliminate procedures that can produce more efficient estimators with less data.

In words, what self-efficiency requires is that an estimation procedure cannot be improved upon itself by merely mixing the complete-data estimator with an estimator from applying the *same*

procedure to part of the data. This does not require that the complete-data estimator, $\hat{\theta}_{com}$, to be the most efficient estimator, but only that it is most efficient "within itself" (hence the term), defined by the linear mixture class $\hat{\Theta} = \{\lambda\hat{\theta}_{obs} + (1-\lambda)\hat{\theta}_{com},\ \lambda \in (-\infty, \infty)\}$, with respect to the missing-data mechanism; here $\hat{\theta}_{obs}$ is the observed-data estimator resulting from user's estimation procedure on W_o rather than W_c. Nielsen's first example provides such an illustration because his complete-data estimation procedure is self-efficient with respect to his MDM, but it is not fully efficient. In general, it is easy to show that any sample average procedure of i.i.d terms, namely, $\sum_{i=1}^{n} g(X_i)/n$, is self-efficient with respect to an MDM that randomly selects a sub-sample of $\{X_1, \ldots, X_n\}$, as in Nielsen's first example. As we also see from Nielsen's verification, the bias in the RMI variance estimator is positive and thus the resulting RMI interval estimator is confidence valid when $m = \infty$.

The above discussion should not be viewed as a suggestion that self-efficiency alone is sufficient to guarantee the validity of RMI inference. This cannot be the case because self-efficiency only regulates the user's complete-data estimation procedure, and as such it puts no restriction on the imputation model whatsoever. Technically, it cannot be a necessary condition either, because the estimator from the user's procedure applied to the incomplete data can be made arbitrarily different from the RMI estimator, $\bar{\theta}_\infty$ (we focus on $m = \infty$ to separate the issue of uncongeniality from the issue of lack of efficiency due to finite number of imputation). Nielsen's second choice of the user's estimation procedure in his second example provides an extreme illustration of this arbitrariness, because his complete-data estimator is $\bar{X}$, but the X sample is completely imaginary and thus there is no coherent way of defining the user's observed-data estimator $\hat{\theta}_{obs}$ without imputation. In such cases, properties of $\bar{\theta}_\infty$ will put little restriction on the relationship between $\hat{\theta}_{com}$ and $\hat{\theta}_{obs}$, which is the key to self-efficiency. Nevertheless, in many practical situations, there is a strong correlation between $\hat{\theta}_{obs}$ and $\bar{\theta}_\infty$ even under uncongeniality, which leads to a strong relationship between self-efficiency and the properties of $\bar{\theta}_\infty$. Indeed, by Lemma 1 of Meng (1994), when $\bar{\theta}_\infty = \hat{\theta}_{obs}$, self-efficiency becomes the sufficient and necessary condition for the RMI variance estimator (with $m = \infty$) to be consistent, under the additional assumption that the imputation model is second-moment proper with respect to the user's complete-data procedure. In addition to his Example 1, Nielsen's other examples also demonstrate this close relationship between self-efficiency and the validity of RMI (interval) estimators, as detailed in the rest of this article.

4 Checking Self-efficiency for Example 2

Nielsen's second example assumes that we have an i.i.d. sample $\{Y_i,\ i = 1, \ldots, n\}$ from $N(\theta, 1)$. We then *imagine* that we could have had another independent sample $\{X_i,\ i = 1, \ldots, n\}$ from $N(\theta, \sigma^2)$ with known σ^2, but they are all missing. Nielsen's first choice of the complete-data estimator is the simple average of all X's and Y's, namely, $\hat{\theta}_{com} = (\bar{X} + \bar{Y})/2$. Given this complete-data estimator, it is natural to assume that the user would estimate θ by $\hat{\theta}_{obs} = \bar{Y}$ if he does not observe any X_i's (which is indeed the case). To check the self-efficiency of this procedure, let us examine the linear mixture $\hat{\theta}_\lambda = \lambda\hat{\theta}_{obs} + (1-\lambda)\hat{\theta}_{com}$, whose sample variance, under Nielsen's specification, is

$$\mathrm{V}(\hat{\theta}_\lambda) = \frac{1}{4n}\left[(1-\lambda)^2\sigma^2 + (1+\lambda)^2\right] = \frac{1+\sigma^2}{4n}\left[\lambda^2 - 2\frac{\sigma^2-1}{\sigma^2+1}\lambda + 1\right]. \tag{2}$$

Consequently,

$$D_v(\lambda) \equiv \mathrm{V}(\hat{\theta}_\lambda) - \mathrm{V}(\hat{\theta}_{com}) = \frac{1+\sigma^2}{4n}\left[\lambda^2 - 2\frac{\sigma^2-1}{\sigma^2+1}\lambda\right], \tag{3}$$

which is guaranteed to be non-negative for all λ if and only if $\sigma^2 = 1$. That is, only when $\sigma^2 = 1$ is the procedure self-efficient, which of course is well known because when $\sigma^2 \neq 1$, we can easily improve the efficiency of $\hat{\theta}_{com}$ by weighting $\bar{X}$ and $\bar{Y}$ by their known precisions, that is, the complete-data

procedure should be $(\bar{Y} + \sigma^{-2}\bar{X})/(1 + \sigma^{-2})$, which would still render the same $\hat{\theta}_{obs} = \bar{Y}$ when all X's are missing. We note that in this case, requiring self-efficiency leads to the fully-efficient MLE, but this is not true in general, as Nielsen's first example shows. Furthermore, if this were for a real-life application, there would really be no reason for not using the MLE as the complete-data estimation procedure, for the missing X's are completely imaginary and thus the actual value of σ^2 is not relevant. Indeed, the RMI procedure (ignoring small m variations) would just reproduce the standard estimation procedure based on $\bar{Y}$ (or its asymptotic equivalent when the prior on θ used in imputation is not the constant prior, such as $N(0, 1)$ used by Nielsen), regardless of the actual value of σ^2.

When $\sigma^2 \neq 1$, the $D_v(\lambda)$ function of (3) has two roots, $\lambda = 0$ and $\lambda = 2(\sigma^2 - 1)/(\sigma^2 + 1)$. Consequently, when $\sigma^2 > 1$, only some positive λ's can make $D_v(\lambda) < 0$, and when $\sigma^2 < 1$, only some negative λ's can do the same. In other words, when $\sigma^2 > 1$, $\hat{\theta}_{com}$ is the most efficient estimator only among the "negative mixture" class

$$\hat{\Theta}^- = \{\lambda\hat{\theta}_{obs} + (1 - \lambda)\hat{\theta}_{com},\ \lambda \leq 0\}, \tag{4}$$

and when $\sigma^2 < 1$, $\hat{\theta}_{com}$ is the most efficient estimator only among the "positive mixture" class

$$\hat{\Theta}^+ = \{\lambda\hat{\theta}_{obs} + (1 - \lambda)\hat{\theta}_{com},\ \lambda \geq 0\}. \tag{5}$$

Since in this case the RMI estimator with $m = \infty$, $\bar{\theta}_\infty$, is (asymptotically) the same as $\hat{\theta}_{obs} = \bar{Y}$, by Lemma 1 and Lemma 2 of Meng (1994), we can conclude that the RMI variance estimator has a positive bias when $\sigma^2 > 1$ and negative bias when $\sigma^2 < 1$. This is in complete agreement with Nielsen's finding with finite m, as it should be.

5 Checking Self-efficiency for Example 3

Nielsen's third example, which is a simulation study, assumes that, in the absence of missing observations, we have an i.i.d sample $\{(X_i, Y_i), i = 1, \ldots, n\}$ from a bivariate normal with equal means $E(X) = E(Y) = \theta$, which is our estimand. Nielsen simulated 5000 data sets of size $n = 200$ using $\theta = 1$, $\sigma_x^2 = 0.5$, $\sigma_y^2 = 1$, and $\sigma_{x,y} = 0.5$, and hence $\rho_{x,y} = \text{Corr}(X, Y) = \sqrt{0.5} = 0.707$ (note that Nielsen's ρ is our $\sigma_{x,y}$, not our $\rho_{x,y}$). The MDM is such that all Y_i's are observed, but X_i is observed only when $Y_i \leq C$. Let $R_i = 1_{\{Y_i \leq C\}}$, then X_i is observed if and only if $R_i = 1$. Consequently, $n_{obs} \equiv \sum_{i=1}^n R_i$ denotes the number of observed X_i's. For Nielsen's simulation, C was set to 1.84, so $E(R) = P(Y \leq C) = 0.7995 \approx 0.8$ and $E(n_{obs}) = E(R)n \approx 160$, implying about 20% of X_i's are missing.

Unlike Example 2, where the MDM is MCAR, the MDM here is only MAR. Although MAR is enough to guarantee the ignorability of the MDM for *likelihood* inference when the *additional* "parameter distinct" (PD) assumption holds (which is the case for this example because the conditional distribution of R given the observed data is free of any unknown parameter), it is not enough for a *design-based* inference (regardless of whether PD holds or not); see Rubin (1976) for a detailed discussion of such issues. This non-ignorability complicates our construction of user's estimation procedure that would naturally lead to Nielsen's complete-data estimator $\hat{\theta}_{com} = (\bar{X} + \bar{Y})/2$ when $R_i = 1$ for all i. For example, neither of the two obvious "design-based" choices,

$$\hat{\theta}_{obs}^{(1)} = \frac{\sum_{i=1}^n (X_i R_i + Y_i)}{n_{obs} + n} \quad \text{and} \quad \hat{\theta}_{obs}^{(2)} = \frac{1}{2}\left[\frac{\sum_{i=1}^n X_i R_i}{n_{obs}} + \bar{Y}\right], \tag{6}$$

is consistent for θ, because the mean of X conditioning on $Y \leq C$ is not θ as long as C is finite, even though the asymptotic bias of either estimator may be negligible for practical purposes when $C_0 = (C - \theta)/\sigma_y$ happens to be large.

The inconsistency of both of these two estimators also highlights the danger of letting a general user

deal with missing data without being provided with the imputations. If this were a real application, and the user is unaware of the actual MDM, it is likely that he would mimic what he would do with complete data, either by averaging over all the observed X's and Y's, namely $\hat{\theta}^{(1)}_{obs}$, or by averaging the sample average of the observed X's, $\bar{X}_{obs}$, and $\bar{Y}$, that is, $\hat{\theta}^{(2)}_{obs}$. Of course, a more sophisticated user may compute the maximum likelihood estimator for θ under the Nielsen's bivariate normal model assuming ignorability (which is a correct assumption), which would produce a consistent (and fully efficient) estimator. But then $\hat{\theta}_{com} = (\bar{X} + \bar{Y})/2$ could be this user's complete-data estimator only by an illogical mathematical construction. Indeed, under the assumption that the covariance matrix is known (this assumption greatly simplifies but has no real consequences for our current discussion), the complete-data MLE is

$$\hat{\theta}_{ML.com} = \frac{(\sigma_y^2 - \sigma_{x,y})\bar{X} + (\sigma_x^2 - \sigma_{x,y})\bar{Y}}{\sigma_x^2 + \sigma_y^2 - 2\sigma_{x,y}}. \tag{7}$$

For Nielsen's simulation configuration, $\sigma_x^2 - \sigma_{x,y} = 0$ and thus $\hat{\theta}_{ML.com} = \bar{X}$, that is, the entire Y sample carries no additional information about θ once the X sample is fully observed. The reason is that when the slope for regressing Y on X is $\beta_{y|x} = \sigma_{x,y}/\sigma_x^2 = 1$, which is Nielsen's choice, the "residual" $Y - \beta X \sim N((1 - \beta_{y|x})\theta, (1 - \rho^2_{x,y})\sigma_y^2)$ provides no additional information about θ. Clearly, for a user who is sophisticated enough to compute the MLE when some X's are missing, it would take an extraordinary argument to persuade him that he should adopt $(\bar{X} + \bar{Y})/2$ instead of the MLE, $\bar{X}$, when no X_i is missing.

Therefore, a more logical and more relevant choice of a user's *consistent* estimation procedure is to assume that, in the absence of MI, the user would attempt to impute the missing X_i's in $(\bar{X} + \bar{Y})/2$ in some convenient but "consistent" way (e.g., not leading to an inconsistent estimator such as with imputing all missing X_i's by $\bar{X}_{obs}$, that is, $\hat{\theta}^{(2)}_{obs}$ given in (6)). There are a number of ways of doing so. One of them is to impute the missing X_i by its fitted value from the regression

$$\hat{X}_i(\theta) = \theta + \beta(Y_i - \theta), \tag{8}$$

where $\beta \equiv \beta_{x|y} = \sigma_{x,y}/\sigma_y^2$ is assumed to be known for simplicity of our presentation (in practice β would be estimated by LSE from all the fully observed pairs of (X_i, Y_i); this would alter our asymptotic variance calculations but would not change conclusions that are relevant for our current investigation). This leads to a "self-consistent" equation for $\hat{\theta}$,

$$\hat{\theta} = \frac{1}{2}\left[\frac{\sum_{i=1}^n [R_i X_i + (1 - R_i)\hat{X}_i(\hat{\theta})]}{n} + \bar{Y}\right], \tag{9}$$

solving which leads to our estimation procedure

$$\hat{\theta}_{obs} = \frac{\sum_{i=1}^n (X_i R_i + Y_i + \beta(1 - R_i)Y_i)}{\sum_{i=1}^n (1 + R_i + \beta(1 - R_i))}. \tag{10}$$

Clearly, $\hat{\theta}_{obs}$ automatically yields $\hat{\theta}_{com} = (\bar{X} + \bar{Y})/2$ when all $R_i = 1$. It is also easy to verify that $\hat{\theta}_{obs}$ is consistent by the law of large numbers and the fact that $E[R(X - \theta)] = \beta E[R(Y - \theta)]$.

To check the self-efficiency of this procedure, we can compute the MSE of $\hat{\theta}_\lambda = \lambda\hat{\theta}_{obs} + (1 - \lambda)\hat{\theta}_{com}$ or asymptotically its variance, which can be obtained via the standard delta method (see Appendix for details). Figure 1 plots $n\mathrm{V}(\hat{\theta}_\lambda)$ as a function of λ, where the horizontal line is

$$n\mathrm{V}(\hat{\theta}_{com}) = \frac{1}{4}(\sigma_x^2 + \sigma_y^2 + 2\sigma_{x,y}) = \frac{5}{8} = 0.625, \tag{11}$$

under Nielsen's setting. Similar to the case of $\sigma^2 < 1$ in Example 2 (recall here we also have $\sigma_x^2 = 0.5 < \sigma_y^2 = 1$), $\hat{\theta}_{com}$ is the most efficient among the positive mixture class $\hat{\Theta}^+$ (defined in (5))

but not among the negative mixture class $\hat{\Theta}^-$ (defined in (4)); note that because $n\mathrm{V}(\hat{\theta}_\lambda)$ is quadratic in λ, its value outside the plotting range in Figure 1 must also be above the horizontal line. If $\hat{\theta}_{obs}$ were (asymptotically) the same as $\bar{\theta}_\infty$, then again by Lemma 1 and Lemma 2 of Meng (1994), we could immediately conclude that the RMI variance estimator would have a negative bias. However, for the current problem, $\hat{\theta}_{obs}$ of (10) is not asymptotically equivalent to $\bar{\theta}_\infty$. Consequently, we cannot directly apply Meng's (1994) Lemmas to conclude the negative bias of the RMI variance estimator.

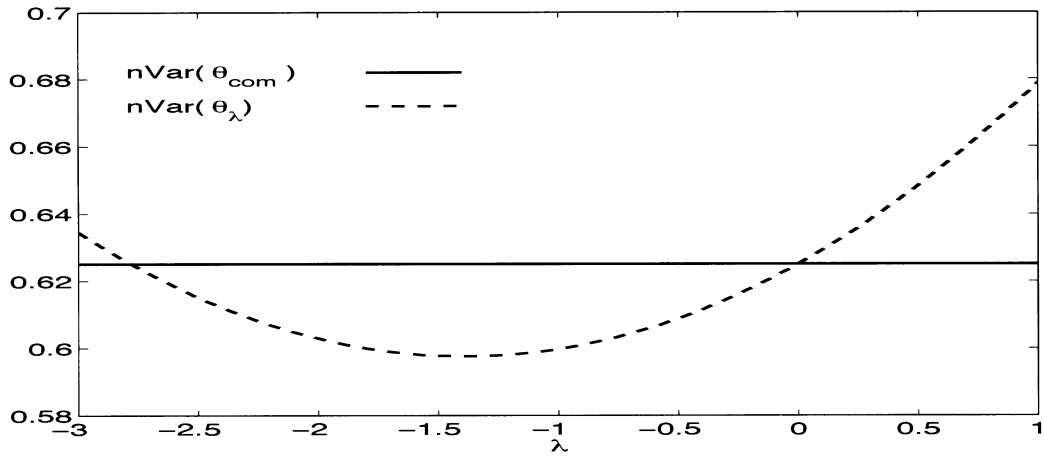

Figure 1. *Compare Variance of* $\hat{\theta}_\lambda = \lambda\hat{\theta}_{obs} + (1-\lambda)\hat{\theta}_{com}$ *with* $\mathrm{V}(\hat{\theta}_{com})$.

Nevertheless, because of the close relationship between $\hat{\theta}_{obs}$ and $\bar{\theta}_\infty$, we would still expect a negative bias in the RMI variance estimator based on our calculation for $\hat{\theta}_{obs}$ before performing an actual check for $\bar{\theta}_\infty$, either analytically or via simulation. The fact that Nielsen's simulation is consistent with this expectation is good indication of the role of self-efficiency in assessing the validity of the RMI inference. It also illustrates how we might investigate the bias issue in the RMI variance estimator of $\bar{\theta}_\infty$ via a corresponding, but more convenient, investigation by constructing a reasonably related $\hat{\theta}_{obs}$. The latter is more convenient because of the flexibility we have in constructing $\hat{\theta}_{obs}$, which does not involve the Bayesian imputation, at least not the full details. Evidently, this indirect approach cannot be 100% effective, but it provides a way of diagnosing/detecting potential bias in the RMI variance estimator caused by user's choice of his complete-data estimator. Indeed, as we will see in the next section, not only does this diagnostic tool provide a qualitative indication (i.e., the sign of bias), but it can also provide a suggestive quantitative assessment of the magnitude of the bias in the RMI variance estimator.

6 Assessing Self-inefficiency and Bias in RMI Variance Estimator

As we discussed in Section 3, the linear-combination formulation of self-efficiency provides a direct theoretical quantification of our desire for excluding estimation procedures that can provide more efficient estimators with less data. Mathematically, checking self-efficiency via directly computing the (asymptotic) variance of the linear combination $\hat{\theta}_\lambda$, as in the previous two sections, can be short cut. Since $\mathrm{V}(\hat{\theta}_\lambda)$ is a simple quadratic form of λ, in particularly,

$$\mathrm{V}(\hat{\theta}_\lambda) - \mathrm{V}(\hat{\theta}_{com}) = \lambda^2\mathrm{V}(\hat{\theta}_{obs} - \hat{\theta}_{com}) - 2\lambda\mathrm{Cov}(\hat{\theta}_{com}, \hat{\theta}_{com} - \hat{\theta}_{obs}),$$

the following equivalence result is trivial to verify.

THEOREM 1. *Let* $\hat{\theta}_{com} = \hat{\theta}(W_c)$ *and* $\hat{\theta}_{obs} = \hat{\theta}(W_o)$. *Then*

(1) $\hat{\theta}(\cdot)$ *is self-efficient if and only if* $\mathrm{Cov}(\hat{\theta}_{com}, \hat{\theta}_{com} - \hat{\theta}_{obs}) = 0$;
(2) $\hat{\theta}(\cdot)$ *is self-inefficient among the "positive" mixture class* $\hat{\Theta}^+ = \{\lambda\hat{\theta}_{obs} + (1-\lambda)\hat{\theta}_{com},\ \lambda \geq 0\}$ *if and only if* $\mathrm{Cov}(\hat{\theta}_{com}, \hat{\theta}_{com} - \hat{\theta}_{obs}) > 0$;
(3) $\hat{\theta}(\cdot)$ *is self-inefficient among the "negative" mixture class* $\hat{\Theta}^- = \{\lambda\hat{\theta}_{obs} + (1-\lambda)\hat{\theta}_{com},\ \lambda \leq 0\}$ *if and only if* $\mathrm{Cov}(\hat{\theta}_{com}, \hat{\theta}_{com} - \hat{\theta}_{obs}) < 0$.

Therefore, in order to check self-efficiency, we only need to compute

$$\Delta = \mathrm{Cov}(\hat{\theta}_{com}, \hat{\theta}_{com} - \hat{\theta}_{obs}) = \mathrm{V}(\hat{\theta}_{com}) - \mathrm{Cov}(\hat{\theta}_{com}, \hat{\theta}_{obs}). \tag{12}$$

The complete-data variance $\mathrm{V}(\hat{\theta}_{com})$ is typically easy to compute (e.g., as in (11)). The calculation of $\mathrm{Cov}(\hat{\theta}_{com}, \hat{\theta}_{obs})$ can be a bit involved (see Appendix), but nevertheless it can be performed without involving MI. It turns out that for the current problem, Δ can be simplified to

$$\Delta = \frac{(1+\beta)(\sigma_x^2 - \sigma_y^2)[1 - \Phi(C_0)]}{4n[(1-\beta)\Phi(C_0) + (1+\beta)]}, \tag{13}$$

where $C_0 = (C - \theta)/\sigma_y$. Interestingly, this calculation reveals that there are two special cases for which $\hat{\theta}_{obs}$ of (10) becomes self-efficient. The obvious one is when $\sigma_x^2 = \sigma_y^2$, in which case the equal-weighted $\hat{\theta}_{com} = (\bar{X} + \bar{Y})/2$ is the complete-data MLE and therefore it is more efficient than any other estimator. The less obvious case is when $\beta = -1$. An inspection of (10) reveals that when $\beta = -1$, $\hat{\theta}_{obs}$ is the same as

$$\hat{\theta}^*_{obs} = \frac{\sum_{i=1}^n R_i(X_i + Y_i)/2}{\sum_{i=1}^n R_i}. \tag{14}$$

In other words, $\hat{\theta}_{obs}$ is simply $\hat{\theta}_{com}$ applied to all the fully observed pair $\{X_i, Y_i\}$, with all the other observed Y_i's discarded. As we mentioned in Section 3, the sample average of i.i.d. terms is self-efficient with respect to an MDM that randomly selects a sub-sample. However, our MDM here is not a random selection of sub-sample. Indeed $\hat{\theta}^*_{obs}$ is inconsistent unless $\beta = -1$ because it converges to

$$\frac{E[R(X+Y)]}{2E(R)} = \theta - \frac{(1+\beta)\sigma_y}{2}\frac{\varphi(C_0)}{\Phi(C_0)},$$

where φ and Φ are respectively the density and cdf of $N(0, 1)$. So the fact $\hat{\theta}_{obs} = \hat{\theta}^*_{obs}$ alone does not explain why $\hat{\theta}_{obs}$ is self-efficient (being consistent is only a necessary condition for self-efficiency).

The real reason is that when $\beta = -1$, $\mathrm{Cov}(X + Y, Y) = (1+\beta)\sigma_y^2 = 0$, and hence $X + Y$ and Y are independent (since they are jointly normal). Consequently, as far as the procedure $\hat{\theta}^*_{obs}$ goes, when $\beta = -1$, the MDM given by $R = 1_{\{Y \leq C\}}$ is the same as an independent random sub-sampling of $\{X_i + Y_i, i = 1, \ldots, n\}$, and therefore $\hat{\theta}^*_{obs}$ is self-efficient. This is a good example illustrating how self-efficiency captures the MDM through the choice of $\hat{\theta}_{obs}$.

For Nielsen's simulation study, $\beta = 0.5$ and $\sigma_x^2 - \sigma_y^2 = -0.5$ and thus $\Delta < 0$, indicating a negative bias in the RMI variance estimator. The Δ quantity not only is useful for detecting the sign of the bias in the RMI variance estimator, it is also suggestive to the magnitude of the bias. This is because, as Kott (1992) observed, mathematically, the bias in the RMI variance estimator is due to the presence of the "cross term" in the general expression

$$\mathrm{V}(\bar{\theta}_\infty) = \mathrm{V}(\hat{\theta}_{com}) + \mathrm{V}(\bar{\theta}_\infty - \hat{\theta}_{com}) + 2\mathrm{Cov}(\hat{\theta}_{com}, \bar{\theta}_\infty - \hat{\theta}_{com}), \tag{15}$$

where the first two terms on the right-hand side are consistently (when $m = \infty$) estimated respectively by the within-imputation variance and between-imputation variance (under the assumption of proper or second-moment proper imputation model for the user's complete-data estimator). Consequently,

the bias in the RMI variance combining rule (with $m = \infty$) is given by

$$BT_\infty = 2\text{Cov}(\hat{\theta}_{com}, \hat{\theta}_{com} - \bar{\theta}_\infty). \tag{16}$$

If we substitute $\hat{\theta}_{obs}$ for $\bar{\theta}_\infty$, we see that BT_∞ is just 2Δ, which can be computed or estimated by the user without MI. As long as we construct $\hat{\theta}_{obs}$ with some care (as in Section 5), the resulting 2Δ serves as a reasonable approximation to BT_∞ or its finite version BT_m.

To illustrate, for Nielsen's Example 3, we can easily calculate from (13) that $2\Delta = -0.000198$. This *suggests* that the *relative bias* in the RMI variance estimator, $RT_\infty = BT_\infty/\text{V}(\bar{\theta}_\infty)$, is about

$$RO \equiv \frac{2\Delta}{\text{V}(\hat{\theta}_{obs})} = \frac{-0.000198}{0.003394} = -5.8\%, \tag{17}$$

where $\text{V}(\hat{\theta}_{obs}) = 0.003394$ is from Appendix. The actual bias in RMI variance estimator with $m = 5$, from Nielsen's Figure 1 as well as from our own independent simulation, is about -0.00026 with a relative bias about -7.2%. Clearly, when $\hat{\theta}_{obs}$ is not equivalent to $\bar{\theta}_\infty$, RO defined in (17) cannot be RT_m whether $m = \infty$ or not. Nevertheless, for practical purposes, it can be quite useful to have such a "ball-park" quantitative assessment at the outset of the actual RMI analysis, if the potential bias in the RMI variance estimator is of real concerns.

7 Concluding Remarks

The key message from Nielsen's paper, in our view, is the further emphasis that in order for RMI to provide valid inference, the user's complete-data procedure cannot be completely arbitrary. We certainly fully agree with this point, as previously investigated in various forms (e.g., Rubin's (1987) randomization validity requirement and Meng's (1994) self-efficiency requirement). However, as we show by both theory and examples, the needed restrictions appear to be more in line with self-efficiency than with full efficiency. Whereas full efficiency implies self-efficiency, the latter in general is a weaker requirement, and it is equivalent to the former only when we require it hold for an arbitrary missing-data mechanism. Although by itself self-efficiency is neither a sufficient nor a necessary condition for the validity of the RMI inference, we believe it is an important building block, providing a theoretical explanation why empirical investigations have shown satisfactory performance of RMI in many cases, where the complete-data estimators are sensible but by no means fully efficient (e.g., the list provided in Rubin, 1996). Of course, much more research is needed regarding the relationship between self-efficiency, full efficiency, as well as robustness. We therefore conclude by thanking Nielsen for providing several stimulating theoretical examples where such issues can be further examined.

Acknowledgements

We thank Soren Feodor Nielsen for helpful exchanges regarding the simulation studies for his Example 3, and Radu Craiu, Donald Rubin and Alan Zaslavsky for their comments. The research was supported in part by US National Science Foundation grants DMS 0072510 and DMS 0204552.

Appendix

Since $\hat{\theta}_{obs}$ of (10) is a ratio estimator, we invoke the following well-known large sample approximation, in obvious notation, for computing $\text{V}(\hat{\theta}_{obs})$ and $\text{Cov}(\hat{\theta}_{com}, \hat{\theta}_{obs})$:

$$\text{Cov}\left(\frac{U_1}{W_1}, \frac{U_2}{W_2}\right) \approx \frac{1}{E(W_1)E(W_2)} Cov\left(U_1 - \frac{EU_1}{EW_1}W_1, U_2 - \frac{EU_2}{EW_2}W_2\right).$$

A bit of algebra then yields

$$n\mathrm{V}(\hat{\theta}_{obs}) = \frac{\mathrm{V}(XR) + (1+\beta)^2\sigma_y^2 + \beta^2\mathrm{V}(RY)}{[1+\beta+\Phi(C_0)(1-\beta)]^2} + 2\frac{(1+\beta)\mathrm{Cov}(XR, Y) - \beta\mathrm{Cov}(XR, YR) - \beta(1+\beta)\mathrm{Cov}(Y, RY)}{[1+\beta+\Phi(C_0)(1-\beta)]^2},$$

and

$$n\mathrm{Cov}(\hat{\theta}_{obs}, \hat{\theta}_{com}) = \frac{\mathrm{Cov}(XR, X) + (1+\beta)\sigma_{x,y} + (1-\beta)\mathrm{Cov}(RY, X)}{2[1+\beta+\Phi(C_0)(1-\beta)]} + \frac{(1+\beta)\sigma_y^2 - \beta\mathrm{Cov}(RY, Y)}{2[1+\beta+\Phi(C_0)(1-\beta)]}.$$

To simplify the algebra, we can assume $E(X) = E(Y) = 0$ as long as we rewrite $R = 1_{\{Y \le C_0\}}$. It is then easy to verify the following identities

$$E(RY) = -\sigma_y\varphi(C_0), \qquad E(RY^2) = \sigma_y^2[\Phi(C_0) - C_0\varphi(C_0)],$$
$$E(RX^2) = \tau^2\Phi(C_0) + \beta^2 E(RY^2), \quad E(RXY) = \beta E(RY^2),$$

where $\tau^2 = \sigma_x^2 - \beta\sigma_{x,y}$.

Combining all the expressions above yields

$$n\mathrm{V}(\hat{\theta}_{obs}) = \frac{\tau^2\Phi(C_0) + (1+\beta)^2\sigma_y^2}{[1+\beta+\Phi(C_0)(1-\beta)]^2}$$

and

$$n\mathrm{Cov}(\hat{\theta}_{obs}, \hat{\theta}_{com}) = \frac{\tau^2\Phi(C_0) + (1+\beta)(\sigma_{x,y} + \sigma_y^2)}{2[1+\beta+\Phi(C_0)(1-\beta)]}.$$

For Nielsen's simulation configurations, we have

$$n\mathrm{V}(\hat{\theta}_{obs}) = 0.678801 \quad \text{and} \quad n\mathrm{Cov}(\hat{\theta}_{obs}, \hat{\theta}_{com}) = 0.644784.$$

Together with $n\mathrm{Cov}(\hat{\theta}_{com}) = 0.625$ as in (11), we plot Figure 1 via

$$\mathrm{V}(\hat{\theta}_\lambda) = \lambda^2\mathrm{V}(\hat{\theta}_{obs}) + (1-\lambda)^2\mathrm{V}(\hat{\theta}_{com}) + 2\lambda(1-\lambda)\mathrm{Cov}(\hat{\theta}_{obs}, \hat{\theta}_{com}).$$

Résumé

A la suite d'une examination des exemples fournis par Nielsen (2003), cet article mène une investigation plus approfondie sur la connection entre auto-efficacité (self-efficiency, Meng, 1994) et la valabilité de la loi de calcul de variances établi par Rubin dans le contexte de l'imputation multiple (RIM). La régle RIM de calcul de la variance s'appuye sur une hypothèse intuitive qui établit que l'efficacité de l'estimateur décroît avec le pourcentage des données manquantes. Il y a, quand même, des procedures d'éstimation qui ne respectent pas l'hypothèse d'en haut. Plus précisement elles produisent des estimateurs plus efficaces avec moins des données. L'auto-efficacité est une notion theorique qui peut nous aider a eviter ces situations paradoxales et incorrectes. Quand un ultilisateur, ignorant de l'auto-inefficacité de son choix, adopte une procédure d'estimation qui n'est pas auto-efficace, la validité théorique de ses resultats est une problème complexe. Nous proposons aussi une méthode de diagnostiquer la potentielle auto-inefficacité et le biais de l'estimateur de variance RMI qui est fondée sur un estimateur alternatif plus convenable que celui produit par RMI.

[*Received August 2003, accepted August 2003*]